문화콘텐츠 분야
인공지능(AI) 윤리지침 연구

2025년 4월

문화체육관광부
신수정

<목차>

국외훈련 개요

1. 훈련국 : 미국

2. 훈련기관명 : CTDS (Center on Technology, Data and Society).
애리조나 주립대학(ASU) 행정대학원 부설

3. 훈련분야 : 문화콘텐츠

4. 훈련기간 : 2024. 5. 20. ~ 2025. 5. 19

훈련기관 개요

기관명	Center on Technology, Data and Society (CTDS) 미국 애리조나 주립대학(ASU), 행정대학원 부설 연구소
소재지	Center on Technology, Data and Society, 411 North Central Ave., Suite 450, Phoenix, AZ 85004
홈페이지	http://techdatasociety.asu.edu
이메일	CTDS@asu.edu
설립목적	ㅇ 2019년, 정보기술의 사용과 접근, 분석을 통해 공공 문제 해결에 기여하기 위하여 애리조나 주립대(ASU; Arizona State University) 행정대학원(SPA; School of Public Affairs) 내에 설립 ㅇ 정보기술이 거버넌스, 공공정책, 사회에 미치는 영향과 디지털 불평등, 인공지능(AI), 사물인터넷(IoT), 전자정부 등 현안이 주요 연구대상
조직	ㅇ 애리조나 주립대학(ASU)을 비롯하여 다양한 대학, 연구소 등의 펠로우 연구진으로 구성 (센터장, 펠로우 교수, 애리조나 주립대(ASU) 소속 연구진, 타 대학 연계 연구진, 대학원생 등) · Anthony Howell, Associate Professor & Director, Center on Technology, Data and Society · Spiro Maroulis, Associate Professor & Interim Director, 행정대학원(SPA) · Yushim Kim, Associate Professor, 행정대학원(SPA)
연구분야	ㅇ 주요 연구 분야 · Digital Citizens and Communities · Digital Governance · AI, IoT and Smart Society · Data and Modeling for Public Good

훈련결과보고서 요약서

<table>
<tr><td>성 명</td><td>신수정</td><td>직 급</td><td colspan="2">행정주사</td></tr>
<tr><td>훈련국</td><td>미국</td><td>훈련기간</td><td colspan="2">2024.5.20.~2025.5.19</td></tr>
<tr><td>훈련기관</td><td>Center on Technology, Data and Society (CTDS)
미국 애리조나 주립대학(ASU), 행정대학원 부설 연구소</td><td colspan="2">보고서 매수</td><td>157매</td></tr>
<tr><td>훈련과제</td><td colspan="4">문화콘텐츠 분야 인공지능(AI) 윤리지침 연구</td></tr>
<tr><td>보고서 제목</td><td colspan="4">문화콘텐츠 분야 인공지능(AI) 윤리지침 연구</td></tr>
<tr><td>내용요약</td><td colspan="4">아래 참조</td></tr>
</table>

요약서

문화콘텐츠 분야 인공지능(AI 윤리지침 연구

ㅇ 최근 인공지능(AI)의 급속한 발전과 확산은 전 세계 경제 전반에 걸쳐 중대한 영향을 미치고 있다. PwC(PricewaterhouseCoopers) 연구에 따르면 2030년 AI는 글로벌 GDP를 약 14% 증가시키며, 약 15.7조 달러의 추가적인 경제적 가치를 창출할 것으로 전망된다. 이러한 변화는 문화콘텐츠 산업에서도 콘텐츠 제작, 배포, 소비 전 과정에서 근본적인 변화를 가져오고 있다. 특히, 생성형 AI는 영화·음악·게임·광고 등의 콘텐츠 제작에 직접 활용되면서, 산업 전반의 운영 방식에 전환을 가져오고 있고, 이에 따라 생성형 AI가 엔터테인먼트 산업의 운영 이익률을 14.4% 포인트까지 향상시킬 것으로 전망되고 있다.

ㅇ 이처럼 AI 활용이 확대되면서, AI가 생성한 콘텐츠의 저작권 문제, AI가 만들어내는 편향된 정보, 알고리즘의 불투명성 등 여러 가지 윤리적 문제도 동반되고 있다. 이로 인해 AI 윤리에 대한 체계적인 논의와 정책적 대응이 함께 요구되고 있다.

ㅇ 이러한 상황에서 각국은 자국의 정책 환경과 이해관계에 기반해 서로 다른 규제 접근 방식을 채택하고 있고, 각국 정부와 AI 기업은 시장 진출 대상 국가별로 서로 다른 규제 환경에 균형을 맞춰 정책 전략을 조정해야 하는 과제에 직면하고 있다.

ㅇ 이에 본 연구는 세계 최초의 포괄적 AI 규제법인 EU 인공지능법(AI Act)의 체계 및 주요 쟁점을 검토하고, 글로벌 AI 산업을 주도하고 있는 미국의 AI 윤리지침 및 행정명령의 원칙과 체계, 주요 쟁점을 분석하여 정책적 함의를 도출하고자 하였다.

이를 바탕으로 미국과 EU의 AI 윤리지침 및 규제 체계를 비교·분석하였고, 국제 표준 정비를 위한 양측의 협력 동향을 검토하여, 그 의미를 살펴보았다. 또한 세계 AI 규제 표준을 선도해 나가고 있는 미국과 EU의 AI 윤리지침 및 규제에 대한 주요 글로벌 AI 기업들의 대응 현황과 전략을 검토하고, 이를 통해 우리 정부와 기업의 향후 정책 방향의 시사점을 도출하는 것을 주요 목표로 하였다.

이를 위해, 연구 방법은 문헌 연구, 법령 및 정책 분석, 사례 분석 등의 방법론을 활용하였고, AI 윤리 지침 및 규제 분석은 문화 콘텐츠 분야에서 가장 광범위하게 활용되고 있는 범용 AI 및 생성형 AI 관련 항목을 중심으로 검토하였다.

ㅇ 연구 분석 및 검토 결과, EU는 위험 기반 접근법을 중심으로 AI 시스템의 위험 수준을 4단계로 구분하여 위험 수준별 규제를 적용하고 있으며, 강력한 법적 의무와 과징금을 통해 AI 기술이 초래할 수 있는 부정적 영향을 통제하고자 하는 정책 방향을 채택하고 있다.

특히 국제 무역 규범과의 충돌 가능성 등 여러 논란에도 불구하고, AI 시스템이 생성하는 결과물이 EU 내에서 사용되기만 해도 법의 적용을 받도록 하는 역외적용 규정, 제3국 범용 AI 모델 공급자의 제품 출시 전 EU 내 공인 대리인 지정 의무 규정, 소스 코드를 비롯한

기술적 수단을 통해 범용 AI 모델 내부 접근 허용 요구, AI 시스템의 투명성 요건 강화 등을 통해 EU의 AI 규제를 글로벌 표준으로 자리 잡도록 유도하고 있다. 이는 법 규제를 통한 기술적·윤리적 통제를 수단으로, EU의 인공지능(AI) 산업 경쟁력을 강화하고, 글로벌 AI 거버넌스에서 주도권을 확보하고자 하는 전략적 접근으로 해석된다.

이에 반해, 미국은 AI 규제에 있어서, 포괄적 규제 방식을 채택하기보다 특정 영역별 규제 접근 방식(Sectorally Specific Approach)을 취하고 있다. 즉, 기존 산업별 법률과 표준을 활용하면서, 윤리지침과 행정명령을 추가 적용하는 방식으로 규율을 적용하고 있다. 이는 세계 AI 기술을 선도하고 있는 미국 빅테크 기업들의 기술혁신을 촉진하면서, 동시에 필요한 경우 정부가 개입할 수 있는 정책적 유연성을 확보하고자 하는 전략으로 해석된다.

이러한 미국과 EU의 AI 윤리지침 및 규제에 대한 글로벌 AI 기업들의 대응 현황 및 전략 분석을 위해 본 연구에서는 스탠포드 대학 파운데이션 모델 투명성 지수(The Foundation Model Transparency Index(FMTI)) 보고서와 주요 AI 모델의 시스템 카드 및 모델카드를 분석하였다.

검토 결과, 글로벌 AI 기업들은 국가별 규제 요건에 맞추어 시장 경쟁력을 유지할 수 있는 수준으로 기술문서를 공개하고, 법적 규제 준수를 위한 기술적 조치와 자율적 내부 정책 강화를 병행하며 전략적으로 대응하고 있는 것으로 나타났다.

ㅇ 나아가, 현재 글로벌 AI 규제 환경은 단순히 AI 기술 규제의 문제가 아니라, 각국의 데이터 보호 및 기술 산업 보호 전략과 밀접하게 연관되어 있다. 국가 안보를 근거로 데이터 보호 및 타국 기술 기업 규제를 강화하고자 하는 미국의 입장을 보여주는 틱톡(TikTok) 강제매각법 사례, 강력한 AI 및 데이터 규제로 기업 활동을 제한하고 있는 EU의 애플 과징금 사례(경쟁법, 디지털시장법), 데이터를 국가 전략 자산

으로 간주하여 데이터 해외 반출을 금지하고 있는 중국 정부 규제에 따라 중국 현지 데이터 저장 정책을 적용한 테슬라의 사례는 이를 직접적으로 보여주고 있다.

ㅇ 이 같은 글로벌 AI 및 관련 데이터 규제 환경 변화 흐름에 따라 우리 정부도 뉴욕구상 ('22.9.21.), 파리 이니셔티브 ('23.6.21.), AI 서울 정상회의 ('24.5.21.) 등을 통해 글로벌 AI 거버넌스에 적극 참여하고 있다. 또한, 과학기술정보통신부의 디지털 권리장전 발표('23.9.25.), AI 기본법 제정('25.1.21.) 등 체계적인 법제화 노력도 병행하고 있다.

ㅇ AI 및 데이터 규제는 기술 발전, 시장 경쟁, 국가 안보 등 다양한 요소가 복합적으로 얽혀 있어, 단기적인 규제 대응을 넘어 장기적인 전략 수립과 지속적인 국제 협력 강화가 필요하다.

따라서, 앞으로 기업은 글로벌 규제 동향을 모니터링하면서 법적 리스크를 최소화하는 전략을 마련해야 할 것이고, 정부는 국제 협상에 적극적으로 참여하여 글로벌 AI 거버넌스 체계 구축 과정에 확고한 기반을 다질 수 있도록 정책을 추진해야 할 것이다.

기업과 정부가 함께 이러한 노력을 이어간다면, AI 기술혁신과 규제 환경 변화 속에서 경쟁력을 유지하며, 글로벌 AI 시장에서 주도적인 역할을 수행할 수 있을 것이다.

Ⅰ. 연구 개요 및 방법론

1. 연구 배경

(1) AI 확산과 문화콘텐츠 산업의 변화 및 전망

○ 인공지능(AI)의 급속한 발전과 확산은 전 세계적으로 경제와 산업 구조 전반에 걸쳐 중대한 변화를 불러오고 있다. PwC(PricewaterhouseCoopers) 연구[1]에 따르면, 2030년 AI는 글로벌 GDP를 14%까지 증가시키며, 약 15.7조 달러의 경제적 가치를 추가 창출할 것으로 전망된다. 이는 AI가 가져올 수 있는 막대한 경제적 잠재력을 보여주는 수치로, 다양한 산업에서 AI의 영향력이 지속적으로 확대될 것임을 시사한다.

(출처: PwC (2017), *Sizing the Prize* | 이미지 제작 도구: Canva)

○ 이러한 변화 속에 문화콘텐츠 산업에서도 AI는 콘텐츠 제작, 배포, 소비 방식에 근본적인 변화를 가져오고 있다. 이제 AI 기술은 단순한 데이터 분석 및 추천 알고리즘 도구를 넘어 콘텐츠 기획과 창작, 소비자 맞춤형 추천, 마케팅 자동화 등 다양한 영역에서 핵심 기술로 자리잡고 있다. 특히, 생성형 AI가 영화·음악·게임·광고 등의 콘텐츠 제작에 직접 활용되면서, 산업 전반의 운영 방식에 전환을 가져오고 있다.

PwC는 "strategy_business"(2024.6.27.)에서, '23년 미디어 엔터테인먼트 산업의 생성형 AI 시장 규모를 14억 달러로 추정하고, '33년까지 연평균 복합성장률(CAGR) 26.3%의 높은 성장률을 기록할 것이라는 Market.us의 전망을 인용하면서, 생성형 AI가 엔터테인먼트 산업의 운영 이익률을 14.4% 포인트 향상시킬 것으로 분석하였다[2].

---

1) PwC.(2017). Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise? Retrieved from https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf.
2) Greenstein, B., Kosar, J., Light, C., & Shelton Rose, M. (2024, June 27). Productivity or pioneering? Your industry's GenAI adoption play. strategy+business. Retrieved from https://www.pwc.com/gx/en/issues/technology/genai-adoption-game-changer.html.

또한, AI 영향 지표(AI Impact Index)와 AI 적응 성숙도(Adoption Maturity) 개념을 제시하여, 기술·통신·엔터테인먼트를 밀접히 연관되는 하나의 카테고리로 묶고, 이 분야의 AI 시장 도입 영향력과 도입 속도를 평가[3]하였다.

이에 따르면, 기술(Technology)·통신(Communication)·엔터테인먼트(Entertainment) 산업의 AI 도입 영향력은 3.1점(5점 만점)으로 헬스케어(3.7), 자동차(3.7), 금융(3.3), 운송 및 물류(3.2) 산업 보다는 낮은 수준이었으나, AI 기술 도입 속도를 평가하는 적응 성숙도의 단기(0-3년) 평가는 이들보다 높은 47%로 나타났다.(헬스케어(37%), 자동차(36%), 금융(41%), 운송 및 물류(41%))

이는 AI의 시장 도입 영향력은 중간 수준이지만, 도입 속도는 빠르다는 것을 의미한다. 특히 추천 알고리즘, 콘텐츠 제작, 맞춤형 광고 등의 분야에서 빠르게 정착될 가능성이 높으며, 엔터테인먼트 산업 전반에 걸쳐 AI 활용도가 급격히 증가할 것으로 예상된다.

(출처: PwC (2017), *Sizing the Prize*)

ㅇ PwC는 기술·통신·엔터테인먼트 분야에서 인공지능(AI)의 잠재력이 가장 큰 영역을 아래 3가지로 분석하였다.

- **(미디어 아카이빙 및 검색)** 분산된 콘텐츠를 모아 추천하는 기능
- **(맞춤형 콘텐츠 제작)** 마케팅·영화·음악 등 다양한 분야에서 맞춤형 콘텐츠 생성
- **(개인화된 마케팅 및 광고)** 소비자 맞춤형 마케팅 전략과 광고 제공

3) PwC.(2017). Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise? Retrieved from https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf.

AI 기술 발전으로 소비자는 더욱 개인화된 콘텐츠를 빠르고 쉽게 접할 수 있게 되고, 사용자의 선호도에 기반한 고품질 콘텐츠의 생성·추천·제공 서비스가 확대될 것으로 예상된다. 이에 따라 양질의 서비스 수요는 지속 증가할 것이며, 방대한 데이터 속에서 유의미한 정보를 효과적으로 추출하고 활용하는 것은 시장 경쟁력 확보에 중요 과제가 될 것이다.

PwC는 엔터테인먼트 분야에서 이미 개인화된 콘텐츠 추천 서비스가 있지만, 기존 데이터에 콘텐츠가 계속해서 새로 생성되면서 엄청난 양이 축적될 것이고, 이를 태깅, 추천, 수익화하는 체계에 난관이 생길 것임을 지적하면서, AI 기술은 이같이 점차 거대해져 가는 데이터를 효율적으로 분류하고 보관할 수 있는 기능을 제공할 수 있고, 보다 정확한 타겟팅과 수익 증대를 가능하게 할 것이므로, 더 빠르고 깊숙히 시장에 확산될 것으로 전망했다.

이처럼, 문화콘텐츠 산업에서 AI는 콘텐츠 생성과 소비 방식 전반을 변화시키는 핵심 기술로 자리 잡고 있다. 특히, AI가 콘텐츠 창작·추천·배포·소비 과정 전반에 직접 활용되면서, 생성형 AI 및 범용 AI가 문화콘텐츠 분야 AI 논의의 중심이 되고 있다.

생성형 AI 및 범용 AI 두 개념은 기술적으로 유사한 기능을 수행하지만, 각국의 규제 및 정책 맥락에서 다소 다르게 사용되고 있는 명칭이다. EU는 인공지능법(AI Act)에서 "범용 AI(General-Purpose AI)"라는 용어를 사용하여, 다양한 목적에 활용될 수 있는 AI 모델과 시스템을 규제 대상에 포함하고 있다. 반면, 미국은 AI 윤리지침과 행정명령 등에서 "생성형 AI(Generative AI)"라는 용어를 사용하여, 콘텐츠 생성 기능이 있는 AI 기술을 강조한다.

즉, 범용 AI는 여러 용도로 활용 가능한 AI를 포괄하는 개념이며, 생성형 AI는 텍스트·이미지·영상 등 결과물 생성 기능이 특화된 AI를 지칭하는 용어로 사용된다. 문화콘텐츠 분야에서는 두 개념의 AI가 모두 핵심적 역할을 하고 있다. 따라서, 본 연구에서는 이 둘을 중심으로 논의한다.

(2) 신뢰할 수 있는 AI 윤리 동향

ㅇ 앞서 살펴본 바와 같이, AI 기술은 콘텐츠 제작의 자동화 뿐 아니라, 디지털 아트, 음악, 영화 등 미디어 제작·배포·소비 과정 전반에서 광범위하게 활용되고 있다. ChatGPT, DALL·E, Midjourney와 같은 생성형 AI는 대규모 데이터를 학습하여, 텍스트 기반 스토리텔링, 이미지 및 영상 제작, 음악 작곡 등 다양한 분야에서 빠르게 확산되고 있다.

그러나, AI 활용이 확대됨에 따라, AI가 생성한 콘텐츠의 저작권 문제, AI가 만들어내는 편향된 정보, 알고리즘의 불투명성 등 여러 가지 윤리적 문제도 야기되고 있다. AI 기술이 콘텐츠의 생산·유통·소비의 효율성을 높이고 창작의 영역을 확장하는 긍정적 효과를 가져오고 있지만, 이와 함께 여러 쟁점도 야기하고 있는 바, 이에 대한 체계적인 논의와 정책적 대응이 필요해지고 있다.

ㅇ 이처럼 AI 기술은 급속히 발전하며 다양한 분야에 영향을 미치고 있지만, 여전히 불확실성이 큰 기술이다. 이에 따라, 국제 사회도 신뢰할 수 있는 AI 윤리 기준을 마련하고, 기술과 인권 보호를 조화롭게 추진하기 위해 노력하고 있다.

OECD (경제협력개발기구) 4)	• **'19.5월 AI 원칙(OECD AI Principles) 채택** - (목표) AI발전과 인권 보호의 조화 - (내용) 신뢰할 수 있는 AI 개발, 인간의 권리 및 민주적 가치 존중 - AI 행위자들이 신뢰할 수 있는 AI를 개발하고, 정책 입안자들에게 효과적인 AI 정책에 대한 권고를 제공하기 위한 지침으로 설계 • **'24.5월 OECD 각료 이사회 AI 원칙 개정안 채택**
유네스코 (UNESCO) 5)	• **'21.11월 AI 윤리 권고(Recommendation on the Ethics of Artificial Intelligence) 채택** - (목표) 인권, 투명성, 공정성을 강조하며 AI 기술이 인간의 존엄성과 사회적 가치를 보호하고 증진하는 방향으로 발전 - (내용) 인권 증진, 데이터 거버넌스, 투명성, 책임성 등을 포함한 다양한 정책적 권고를 제시 - 최초의 국제 AI 윤리지침으로 회원국의 AI 윤리정책 수립에 중요한 기준점 역할

4) OECD. (2024, May 3). OECD updates AI principles to stay abreast of rapid technological developments [Press release]. Retrieved from https://www.oecd.org/en/about/news/press-releases/2024/05/oecd-updates-ai-principles-to-

ㅇ 이러한 국제적 협력 노력과 함께, 각 국가들도 AI 윤리 문제 대응을 위한 정책을 추진하고 있다. 이러한 개별 국가의 움직임은 각국의 서로 다른 입장을 반영하고 있어, 접근 방식에도 차이를 보이고 있다.

유럽연합 (EU)	• 세계 최초로 포괄적 인공지능법(AI Act)을 제정, AI 시스템을 위험도 기반으로 분류하고, 규제를 강화하는 방향으로 정책 추진
미국 (US)	• 다양한 AI 윤리지침과 정부 행정명령을 통해 자율적 규율과 규제의 균형을 조율하는 방향으로 정책 추진

ㅇ AI 윤리를 논할 때, "Ethics(윤리)"의 개념은 단순한 도덕적 개념이 아니라 사회적 맥락을 고려한 바람직한 행동을 의미한다, 따라서, AI 윤리는 각국의 사회적 상황과 가치 체계에 따라 다르게 적용될 수 있다.

예를 들어, 딥러닝 알고리즘의 불투명성 문제로 인해 AI의 설명 가능성(Explainability)과 투명성(Transparency)이 강조되고 있지만, 투명성을 최우선으로 강조할 경우, AI 시스템의 효율성이 저하될 수 있는 문제가 발생할 수 있다. 이 때문에, AI의 활용 영역과 문제에 따라 우선하는 윤리적 가치가 달라질 수 있다.

즉, 사회적·문화적·경제적 배경과 상황의 차이에 따라, 국가별로 AI 윤리를 대하는 입장이 상충될 수 있는 것이다. 이는 AI 기술의 적용 방식과 규제 접근 방식에도 영향을 미친다. 결국, AI 윤리 논의는 다양한 가치의 충돌을 조정하고, 해결해 나가는 과정이며, 이를 통해 국제 사회가 합의된 원칙을 도출하고 제도화해 나가야 한다.

이에 따라, AI 기술 발전과 글로벌 규제 정립이 동시에 이루어지고 현 시점에는 각국의 대응 현황을 면밀히 검토하여 우리 정부의 대응 방향을

---

stay-abreast-of-rapid-technological-developments.html; OECD. (n.d.). AI principles. Retrieved from https://www.oecd.org/en/topics/sub-issues/ai-principles.html

5) UNESCO. (n.d.). Forum on the ethics of AI. Retrieved from https://www.unesco.org/en/forum-ethics-ai; YTN. (2021, November 26). 유네스코, 세계 첫 'AI 윤리적 사용 지침' 마련. Retrieved from https://www.ytn.co.kr/_ln/0104_202111261132522639

정립하고, 국제 논의에 적극적으로 참여하여 AI 윤리 기준을 조율해 나가는 정책적 노력이 요구되고 있다.

2. 연구 목적

ㅇ AI 기술의 급속한 발전으로 산업 전반이 빠르게 진화하고 있으며, 이에 따라 시의성 있는 AI 윤리 기준 정립이 요구되고 있다. 이러한 상황에서 각국은 자국의 정책 환경과 이해관계에 기반해 서로 다른 규제 접근 방식을 채택하고 있고, 이에 따라 각국 정부와 AI 기업은 시장 진출 대상 국가별로 서로 다른 규제 환경에 균형을 맞춰 정책 전략을 조정해야 하는 과제에 직면해 있다.

ㅇ 이에 본 연구는 문화 콘텐츠 분야에서 가장 광범위하게 활용되고 있는 범용 AI 및 생성형 AI를 중심으로, 세계 최초의 포괄적 AI 규제법인 EU 인공지능법의 체계 및 주요 쟁점을 검토하고, 글로벌 AI 산업을 주도하고 있는 미국의 AI 윤리지침 및 행정명령의 주요 원칙과 체계, 주요 쟁점을 분석하여 정책적 함의를 도출한다.

이를 바탕으로 미국과 EU의 AI 규제 체계를 비교·분석하고, 국제 표준 정비를 위한 양측의 협력 동향을 검토하여 그 의미를 파악하고자 한다. 그리고 이러한 AI 윤리지침과 규제에 대한 주요 글로벌 AI 기업들의 대응 현황과 전략을 검토하여, 그 시사점을 도출하는 것을 주요 목표로 한다.

ㅇ 끝으로, AI 규제가 단순히 기술 규제 만이 아니라 각국의 데이터 주권 강화와 국제적 영향력 확대 전략과 연계됨을 고려하여, 미국 틱톡(TikTok) 강제매각법 사례, EU 애플 과징금 사례, 중국 테슬라의 중국 내 데이터 저장 정책 사례 등 최신 사례를 분석하여, 우리 정부와 기업이 향후 취해야 할 전략적 방향을 간략히 제언하고자 한다.

3. 연구 방법론

ㅇ 본 연구는 EU 인공지능법(AI Act) 및 미국의 AI 윤리지침 및 행정명령을 중심으로 법령·정책 분석을 수행하고, 글로벌 AI 기업들의 대응 현황 및

전략을 조사하여 AI 규제 환경에 대한 종합적 시사점을 도출하고자 한다. 이를 위해 다음과 같은 연구 방법을 활용하였다.

❶ AI 규제 법·정책 분석

❶-1. EU AI 규제 분석

분석 대상	• EU 인공 지능법(AI Act)
연구 방법	• 법령 분석(Legal Analysis), 문헌 연구(Literature Review)
분석 내용	• (법 구조 분석) 범용 AI 관련 내용을 중심으로 법 체계 및 근거 검토 • (주요 쟁점 분석) 법 조항별 주요 쟁점 검토

❶-2. 미국 AI 윤리지침·규제 분석

분석 대상	• AI 권리장전 • NIST AI 리스크 관리 프레임워크(NIST AI RMF) • NIST 생성형 AI 프로파일(NIST GAI 프로파일) • 행정명령 제14110호
연구 방법	• 문서 분석(Document Analysis), 내용 분석(Content Analysis), 정책 분석(Policy Analysis), 문헌 연구(Literature Review)
분석 내용	• (AI 권리장전 분석) 전체 체계·주요 내용·각 원칙의 의미·실질적 적용 및 기존 법률과 연계성 검토 • (NIST AI RMF 분석) 프레임워크의 구조·핵심 개념·주요 리스크 요소·구체적 대응 지침 분석 • (NIST GAI 프로파일 분석) 프로파일의 전체 체계·목적·주요 리스크 항목·대응 방안 검토 • (행정명령 제14110호 분석) 행정명령 전체 구조·문화 콘텐츠 분야 관련 주요 조항 세부 내용 분석 및 주요 쟁점 검토

❶-3. 미국-EU AI 규제 비교 연구

분석 대상	• 미국과 EU의 AI 규제 체계 • 미국과 EU의 국제 표준 정비 협력 동향
연구 방법	• 문서 분석(Document Analysis), 내용 분석(Content Analysis), 문헌 연구(Literature Review)
분석 내용	• (규제 체계 비교·분석) 미국과 EU의 AI 규제 체계 주요 개념 및 원칙의 공통점과 차이점 도출 • (국제 표준 정비 동향 검토) 미국과 EU의 최근 국제 표준 정비 협력 동향 분석 및 향후 전망 제시

❷ 글로벌 기업의 규제 대응 및 전략 분석

❷-1. 주요 AI 기업의 투명성 및 규제 준수 현황 분석

분석 대상	• 스탠포드 대학 파운데이션 모델 투명성 지수(The Foundation Model Transparency Index(FMTI)) 보고서
연구 방법	• 문서 분석(Document Analysis)
분석 내용	• (투명성 지수 평가 분석) 주요 파운데이션 모델의 투명성 수준 및 규제 준수 현황 분석

❷-2. 개별 AI 기업의 규제 대응 전략 검토

분석 대상	• AI 모델 시스템 카드(System Cards), 모델 카드(Model Cards)
연구 방법	• 문서 분석(Document Analysis)
분석 내용	• (규제 대응 전략 검토) 공개된 모델 정보를 분석하여, 미국과 EU의 규제 대응 항목과 과제 분석, 기업별 대응 전략 검토

❸ 정부 및 기업 정책 제언을 위한 최신 동향 분석

분석 대상	• 미국·EU·중국의 AI 관련 기타 규제 최신 사례 - 미국 틱톡(TikTok) 강제매각법 사례 - EU 애플 경쟁법 및 디지털시장법 과징금 사례, - 테슬라의 중국 내 데이터 저장 정책 사례
연구 방법	• 사례 연구(Case Study), 문서 분석(Document Analysis), 내용 분석(Content Analysis), 정책 분석(Policy Analysis)
분석 내용	• (사례별 규제 환경 분석) AI 관련 데이터 규제 및 국가별 정책 검토 • (정책적 시사점 도출) 사례별 기업 대응 분석 및 정책 방향 제언

Ⅱ. EU AI Act와 미국 AI 윤리지침

1. 유럽연합 인공지능법(EU AI Act)

본 절(Section)에서는 EU 인공지능법(AI Act)의 법적 구조와 근거를 검토하기 위해 법령 분석(Legal Analysis)을 적용하였으며, 문헌 연구(Literature Review)를 통해 관련 쟁점을 분석하였다.

(1) 도입 배경

ㅇ EU 인공지능법(AI Act)을 구체적으로 검토하기에 앞서, EU 집행위원회(EC)가 설명하는 '인공지능 규제 프레임워크 제안'[6]을 통해 법 도입 배경과 체계를 개괄적으로 살펴보고자 한다.

유럽연합 집행위원회는 EU 인공지능법(AI Act)을 인공지능과 관련한 세계 최초의 종합적인 법적 틀로 지칭하고, ①기본권·안전·윤리원칙 존중, ②강력하고 영향력 있는 AI 모델에 대한 조치, ③신뢰할 수 있는 AI 촉진 3가지를 법안의 기본 목적으로 설명하고 있다.

ㅇ 또한, 기존 법률이 일부 보호 기능을 하고는 있지만, AI 시스템이 야기할 수 있는 특정 문제를 다루기에는 부족하다고 보고, 법안의 구체적 틀을 아래와 같이 안내하고 있다.

- AI 응용프로그램으로 생성되는 위험에 대한 조치
- 수용할 수 없는 위험을 초래하는 AI 관행 금지
- 고위험 응용프로그램 목록 결정
- 고위험 응용프로그램 AI 시스템 요건 정립
- 고위험 AI 응용프로그램 배포자 및 공급자의 구체적 의무 정의
- 특정 AI 시스템의 시장 출시 혹은 서비스 제공 전 적합성 평가 요구
- 특정 AI 시스템 시장 출시 이후 시행조치 마련
- 유럽연합 및 국가 차원에서 거버넌스 구조 확립

6) 유럽연합 집행위원회(European Commission), "Regulatory framework proposal on artificial intelligence“, "https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai"

ㅇ 각 AI 시스템의 잠재적 위험에 기반하여, 4가지 수준으로 AI 시스템 위험 수준을 구분하고 있다.

[출처: 유럽연합 집행위원회 누리집 "Regulatory framework proposal on artificial intelligence"]

- 인권·생계·안전에 명백한 위협이 되는 모든 AI 시스템은 "수용할 수 없는 위험"으로 금지되며, 여기에는 정부에 의한 사회적 점수 부과부터 위험 행동을 조장하는 음성 보조 기능이 있는 장난감까지 모두 포함되는 것으로 보고 있다.

- "고위험" AI 시스템에는 아래 항목들이 포함된다(본 보고서는 문화콘텐츠 분야 AI 윤리지침 연구인 바, 법안 세부 검토에서는 범용 AI 모델을 중심으로 분석한다.)

- 시민의 생명과 건강을 위협할 수 있는 중요 기반 시설 (예: 교통)
- 타인 인생에서 교육과 직업적 경로의 기회를 결정할 수 있는 교육 또는 직업 훈련 (예: 시험 점수화)
- 제품의 안전 구성 요소(예: 로봇 보조수술에서의 AI 응용프로그램)
- 채용, 근로자 관리 및 자영업 창업 준비(예: 채용 절차를 위한 이력서 분류 소프트웨어)
- 민간 및 공공 필수 서비스 (예: 대출 기회를 거부할 수 있는 신용 점수화)
- 기본권을 침해할 수 있는 법 집행 (예: 증거의 신뢰성 평가)
- 이주·난민·국경 통제 관리 (예: 비자 신청 자동 심사)
- 사법 및 민주적 절차의 관리 (예: 법원 판결 검색을 위한 AI 솔루션)

이 고위험 AI 시스템에 해당하는 경우, 시장에 출시 되기 전에 아래와 같은 의무사항들을 엄격히 따르도록 하고 있다.

- 적절한 위험 평가 및 완화 시스템
- 리스크 및 차별적 결과를 최소화하기 위해 고품질의 데이터셋을 시스템에 제공
- 결과물에 대해 추적 가능하도록 활동 로그 기록
- 당국이 규정 준수 여부를 평가할 수 있도록 시스템에 대한 모든 필요 정보를 제공하는 상세한 세부 문서
- 배포자에게 명확하고 적절한 정보 제공
- 리스크 최소화를 위해 인간이 감독하는 적절한 조치
- 높은 수준의 견고성, 보안 및 정확성

또한, 모든 원격 생체 신원확인 시스템은 고위험으로 간주되고, 엄격한 요건을 준수해야 한다. 법 집행 목적이라고 하더라도 공공장소에서는 사용이 원칙적으로 금지되고, 실종 아동 수색, 특정된 즉각적인 테러 위협, 또는 중범죄 가해자나 용의자를 탐색·추적·신원확인 또는 기소하려는 목적 등으로 예외 사항을 엄격히 정의하여 규제한다.

- "제한된 위험"은 AI 사용에 있어서의 투명성 문제와 관련된 위험을 말하며, 신뢰 촉진을 위해 필요한 경우, 사람들이 AI와 상호작용 정보를 제공받을 수 있도록 하는 특정한 투명성 의무를 도입하였다.

예를 들어, 챗봇과 같은 AI 시스템을 사용하는 경우, 사람들이 기계와 상호작용하고 있다는 것을 인식할 수 있도록 하여, AI에게 계속 정보를 제공받을지 여부를 결정할 수 있도록 해야 한다.

공급자들은 AI 생성 콘텐츠임을 기술적으로 식별가능하도록 해야 하며, 이외에도, 공익적 목적의 정보 제공을 위해 대중에게 게시되는 AI 생성 텍스트는 인공적으로 생성된 것임을 표시해야 하고, 이는 딥페이크를 구성하는 오디오·비디오에도 적용된다.

- 최소 또는 무위험 단계 AI는 자유로운 사용이 허용되며, 여기에는 AI 기반 비디오 게임이나 스팸 필터와 같은 응용프로그램이 포함된다. 현재 EU내에서 사용되는 대다수 AI 시스템이 이에 해당하는 것으로 보고 있다.

ㅇ 유럽연합 집행위원회는 범용 AI 모델이 점점 더 고성능의 강력한 AI 솔루션을 가능하게 하고 있고, 이러한 기능을 모두 감독하는 것이 상당히 어렵기 때문에, 인공지능법(AI Act)을 통해 모든 범용 AI 모델에 투명성 의무를 도입한 것이라고 설명한다.

이를 통해 해당 모델들을 더 잘 이해하고, 강력하고 영향력 있는 모델에 대한 추가적인 위험 관리 의무화가 더 잘 될 것으로 보고 있다. 이 추가적인 의무에는 시스템적 리스크에 대한 자가 평가 및 완화 조치, 중대한 사고의 보고, 테스트·모델 평가 시행, 사이버 보안 요건 등이 포함된다.

하지만 이러한 투명성 요건이 기업에게 기밀정보 노출과 같은 부담으로 작용할 수 있다는 우려도 제기된다. 예를 들어 민감한 알고리즘 구조나 데이터 세트가 잘못 공개될 경우, 기업에는 경쟁력 손실뿐 아니라 보안 위협까지 발생할 수 있다. 따라서, AI 규제의 실효성과 기업 혁신 간의 균형을 어떻게 유지할 것인지에 대한 지속적인 논의가 필요하다.

ㅇ ’24.7.12. 유럽연합(EU) 관보 게재[7]로 공포된 EU 인공지능법(AI Act)의 시행일은 아래와 같다. 유럽연합은 본 법 공포 후, 기업 및 관련기관의 준비 기간을 고려하여 조항별 시행 시기를 상이하게 적용하고 있다.

일정	내용
‘23.12월	▪ 유럽 의회 및 유럽 이사회 AI 법안에 대한 정치적 합의
‘24.5.21.	▪ 유럽연합 이사회 EU AI Act 공식 채택
‘24.7.12.	▪ 공포(유럽연합 관보 게재)

7) 유럽연합 관보, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)" (2024.7.12.), "https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ",

'24.8.1.	▪ 발효(제113조)	
'26.8.2.	▪ 시행(단, 제113조에 따라 아래 조항은 시행일 적용 상이)	
***조항별 시행일**	**'25.2.2.**	제1장(총칙) 및 제2장(금지된 AI 관행) 시행
	'25.8.2.	제3장 제4절(통지당국 및 통지기관), 제5장(범용 AI 모델), 제7장(거버넌스), 제12장(벌칙), 제78조(기밀유지) 시행 (*단 제12장 제101조(범용 AI 공급자 벌금) 제외)
	'27.8.2.	제6조제1항 및 본 법률에 따른 해당 의무

[출처: 유럽연합 집행위원회(European Commission) 누리집, "Regulatory framework proposal on artificial intelligence"; 유럽연합 관보, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)" (2024.7.12.)]

이는 본 법이 상당한 준비 기간과 세부적인 기술적·윤리적 대응을 요구하고 있음을 시사한다. 법 시행 시기에 맞춰 대응하지 못할 경우, 높은 수준의 과징금 등 심각한 제재를 받을 가능성이 크다는 점에서, 시장 적응력 강화를 위한 선제적 대응이 필요하다.

특히, 시장 출시 전 적합성 평가와 투명성 요건 준수는 필수적이므로, 본 법이 글로벌 기업들의 기술적·법적 대응 역량의 하나의 기준이 될 것으로 예상된다.

(2) 법적 구조 및 주요 쟁점

ㅇ '24.7.12. 공포된 유럽연합 인공지능법(AI Act)의 구조는 아래와 같이 전체 13장으로 구성되어 있다.

구분	내용
제1장	총칙 (General Provisions)
제2장	금지된 AI 관행 (Prohibited AI Practice)
제3장	고위험 AI 시스템 (High-Risk AI Systems)
제4장	특정 AI 시스템 공급자 및 배포자의 투명성 의무 (Transparency Obligations for Providers and Deployers of Certain AI Systems)

구분	내용
제5장	범용 AI 모델 (General-Purpose AI Models)
제6장	혁신 지원 조치 (Measures in Support of Innovation)
제7장	거버넌스 (Governance)
제8장	고위험 AI 시스템의 EU 데이터베이스 (EU Database for High-Risk AI Systems)
제9장	사후시장 모니터링, 정보 공유, 시장 감독 (Post-Market Monitoring, Information Sharing, Market Surveillance)
제10장	행동강령 및 지침 (Codes of Conduct and Guidelines)
제11장	권한 위임 및 위원회 절차 (Delegation of Power and Committee Procedure)
제12장	벌칙 (Penalties)
제13장	최종조항 (Final Provisions)

ㅇ 위와 같이 유럽연합 인공지능법(AI Act)은 광범위한 규제 구조를 기반으로 다양한 AI 시스템을 포괄적으로 다루고 있다. 본 보고서는 이 같은 구조를 문화콘텐츠 분야에서 특히 그 역할이 점차 확대되고 있는 범용 AI(General Purpose AI)에 초점을 맞추어 아래와 같이 8개의 구조로 재구성하고, 이를 기반으로 세부 내용과 법적 쟁점을 심층 분석하고자 한다.

분석 구조	주요 내용
1. 목적·범위·정의	법의 목적과 범위, 주요 개념 정의를 중심으로 분석
2. 금지행위 및 과징금	범용 AI와 관련된 금지행위와와 위반 시 과징금 체계 분석
3. 투명성 의무 및 과징금	범용 AI와 관련된 투명성 의무와 규정 위반 시, 과징금 체계 분석
4. 범용 AI 모델 공급자의 의무 및 과징금	범용 AI 모델 공급자가 준수해야 할 법적 의무와 위반 시 부과되는 제재를 분석
5. 거버넌스	범용 AI 개발 및 배포를 감독하기 위한 거버넌스 구조 분석
6. 범용 AI 시스템의 시장 감시 등	AI 시스템의 시장 감시 및 사후 평가 체계 분석
7. 기밀 유지 및 구제	기업 기밀 보호와 규제 준수 간의 균형 및 구제 절차 분석
8. 혁신 지원 조치	규제 샌드박스를 통한 혁신과 규제 조화 방안 분석

1) 목적·범위·정의

(i) 목적 및 범위

ㅇ 유럽연합 인공지능법(EU AI Act)은 제1장 총칙 부분에 법의 목적, 범위, 정의를 규정하여, 법 조항 전체에 적용되는 개념적 틀을 제시하고 있다. 먼저, 목적을 살펴보면, 법의 취지는 앞서 유럽연합집행위원회의 설명에서도 본 바와 같이, 제1조에서 아래와 같이 규정되어 있다.

> **제1조 목적(Subject Matter)**
> 1. 본 법의 목적은 내부 시장의 기능을 개선하여, 인간 중심적이고 신뢰할 수 있는 AI의 활용을 촉진하면서, 유럽연합 내에 AI 시스템의 유해한 영향으로부터 민주주의, 법치주의, 환경보호를 포함해 헌장에서 보호하고 있는 기본권·안전·건강에 대한 높은 수준의 보호를 보장하고, 혁신을 지원하는 것임.
>
> 2. 본 규정은 다음을 규정한다.
> (a) 유럽연합 내에서 AI 시스템의 시장 출시, 서비스 제공 및 사용에 대한 조화된 규칙;
> (b) 특정 AI 관행의 금지;
> (c) 고위험 AI 시스템에 대한 구체적인 요구사항과 그러한 시스템 운영자의 의무;
> (d) 특정 AI 시스템에 대한 조화된 투명성 규칙;
> (e) 범용 AI 모델의 시장 출시를 위한 조화된 규칙;
> (f) 시장 모니터링, 시장 감시, 거버넌스 및 집행에 관한 규칙;
> (g) 특히 중소기업(SME) 및 스타트업에 중점을 둔 혁신 지원을 위한 조치.

따라서, 본 법의 본문에서 규정하고 있는 목적은 ①내부 시장 기능을 개선하여, **인간중심적이고 신뢰할 수 있는 AI 활용 촉진**, ②**AI 시스템의 유해한 환경으로부터 기본권·안전·건강 보호,** ③**혁신 지원** 3가지로 크게 볼 수 있다.

ㅇ 법에 적용을 받는 범위는 아래와 같이 제2조에서 규정하고 있는데, **EU 시장에 진출하려는 기업은 법인의 소재지와 상관없이 본 법에 적용**을 받도록 하고 있다.

> **제2조 범위(Scope)**
> 1. 적용범위
> **(a) 유럽연합 내에서 AI 시스템을 시장에 출시하거나 서비스를 제공하는 공급자 또는 범용 AI 모델을 시장에 출시하는 공급자 (공급자의 설립장소나 위치는 EU 내이든 제3국이든 관계없음);**

(b) EU내에 설립장소가 있거나 소재하고 있는 AI 시스템 배포자;
(c) 제3국에 설립된 장소가 있거나 위치한 AI 시스템의 공급자와 배포자로 해당 AI 시스템이 생성하는 결과물이 유럽연합 내에서 사용되는 경우;
(d) AI 시스템의 수입자 및 유통업자;
(e) 자사의 사명과 상표로 출시하는 제품과 함께 AI 시스템을 시장에 출시하거나 서비스를 제공하는 제품 제조업자
(f) 설립장소가 EU 내가 아닌 공급자의 공인 대리인;
(g) EU내에 소재한 영향을 받는 사람들

특히, 제2조제1항(c)목에서는 AI 시스템이 생성하는 결과물이 유럽연합 내에서 사용되기만 해도 본 법의 적용을 받도록 규정하고 있어, 적용 범위를 역외로까지 확대하고 있다.

이는 규제 준수 의무 범위를 불명확하게 하여, 법적 안정성을 저해하고, 기업에게 추가적인 법적 부담을 가중시킬 수 있다는 우려를 초래한다. 이러한 제2조제1항(c)목의 주요 쟁점을 관련 문헌을 토대로 아래와 같이 분석하고자 한다.

❑ **주요쟁점: 제2조제1항(c)목의 역외적용 문제[8)]**

본 조항은 법 적용의 역외성을 강조한 조항으로, 아래와 같이 AI 기업에 과도한 부담을 줄 수 있다는 쟁점이 제기되고 있다.

① 법적 책임의 확대 – 계약상의 제한 필요

- 제2조제1항(c)목은 EU 내에서 AI 시스템 결과물을 사용하는 모든 행위를 규제 대상으로 포함시키고 있다. 이 때문에 EU 시장에 직접 진출하지 않은 기업에도 법적 책임이 부과될 수 있다는 우려가 제기된다.

  예를 들어, EU와 무관하게 운영되고 있는 기업이라고 하더라도, 계약 상대방에게 공급한 AI 시스템을 통해 생성된 결과물이 EU 내에서 사용될 경우, 법적 책임을 질 수 있기 때문에, 기업의 글로벌 비즈니스 전략에 복잡성을 심화시키는 요인으로 작용할 수 있다.

8) DLA Piper. (2024, February). Extra-territorial Application of the AI Act: How Will It Impact Australian Organisations?, https://www.dlapiper.com/en/insights/publications/2024/02/extra-territorial-application-of-the-ai-act-how-will-it-impact-australian-organisations; William Fry. (2024, July 23). A Practical Guide to the Extraterritorial Reach of the AI Act., https://www.williamfry.com/knowledge/a-practical-guide-to-the-extraterritorial-reach-of-the-ai-act/

- 결국, 기업은 계약 상대방의 EU 진출 가능성을 사전에 명확히 파악하고, 필요한 경우, EU에서의 사용 제한 조건을 명시하는 등 거래 조건에 대한 세분화를 검토해야 하는 부담이 발생한다.

  특히, 범용 AI 활용 등 해당 결과물이 EU내에서도 사용될 가능성이 있는 경우, 계약서에 제한 사항을 명확히 기재함으로써 불필요한 규제 준수 의무를 최소화하는 방안을 고려해야 한다.

② 적용 범위의 모호성 – 적용 범위의 명확성 개선 필요

- 제2조제1항(c)목의 "유럽연합 내에서 사용"이라는 표현은 직·간접적 사용 여부를 어떻게 판단할 것인가에 대한 해석상의 문제가 발생할 수 있다. 이 때문에 AI 시스템 결과물의 유통 경로나 활용 방식이 여러 단계를 거치게 되는 경우, 규제 대상 여부가 불분명해질 수 있다.

  예를 들어, 어떤 디지털 콘텐츠가 수 차례의 중간 단계를 거쳐 유통되는 경우, 최초 공급 기업은 자사 AI 시스템 결과물이 EU 시장에 진입할 가능성을 사전에 예측하거나 통제하기 어렵다. 따라서, 기업들은 계약 체결 시, EU 시장에서의 사용 가능성을 명확히 정의하거나, 결과물의 유통 경로 추적 시스템 마련 등의 부담을 지게 된다. 이 때문에 법 적용 범위의 명확성이 개선되어야 한다는 지적이 제기되고 있다.

③ 역외 적용의 글로벌 영향

- 위에서 살펴본 바와 같이 본 법의 역외적용 조항은 EU 시장에 접근하려는 계약 상대방과 거래하고자 하는 비EU 진출 기업들에게도 추가적인 규제 준수 비용을 발생시킬 수 있다.

  이는 AI 기술의 급속한 발전과 글로벌 비즈니스 환경에서 EU의 규제가 글로벌 표준으로 자리잡을 가능성을 높여 줄 수는 있지만, 기업의 혁신과 시장 접근성에 제약으로 작용할 수 있다는 우려도 제기되고 있다. 이 같은 우려는 결국 EU 시장에 제한 요소로 부정적 영향을 끼칠 수 있는 바, 유럽연합은 국제 사회화의 협의를 통해 명확한 적용 범위의 해석과 조치를 마련할 필요가 있다.

또한, **제2조제1항(g)목**은 "EU내에서 영향을 받는 사람들"이라는 표현을 사용하고 있는 데, 이 역시 그 정의가 모호하여 적용 범위에 대한 해석이 상이할 수 있다. 즉, AI 시스템의 사용자·소비자와 같이 직접적으로 영향을 받는 사람들을 지칭하는 것인지, 해당 시스템의 간접적인 영향을 받는 사람들까지 확대되는 것인지 불문명하다. 이 때문에 관련 기업들이 추가적인 법적 리스크를 부담하게 될 가능성이 있는 바, 이에 대한 유럽연합의 명확한 해석과 지침이 요구된다.

(ii) 정의

본 법에서 다루는 주요 개념은 제3조 정의 부분에서 구분하여 다루고 있는데, 문화콘텐츠 분야에서 주로 사용되는 범용 AI의 개념과 관련 용어를 중심으로 살펴보면 아래와 같다.

제3조 정의(Definition)

(1) 'AI 시스템(AI System)' : 다양한 수준의 자율성을 갖고 작동하도록 고안되었고, 배포 이후 적응성(변화)을 보여줄 수 있으며, 명시적 또는 암시적 목적을 위해 입력된 정보로부터, 물리적 또는 가상 환경에 영향을 줄 수 있는 예측·콘텐츠·추천 또는 결정과 같은 결과물을 어떻게 생성해 낼지 추론해 내는 기계 기반 시스템

(63) '범용 AI 모델(general-purpose AI model)' : 대규모 데이터로 자기 주도 학습을 이용해 훈련한 AI 모델을 포함하여, 상당한 범용성을 보여주는 것으로, 해당 모델이 시장에 어떤 방식으로 출시되었는 지와 상관없이 광범위한 특정 과제를 능숙하게 수행할 수 있는 AI 모델을 말하며, 다양한 다운스트림 시스템과 애플리케이션에 통합될 수 있는 것을 말함. 단, 연구, 개발 또는 시장 출시 전 시험(prototyping) 활동에 사용되는 AI 모델은 제외

(66) '범용 AI 시스템(general-purpose AI system)' : 범용 AI 모델에 기반한 AI 시스템으로, 타 AI 시스템과 통합뿐 아니라 직접 사용하여 다양한 목적을 수행할 수 있는 AI 시스템

(64) '높은 영향력(high-impact capabilities)' : 가장 진보된 범용 AI 모델에 기록된 성능에 부합하거나 초과하는 성능

(65) '시스템적 위험(systemic risk)' : 영향력이 미치는 범위 때문에 또는 공공보건, 안전, 공공 보안, 기본권 또는 사회 전체에 실제적 또는 합리적으로 예견 가능한 부적적 영향 때문에 유럽연합 시장에 상당한 영향을 갖고 있는 범용 AI 모델의 높은 영향력에 특정한 위험

(**67) '부동 소수점 연산(floating-point operation)'** : 일반적으로 고정된 기반의 정수 지수로 스케일된 고정된 정밀도 정수, 컴퓨터에 표시되는 실수의 부분 집합인 부동 소수점 숫자와 관련한 모든 수학적 연산 및 할당을 말함.

(68) '다운스트림 공급자(downstream provider)' : AI 모델을 통합한 AI 시스템(범용 AI 시스템 포함)의 공급자로, 해당 AI 모델이 자체 공급 AI 모델인지 수직적으로 통합된 AI 모델인지 또는 계약관계에 기반해 타인이 제공하는 AI 모델인지 상관없음

○ **'AI 시스템'**은 입력 정보로부터 결과물을 어떻게 생성해 낼지 추론해 내는 기계 기반 시스템으로 정의하고 있는데, **'AI 시스템'**의 정의는 본 법을 이해하고 적용시키는 데 가장 핵심적인 개념인 바, 규정 경위와 의미를 아래와 같이 상세히 살펴보고자 한다.

※ **'AI 시스템'** 정의의 배경 및 의미

- **'AI 시스템'의 정의는 법의 대상과 입법 범위를 결정**하는 핵심 개념이므로, 법안 논의 과정에서 많은 협의가 진행되었다. 유럽연합은 2023.12.8. 삼자협의[9]를 통해, AI 개념이 지나치게 광범위하지 않아야 하고, 국제적 합의와도 상호 인정되는 개념이어야 한다는 의견을 반영하여, **경제협력개발기구(OECD)의 'AI 시스템' 정의[10]를 채택**하였다[11].

- 2003.11.8. 경제협력개발기구(OECD)는 인공지능(AI) 원칙(권고안)의 'AI 시스템' 정의를 아래와 같이 수정하였고, 설명각서(EXPLANATORY MEMORANDUM)[12]를 통해, AI 시스템 구축과 운영에 대한 정보를 제공하여, 이를 해석하는 데 도움을 주고 있다. **설명각서에 상술된 'AI 시스템' 정의의 주요 개념**은 아래와 같다.
  (※설명각서는 경제협력개발기구(OECD) 인공지능(AI) 원칙(권고안)의 일부는 아님)

'AI 시스템' 정의, 경제협력개발기구(OECD) AI 원칙(권고안) (2023.11.8. 수정)
An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.
*(AI 시스템은 **명시적 또는 암시적 목표**를 위해, 입력받은 정보로부터 물리적 또는 가상 환경에 영향을 미칠 수 있는 **예측·콘텐츠·추천 혹은 결정**을 **생성하는 방법을 추론**하는 기계기반 시스템이다. 각기 다른 AI 시스템은 **자율성**과 배포 이후 **적응성** 수준이 다르다.)*

① '명시적 또는 암시적 목표'

'AI 시스템' 의 목표는 명시적 또는 암시적이 될 수 있다. 이에 대해서는 아래와 같이 예시를 들어 설명하고 있다.

▶ **인간이 정의한 명시적 목표**: 배포자가 직접적으로 시스템에 인코딩한 경우(목적 함수); 간단한 분류기, 게임 플레이 시스템, 강화 학습 시스템, 결합 문제해결 시스템, 계획 알고리즘 및 동적 프로그램 알고리즘

---

9) 유럽의회·유럽연합 이사회·유럽연합 집행위원회가 참여하는 협의체로, 유럽연합 입법절차의 중요 단계. 유럽의회와 유럽 이사회가 각기 다른 형태로 통과시킨 법안을 조정하고, 최종 합의에 도달하여 법안을 확정.

10) 2023.11.8. 수정된 OECD AI 원칙(OECD AI Principles)의 'AI 시스템' 정의(경제협력개발기구(OECD), "OECD AI Principles Revised"(2023, November 8) 참조)

11) Credo AI, "EU AI Act Political Agreement: What You Need to Know for 2024"(2023, December),

12) 경제협력개발기구(OECD), "Explanatory Memorandum on the Updated OECD Definition of an AI System"(2023, November 8)

▶ **(일반적으로 인간이 특정한) 규칙의 암시적 목표:** 규칙으로 현재 상황에 따라 AI 시스템이 취하는 행동을 지시(예: '빨간 신호등에서는 멈춰라'는 운전 시스템 규칙) 인간이 규칙을 특정하지만, 시스템의 근본 목적은 법 준수나 사고 방지와 같이 명시적이지 않음.

▶ **훈련 데이터의 암시적 목표:** 궁극적인 목표가 명시적으로 프로그램되지 않지만, 데이터 학습과 데이터 모방 학습 시스템 아키텍처를 통해 통합하여 달성하는 경우(예: 응답 생성 대형 언어 모델의 경우, 명시적인 목표가 코딩되지는 않지만, 그럴듯한 응답을 생성 시, 보상을 받으면서 목표달성)

▶ **사전에 충분히 알려지지 않은 목표:** 사용자의 선호도를 점진적으로 좁혀나가기 위해 강화학습을 사용하는 추천 시스템 등

② 'AI 시스템 생성 출력물(예측·콘텐츠·추천 혹은 결정)'

'AI 시스템' 생성 출력물은 AI 시스템이 수행하는 다양한 기능을 반영하여, 그 범위가 추천·예측·결정 등으로 광범위하고, 인간이 출력에 관여하는 수준도 다양하다. AI의 자율성이 가장 높은 출력은 "결정"이며, "예측"이 가장 자율성이 낮다고 볼 수 있다. (예: 운전 보조 시스템에서 카메라 입력값 중 특정 픽셀 영역을 보행자로 **"예측"**하여, 브레이크를 작동을 **"추천"**하거나 브레이크 작동 적용을 **"결정"**)

생성형 AI 시스템은 텍스트, 이미지, 오디오 및 비디오 등의 "콘텐츠"를 출력물로 생성한다. 텍스트 생성을 특정 단어를 출력하는 일련의 "결정"으로 보거나, 특정 맥락에 쓰일 법한 단어를 "예측"하는 것으로 볼 수도 있지만, 콘텐츠 생성 시스템이 AI 시스템의 중요 범주에 속하게 되면서, **콘텐츠를 별도의 출력물 범주로 구분**하고 있다.

③ '출력을 생성하는 방법을 추론'

'추론'은 일반적으로 배포 이후에 시스템이 입력값에서 출력물을 생성하는 단계를 말하지만, 'AI 시스템' 정의에서 '추론'은 입력/데이터로부터 모델이 도출되는 AI 시스템의 빌드 단계(Build Phase)도 포함한다. 즉, AI 시스템이 데이터를 통해 패턴을 학습하고, 이를 바탕으로 모델을 생성하는 과정 전체가 추론의 일부로 이해될 수 있다. 이는 단순히 학습된 모델을 통해 예측을 생성하는 것뿐만 아니라, 모델을 학습시키는 초기 과정도 포함된다는 것을 의미하는 것으로 볼 수 있다.

④ 'AI 시스템의 자율성'

인간이 자율성을 위임하고, 프로세스가 자동화된 후, 인간의 개입 없이 시스템이 학습하거나 행동할 수 있는 정도를 의미한다. 일부 AI 시스템은 결과물에 대한 목표가 명시적으로 기술되지 않은 경우에도, 인간의 구체적인 지시 없이 결과물을 생성할 수 있다. 인간은 AI 시스템 설계·데이터 수집 및 처리·개발·검증·유효성 검사·배포 또는 운영 및 모니터링 등 AI 시스템 라이프사이클 어떤 단계에서든 감독할 수 있다.

⑤ 적응성

일반적으로 초기 배포 이후 계속해서 진화할 수 있는 머신러닝 기반 AI 시스템과 관련된 개념이다. 이 시스템은 배포 전 또는 이후 입력값 및 데이터와 직접적인 상호작용을 통해 행동을 수정한다(예: 개인의 목소리에 맞춰 적응하는 음성 인식 시스템, 개인 맞춤형 음악 추천 시스템 등). 한 번 또는 정기적·지속적으로 AI 시스템을 훈련하고, 시스템은 운영 과정에 데이터의 패턴과 관계성을 추론하게 되면서, 프로그래머가 처음에는 생각하지 못했던 새로운 형태의 추론 수행 능력이 개발된다.

ㅇ AI 시스템을 구성하는 모델 중 범용성을 갖고 있는 모델을 **'범용 AI 모델'**로 구분하고, 이 범용 AI 모델에 기반한 AI 시스템을 **'범용 AI 시스템'**으로 구분하여, 타 AI 시스템과 통합하거나 직접 사용하여 다양한 목적을 수행할 수 있는 AI 시스템으로 정의하고 있다.

ㅇ 범용 AI 모델 중 **'높은 영향력'**을 지닌 AI 모델을 **'시스템적 위험'**이 있는 AI 모델로 구분하여 특정 의무사항을 부과하고 있는데, 이 **'시스템적 위험'이 있는 범용 AI 모델** 판단 기준 중 하나가 **'부동 소수점 연산'**을 통한 학습 사용 누적 계산량이 특정 수치(10^{25}) 이상[13])에 해당하는 경우이다.

여기서 **'높은 영향력(high-impact capabilities)'**은 가장 진보된 범용 AI 모델에 기록된 성능에 부합하거나 초과하는 성능을, **'시스템적 위험(systemic risk)'**은 공공보건·안전·공공 보안·기본권 등에 영향을 미칠 수 있는 위험으로 다소 포괄적으로 정의하고 있고, 각 요건은 개별 조항에서 별도로 규정하고 있다.

ㅇ 또한, **범용 AI 시스템을 포함하여, AI 모델을 통합한 AI 시스템 공급자를 '다운스트림 공급자'**로 정의하고 있다.

2) 금지행위 및 과징금

ㅇ 본 법은 **제5조에서 AI로 금지되는 행위**를 규정하고 있으며, 이를 위반할 경우**, 제99조(벌칙)제3항**에 따라, **최대 3,500만 유로 또는 이전 회계연도 전세계 연간 총 매출액의 최대 7%중 높은 금액**의 과징금이 부과된다.

13) 유럽연합 인공지능법(EU AI Act) 제51조제2항 참조

이러한 과징금 수준은 본 법이 매우 엄격한 규제를 도입하고 있음을 보여준다. 이에 따라, AI 활용이 점차 광범위하게 확대되고 있는 문화콘텐츠 분야에서도 세심한 주의가 요구된다.

특히 아래에 열거된 금지행위들은 문화콘텐츠 분야에서 AI 기술을 활용 시, 그 잠재적 위험 요소를 검토할 필요가 있다.

예를 들어, 게임 등의 가상 콘텐츠 개발, 어린이 콘텐츠 또는 맞춤형 콘텐츠(영상·음악·이미지 등) 제작, 알고리즘 추천, 딥페이크 기술을 활용한 콘텐츠 제작 등에 있어, 개인과 사회에 미치는 영향과 그 파급효과를 신중히 고려해 봐야 한다.

- 개인이 충분한 정보를 바탕으로 결정을 내릴 능력을 현저히 저하시켜, 당사자나 타인 또는 단체에 심각한 손해를 끼치거나 그럴 가능성이 있는 결정을 내리도록 할 목적으로 또는 그 같은 결과를 초래하는 사람의 의식 너머의 잠재적인 기법 또는 의도적인 조작 혹은 기만적 방식의 기법을 배포하는 AI 시스템을 시장에 출시·서비스 제공 또는 사용하는 경우,
- 나이·장애 및 특정 사회적·경제적 상황의 취약성을 악용하는 경우,
- 개인의 사회적 행동, 개인적·성격적 특성에 기반해 사회적 점수로 평가하는 경우로, 데이터 수집 맥락과 무관하게 특정 개인이나 단체를 불리하게 대우하는 경우 및 개인 또는 단체의 사회적 행동이나 그 중요성에 정당하지 않게 불리하게 대우하는 경우
- 직장 또는 교육기관에서 개인의 감정을 추론하는 AI 시스템을 사용하는 경우
- 종교·정치적 견해·노동조합 가입·종교적 또는 철학적 신념·성생활 또는 성적 지향을 추측하거나 추론하려는 목적으로 생체 데이터에 기반해 개인을 분류하는 경우,

아래에서는 문화콘텐츠 분야에서 상기 금지행위와 관련하여 제기되고 있는 구체적인 위험요소를 관련 문헌을 기반으로 살펴보고자 한다.

❑ **문화콘텐츠 분야에서 주의해야 할 금지행위 위험 요소[14)]**

① AI 기술이 잘못된 정보와 조작을 초래할 가능성

▶ 소셜 미디어나 온라인 커뮤니티 등에서 AI 알고리즘이 클릭 등의 사용자 참여를 유인하기 위해 보다 자극적이거나 분열적 콘텐츠(정치적 극단주의 내용, 가짜 뉴스 및 허위 정보, 혐오 발언, 음모론 등)를 우선시하여, 이를 더 많이 노출함으로써, 인간의 판단을 흐리고 민주적 의사결정을 방해하거나 사회적 불신을 조장할 가능성이 있음

▶ 딥페이크 기술을 통한 정치적 메시지 조작, 가짜 뉴스, 배우의 AI 무단복제 등 기만적 방식으로 기술을 악용하여 사회에 부정적 영향을 미칠 가능성이 있음

② 문화적 획일화 및 사회적 단절을 야기할 가능성

▶ 영화·음악·게임 등의 콘텐츠에서 AI 알고리즘 추천 시스템이 사용자에게 선호하는 콘텐츠만을 반복적으로 제공하는 필터 버블(Filter Bubbles)[15)]과 반향실(Echo Chambers)[16)] 현상으로 인해 인간의 다양한 관점을 제한하고, 문화적 획일화 및 사회적 단절을 초래할 가능성이 있음

③ 콘텐츠 제작·배포 과정에 인간의 무의식 속에 부정적 영향을 미칠 가능성

▶ AI 기술은 기본적으로 개발자가 설계한 방식으로 데이터를 학습하고 사용자의 요청에 기반하여 결과물을 생성하므로, 개발자·사용자 등의 세계관·가치관이 학습 데이터 및 결과물에 내재될 수 있음. 이를 통해 AI가 개인의 성별·인종·정치적 견해 등이 반영된 불공정하거나 왜곡된 콘텐츠를 생성할 가능성이 있음

특히 문화콘텐츠 분야에서 생성형 AI를 활용해 예술 작품·영화 제작·음악 작곡 등의 창작활동을 할 경우, AI가 콘텐츠 제작자의 편향된 세계관 또는 학습 데이터의 편향성으로 인해 인종·성별·종교에 대한 차별적 표현을 포함한 콘텐츠 생성, 특정 집단의 극단적 메시지를 암시, 역사적 맥락을 고려하지 않은 잘못된 혹은 부적절한 문구·이미지 생성으로, 대중 관객의 잠재적 인식 속에 왜곡된 이미지나 고정관념을 강화할 가능성이 있음

---

14) ENCATC. (2023). Artificial Intelligence Embraced: the future of the cultural and creative sector. In Book of Proceedings, 2023. Retrieved from https://encatc.org/media/7193-2023-book-of-proceedings.pdf#page=59.; Kulesz, O. (2024). Artificial Intelligence and International Cultural Relations: Challenges and Opportunities for Cross-Sectoral Collaboration. ifa – Institut für Auslandsbeziehungen. Retrieved from https://www.ifa.de/en/publications

15) 필터 버블(Filter Bubbles): 이용자의 관심사에 맞춰 필터링된 인터넷 정보로 인해 편향된 정보에 갇히는 현상, 대형 인터넷 정보기술(IT) 업체가 개인 성향에 맞춘(필터링된) 정보만을 제공하여 비슷한 성향 이용자를 한 버블 안에 가두는 현상을 지칭 (출처: NAVER 지식백과)

16) 반향실(Echo Chambers): 자신의 신념과 일치하는 정보만을 반복적으로 수용·소비함으로써 기존의 신념이 더욱 강화되는 현상, 선호에 맞지 않는 정보는 차단되기 때문에 확증적 편향을 강화시키며, 양극화 현상을 부추길 수 있어 최근 미디어 환경의 과제로 떠오르고 있음 (출처: NAVER 지식백과)

④ AI 알고리즘 추천 등에 생체 데이터 활용 가능성

▶ 생체 인식 데이터를 기반으로 사용자의 감정을 분석하고 분류하는 시스템은 개인 프라이버시 침해 등 시민권을 위협하는 것으로 간주될 수 있음. 인간의 특정 감정(슬픔, 기쁨, 분노 등)을 자극하거나 이를 분석하여 콘텐츠를 추천하는 시스템은 인간의 정서를 정치적·상업적 목적으로 조작할 수 있다는 위험성을 내포하고 있으므로 각별한 주의가 필요

⑤ 사회·경제·심리적 취약성을 이용해 특정 방향으로 행동을 유도할 가능성

▶ AI 알고리즘 추천 시스템이 사용자 데이터를 기반으로 의도적으로 소비 행동을 유도하거나, 게임 스토리와 광고를 통해 특정 메시지를 무의식적으로 주입하여 개인의 의사결정을 조작한다고 판단되는 경우 규제 대상이 될 수 있음

ㅇ 문화콘텐츠는 기본적으로 인간의 사상·감정과 상호작용을 기반으로 한다. 따라서, 어떤 콘텐츠이든 인간의 무의식에 영향을 미칠 수 있다. 그러므로, 문화콘텐츠 업계에서는 AI를 활용할 때, 이러한 영향을 고려해야 한다. 하지만, 이를 어느 정도까지 금지행위로 볼 지에 대해서는 추가적인 사회적 논의가 필요할 것으로 보인다.

특히, 사람이 창작한 스릴러 및 범죄 영화, 극단적 감정을 자극하는 음악 콘텐츠가 자살 등의 이상행동을 유발해 사회적 문제가 된 사례들이 다수 있음에도, AI를 활용한 창작 콘텐츠는 초기 창작 의도와 다르게 사후에 부정적인 사회적 영향을 미치게 되면, 모두 금지행위에 해당하는 지, 어느 수준까지 금지행위로 볼 지에 대한 경계가 명확하지 않다.

따라서, 창작 시점의 의도와 사후적 파급 효과에 대한 책임 소재 및 그 수준과 관련한 명확한 기준이 필요하다. 즉, 제한이 거의 없는 인간의 콘텐츠 창작이 AI 활용 시에는 금지행위로 간주된다면, 명확한 구분과 경계에 대한 사회적 합의가 필요할 것이다.

ㅇ 또한, **동조 제8항**에서는 다른 유럽연합 법을 침해하는 경우에 적용되는 금지 사항에는 영향을 미치지 않는다고 명시하고 있어, 기존 **유럽연합 법에서 금지된 AI 관행이 있다면, 본 조항과 함께 중첩되어 규제 대상**이 된다. 따라서, 제5조 금지행위 조항은 기존 규제를 더 강화하고 있는 것으로 볼 수 있다.

3) 투명성 의무 및 과징금

ㅇ 제4장 제50조에서는 투명성 의무가 부과되는 특정 AI 시스템을 구분하고, 관련 공급자·배포자의 의무사항을 아래와 같이 규정하고 있다. 이를 위반할 경우, **제99조(벌칙)제4항(g)목**에 따라, **최대 1,500만 유로 또는 이전 회계연도 전세계 연간 총 매출액의 최대 3%** 중 높은 금액의 과징금이 부과된다.

특정 AI 시스템 공급자·배포자 구분	의무사항
▪ 자연인과 직접 상호작용 하도록 의도된 AI 시스템 공급자	- 자연인이 AI 시스템과 상호작용하고 있음을 알 수 있도록 할 것.
▪ 합성 오디오·이미지·비디오 또는 텍스트 콘텐츠를 생성하는 AI 시스템 (범용 AI 시스템 포함) 공급자	- 인공적으로 생성된 것임을 기계가 감지할 수 있는 형식으로 표시할 것
▪ 감정 인식 시스템 또는 생체 분류 시스템 배포자	- 노출되는 자연인에게 시스템 작동을 알릴 것
▪ 딥페이크를 구성하는 이미지·오디오 또는 비디오 콘텐츠를 생성 혹은 조작하는 AI 시스템 배포자	- 콘텐츠가 인공적으로 생성 또는 조작되었음을 공개할 것
▪ 공익 관련 내용을 대중에게 알리기 위한 텍스트를 생성하거나 조작하는 AI 시스템 배포자	- 텍스트가 인공적으로 생성 또는 조작되었음을 공개할 것

종합해서 보면, AI 시스템이 작동 중임을 알리고, 생성된 콘텐츠가 AI에 의해 생성되고 조작된 것임을 투명하게 공개하도록 의무화하고 있다. 정보의 제공 방식 및 시기는 자연인에게 명확하고 식별가능한 방법으로 늦어도 첫 상호작용의 시점 또는 첫 노출 시점에 제공해야 한다[17].

또한, 본 조항의 투명성 의무는 제3장(고위험 AI 시스템)에 규정된 요건·의무사항 및 유럽연합 내 또는 AI 시스템 배포자 자국법에 명시된 다른 투명성 의무에 영향을 미치지 않는다고 규정[18]하고 있다. 이는 기존 규제와 중첩되어 규제가 더 강화된 것으로 해석될 수 있다.

17) 유럽연합 인공지능법(EU AI Act) 제50조제5항 참조
18) 유럽연합 인공지능법(EU AI Act) 제50조제6항 참조

하지만, 본 의무의 실제적인 이행과 관련한 구체적인 사항은 별도의 실천 규범(Codes of Practice)을 작성하여 시행하는 것으로 규정[19]하고 있어, 이행 방식 등의 구체적인 사항은 실천 규범이 마련되는 과정에 추가 논의가 예상된다.

4) 범용 AI 모델 공급자의 의무 및 과징금

(i) 범용 AI 모델 공급자의 의무

ㅇ 본 법 제53조는 범용 AI 모델 공급자에게 다음과 같은 의무를 규정하고 있다.

- 훈련·테스트과정·평가 결과를 포함한 기술문서 작성 및 최신상태 유지
- 범용 AI 모델을 시스템에 통합하고자 하는 AI 시스템 공급자에게 정보 및 문서를 작성하여 제공하고 최신 상태로 유지
- 저작권 및 관련 권리에 대한 유럽연합법을 준수하고, 권리의 유보를 식별하고 준수하는 정책을 수립
- AI 사무국이 제공하는 템플릿에 따라 범용 AI 모델의 훈련에 사용된 콘텐츠에 대한 충분히 상세한 요약을 작성하여 대중에게 공개

ㅇ 상기 의무들은 AI 모델 공급자에게 상당한 기술적·법적 부담을 초래할 수 있다. 특히, 저작권 준수 의무와 훈련 콘텐츠의 대중 공개는 실행 과정에서 현실적인 어려움이 예상된다. 이에 따라, 아래에서는 범용 AI 모델 공급자가 준수해야 할 의무와 관련된 주요쟁점을 관련 문헌 연구를 통해 구체적으로 살펴보고, 그 문제점 및 대안을 검토해 보고자 한다.

❑ 주요쟁점 1: 저작권 준수 관련

① AI 학습 데이터[20]

AI 모델은 인터넷 등에서 방대한 데이터를 학습하며, 이 학습 데이터의 양과 질이 성능을 결정한다. 데이터 필터링[21] 및 디지털 해시 태깅[22]과 같은 기술은 저작권

19) 유럽연합 인공지능법(EU AI Act) 제50조제7항 참조
20) Fernandez. (2024, May 14). The Luxembourg Model of AI Governance: Toward Human-Centric AI Regulation. University of Luxembourg. Retrieved from https://orbilu.uni.lu/handle/10993/62607
21) 데이터 필터링(Data filtering): AI 모델 학습 데이터 중 저작권 침해 가능성이 있는 데이터를

준수를 지원하지만, 학습 데이터의 규모를 축소시켜 성능 저하로 이어질 수 있다. 이 같은 성능 저하는 특히 자연어 처리 모델과 같이 대규모 데이터에 의존도가 높은 경우 두드러지게 나타난다.

또한, 실제로 사람이 일일이 확인해 봐도 저작권 자유 이용 가능 여부가 명확히 표시된 데이터가 극히 드물다는 현실은 기술적·법적 제약을 더욱 심화시킨다.

② AI 생성 결과물[23)]

AI 결과물은 사용자의 입력값에 따라 AI가 자체적으로 학습한 결과를 각기 다르게 생성하므로, AI 모델 공급자가 사전에 이에 대한 저작권 침해 여부를 파악하는 것은 현실적으로 거의 불가능하다.

따라서, **사용자와 공급자 간에 책임을 명확히 구분하는 기준(**예를 들어, 사용자가 특정 데이터 입력을 통해 AI에게 명확히 저작권 침해 가능성이 있는 결과물을 요청한 경우라면, 책임을 사용자에게, 공급자가 적법하지 않은 방식으로 데이터를 학습시켜 특정 데이터 사용이 불법이라는 사실을 인지하고도 이를 방치한 경우라면, 책임을 공급자에게 귀속 등) 마련이 필요하다.

무엇보다 공급자가 준수 의무 범위를 명확히 파악할 수 없어 과도한 책임을 부담하지 않도록 데이터 필터링 등과 같이 **적법성 확인을 위해 적용해야 하는 기술 및 절차·기준 등 의무사항에 대한 명확화가 필요**할 것이다.

③ 권리귀속과 책임 소재[24)]

일반적인 권리 귀속과 책임 소재 관계를 고려하면, AI 생성 결과물의 저작권자가 결과물의 침해에 대해 책임지는 것이 타당할 것이다. 하지만, 현재 사람이 아닌 AI가 생성한 결과물의 저작권은 인정되지 않고 있으므로, 공급자에게 권리없이 과도한 책임만 부과하는 것은 불합리하다는 지적이 제기된다.

사전에 걸러내는 기술. 메타데이터 분석이나 특정 알고리즘을 통해 수행되며, 저작권 상태가 불명확한 데이터를 제거하거나 적법한 데이터만 선택적으로 학습에 활용하도록 설계됨(출처: Oppedal, N. M. (2024). Balancing Innovation and Copyrights: The Legal Framework for AI Training in the European Union. Retrieved from https://arno.uvt.nl/show.cgi?fid=176939)

22) 디지털 해시 태깅(Digital hash tagging): AI 모델 학습 데이터의 적법성을 보장하기 위해 데이터에 고유한 해시 값을 부여하거나, 저작권 및 데이터 출처 정보를 추가로 포함하여 추적 및 관리를 가능하게 하는 기술(출처: Oppedal, N. M. (2024). Balancing Innovation and Copyrights: The Legal Framework for AI Training in the European Union. Retrieved from https://arno.uvt.nl/show.cgi?fid=176939)

23) Margoni. (2024, May 30). Regulating AI through Global Frameworks: Challenges and Opportunities. SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5036164

24) Guadamuz. (2024, November 10). AI Regulation in Global Perspective: Challenges for Intellectual Property Law. The Journal of World Intellectual Property, Wiley Online Library. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/jwip.12330

이에 대한 대안으로 사용자와 공급자 간에 책임 소재를 분담할 수 있도록 약관이나 계약 조건을 명확히 하는 방안이 제안되고 있다. 즉, 사용자 입력이 AI 결과물 생성에 출발점이 이므로, 공급자는 제공한 AI 시스템의 한계점을 명확히 사용자에게 밝히고, 사용자가 이를 벗어나는 사용을 하는 경우, 법적 책임이 사용자에게 귀속됨을 명시하자는 것이다.

실제 OpenAI의 ChatGPT는 아래와 같이 이용약관에 사용자가 생성물의 사용에 대한 책임을 지도록 명시하고 있다.[25)]

OpenAI
(Published: December 11, 2024)

Terms of Use

Content

Your Content. You may provide input to the Services("Input"), and receive output from the Services based on the Input("Output"). Input and Output are collectively "Content". **You are responsible for Content, including ensuring that it does not violate any applicable law or these Terms.** You represent and warrant that you have all rights, licenses, and permissions needed to provide Input to our Services.

(**귀하의 콘텐츠.** 귀하는 서비스에 입력을 제공할 수 있으며(이하 "입력"이라 칭함), 해당 입력을 기반으로 서비스에서 출력을 받을 수 있습니다(이하 "출력"이라 칭함). 입력과 출력은 통칭하여 "콘텐츠"로 정의합니다. **귀하는 콘텐츠에 책임을 지며, 해당 콘텐츠가 관련 법이나 본 약관을 위반하지 않도록 보장**해야 합니다. 귀하는 서비스에 입력 제공 시 필요한 모든 권리, 라이선스 및 허가를 취득하고 있음을 진술하고 보증해야 합니다.

하지만, 본 법에서는 공급자의 의무만을 규정하고 있어, 양자간 계약과는 상관없이 법 준수 의무가 공급자에게 있다. 따라서, 사용자와 공급자 간 역할과 책임을 명확히 하는 추가적인 제도적 장치와 법적 기준이 필요할 것이다.

• 이와같이, 공급자의 저작권 준수 의무에 대해서는 효과적인 검증 및 관리, 공급자가 수용 가능한 수준의 표준화된 필터링 기술 개발, 공급자와 사용자 간의 공동 책임 체계 구축 등 다각적인 접근이 필요하다.

예를 들어, 본 법의 세부 조항을 기반으로 필터링 기술의 표준화를 추진하고, 국제적으로 합의된 책임 분담 가이드라인을 마련한다면, 기술 발전과 저작권 보호 간의 균형을 보다 효과적으로 유지할 수 있을 것이다.

---

25) OpenAI. (2024, December 11). Terms of Use. Retrieved from https://openai.com/policies/terms-of-use/

❑ 주요쟁점 2: 훈련 콘텐츠 정보 공개 관련

① AI 학습 특성의 복잡성[26)]

AI 모델은 초기 학습 데이터를 기반으로 스스로 다양한 패턴과 지식을 확장하면서 학습해 나가는 특성을 가진다. 특히 최신 AI 모델은 수십억개의 매개변수를 포함할 정도로 복잡성이 높아, 훈련에 사용된 데이터를 추적하거나 요약하는 것이 현실적으로 쉽지 않다.

더욱이 현재 AI 시스템은 법적 규제가 없는 상황에서 개발되었기 때문에, 개발 단계에서 학습 데이터 추적을 염두에 두고 개발되지 않았다. 이 때문에 특히 중소기업의 경우 사후적으로 데이터를 정리하고 문서화하는 것 자체가 부담이 될 수 있다.

따라서, 상세한 데이터 공개 요구는 기술적으로 AI 공급자들에게 과도한 부담을 줄 가능성이 있다.

② 표준화된 데이터 이력 추적 시스템의 부재[27)]

현재 AI 모델 개발에 있어 훈련 데이터의 이력을 추적할 수 있는 국제적 표준이 부재하다. AI 모델 개발자들이 사전 훈련 및 미세 조정 데이터를 체계적으로 문서화할 필요성은 인식하고 있지만, 이에 대한 표준화된 구체적 기준이 마련되지 않아 혼란을 야기시키고 있다.

또한 유럽연합이 요구하는 훈련 데이터 요약의 구체적 수준이 아직 명확하지 않은 상황이어서, 새로운 데이터 이력 추적 시스템을 개발하거나 도입하는 것은 AI 공급자에게 상당한 기술적·재정적 부담을 가중시킬 가능성이 있다.

---

26) V7 Labs. (2022, July 11). Quality Training Data for Machine Learning: A Guide. Retrieved from https://www.v7labs.com/blog/quality-training-data-for-machine-learning-guide; Bloomberg Law. (2024, July 31). European AI Act Training Disclosures Expose US Copyright Risks. Retrieved from https://news.bloomberglaw.com/ip-law/european-ai-act-training-disclosures-expose-us-copyright-risks.

27) MIT GenAI. (2024, March 27). Towards Standards for Data Transparency for AI Models. Retrieved from https://mit-genai.pubpub.org/pub/uk7op8zs/release/2; U.S. AI.gov. (2024, May). Proceedings: Towards Standards for Data Transparency for AI Models. Retrieved from https://ai.gov/wp-content/uploads/2024/06/PROCEEDINGS_Towards-Standards-for-Data-Transparency-for-AI-Models.pdf.

③ 기업 기밀 유지[28)]

AI 모델의 훈련 콘텐츠는 기업의 핵심 기술력 및 경쟁력과 직결된다. 따라서, 기업들은 점점 더 훈련 데이터를 비공개하는 입장으로 전환하고 있다. 예를 들어, Open AI는 GPT-3까지는 훈련 데이터 출처를 공개했으나, GPT-4부터는 비공개로 전환했고, Stability AI의 경우 초기에 AI 모델 "Stable Diffusion"의 훈련 데이터를 공개했으나, 2023년 저작권 등 법적 문제로 미국과 영국에서 소송을 당하면서, 2023년 11월 출시한 AI 모델 "Stable Video Diffution"의 훈련 데이터는 공개하지 않고 있다.

이와 같이, 훈련 데이터 공개 의무가 명확한 기준 없이 요구될 경우, 기업들은 경쟁 우위 확보와 법적 리스크 해소를 위해 최소한의 정보만 공개하는 방식으로 대응할 가능성이 높다. 따라서, 유럽연합은 기업들이 핵심 기술 공개 부담이나 법적 리스크 없이 훈련 데이터 공개 의무를 준수할 수 있도록, 현실적이고 수용 가능한 표준화된 데이터 추적 및 관리 시스템을 제시할 필요가 있다. 예를 들어, 데이터 요약의 범위를 제한하고, 공개되는 템플릿 정보를 최소화하여 공급자의 부담을 완화할 필요가 있다.

ㅇ 범용 AI 모델의 위와 같은 의무 준수는 조화된 표준이 공표되기 전까지 실천 규범(Codes of Practice)에 기반해 이루어질 예정이므로, 해당 규범이 확정된 이후 구체적인 사항을 확인할 수 있을 것이다. 본 법 Recital 107에서도 AI 공급자의 영업 비밀과 기업 기밀 정보 보호의 필요성을 인정하고 있는 만큼, 유럽연합이 구체적인 실천 규범(Code of Practice)을 어떻게 확정하는 지, 이를 통해 기업들이 실질적으로 이행할 수 있는 범위 내에서 훈련 데이터 공개의 표준이 마련되는지 점검할 필요가 있다.

(ii) 공인 대리인 지정 의무

ㅇ 본 법 제54조에서는 제3국에 설립된 범용 AI 모델 공급자의 경우, **유럽연합 시장에 제품을 출시하기 전**에 **유럽연합 내 공인 대리인을 지정**

28) Center for Art Law. (2024, May 21). Generative AI and Transparency of Databases and Their Content: From a Copyright Perspective. Retrieved from https://itsartlaw.org/2024/05/21/generative-ai-and-transparency-of-databases-and-their-content-from-a-copyright-perspective/; Baker Botts. (2024, April 8). The EU AI Act: Uncharted Territory for General Purpose AI. Retrieved from https://www.bakerbotts.com/thought-leadership/publications/2024/april/the-eu-ai-act-uncharted-territory-for-general-purpose-ai; Tech Policy Press. (2024, December 9). How the EU AI Act Can Increase Transparency Around AI Training Data. Retrieved from https://www.techpolicy.press/how-the-eu-ai-act-can-increase-transparency-around-ai-training-data/.

하도록 의무화하고, 공인 대리인에게 본 법 준수와 관련한 모든 문제를 처리할 수 있는 권한을 위임하여야 한다고 규정하고 있다. 이는 단순한 연락 창구가 아니라, 실질적인 법적 권한을 가진 현지 법인을 요구하는 것으로 해석될 수 있다.

ㅇ 해당 법 조항은 역외 AI 공급자에게도 동일한 법적 책임을 부과하여, 법 준수를 보장하고자 하는 것으로 보이나, 역내 서비스 공급자와 비교했을 때, 역외 공급자에게만 추가적인 법적·행정적 부담을 지우는 것이 될 수 있고, 유럽연합 시장 진입을 제한하는 요인으로 작용할 가능성이 있다. 따라서, 본 조항의 적용 방식에 따라, WTO GATS(서비스 무역에 관한 일반협정)의 시장접근(제16조) 및 내국민대우(제17조) 조항[29] 위반 소지가 있을 수 있다[30].

ㅇ 다만, WTO GATS의 시장접근 및 내국민대우 원칙은 회원국이 시장 개방을 약속한 서비스 부문에서만 적용되므로, 해당 범용 AI 모델 서비스 분야가 유럽연합의 WTO GATS 개방 약속 목록[31]에 포함된 서비스 영역에 해당하는 지 확인하고, WTO GATS 원칙을 준수하도록 조정하는 협상 과정이 필요할 것이다.

(iii) 시스템적 위험이 있는 범용 AI 모델 공급자의 의무

ㅇ 본 법 제51조는 아래와 같이 특정 범용 AI 모델을 "시스템적 위험이 있는 범용 AI 모델"로 구분하고, 이에 해당하는 경우, 제55조에 따라 일반적인 범용 AI 모델 공급자의 의무 외에 추가적인 의무를 부과하고 있다.

- "시스템적 위험이 있는 범용 AI 모델" 구분 기준

본 법 제51조에서는 아래 두 가지 기준 중 하나에 해당하는 AI 모델을 "시스템적 위험이 있는 범용 AI 모델"로 분류하고 있다.

---

29) 세계무역기구(WTO). "General Agreement on Trade in Services (GATS). "https://www.wto.org/english/docs_e/legal_e/26-gats_01_e.htm

30) 대외경제정책연구원(KIEP). (2018, December 31). 국제 통상 환경 변화와 한국의 대응 전략. Retrieved January 29, 2025, from https://www.kiep.go.kr/gallery.es?mid=a10101040000&bid=0001&list_no=2328&act=view

31) 세계무역기구(WTO). "EU Schedule of Specific Commitments under GATS. "https://www.wto.org/english/tratop_e/serv_e/serv_commitments_e.htm.

- '높은 영향력'을 가진 것으로 평가되는 것, 집행위원회 직권으로 또는 과학 패널이 이와 동등한 영향력을 갖고 있다고 보는 것
- **부동 소수점 연산으로 측정한 학습 사용 누적 계산량 10^{25} 이상인 경우**

따라서, 부동 소수점 연산 계산량 기준으로 보면, 대규모 연산을 필요로 하는 ChatGPT, DALL-E 등의 생성형 AI는 대부분 시스템적 위험이 있는 것으로 구분될 수 있다.

하지만, 이 부동 소수점 연산 수치는 시스템마다 고정적인 것이 아니라, AI 모델 사용 방식에 따라 그 계산량이 변동될 수 있다. 예를 들어, 일반적으로 대규모 연산이 필요하지 않은 이미지 생성 AI라도, 고해상도 이미지 처리를 반복적으로 수행하면 누적 수치가 급증할 수 있다. 따라서, 부동 소수점 누적 계산량 기준만으로 AI 모델의 시스템적 위험 여부를 단정하기에는 한계가 있을 것으로 보인다. 이에 대해 관련 문헌을 기반으로 아래와 같이 살펴보았다.

❑ 주요쟁점 : "시스템적 위험이 있는 범용 AI' 부동 소수점 연산 기준

- 본 법에서 10^{25} 부동소수점연산(FLOP) 수치를 기준으로 한 것은 GPT-3, GPT-4, Gemini 등 주요 AI 모델들의 학습 필요 연산량 분석 수치를 기반으로 설정된 것으로 보인다.

모델명	**연산량(FLOP)** - 추정치
GPT-3.5 (OpenAI, 2022-11-28)	$2.58x10^{24}$
GPT-4 (OpenAI, 2023-03-15)	$2.10x10^{25}$
Gemini Ultra (Google DeepMind, 2023-12-06)	$5.0x10^{25}$
Claude 2 (Anthropic, 2023-07-11)	$3.87x10^{24}$

(출처: Epoch AI. (2024, April 5). Tracking Large-Scale AI Models. Retrieved from https://epoch.ai/blog/tracking-large-scale-ai-models)

- 하지만, AI 모델마다 부동소수점연산(FLOP) 계산 방식이 상이하고[32], 동일한 부동소수점연산(FLOP) 수치를 보여주는 모델들도 실제 사용 환경에서, 학습 데이터 크기, GPU/TPU 최적화 방식 등에 따라 연산량에 차이가 발생할 수 있다[33]. 이에

32) Casson, A. (2023, May 16). Transformer FLOPs. Retrieved from https://www.adamcasson.com/posts/transformer-flops.

따라, 일부 기업들은 "FLOP 최적화"를 통해 법적 수치 기준을 피하는 방식으로 우회할 가능성도 있다. 때문에 변동성이 큰 FLOP 수치를 정량적 규제 기준으로 삼는 것은 한계가 있다는 지적에 제기되고 있다[34].

- 부동소수점연산(FLOP) 수치의 증가가 AI 모델의 성능·사회적 영향·환경적 비용 증가 등과 직접적으로 연결된다는 점을 고려할 때, 현재 거대 AI 모델의 연산량을 기반으로 설정된 해당 수치는 정책적 정당성과 함께 "시스템적 위험이 있는 범용 AI 모델"을 규제하기 위한 현실적인 기준으로 볼 수 있다.

  그러나, 해당 연산이 변동성이 크고, 연산 최적화 기법을 적용하면 해당 수치를 낮춰 이를 회피할 가능성도 존재하므로, 이를 보완적 요소로 활용하고, 향후 AI 모델의 실질적인 위험성을 보다 정밀하게 평가할 수 있는 추가적인 기준 마련이 필요할 것으로 보인다.

- "시스템적 위험이 있는 범용 AI 모델" 추가 의무 사항

본 법 제55조는 "시스템적 위험이 있는 범용 AI 모델"로 분류된 경우, 앞서 살펴본 범용 AI 모델 공급자의 의무(제53조 내지 제54조) 외에도 아래와 같은 추가적인 의무를 부과하고 있다.

- 최신 기술을 반영한 표준화된 프로토콜 및 도구에 따른 모델 평가 수행(시스템 위험을 식별·완화하기 위한 적대적 테스트(adversarial testing) 시행 및 문서화 포함)
- 유럽연합 차원에서 발생 가능한 시스템적 위험을 평가·완화
- 중대사고 및 시정조치 관련 정보를 지속 추적·문서화하여 보고
- 시스템적 위험이 있는 범용 AI 모델 및 인프라에 대한 사이버보안

위에서 언급한 추가 의무 사항 중, 특히 최신 기술을 반영한 표준화된 프로토콜 및 도구에 대한 구체적인 내용이 명확하지 않고, 적대적 테스트의 수준과 범위에 대해서도 명확한 세부 사항이 부족하다.

---

33) Epoch AI. (2021, November 29). Measuring FLOPs Empirically. Retrieved from https://epoch.ai/blog/measure-FLOP-empirically.
34) Engine. (2024, October 24). AI Essentials: What Is Compute and How Is It Measured? Retrieved from https://www.engine.is/news/category/ai-essentials-what-is-compute-and-how-is-it-measured.

법문에는 조화된 표준(Harmonized Standard)이 발표되기 전까지, 제56조의 실천 규범(Codes of Practice)을 따를 수 있도록 규정[35]하고 있고, 시스템적 위험의 유형 및 특성을 식별하는 내용도 실천 규범에 포함하도록 하고 있다.

하지만, 해당 실천 규범의 구체적인 내용도 아직 확정되지 않은 상황이다. 이에 따라, AI 모델 공급자는 명확한 기준이 부재한 상태에서 시스템적 위험 여부를 판단하고, 이를 평가 및 완화해야 하는 과정에 실질적인 어려움을 겪을 가능성이 크다.

따라서, 이러한 법적 불확실성을 해소하고 효율적인 법 준수가 가능하도록 하기 위해서는 유럽연합이 적시에 실천 규범을 정비하고, 적대적 테스트 및 최신 기술을 반영한 프로토콜의 구체적 지침을 제공해야 한다. 또한, 기업들이 현실적으로 준수할 수 있는 명확한 표준을 제시하고, 충분한 준비기간을 고려한 단계적 이행 계획을 마련하는 것이 필요할 것이다.

(iv) 범용 AI 모델 공급자에 대한 과징금

ㅇ 본 법 제101조는 범용 AI 모델 공급자에 대한 과징금 부과를 규정하고 있다. 본 조항은 **고의 또는 과실을 요건**으로 하며, **위반 판단의 범위를 "본 법의 관련 조항 위반(infringed the relevant provisions of this Regulation)"으로 폭넓게 규정**하고 있다. 이를 위반하는 경우, **최대 1,500만 유로 또는 이전 회계연도 전세계 연간 총매출액의 최대 3% 중 높은 금액**의 과징금이 부과된다.

본 법 제5장에서 범용 AI 모델을 별도로 구분하고 있고, 범용 AI 모델 공급자의 의무사항이 제53조 내지 제55조에 규정되어 있음에도 과징금 부과의 대상 조항을 특정하지 않고, "본 법에 관련 조항 위반(infringed the relevant provisions of this Regulation)"으로 포괄적으로 규정한 것은 범용 AI 모델 공급자가 본 법의 어느 조항이든 관련 규정을 위반할 경우, 본 조항의 과징금 대상이 될 수 있다는 의미로 해석될 수 있다.

35) 유럽연합 인공지능법(EU AI Act) 제55조 참조

AI 기술의 발전과 정책 변화에 따라 향후 범용 AI 모델 공급자에 대한 의무가 본 법의 다른 조항으로 추가될 가능성을 열어둔 것으로 보이며, 이는 미래 입법의 유연성을 고려한 설계로 볼 수 있으나 법적 안정성과 예측 가능성을 저하시켜 아래와 같이, 실질적인 법 준수에 어려움을 초래할 수 있다.

- 과징금 대상 "관련 조항(the relevant prvisions)"이 구체적으로 명시되지 않아, 어떤 의무 위반이 과징금 부과 대상에 해당하는 지 사전에 명확히 판단하기 어려움

- 법적 불확실성으로, 본 법 준수를 위한 내부 법무·컴플라이언스 대응에 기술적·경제적 부담 증가

- 향후 구체적인 적용 기준 부재로 사례별로 규제 당국의 해석이 상이할 수 있고, 이에 따라, 과징금 적용 범위의 객관성 확보가 어려워질 수 있음

물론 본 법이 빠르게 변하는 다양한 유형의 AI 관련 규제를 담고 있는 만큼, 과징금 부과 대상을 특정 조항으로 한정하면 새로운 위험 요소 발생 시, 대응이 어려울 수 있고, 법 적용이 경직될 가능성이 높다. 하지만, 법을 준수해야 하는 AI 모델 공급자의 입장에서는 자체적으로 전체 법 조항을 검토하고 적용되는 의무 조항을 스스로 파악해야 한다는 부담이 발생한다.

따라서, 과징금 부과 대상이 되는 조항을 보다 구체적으로 특정하거나, 시행령 또는 세부 지침을 통해 법 적용의 명확성을 높이는 보완 조치가 필요할 것으로 보인다.

5) 거버넌스

ㅇ 본 법 제7장(거버넌스)은 앞서 언급한 범용 AI 모델의 의무이행을 위한 관리·감독 체계를 아래와 같이 유럽연합 차원과 개별 회원국 차원으로 구분하여, 중앙 및 국가 기관 간 역할을 명확히 규정하고 있다.

유럽연합 차원	▪ **유럽연합 AI 이사회(European AI Board)** - 유럽연합 집행위원회와 각 회원국의 자문 역할 수행 - 유럽연합 인공지능법(EU AI Act)의 이행, 특히 범용 AI 모델 규정 집행과 관련하여 자문 제공 - 유럽연합 인공지능법(EU AI Act) 적용 관련 일체에 대한 권고(집행위원회 지침·유럽연합 인공지능법(EU AI Act)의 실천 규범 및 행동 강령 개발 및 적용 등 포함) - 회원국 간 협력 및 국제 협력 등
	▪ **AI 사무국(AI Office)** - 유럽연합 AI 이사회(European AI Board)의 사무국 역할 (회의소집·의제 준비 등) - AI 분야 유럽연합 전문성 및 역량 개발 등
	▪ **자문 포럼(Advisory Forum)** - 산업·스타트업·중소기업·시민사회 및 학계를 포함하여 이해관계자간 균형을 고려해 유럽연합 집행위원회가 선정 - 기본권청(The Fundamental Rights Agency), 유럽 네트워크 정보 보호원(ENISA), 유럽연합 표준화 위원회(CEN), 유럽연합 전자기술 표준화 위원회(CENELEC), 유럽연합 통신 표준 협회(ETSI)는 자문 포럼의 상임 구성원 - 기술적 전문성 제공, 유럽연합 AI 이사회 및 집행위원회에 자문 등
	▪ **독립 전문가 과학 패널(Scientific Panel of Independent Experts)** - AI 분야 과학적·기술적 전문성, AI 시스템·범용 AI 모델 공급자로부터의 독립성 등을 고려해 유럽연합 집행위원회가 선정 - 범용 AI 모델의 시스템적 위험 가능성, 범용 AI 모델과 시스템의 평가 방법론 등 AI 사무국에 자문, 시장 감시 당국 업무 지원 등
개별 회원국 차원	▪ **국가관할 당국(National Competent Authority)** : 회원국은 통지 당국과 시장 감시 당국을 국가 관할 당국으로 지정 - 통지 당국(Notifying Authorities): AI 시스템 적합성 평가 기관통지·관리 - 시장 감시 당국(Market Surveillance Authorities): 시장 감시활동 수행

즉, 유럽연합 차원의 기관은 법의 이행 및 규제 방향 설정을 담당하며, 각 회원국은 이를 국내 시장에서 실행하는 역할을 담당한다. 이를 통해, 유럽연합 차원에서 정책의 일관성을 유지하면서 동시에, 각 회원국의 국가별 실무 적용을 가능하게 하는 거버넌스 구조를 채택하고 있다.

6) 범용 AI 시스템 및 범용 AI 모델 공급자 시장 감시 등

ㅇ 본 법은 AI 모델이 다양한 AI 시스템에서 활용되면서 발생할 수 있는 법적 책임과 안정성을 확보하기 위해 시장 감시 등의 규제 대상을 "범용 AI 시스템(제75조)"과 "범용 AI 모델 공급자(제88~94조)"로 분리하여 규정하고 있다.

- **특정 조건 및 고위험 해당 범용 AI 시스템(제75조)**

본 법 제75조는 범용 AI 모델이 AI 시스템을 구성하면서 발생할 수 있는 법적 책임을 명확히 하고, 특히 주의가 필요한 범용 AI 시스템의 실제 시장 적용을 감시하기 위해, 아래와 같이 특정 조건에 해당하는 범용 AI 시스템의 시장 감시 등을 규정하고 있다.

• AI 시스템이 범용 AI 모델을 기반으로 하면서, 해당 AI 시스템과 모델을 동일한 공급자가 개발한 범용 AI 시스템

▶ **AI 사무국**이 해당 범용 AI 시스템의 본 법 의무 준수 여부를 **감독**

: 범용 AI 모델 공급자가 범용 AI 시스템까지 개발한 경우, AI 모델과 AI 시스템 간의 연계성이 더 강화될 수 있다. 이에 따라, 규제 공백이 발생하지 않도록, AI 시스템을 AI 사무국이 직접 감독할 수 있는 법적 근거를 마련한 것으로 해석된다.

• 최소 하나 이상의 고위험으로 분류되는 목적으로 배포자가 직접 사용할 수 있는 범용 AI 시스템

▶ 고위험 목적의 범용 AI 시스템의 경우, 유럽연합 차원에서 통합적 규제 대응이 이루어지도록 필요 시, **AI 사무국과 시장 감시 당국이 협력**하여 시스템의 법적 요건 준수 평가를 수행하고, 조사 결과를 AI 이사회 및 기타 시장 감시 당국에 보고

▶ 시장 감시 당국이 AI 모델과 관련한 특정 정보에 접근이 불가하여, 고위험 AI 시스템 조사를 완료하지 못할 경우, **AI 사무국에 정보 접근을 강제할 수 있는 요청** 가능, AI 사무국은 30일 내 모든 관련 정보를 제공해야 함. (시장 감시 당국은 제78조에 따라, 획득한 정보에 대한 기밀 보호 의무가 있음)

: 범용 AI 시스템이 고위험 AI 시스템으로 사용될 경우, 철저한 감독과 강제 조사가 필요하다는 점을 반영한 것으로 보인다.

- **범용 AI 모델 공급자의 시장 감시(제88~94조)**

본 법 제88~94조에서 범용 AI 모델 공급자에 대한 감독·조사·집행 및 모니터링에 대해 규정하고 있으며, 아래와 같이 AI 사무국이 유럽연합 집행위원회로부터 권한을 위임(제88조)받아 이를 수행하도록 하고 있다.

· 범용 AI 공급자가 실천 규범을 비롯하여 본 법 이행 여부를 모니터링하고, 시스템적 위험 가능성이 있는 AI 모델에 대한 과학 패널의 경고 접수 시, 이에 대한 평가를 수행

· 본 법 제54조와 제55조(범용 AI 모델 공급자와 시스템적 위험이 있는 범용 AI 모델 공급자의 의무 규정) 준수를 비롯하여, 범용 AI 모델 공급자의 본 법 준수 평가에 필요한 모든 정보 요구 가능

이때 공급자가 부정확하거나 불완전, 오도하는 내용의 정보를 제공하는 경우, 제101조에 규정된 **과징금(최대 1,500만 유로 또는 이전 회계연도 전세계 연간 총매출액의 3% 중 높은 금액)이 부과**[36]됨

· 수집한 정보가 불충분하다고 판단되거나, 과학 패널의 경고에 따라 시스템적 위험이 있는 범용 AI 모델의 시스템적 위험을 조사하기 위해, 범용 AI 모델 평가를 수행할 수 있고, 독립 전문가를 임명하여 평가를 수행하도록 할 수 있음

이때 **AI 사무국이 API 또는 소스코드를 포함한 추가 적절한 기술적 수단을 통해 해당 범용 AI 모델에 접근할 수 있도록 요청할 수 있고, 접근 실패 시에는 제101조에 규정된 벌금(최대 1,500만 유로 또는 이전 회계연도 전세계 연간 총매출액의 3% 중 높은 금액)이 부과**됨.

AI 사무국이 소스코드를 포함한 기술적 수단을 통해 범용 AI 모델 내부에 직접 접근할 수 있도록 요구하는 것은 AI 모델의 투명성을 높이고 법적 준수를 보장하기 위한 조치로 볼 수 있다. 하지만, 이는 기업의 기밀 보호에 문제를 야기할 수 있으며, 국제 무역 규제(WTO TRIPS, TBT 등)와도 충돌할 가능성이 있다. 이에 대해서는, 아래에서 관련 문헌을 통해 보다 구체적으로 살펴보고자 한다.

---

36) 유럽연합 인공지능법(EU AI Act) 제101조제1항(b)목 참조

❑ **주요쟁점 : AI 사무국의 소스코드 등을 통한 범용 AI 모델 접근 요청 관련**

- 세계무역기구(WTO) '무역관련지식재산권에 관한 협정(TRIPS)'[37] 관련

- **제39조제2항(영업비밀 보호)**에 따르면, 개인 또는 법인은 영업비밀이 정당한 상업 관행(honest commercial practices)에 반하는 방식으로 타인에게 공개되거나 취득되거나 사용되지 않도록 보호할 수 있음을 명시하고 있다. 소스코드는 잘못 공개될 경우 기업 생존에 위협이 될 수 있는 중요 기술 정보로 영업비밀에 해당할 수 있는 바, 강제적인 소스코드 공개 요구는 영업 비밀 보호에 반할 수 있다[38].

  따라서, 유럽연합은 취득한 정보 보호를 위한 구체적 절차와 유출 시 배상 규정을 명확히 마련해야 한다. 예를 들어, 정보 접근 시 엄격한 보안 프로토콜을 채택하거나, 정보 누출 시 손해 배상 기준·범위·방식·절차 등을 설정할 필요가 있다.

- 세계무역기구(WTO) '무역기술장벽에 관한 협정(TBT)'[39] 관련

- **제2조제2항(무역 장벽 방지)**에 따르면, 회원국은 기술 규정을 준비·채택·적용 시, 국제무역에 불필요한 장벽을 초래하지 않아야 하고, 정당한 목적 달성을 위해 필요한 범위 내(국가 안보, 인간의 건강·안전보호 등)로 제한하고 있다.

  AI 사무국이 소스코드 등을 통해 범용 AI 모델 내부에 직접 접근할 수 있도록 해야 한다는 조항이 정당한 목적 범위로 인정될 수준을 넘어서 과도하게 적용된다면, 앞서 언급한 기술 노하우나 영업비밀 노출 등의 부담이 유럽연합 시장 진출의 무역 장벽으로 작용[40]해 이 조항을 위배할 가능성이 있다.

  따라서, 유럽연합은 해당 요구 사항이 '정당한 목적' 범위 내임을 명확히 하고, 업계와 상호협력하여 필요한 최소한의 정보만 요구하는 방식으로 조율해야 할 것이다.

---

37) 세계무역기구(WTO), "무역관련 지식재산권에 관한 협정, Agreement on Trade-Related Aspects of Intellectual Property Rights(TRIPS)"(1994), "https://www.wto.org/english/docs_e/legal_e/trips_e.htm#art39."

38) Mitchell, A. D., Let, D., & Tang, L. (2023, December 15). AI Regulation and the Protection of Source Code. International Journal of Law and Information Technology, 31(4), 283. Retrieved from https://academic.oup.com/ijlit/article/31/4/283/7475778?login=false.; World Trade Organization (WTO). (2017). TRIPS Agreement: Article 39 and Other Provisions on the Protection of Undisclosed Information. Retrieved from https://www.wto.org/english/res_e/publications_e/ai17_e/trips_art39_oth.pdf

39) 세계무역기구(WTO), "무역기술장벽에 관한 협정, Agreement on Technical Barriers to Trade(TBT)"(1995), "https://www.wto.org/english/docs_e/legal_e/17-tbt_e.htm"

40) Derechos Digitales. (2023, April 23). Digital Trade Agreements Cannot Prevent AI Transparency. Retrieved from https://www.derechosdigitales.org/22304/digital-trade-agreements-cannot-prevent-ai-transparency/

- **제2조제4항(국제 표준 준수)**에 따르면, 회원국은 국제 표준이 존재하거나 곧 완성될 예정인 경우, 기술 규정에 이를 채택해야 한다. 따라서, 시스템적 위험 여부를 검증할 국제 표준이 존재한다면, AI 사무국이 이를 확인하기 위해 소스코드 등을 통해 범용 AI 모델 내부에 직접 접근할 수 있도록 해야 한다는 것은 본 조항 위반이 될 수 있다. 하지만, 아직 시스템적 위험의 유형과 특성조차 명확히 정립되지 않은 상황이어서, 이에 대한 검증의 국제표준을 논하기도 어렵다.

  본 법은 제56조의 실천 규범에 시스템적 위험의 유형과 특성을 규정하도록 하고 있으므로, 유럽 연합은 이를 통해 시스템적 위험 유형과 특성을 구체적으로 정립하고, 이를 기반으로 국제 표준 개발을 지원해야 할 것이다.

- 세계무역기구(WTO) '서비스무역에 관한 일반협정(GATS)'[41)]

- **제6조(국내규제)**에서는 시장개방을 약속한 분야[42)]에 있어서 각 회원국이 서비스무역과 관련한 모든 조치를 합리적이고 객관적이며 공평한 방식으로 시행해야 한다고 규정하고 있다. 따라서, 소스코드 등을 통한 범용 AI 모델 접근 요구가 무역 당사자들에게 불합리한 부담으로 작용된다면 이 조항에 위배될 수 있다.
- **제16조(시장접근)**에서는 시장개방을 약속한 분야에 있어서 각 회원국이 공급자 수 제한이나 공급자의 형태를 제한하는 등으로 시장 진입을 제한하는 것을 금지하고 있다. 소스코드 등을 통한 범용 AI 모델 접근 요구가 해당 정보를 기밀 정보로 관리하고 있는 공급자에게만 부담으로 적용된다면, 이는 실질적으로 공급자의 수 제한이나 공급자 형태를 제한하는 결과가 될 수 있다.

  따라서, 유럽연합은 규제의 합리성과 공평성을 보장하기 위해 공급자의 규모와 역량을 고려하여 시장 접근 제한 요소를 최소화할 수 있는 방안을 마련해야 할 것이다.

이와 같이, 소스코드 등을 통한 모델 접근 요구는 AI 업계의 부담을 초래하고, 각국의 이해 충돌로 인한 통상 마찰 등으로 이어질 수 있는 바, 유럽연합은 기업의 기술 기밀 보호와 규제의 필요성 간의 균형을 맞출 수 있도록 다양한 의견을 수렴하여 신중히 검토하고 세부사항을 조정할 필요가 있다.

---

41) 세계무역기구(WTO), "서비스무역에 관한 일반협정, General Agreement on Trade in Services (GATS)"(1995), https://www.wto.org/english/docs_e/legal_e/26-gats_01_e.htm

42) 세계무역기구(WTO)의 '서비스무역에 관한 일반협정(GATS)'은 서비스 무역에 대한 기본 원칙과 규제를 설정하여, 공정하고 투명한 무역 환경을 조성하기 위한 다자간 협정으로, 일부 조항(제16조 시장접근, 제17조 내국민대우 등)들은 특정 서비스 부문에 대한 시장 개방 약속과 관련하여 준수해야 할 의무를 규정하고 있음.

7) 기밀유지

ㅇ 본 법 제78조는 집행위원회, 시장 감시 당국 등이 법 이행 과정에서 획득한 소스코드, 지식재산권, 기업 기밀정보, 영업비밀 등의 보안 정보 및 데이터에 대한 기밀성 유지 의무를 규정하고 있다.

ㅇ 이에 따르면, **법 적용 당국은 엄격히 필요한 경우에만 데이터를 요청**해야 하며, **획득한 정보 및 데이터의 보안과 기밀성 유지를 위해 적절한 사이버보안책을 마련**해야 하고, 획득 **목적 달성 후에는 가능한 빨리 데이터를 삭제**해야 한다고 명시하고 있다.

그러나, 이 같은 규정에도 불구하고, 실제 적용 과정에서 아래와 같은 문제가 제기될 수 있다.

- 데이터 요청을 **"엄격히 필요한 경우"**로 제한하고 있으나, 데이터 요청의 적절성을 판단할 기준이 명확하지 않으며, 이에 따라 **기업이 과도한 데이터 요청을 받았을 때 이의를 제기할 수 있는 절차**(이의 제기기관·절차·처리 소요 기간 등)가 구체적으로 제시되지 않았다.
- 법 적용 당국이 수집한 데이터를 보호하기 위해 **"적절한 사이버 보안책"**을 마련해야 한다고 규정하고 있으나, 어떠한 기준에서 보안 조치가 적절한 것으로 판단되는 지에 대한 구체적 내용이 없어, **기밀 정보 제공 의무자가 수용할 수 있는 수준의 보안 대책이 충분히 보장되지 않을 위험**이 있다.
- 획득 목적 달성 후, **가능한 빨리 데이터를 삭제**해야 한다고 규정되어 있으나, **관련 세부사항(폐기 기한·목적 달성 여부 확인 방식·공급자의 폐기요청 가능 여부·폐기 확인 방법 등)이 불명확**하여, 기업이 제공 정보가 적시에 폐기되었는 지 확인하기 어렵다. 이는 기업들에게 불필요한 법적 불확실성을 초래할 수 있다.

이에 따라, 본 기밀 유지 조항이 기업의 **기밀 보호에 대한 법적 강제력을 충분히 갖추지 못하는 경우, 결국 기업이 불리한 입장**에 놓일 가능성이 크다는 우려가 제기[43]되고 있다. 따라서, 기존 유럽연합의 영업비밀 보호

43) Mylly, U.-M. (2023). Transparent AI? Navigating Between Rules on Trade Secrets and Access to Information. IIC - International Review of Intellectual Property and Competition

지침(Trade Secrets Directive)등 관련 법과의 균형을 고려한 세부 지침 마련이 필요하다.

ㅇ 또한, 해당 조항은 국제 및 무역협정 조항에 따라, 적절한 수준의 기밀성을 보장하는 **양자 및 다자간 기밀 협정을 체결한 제3국의 규제 당국과 기밀 정보를 교환**할 수 있도록 규정하고 있다.

그러나, 이러한 정보 공유 조항은 유럽연합의 규제 당국을 신뢰하고 정보를 제출한 기업들에게, 해당 정보가 제3국의 규제 당국과 공유될 가능성이 있다는 우려를 야기할 수 있다. 특히, 제3국의 규제 환경은 유럽연합과 동일한 수준의 기밀 유지 기준을 준수하지 않을 수 있고, 이 같은 경우라면, 기업들은 더 이상 정보 보호를 신뢰할 수 없게 된다.

이에 따라, 해당 규제 내용이 유럽연합 시장 진입에 장애가 되지 않도록 해야 한다는 지적[44]이 제기되고 있다. 따라서, 유럽연합은 해당 조항을 도입한 경위를 설명하고, 이 같은 우려를 해소할 수 있도록 기밀 정보 공유 및 보호 방안과 관련한 세부적인 지침을 마련할 필요가 있을 것이다.

8) 시장 감시 당국에 대한 이의 제기

ㅇ 본 법 제85조는 법 위반 당사자의 시장 감시 당국에 대한 이의 제기를 규정하고 있다. 이에 따르면, 위반사항에 대해 **당사자는 시장 감시 당국에 이의를 제기**할 수 있고, **이는 시장 감시 당국의 전용 절차에 따라 처리**된다.

그러나, 시장 감시 당국은 개별 회원국 차원의 국가 관할 당국에 속하므로, 회원국마다 이의 제기 절차 및 요건이 상이할 가능성이 있다. 이는 특히 제3국 공급자에게 법적 불확실성을 초래할 수 있으므로, 통일된 해석 및 집행 지침이 필요하다는 의견[45]이 제기되고 있다.

Law, 54, 1013–1043. Retrieved from https://doi.org/10.1007/s40319-023-01328-5.
44) Japan Business Council in Europe (JBCE). (2021, August 6). JBCE Views on the European Commission's AI Act. Retrieved from https://www.jbce.org/images/JBCE-comments-on-the-AI-Act-FINAL.pdf.
45) Schröder, C., & Ashkar, D. (2024, October 22). The Artificial Intelligence Act of the

따라서, 제3국 공급자의 법적 안정성을 강화하고, 유럽연합 시장의 규제 일관성을 확보하기 위해, 유럽연합은 개별 회원국 차원의 시장 감시 당국간 절차의 차이점 및 공통점을 분석하고, 이의 제기 절차를 보다 통합적이고 일관된 방식으로 운영할 필요가 있다.

9) 혁신지원 조치

ㅇ 유럽연합은 AI 기술 발전과 규제 준수의 균형을 유지하기 위해, 본 법 제6장(제57조 내지 제63조)에서 혁신 지원 조치를 규정하고 있다. 주요 내용은 "AI 규제 샌드박스(AI Regulatory Sandbox)"를 중심으로, 기업이 규제 준수 의무 및 요건을 시장 진출 전에 검증할 수 있도록 하는 것이다.

본 조항은 중소기업 등에 대한 전반적인 혁신 지원 체계를 규정한 것으로, 본 법의 목적에 중요한 항목이다. 다만, 그 주요 적용 대상이 본 보고서에서 다루는 범용 AI보다는 고위험 AI 시스템이므로, 아래와 같이 중요 내용만 간략히 정리한다.

- 'AI 규제 샌드박스(AI regulatory sandbox)'

- **정의[46] : 규제력이 있는 ①감독**하에 **②실제 환경 조건**에서, **③제한된 시간** 동안, **④샌드박스 계획[47]**에 따라, **⑤혁신적인 AI 시스템을 개발·훈련·검증 및 테스트해 볼 수 있는 기회를 제공**해 주는 것
- **설립목적[48]**

① **법적 확실성** 개선: AI 기업들이 규제 요구사항을 명확히 파악하여, 법적 리스크 최소화하고, 규제 당국과 분쟁 방지 및 규제 환경의 신뢰를 구축하면서 AI 시스템 개발·발전을 촉진

European Union (AI Act). Orrick. Retrieved from https://www.orrick.com/en/Insights/2024/10/The-Artificial-Intelligence-Act-of-the-European-Union.

46) 본 법 제3조 참조

47) 제3조 정의조항 (55) 샌드박스 계획(sandbox plan): 샌드박스에 참여하는 공급자와 권한 당국간의 합의된 문서로, 샌드박스 내에서 수행되는 활동의 목적·조건·기간·방법 및 요건을 기술한 문서

48) 본 법 제57조 참조

② **모범사례** 공유: 기업들의 모범 선례 학습과 상황별 전략 수립 지원, 규제 조항의 효과성을 평가 및 필요 시 규제 방침 조정

③ **혁신·경쟁력·AI 생태계 발전** 지원: 기업들에게 자유로운 기술 실험·개발 환경을 제공하여 시장 경쟁력 향상, AI 생태계 발전 지원

④ **증거기반 규제 학습** 기여: 실제 데이터와 경험 수집을 통해 효과적인 규제 방안을 학습·개발하여, 안전하고 효과적인 기술 도입 지원

⑤ **스타트업·중소기업 지원**: 법적·행정적 체계 부족으로 규제 준수·기술 개발이 어려운 스타트업·중소기업을 대상으로 기술 개발·시장 진입 지원

• **운영 및 참여 혜택·요건**

① **회원국 의무[49] :** 각 회원국은 2026.8.2.까지 최소 1개 이상의 AI 규제 샌드박스를 운영해야 함.

② **참여 혜택:**

- **적합성 평가:** AI 시스템 공급자는 규제 샌드박스 활동 종료 후, 각 회원국의 관할 당국이 제공하는 종료 보고서로 적합성 평가 절차 등의 준수를 입증 가능
- **법적 부담 완화:** 구체적 계획 및 참여 조건을 준수하고, 당국의 지침을 충실히 따른 경우, 법 위반에 대한 과징금 면제

③ **참여 요건[50]:** AI 규제 샌드박스의 참여 자격·선정 기준, 운영·종료 절차 등에 세부사항 및 원칙은 유럽연합 집행위원회가 채택하는 시행령에서 규정, 신청 후, 결과 통보까지 최대 3개월 소요

- 스타트업·중소기업 지원[51]

• **샌드박스 접근 우선권 :** 유럽연합 내 사무소·지점이 소재한 스타트업·중소기업이 자격조건과 선정기준 충족 시, 샌드박스 참여 우선권 제공 (단, 타 중소기업 등을 배제하는 것은 아님)

• **전용 소통 채널 :** 스타트업·중소기업 대상 질의·응답·자문 제공을 위한 소통 채널 구축 가능(기존 전용 채널 활용 및 신설)

• **적합성 평가 비용[52] 감면 :** 기업 규모에 따라 수수료 감면 가능

49) 본 법 제57조 참조
50) 본 법 제58조 참조
51) 본 법 제62조 참조

ㅇ 지금까지 최초의 포괄적 인공지능(AI) 규제법이라고 할 수 있는 유럽연합 인공지능법(AI Act)을 문화콘텐츠 분야에서 주로 사용되는 범용 AI를 중심으로 살펴보았다. 이어서, 글로벌 인공지능(AI) 기업을 다수 보유하고 있는 미국의 주요 AI 윤리지침을 살펴보고, 이를 유럽연합의 규제와 비교·분석해 보고자 한다.

52) 본 법 제43조 참조

2. 미국의 주요 AI 윤리지침

ㅇ 미국의 AI 윤리지침은 **AI 권리장전 청사진(Blueprint for an AI Bill of Rights)**을 통해 **기본권 보호 원칙과 실천방향을 제시**하고, 이를 기반으로 **NIST AI 위험 관리 프레임워크(NIST AI Risk Management Framework)**를 통해 AI 시스템 개발·사용의 **위험 관리를 위한 실제적인 기술적 지침**을 제공해 주고 있다.

이와 더불어, **행정명령 제14110호**('23.10.30. 제정, '23.11.01. 공포, Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence-안전하고 신뢰할 수 있는 인공지능 개발 및 사용)를 통해 **미국 AI 윤리지침 정책 및 규제의 방향성**을 확인할 수 있다.

(1) AI 권리장전 청사진(Blueprint for an AI Bill of Rights)

ㅇ 미국 AI 권리장전 청사진(The Blueprint for an AI Bill of Rights, 이하 'AI 권리장전'이라 함)은 2022년 10월 미국 백악관 과학기술정책실(OSTP)에서 발표한 백서로, 법적 구속력은 없지만[53], 기존 법률이나 정책으로 명확한 지침을 제공할 수 없는 경우, 정책 결정에 활용하기 위한 자료로 기능한다.

과학기술정책실(OSTP)은 1년여간의 공공 의견 수렴과 연구자, 기술 전문가, 시민사회, 정책 결정자 등의 참여를 거쳐 AI 권리장전을 마련하였다. 이 과정에서 AI 시스템이 초래할 수 있는 차별, 프라이버시 침해, 무분별한 데이터 및 활동기록 수집 등 다양한 문제들을 논의하였으며, 이러한 문제를 방지하고, 시민의 권리를 보호하기 위해 5개 원칙을 설정하였다.

ㅇ 전체 구조는 ①서문, ②5가지 원칙, ③AI 권리장전의 적용, ④정부와 기업이 실천 가능한 구체적 절차를 담은 기술적 보충 자료로 구성되어 있다.

---

53) 본 백서의 법적 고지(legal disclaimer) 부분에서 구속력이 없고, 기존 법령을 대체하는 것이 아니며, 국제협상에서 미국의 입장을 결정하는 것이 아니라고 밝히고 있음. (출처: 백악관 과학기술정책실(OSTP), "What is the Blueprint for an AI Bill of Rights?" (2022.10.4.), https://www.whitehouse.gov/ostp/ai-bill-of-rights/what-is-the-blueprint-for-an-ai-bill-of-rights/)

본 항(Subsection)에서는 문서 분석(Document Analysis), 내용 분석(Content Analysis), 정책 분석(Polocy Analysis)을 통해, AI 권리장전의 주요 내용과 체계를 정리하고, 각 원칙의 의미와 실질적 적용 가능성 및 기존 법률과의 연계성을 검토하였다.

1) AI 권리장전의 5가지 원칙(The five Principles)

① 안전하고 효과적인 시스템(Safe And Effective System)
• 예측 가능한 비의도적인 사용이나 영향으로 인한 피해를 사전 대비하도록 설계 • 설계·개발·배포 과정에 부적절하거나 관련 없는 데이터가 사용되지 않도록 보호, 이와 같은 데이터의 재사용으로 인한 복합적 피해 보호 • 사전 배포 테스트, 위험 식별 및 완화, 지속적인 모니터링 시행 • 사용 의도, 이를 벗어난 안전하지 않은 결과 완화, 특정 분야 표준 준수를 기준으로 안전성과 효과성 입증 • 독립적 평가 수행 후, 보고, 가능하면 이를 대중에게 공개
② 알고리즘 차별 보호(Algorithmic Discrimination Protections)
• 인종, 종교, 국적, 장애 등 기타 법이 보호하는 분류에 따른 차별 보호 • 사전 공정성 평가, 대표적 데이터 사용 및 인구 통계적 특징에 대한 대리변수 보호, 장애인 접근성 보장, 불균형 테스트 및 완화 조치 등 포함 • 독립적 평가 및 보고, 가능하면 대중에게 공개
③ 데이터 프라이버시(Data Privacy)
• 프라이버시 침해 보호 및 데이터 사용에 대한 개인의 자율권 보장 • 프라이버시 보호 선택을 기본값으로 설정, 필요 데이터만 수집, 개인의 동의 및 자율성 보장, 무분별한 감시 방지
④ 고지 및 설명(Notice And Explanation)
• 자동화 시스템 사용을 개인이 인지하고, 시스템 결정을 이해할 수 있어야 함 • 시스템 기능에 대한 문서화, 시스템의 역할, 시스템 사용 고지, 평이한 용어로 기술한 설명 문서 제공
⑤ 인간 대안 및 검토, 대책 (Human Alternatives, Consideration, and Fallback)
• 사용자가 시스템 외 인간이나 다른 대안을 선택할 수 있도록 보장 • 자동화 시스템이 작동하지 않거나 오류 등이 발생 시, 또는 항의·이의 제기 시, 사용자가 인간의 검토와 시정 조치를 받을 수 있도록 해야 함.

2) AI 권리장전의 적용(Applying The Blueprint For An AI Bill of Rights)

(i) 적용 대상

ㅇ AI 권리장전의 적용 대상은 ①자동화 시스템, ②중요 자원이나 서비스에 대한 미국 대중의 권리, 기회 및 접근*에 유의미한 영향을 끼칠 가능성이 있는 것의 2가지를 요건으로 한다.

*** 권리, 기회 또는 접근(Rights, Opportunities, or Access)**
- **시민권, 시민 자유 및 프라이버시:** 언론의 자유, 선거권, 차별·과도한 처벌·불법 감시·공공 및 민간 부문에서의 프라이버시 및 기타 자유 침해에 대한 보호 등
- **동등한 기회:** 교육·주거·신용·고용 및 기타 프로그램에 대한 공평한 접근 등
- **중요 자원 또는 서비스에 대한 접근:** 의료, 금융서비스, 안전, 사회 서비스, 상품 및 서비스에 대한 비기만적 정보, 정부 혜택 등

(ii) 기존 법·정책과의 관계

① 기존 법률과의 연계
• AI 권리장전은 자동화 시스템의 잠재적 피해로부터 미국민을 보호하고, 그 혜택을 보장하기 위한 원칙을 제시한다. 일부 보호 조치는 아래와 같이 미국 헌법과 기존 법률에 의해 시행되고 있다. - 정부 감시 및 데이터 수집: 법적 요건과 사법적 감독 대상 - 범죄 수사에 대한 인간 검토: 헌법적 요건, 사법적 검토 요건 - 미국민 차별 보호: 시민권법(Civil rights laws)
② 추가 보호 조치 및 조정
• AI 권리장전에서 제안하는 추가 보호를 보장하기 위해서는 새로운 법률 제정이나 정책 변경이 필요할 수 있다. 일부 경우, 기존 법과의 충돌을 고려하여 예외나 조정이 필요할 수 있다. - 신규 법률 및 정책 필요성: AI 시스템 발전으로 기존 법률로 불충분한 경우 - 법적 예외 필요성: 기존 법 준수·특정 상황에서의 실용성·공공 이익과 균형 고려 - 대체 매커니즘 필요성: 법 집행 및 기타 규제 맥락에서 대안적 매커니즘을 통한 시민권 등 보호 필요성
③ 특정 맥락을 고려한 청사진 제시
• 기술적 보충 자료(Technical Companion)는 특정 맥락에 맞게 조정이 필요한 추가적 기술 표준 및 실천 사항 개발을 위한 청사진 역할을 한다. - 법률과의 차이: AI 권리장전은 기존 법률을 구체적으로 다루지 않는 대신, 기존 보호 조치의 사례와 자동화 시스템 개발·운영의 미래 비전을 제시 - 공공 및 민간 부문의 활용: 시·주·연방정부·외국의 법과 규제 제안을 분석하거나, 이에 대한 입장을 제시하는 대신, 민간 및 공공의 참여 시 정보 제공

④ 최근 법·정책 변화 및 국제·연방 차원의 연계
• AI 관련 법·정책 환경 변화에 따라 아래와 같이 법 제정 및 기존 법적 보호 확장, 국제 및 연방 차원의 정책과 조화 등 점진적 진전이 이루어지고 있다. - 주 및 지방 정부 입법: AI 관련 법률 제정 - 법원의 법적 보호 확장: 기존 법적 보호를 신규 기술로 확장 - 기업 및 연구자들의 대응: 혁신적인 보호 조치 개발, 윤리적 사용 원칙 제안 - OECD AI 권고: 2019년 AI 권고를 통해 AI의 책임 있는 관리 원칙 제시 - 행정명령 제13960호: 연방정부 내 신뢰할 수 있는 AI 사용 촉진에 대한 행정명령, 연방정부의 AI 사용을 관장하는 원칙 제시 - 행정명령 제13985호: 연방정부를 통한 인종 평등과 소외된 지역사회 지원에 대한 행정명령, AI 권리장전과 일치 - 공정 정보 관행 원칙(FIPPs): 1973년 미국 보건교육복지부 자문위원회 보고서, AI 권리장전은 공정 정보 관행의 요소를 포괄하며, 해당 원칙 실행을 위한 방향성 제시

3) AI 권리장전의 기술적 보충자료(A Technical Companion to The Blueprint For An AI Bill of Rights)

ㅇ AI 권리장전의 기술적 보충자료(Technical Companion)는 업계·정부 등 관련자들이 개별 원칙 보호를 정책 및 실천·기술 설계 과정에서 구축할 수 있도록, 아래와 같이 구현 기술과 예시를 제시하고 있다.

① 안전하고 효과적인 시스템(Safe And Effective System)	
원칙의 중요성	• 의도하지 않은 부적절한 사용으로 사회적 위험 초래 (예) 인종차별 게시물을 비판하는 메시지를 게시물 원본과 혼동하여 삭제, AI 기반 위치 추적 장치가 스토킹에 악용되는 사례 등
기대되는 기술구현	• 세부 분야별 전문가 자문 테스팅, 위험 식별 및 완화 조치, 지속적인 모니터링, 명확한 거버너스 구축, 파생 데이터 출처 추적 및 검토, 민감 영역에서의 데이터 재사용 제한, 독립적 평가 등
적용 사례	• 연방정부의 신뢰할 수 있는 인공지능 사용 촉진에 관한 행정명령 13960, 기업의 AI 리스크 관리 솔루션, 국립표준기술연구소(NIST)의 AI 리스크 관리 프레임워크 등
② 알고리즘 차별 보호(Algorithmic Discrimination Protections)	
원칙의 중요성	• 특정 인구 집단 등에 불평등한 결과를 생산·증폭 시킬 위험 (예) 회사 채용툴이 대다수가 남성인 직원 특징을 학습 후, 차별적인 이유로 여성 지원자를 탈락, 광고 클릭 가능성 예측 시스템이 결과적으로 인종·성 고정관념을 강화 – 마트 캐셔 광고를 여성에게, 택시 일자리를 흑인에게 주로 전달

기대되는 기술구현	• 설계 단계에 공정성에 대한 사전평가(양적·질적), 대표성이 있는 견고한 데이터, 프록시(proxies) 경계, 설계·개발·배포 시, 접근성 보장, 격차 평가(Disparity Assessment), 격차완화(Disparity Mitigation), 지속적인 모니터링, 독립적 평가, 알고리즘 영향 평가 보고·공개(명확하고, 기계 해독 가능한 방식, 평이한 용어 사용으로 책임성 확보) 등
적용 사례	• 연방정부의 주책담보대출 차별 방지 조치, 격차평가를 통한 흑인 환자의 의료 접근성 피해 식별, 대기업 채용 시 모범사례 개발, 표준화 기구 기술 설계 프로세스에 접근성 기준 통합 지침 개발, NIST 인공지능 편견 식별·관리 표준에 대한 특별 간행물 1270 발표
③ 데이터 프라이버시(Data Privacy)	
원칙의 중요성	• 과도하고 무분별한 데이터 수집·오용 등으로 프라이버시 침해 가능성 증가 (예) 보험사가 개인 소셜 미디어 활동 데이터 수집하여 생명 보험료 결정 과정에 사용, 데이터 브로커가 대량 수집한 개인정보에 보안 침해로 수십만 명이 신원 도용 위험에 노출, 기업들이 감시 소프트웨어를 사용하여 노조 활동 추적·감시하는 사례 등
기대되는 기술구현	• 프라이버시 보호를 기본 설정으로 설계·구축, 데이터 수집·사용 범위 제한, 위험 식별 및 완화 조치, 암호화 등 프라이버시 보호 보안, 무분별한 감시에 대한 감독 강화, 제한적·필요에 비례한 감시 및 감시 범위 제한, 구체적·간결하고 직접적인 사용자 동의 요청(동의 거부 허용), 본인 데이터에 대한 접근 및 정정·동의 철회 및 데이터 삭제 절차 마련, 민감 데이터·파생 데이터 접근 제한, 독립적인 평가, 보고·공개(명확하고, 기계 해독 가능한 방식, 평이한 용어 사용) 등
적용 사례	• 프라이버시법(The Privacy Act of 1974), NIST 프라이버시 프레임워크, 스마트폰 개인정보 선택 등
④ 고지 및 설명(Notice And Explanation)	
원칙의 중요성	• 시스템의 의사 결정 과정과 결과를 알지 못해 오류 정정·이의 제기 등 대처에 필요한 정보를 얻지 못할 가능성 (예) 새로운 알고리즘 도입이 고지·설명되지 않아 메디케이드 자금 지원 가정의 건강 관리 서비스가 중단된 후, 적시에 설명·이의 제기에 어려움을 겪은 사례, 부모를 대상으로 한 알고리즘 기반 아동 복지 조사 개시가 고지·설명 부족으로 평가 검증이 어려워지고, 이의제기 정보를 제공받지 못한 사례
기대되는 기술구현	• 일반적으로 접근 가능한 평이한 언어로 작성된 문서 공개, 책임 소재 식별, 적시의 최신 정보, 간결성·명확성, 목적·대상·위험 수준에 맞게 조정, 보고(명확하고, 기계 해독 가능한 방식, 평이한 용어 사용) 등
적용 사례	• 비영리 단체·기업 등이 협력하여 머신러닝 시스템 투명성 프레임워크를 개발하고 대중과 소통, 연방법에 따른 대출기관의 특정 결정에 대한 소비자 통지 의무, 연방 정부의 설명 가능한 AI 시스템 연구 수행 및 지원

⑤ 인간 대안 및 검토, 대책 (Human Alternativies, Consideration, and Fallback)	
원칙의 중요성	• 자동화 시스템이 기능하지 못하는 상황 발생 시, 인간 검토를 통해 보호받을 필요성 (예) 실업 보험 급여 대상 사기 탐지 시스템의 항목 표기 오류 발생 시, 인간 검토 부재로 급여 보류·세금 환급 압류, 병원 소프트웨어가 환자 약물 이력을 개인 이력과 혼동하여 통증 완화제 거부 후 당사자 설명에도 시스템 의존으로 통증 완화제 미처방, 기업 인사 기능 자동화 후, 인간 검토·기타 구제 수단 부재로 해고된 사례 등
기대되는 기술구현	• 간결하고, 명확하고, 접근 가능한 통지 및 지침, 적시에 인간 대안 선택 가능한 메커니즘 제공, 비례성·접근 가능성·편의성·공정성·적시성·효과성·유지성을 반영한 인간 검토 및 구제 방안 제공, 민감분야 관련 추가적 인간 감독·안전장치 구현 등
적용 사례	• 고객 서비스 산업의 챗봇·AI 기반 전화 응답 자동화 서비스와 인간 지원팀 통합 등

ㅇ 이와 같이, AI 권리 장전은 기술적 보충자료(Technical Companion)를 통해 AI 시스템의 안전성과 윤리성을 보전하면서 법적·정책적 환경 변화에 대응하기 위한 지침을 제시하고, 이를 통해 기술 혁신과 개인 보호의 균형을 맞추는 접근 방식을 지원하고 있다.

(2) NIST AI 리스크 관리 프레임워크(Risk Management Framework)

ㅇ NIST[54] AI 리스크 관리 프레임워크(Risk Management Framework)는 AI 권리장전의 원칙을 실행 가능한 위험 관리 체계로 구체화하기 위해 국가 인공지능 구상법 2020(National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283))에 따라 개발된 가이드 라인이다.

NIST는 OECD AI 권고안, ISO/IEC 표준, AI 위험 평가 사례 등을 검토하여, 본 프레임워크를 시스템 설계·개발· 배포·운영하는 모든 조직이 활용할 수 있도록 유연하게 설계하였다.

ㅇ 본 항(Subsection)에서는 문서 분석(Document Analysis) 및 내용 분석(Content Analysis)을 통해 본 프레임워크의 핵심 개념과 구조를 파악하고, AI 위험 요소와 구체적 대응 지침을 검토하였다.

1) 파트 1: 기본 정보(Foundation Information)

ㅇ 본 프레임워크의 본문은 파트 1(기본정보, Foundation Information)과 파트 2(핵심기능 및 실행방안, Core and Profiles)으로 구성되어 있다.

파트 1에서는 **①AI 리스크의 구조, ②대상청중, ③AI 리스크 및 신뢰성 분석**을 통해 AI 리스크의 개념과 주요 관리 원칙을 다루고, **파트 2**에서는 위험 관리를 실질적으로 수행하기 위한 **4가지 핵심 기능(①거버넌스, ②맵핑, ③측정, ④관리)과 기능별 구체적 실행 방안**을 제시한다.

(i) AI 리스크의 구조(Framing Risk)

i) 리스크 개념

- 발생 가능한 사건의 영향이 부정적일 때, 리스크는 "발생 사건(환경)으로 인한 결과 값"과 "사건의 발생 가능성"의 함수로 정의(리스크 관리 과정은 일반적으로 부정적 영향을 다룸)

54) NIST(National Institute of Standards and Technology:)미국 국립표준기술원, 1901년에 설립된 미국 상무성 산하 비규제 정부연구기관으로 측정 표준을 촉진하고 소급성을 유지하며 측정표준의 CRM(인증표준물질) 등의 표준을 개발하는 측정표준 대표기관 (출처: 산업통상자원부 지식경제용어사전)

- 본 프레임워크는 AI 시스템의 부정적 효과를 최소화하고, 긍정적 영향을 최대화할 수 있는 기회를 식별하는 접근법을 제시

ii) 리스크 관리의 주요 과제

① 위험 측정(Risk Measurement)

- **제3자 소프트웨어·하드웨어·데이터 관련 리스크**: ▲리스크 지표 및 방법론의 일관성 부족(제3자 소프트웨어·하드웨어·데이터 자체의 리스크), ▲고객 사용 방식에 따른 복잡성(제3자 소프트웨어, 하드웨어, 데이터 사용 방식에 따른 리스크)
- **신생 리스크 추적**: 특정 맥락에서 AI 시스템의 신생 리스크에 대한 잠재적 영향 평가 접근법 필요
- **신뢰할 수 있는 측정 지표의 부재**: AI 리스크 및 신뢰성 측정 방법에 대한 합의 및 다양한 AI 사용 사례의 적용 가능성 부족
- **AI 생애 주기별 리스크 상이**: ▲리스크의 변화 가능성, ▲AI 관련 행위자간 리스크 관점 차이 및 공동 책임 필요성
- **실험실 측정값과 현실 환경의 리스크 상이 가능성**
- **AI 시스템의 불가해성**: ▲AI 시스템의 불투명성, ▲개발·배포 과정의 투명성 또는 문서화 부족, ▲AI 시스템의 내재적 불확실성
- **인간 기준선**: AI와 인간의 수행 업무 및 수행 방식의 차이로 비교 곤란

② 위험 감수 수준(Risk Tolerance)

- 조직 또는 AI 행위자가 목표 달성을 위해 허용 가능한 리스크 수준 (특정 조직의 우선순위·특정 사용 사례·맥락·법적 및 규범적 요건에 따라 상이)

③ 위험 우선순위 결정(Riks Prioritization)

- 리스크 수준 평가 및 잠재적 영향에 기반해 우선순위 결정
- 인간과 상호 작용에 기반해 우선순위 차등
- 잔여 리스크 문서화

④ 위험의 조직적 통합·관리(Organizational Integration and Management of Risk)

- 사이버 보안·개인정보보호 등 기타 주요 리스크와 통합적으로 처리
- 조직 내 책임 체계·역할 및 책임·문화·인센티브 구조를 확립 및 유지

(ii) 대상 청중(Audience)

ㅇ 본 프레임워크는 AI 라이프사이클 전반의 AI 행위자를 대상으로 하며, 각 단계별 주요 활동 및 대상을 아래와 같이 분류하여 설명하고 있다.

라이프 사이클 단계	주요 활동	주요 대상
① 기획·설계	• 법적·규제적 요건 및 윤리적 고려사항에 비추어 시스템의 개념과 목표 기본적인 가정 및 맥락을 명확히 하고 문서화	• 시스템 운영자, 최종사용자, 세부 전문가, AI 설계자, 영향 평가자, TEVV(테스트, 평가, 검증, 유효성 확인) 전문가, 제품 관리자, 컴플라이언스 전문가, 감사, 거버넌스 전문가, 조직 관리자, 최고 경영진, 영향을 받는 개인 및 커뮤니티 평가자
② 데이터 수집 및 처리	• 데이터를 수집·검증·정제하고, 데이터 세트의 메타데이터 및 특성을 목표·법적 요건·윤리적 고려사항에 비추어 문서화	• 데이터 과학자, 데이터 엔지니어, 데이터 제공자, 세부 분야 전문가, 사회문화 분석가, 인간적 요소 관련 전문가, TEVV 전문가
③ 모델 구축 및 사용	• 알고리즘 생성 또는 선택, 모델 훈련	• 모델러(modelers), 모델 엔지니어, 데이터 과학자, 개발자, 세부 전문가(적용 맥락을 잘 아는 사회 문화 분석가 및 TEVV 전문가의 자문 포함)
④ 검증 및 유효성 확인	• 모델 출력의 검증 및 유효성 확인·조정·해석	
⑤ 배포 및 사용	• 파일럿 실행, 기존 시스템과 호환성 체크, 규제 준수 확인, 조직 변화 관리, 사용자 경험 평가	• 시스템 통합자, 개발자, 시스템 엔지니어, 소프트웨어 엔지니어, 세부 전문가, 조달 전문가, 제3자 공급자, 최고 경영진(인간적 요소 관련 전문가, 사회문화 분석가, 거버넌스 전문가, TEVV 전문가 자문 포함)
⑥ 운영 및 모니터	• AI 시스템을 운영하고, 목표·법적 및 규제적 요건·윤리적 고려사항에 비추어 의도된 영향과 의도되지 않은 영향 및 AI 시스템의 추천을 지속적으로 평가	• 시스템 운영자, 최종 사용자, 실무자, 전문가, AI 설계자, 영향 평가자, TEVV 전문가, 시스템 자금 제공자, 제품 관리자, 컴플라이언스 전문가, 감사, 거버넌스 전문가, 조직 관리자, 영향을 받는 개인/커뮤니티, 평가자
⑦ 사용 또는 영향을 받는 단계	• 시스템/기술 사용, 영향 모니터 및 평가, 영향 완화 모색, 권리 옹호	• 최종 사용자, 운영자 실무자, 영향을 받는 개인 및 커뮤니티, 일반 대중, 정책 입안자, 표준화 조직, 무역 협회, 권리 옹호 단체, 환경 단체, 시민 사회 조직, 연구원

(iii) AI 리스크 및 신뢰성(AI Risks and Trustworthiness)

ㅇ 신뢰할 수 있는 AI 시스템의 특성에는 ①유효성 및 신뢰성, ②안전성, ③보안성 및 회복성, ④책임성 및 투명성, ⑤설명 가능성 및 해석 가능성, ⑥프라이버시 강화, ⑦유해한 편향이 관리된 공정성이 포함된다.

이 특성들은 서로 상충될 수 있으므로, 특성 간 균형을 찾아서 우선순위를 조정해야 한다. (예: 프라이버시 강화는 데이터 사용 제한으로 정확성이 저하될 수 있고, 설명 가능성을 높이면 성능이 저하될 수 있음)

①유효성 및 신뢰성(Valid & Reliance)	•배포된 AI 시스템의 유효성 및 신뢰성은 지속적인 테스트 및 모니터링을 통해 시스템이 의도된 성능을 수행하는지 평가 •유효성(Validation)·신뢰성(Reliability)·정확성(Accuracy)·견고성(Robustness)의 측정은 AI 시스템 신뢰성(trustworthiness)을 향상하는데 기여하며, 오류를 감지하거나 수정 불가 경우에는 인간의 개입이 필요 - 유효성(Validation): 의도된 목적의 부합성 확인 과정, 특정 요구사항(정확도, 기능적 요구 등) 충족 여부를 객관적 증거로 확인 - 신뢰성(Reliability): 전체 수명 주기 동안의 신뢰성 평가를 위한 지속적인 검토 과정, 의도된 사용 조건에서 특정 시간 동안 실패 없이 작동 가능한 안정성과 일관성 평가 - 정확도(Accuracy): 관찰·계산·추정의 결과가 실제 값 또는 실제 받아들여지는 값에 가까운 정도 - 견고성(Robustness) 또는 일반화 가능성(Generaliability): 다양한 상황에서 성능 수준을 유지하는 능력,
②안전성(Safe)	•AI 시스템은 "정의된 조건에서 인간의 생명·건강·재산 또는 환경을 위험에 빠뜨리는 상태를 초래해서는 안됨(출처: ISO/IEC TS 5723:2022) •AI 시스템의 안전한 운영 방법 - 책임있는 설계·개발 및 배포 관행 - 배포자에게 시스템의 책임 있는 사용에 대한 명확한 정보 제공 - 배포자 및 최종 사용자의 책임 있는 의사 결정 - 사건의 경험적 증거에 기반한 위험 설명 및 문서화
③보안성 및 회복성(Secure and Resilient)	•보안성: 회복성을 포함하는 개념, 공격을 회피·방어·대응 또는 복구할 수 있는 프로토콜(인증절차, 방화벽, 침입 탐지 시스템, 데이터 백업 및 복원 등)을 포함 •회복성: 예상하지 못한 부정적 사건 발생 후, 정상 기능으로 돌아갈 수 있는 능력, 모델이나 데이터를 예상하지 못하게 또는 적대적으로 사용하는 경우(남용 또는 오용)까지 포괄하여 회복력이 유지되도록 해야 함.

④책임성 및 투명성(Accountable and Transparent)	•신뢰할 수 있는 AI는 책임성(Accountability)에 의존하고, 책임성은 투명성(Transparency)을 전제로 함. - 투명성: AI 시스템과 그 출력 정보 제공 정도 - 책임성: AI 행위자의 역할을 고려하고, 리스크 관리 같이 손해 감소를 위한 조직적 관행과 거버넌스 구조 유지를 통해 지원
⑤설명 가능성 및 해석 가능성(Explainable and Interpretable)	•투명성·설명가능성·해석가능성은 상호 보완적 신뢰성 향상 개념 - 투명성(Transparency): 시스템 내부에서 어떤 일이 일어났는 지 기록하고 공개(What happened in the system) 예) 대출시스템의 경우, "입력값: '신용점수 500, 소득 40,000 달러, 부채비율 45%'; 출력값: '대출거절') - 설명가능성(Explainability): 결과값을 도출한 논리적·기술적 과정을 설명(How a decision was made in the system) 예) 대출시스템의 경우, 신용점수 400 미만이면 대출 승인에 부정적으로 평가되고, 부정적 평가에 소득 대비 부채 비율이 40% 초과 시 대출 승인 불가; 신용점수와 부채 비율 기준으로 '대출거절' 결정 - 해석가능성(Interpretability): 결과가 왜 그렇게 나왔는지, 사용자 맥락에서 설명(Why a decision was made by the system) 예) 대출시스템의 경우, '대출거절'은 신용 점수에서 부정적 평가를 받았고, 부정적 평가를 받은 대상의 소득 대비 부채 비율이 대출 승인 불가에 해당된 결과임
⑥프라이버시 강화(Privacy-Enhanced)	•익명성·기밀성·통제권과 같은 프라이버시 가치는 AI 시스템의 설계·개발·배포과정의 주요 결정에 기준이 되어야 하며, 보안성·편향성·투명성 등 타 특성들과 절충 필요(데이터 희소성 같은 특정 조건에서 프라이버시 강화는 정확도 저하를 초래 가능) - 프라이버시(Privacy): 인간 자율성·정체성·존엄성 보호에 도움을 주는 규범·관행(침해로부터의 자유·관찰의 제한·개인의 정체성 관련 정보의 공개·통제에 대한 동의 여부 결정 권한을 다룸)
⑦유해한 편향이 관리된 공정성(Fair – with Harmful Bias Managed)	•공정성(Fairness): 해로운 편견이나 차별과 같은 문제들을 처리하면서 평등(equality)과 형평성(equity)에 대한 우려를 다룸 •편향(Bias): 인구통계학적 균형 및 데이터 대표성에 국한되지 않으며, 사회의 공정성 및 투명성 개념과 밀접히 연관됨 - NIST에서 구분한 AI 편향 주요 범주 ❶ 시스템적 편향(systemic bias) : AI 데이터셋, 생애주기 전반에 걸친 조직의 규범·관행·프로세스, AI 시스템이 사용되는 넓은 범위의 사회적 맥락에서 발생 ❷ 계산적 및 통계적 편향(computational and statistical bias) : AI 데이터셋과 알고리즘 처리 과정에서 발생, 특히, 대표성이 없는 샘플로 인한 시스템적 오류에서 발생 ❸ 인간-인지적 편향(human-cognitive bias) : 개인이나 집단이 AI 시스템 정보를 인식해서 결정을 내리거나 누락된 정보를 채워넣는 방식 또는 인간이 AI 시스템의 목적과 기능에 대해 어떻게 생각하는 지와 관련됨, AI 라이프사이클 및 시스템 사용 전반의 의사결정 과정에 나타남.

2) 파트 2: 핵심 기능 및 실행 방안(Core and Profiles)

○ **파트 2**에서는 AI 리스크 관리를 실질적으로 적용할 수 있도록 **4가지 핵심 기능(①거버넌스, ②맵핑, ③측정, ④관리)**과 각 기능을 하위 범주로 세분화하여 구체적 실행 방안을 제시한다. 이는 순서대로 따라야 하는 절차가 아니고, 유연하게 조정 가능하다고 설명하고 있다.

가버넌스 (Govern)	• 리스크 관리 문화 조성 및 유지 ※ 교차적 기능으로 다른 3가지 기능 전반에 통합될 수 있음
맵핑(Map)	• 맥락 인식 및 해당 맥락 관련 리스크 식별
측정(Measure)	• 식별된 리스크를 평가·분석·추적
관리(Manage)	• 리스크 우선순위 책정 및 예상 영향에 기반해 조치

○ 즉, 각 기능별 실행방안인 프로파일은 사용자 각자의 요구사항·리스크 허용 수준·자원 등을 고려하여 아래 예시*와 같이 특정 상황에 맞게 유연하게 적용할 수 있도록 각 기능의 범위 및 하위범위의 기본 구조와 지침을 제공하는 것을 목적으로 한다.

*예) ▲ (대상에 따라) 채용·공정 주택 등의 사용 사례 프로파일,
▲ (시간적 흐름에 따라) 현재 프로파일(Current Profile), 목표 프로파일(Target Profile)

(i) 거버넌스(Govern)

기능의 범위	하위범위
• **거버넌스 1:** AI 리스크를 맴핑·측정·관리하는 것과 관련한 조직 전반의 **정책, 프로세스, 절차 및 관행을 수립**하고, 투명하고 효과적으로 시행	•1.1: AI 관련 법·규제적 요건을 이해·관리·문서화 •1.2: 신뢰할 수 있는 AI의 특성을 조직 정책, 프로세스, 절차 및 관행에 통합 •1.3: 조직의 리스크 허용 범위 기반 리스크 관리의 필요 수준 결정을 위한 프로세스·절차·관행 수립 •1.4: 조직 리스크 우선순위 기반 투명한 정책·절차·기타 통제를 통해 리스크 관리 프로세스와 결과 확립 •1.5: 리스크 관리 프로세스의 지속적인 모니터링 및 정기적 검토 기획·정기적 검토 주기 결정을 포함해 조직적 역할 및 책임 명확히 정의 •1.6: AI 시스템을 목록화하기 위한 매커니즘 확립 및 조직 리스크 우선순위에 따른 자원 제공 •1.7: AI 시스템을 안전하고, 리스크를 증가시키지 않고 조직의 신뢰성을 저하시키지 않는 방식으로 해체 또는 단계적으로 폐기하기 위한 프로세스 및 절차 확립

• **거버넌스 2:** 팀과 개인에 AI 리스크의 맵핑·측정·관리를 위한 훈련·책임·권한을 적절히 부여하기 위해 **책임성 구조 확립**	•2.1: AI 리스크 맵핑·측정 및 관리와 관련한 역할과 책임 및 의사소통 체계를 문서화하고 조직 전반의 개인 및 팀에 이를 명확화 •2.2: 조직의 직원 및 파트너가 관련 정책·프로세스·합의에 일치해서 임무와 책임을 수행할 수 있도록, 이들을 대상으로 AI 리스크 관리 훈련 시행 •2.3: AI 시스템 개발 및 배포 관련 리스크에 대한 결정 책임은 조직의 최고 경영진에 귀속
• **거버넌스 3:** AI 라이프사이클 전반에 리스크 맵핑·측정·관리에 있어, **인력의 다양성, 형평성, 포용성 및 접근성 프로세스**를 우선시	•3.1: 다양한 팀(인구통계학적 특성·학문분야·경험·전문성·배경 등)의 정보를 바탕으로 AI 라이프사이클 전반의 리스크 맵핑·측정·관리 관련 의사결정 •3.2: 인간과 AI의 역할 설정(configuration) 및 AI 시스템 감독을 위해 역할과 책임을 정의하고 구분하기 위한 정책 및 프로세스 수립
• **거버넌스 4:** 조직 팀 내 AI 리스크를 고려하고, 의사소통하는 **문화 조성**	•4.1: 잠재적인 부정적 영향을 최소화하기 위해 AI 시스템의 설계·개발·배포 및 사용에 비판적 사고 및 안전 우선 사고방식을 조성하는 조직의 정책과 관행 수립 •4.2: 조직의 각 팀은 그들이 설계·개발·배치·평가 및 사용하는 AI 기술의 리스크 및 잠재적 영향을 문서화하고, 이러한 영향을 더 폭넓게 소통 •4.3: AI 테스트, 사고 식별 및 정보 공유를 가능하게 하는 조직 관행 수립
• **거버넌스 5:** 관련 **AI 행위자들과 견고한 협력 프로세스 수립**	•5.1: AI 리스크 관련 잠재적인 개인 및 사회적 영향과 관련하여 AI 시스템을 개발 또는 배포한 외부 팀의 피드백을 수집·검토·우선순위 책정 및 통합하기 위한 조직의 정책 및 관행 수립 •5.2: AI 시스템을 개발 및 배포한 팀이 정기적으로 관련 AI 행위자들의 조정된 피드백을 시스템 설계 및 구현에 통합할 수 있도록 하는 매커니즘 확립
• **거버넌스 6:** 제3자 소프트웨어 및 데이터·기타 공급 체인 문제에서 야기된 AI 리스크와 이점을 다루기 위한 정책 및 절차 수립	•6.1: 제3자 지식재산 및 기타 권리의 침해 리스크를 포함하여 제3자와 관련한 AI 리스크를 다루기 위한 정책 및 절처 수립 •6.2: 고위험에 해당하는 제3자 데이터나 AI 시스템의 오류 또는 사고를 처리하기 위한 비상 대비 절차 수립

(ii) 맵핑(Map)

기능의 범위	하위범위
맵핑 1: 맥락 설정 및 이해	•1.1: 의도된 목적, 잠재적으로 유익한 사용 사례, 특정 맥락의 법률·규범·기대사항 및 AI 시스템이 배포될 가능성이 있는 환경을 이해하고 문서화

	•1.2: 맥락 설정을 위한 다학문적 AI 행위자들, 이들의 역량, 기술 및 능력에 인구통계학적 다양성, 광범위한 전문영역, 사용자 경험 전문성을 반영하고, 이들의 참여를 문서화, 다문학적 협력 기회를 우선시 •1.3: AI 기술에 대한 조직의 미션·목표를 이해·문서화 •1.4: 사업 가치 및 사업 활용 맥락을 명확히 정의 또는 기존 AI 시스템을 평가하는 경우 이를 재평가 •1.5: 조직의 리스크 수용성 결정 및 문서화 •1.6: 시스템 요건(예: "본 시스템은 사용자 프라이버시를 존중해야 한다")을 관련 AI 행위자들로부터 도출하고, AI 행위자들은 이를 이해, 설계 의사결정에서는 사회-기술적 영향을 고려하여 AI 리스크를 처리
맵핑 2: AI 시스템의 범주화 수행	•2.1: AI 시스템이 지원할 특정 작업과 해당 특정 작업 수행에 사용되는 방법을 정의(예: 분류기, 생성 모델, 추천 시스템) •2.2: AI 시스템 지식의 한계, 시스템 결과 활용법 및 인간이 시스템을 감독하는 방식을 문서화, 해당 문서는 의사 결정 및 후속 조치 시행 시 관련 AI 행위자를 지원하는데 충분한 정보를 제공해야 함 •2.3: 실험 설계, 데이터 수집 및 선정(예: 가용성, 대표성, 적합성), 시스템 신뢰성, 구성 타당성을 포함하여, 과학적 무결성 및 TEVV(테스트, 평가, 검증, 유효성 확인) 고려 사항을 식별하고, 문서화
맵핑 3: AI 역량, 목표 활용도, 목표, 기대되는 이익 및 비용을 적절한 벤치마크와 비교하여 이해	•3.1: AI 시스템의 의도한 기능과 성능의 잠재적 이익을 검토하고, 문서화 •3.2: 조직의 리스크 수용력과 연계해서, 예상되는 또는 발생한 AI 오류·시스템 기능·신뢰성으로 발생할 수 있는 잠재적 비용(비금전적 비용 포함)을 검토하고, 문서화 •3.3: 시스템 역량, 설정 맥락 및 AI 시스템 범주에 기반해 목표한 적용 범위를 구체화하고 문서화 •3.4: AI 시스템 성능 및 신뢰도와 관련 기술 표준 및 인증과 더불어 운영자와 실무자의 숙련도 유지·개선·향상을 위한 프로세스를 정의하고, 평가하며, 문서화 •3.5: 거버넌스 기능의 조직 정책에 따라 인간 감독을 위한 프로세스를 정의·평가·문서화
맵핑 4: : 제3자 소프트웨어 및 데이터를 포함하여 AI 시스템의 모든 구성요소를 대상으로 리스크와 이익을 맵핑	•4.1: AI 기술과 그 구성요소의 법적 리스크(제3자 데이터 또는 소프트웨어 포함)와 제3자 지식재산권 및 기타 권리의 침해 리스크를 맵핑하는 접근법을 수립·시행·문서화 •4.2: 제3자 AI 기술을 포함해서 AI 시스템의 구성요소를 위한 내부 리스크 통제를 식별하고 문서화

맵핑 5: 개인·단체·커뮤니티·조직 및 사회에 미치는 영향의 특성 분석	•5.1: 예상되는 사용 사례, 유사 맥락에서의 AI 시스템 과거 사용 사례, 공개된 사건 보고서, AI 시스템을 개발 및 배포하는 팀 외부에서 받은 피드백 또는 기타 데이터에 기반해 각 식별된 영향(잠재적으로 이로운 영향 및 해로운 영향 둘다)의 발생 가능성과 규모를 식별하고 문서화 •5.2: 관련 AI 행위자와 정기적 협력을 지원하고, 긍정적·부정적·예상치 못한 영향에 대한 피드백을 통합하기 위한 관행 및 인력을 수립하고 문서화

(iii) 측정(Measurement)

기능의 범위	하위범위
측정 1: 적절한 방법 및 지표를 식별하고 적용	•1.1: MAP 기능에서 열거된 AI 리스크의 측정을 위한 접근법과 지표를 선정하고, 가장 중요한 AI 리스크부터 시행, 측정하지 않거나 측정이 불가능한 리스크와 신뢰성 특성은 적절히 문서화 •1.2: 오류 보고 및 영향받는 커뮤니티에 미칠 수 있는 잠재적 영향을 포함하여, AI 지표의 적절성·기존 통제의 효과성을 정기적으로 평가·업데이트 •1.3: 시스템 개발 최전선 개발자가 아닌 내부 전문가 및 독립 평가자가 정기적 평가와 업데이트에 참여 등
측정 2: 신뢰성 특성을 대상으로 AI 시스템 평가	•2.1: 테스트 세트, 지표, TEVV(테스트, 평가, 검증, 유효성 확인)에 사용되는 도구의 세부사항 문서화 •2.2: 인간을 대상으로 포함하는 평가는 관련 요건(대상이 되는 인간 보호를 포함)을 충족해야 하고, 평가 대상이 해당 인구 집단에 대표성이 있어야 함. •2.3: AI 시스템 성능 또는 보증 기준을 정성적 또는 정량적으로 측정하고, 배치 환경에 유사한 조건에서 입증하며, 측정을 문서화 •2.4: 맵핑 기능에서 확인된 AI 시스템과 그 구성요소의 기능과 동작을 운영 시 모니터링 측정 •2.5: 배포된 AI 시스템의 유효성과 신뢰성을 입증하고, 기술 개발 조건을 넘는 일반화의 한계를 문서화 •2.6: 맵핑 기능에서 확인된 바에 따라 AI 시스템 안정성을 정기적으로 평가, 배치된 AI 시스템은 안전성이 입증되어야 하고, 남아있는 부정적 리스크는 리스크 수용성을 초과하지 않아야 함. •2.7: 맵핑 기능에서 확인된 AI 시스템 보안 및 회복성을 평가하고 문서화 •2.8: 맵핑 기능에서 확인된 투명성 및 책임성과 관련된 리스크를 검토하고 문서화

	•2.9: AI 모델을 설명·검증·문서화하고, 맵핑 기능에서 확인된 맥락에서 AI시스템 결과를 해석하여 책임 있는 사용 및 가버넌스에 정보 제공 •2.10: 맵핑 기능에 확인된 AI 시스템의 프라이버시 리스크를 검토 및 문서화 •2.11: 맵핑 기능에서 확인된 바에 따라, 공정성 및 편향을 평가하고 그 결과를 문서화 •2.12: 맵핑 기능에서 확인한 바에 따라 AI 모델 훈련 및 관리 활동의 지속성과 환경적 영향을 평가하고 문서화 •2.13: 측정 기능에서 채택한 TEVV(테스트, 평가, 검증, 유효성 확인) 지표 및 프로세스의 효과성을 평가하고 문서화
측정 3: 확인된 AI 리스크를 시간의 흐름에 따라 지속적으로 추적하는 메커니즘 구축	•3.1: 배포 맥락에서 의도된 성능·실제 성능 등의 요소에 기반해 기존의 예상 못한, 새로운 리스크를 정기적으로 확인하기 위한 접근법·인력·문서화 구축 •3.2: AI 리스크가 현재 이용가능한 측정 기술을 사용해서 평가하기 어렵거나 지표가 아직 이용할 수 없는 경우에 리스크 추적 접근법을 고려 •3.3: 최종 사용자 및 영향받는 커뮤니티가 문제를 보고하고 시스템 결과에 이의를 제기할 수 있는 피드백 프로세스 확립·AI 시스템 평가 지표에 통합
측정 4: 측정의 효율성에 대한 피드백을 수집 및 평가	•4.1: AI 리스크를 식별하기 위한 측정 접근법은 배포 맥락과 연결되며, 특정 분야 전문가 및 기타 최종 사용자와의 협의를 통해 정보 보완·문서화 •4.2: 배포 맥락 및 라이프사이클 전반에 걸쳐 신뢰성 관련 측정 결과는 특정 전문가 및 관련 AI 행위자들의 의견을 반영하여, 시스템이 의도대로 일관되게 작동하였는 지 여부를 검증·결과 문서화 •4.3: 영향 받는 커뮤니티 포함 관련 AI 행위자들과의 협의를 바탕으로 측정가능한 성능 개선 및 저하, 맥락과 관련된 리스크 및 신뢰성 관련 현장 데이터를 식별하고 문서화

(iv) 관리(Manage)

기능의 범위	하위범위
관리 1: 맴핑과 측정 기능을 통한 평가 및 기타 분석적 결과에 따라 AI 리스크의 우선순위를 책정하고, 대응·관리	•1.1: AI 시스템이 의도된 목적 및 명시된 목표를 달성했는지 여부 및 개발 또는 배치를 진행해야 할지 여부를 결정 •1.2: 문서화한 AI 리스크를 영향·발생 가능성 및 가용 자원이나 방법에 따라 우선순위 책정

	•1.3: 맵핑 기능에서 확인된 높은 우선순위로 여겨지는 AI 리스크에 대한 대응을 개발·기획·문서화, 리스크 반응 선택사항에는 완화·이전·회피 또는 수용이 포함될 수 있음. •1.4: AI 시스템의 후속 사용자와 최종 사용자에게 남아있는 부정적인 잔여 리스크(미완화 리스크의 총합으로 정의)를 문서화
관리 2: AI 이익을 극대화하고 부정적인 영향을 최소화하기 위한 전략을 기획·준비·시행·문서화하고, 관련 AI 행위자들로부터 받은 정보를 반영	•2.1: 잠재적 영향의 정도나 발생 가능성을 줄이기 위해 AI 리스크 관리에 필요한 자원을 AI가 아닌 대체 시스템, 접근법 또는 방법과 함께 고려 •2.2: 배포된 AI 시스템의 가치 유지를 위한 메커니즘을 수립하고 적용 •2.3: 이전에 미확인 리스크가 식별되면 이에 대응하고 복구하기 위한 절차를 추진 •2.4: 의도된 사용과 불일치하게 운영되거나 결과물이 생성된 것으로 입증된 AI 시스템을 대체·회수·비활성화시키기 위한 메커니즘을 수립하여 적용하며 책임을 할당하고 이를 이해
관리 3: 제3자로부터 발생하는 AI 리스크 및 이익 관리	•3.1: 제3자 자원으로부터 발생하는 리스크와 이익을 정기적으로 모니터·리스크 통제 적용 및 문서화 •3.2: 개발에 사용되는 사전 훈련된 모델을 AI 시스템 정기 모니터링 및 유지의 한 부분으로 모니터
관리 4: 반응과 복구를 포함한 리스크 처리와 리스크 식별 및 측정을 위한 의사소통 계획을 정기적으로 문서화·모니터	•4.1: 배포 후 AI 시스템 모니터링 계획을 실행(사용자 및 기타 관련 AI 행위자로부터 정보를 수집하고 평가하는 메커니즘, 이의제기 및 재조정, 해체, 사고 대응, 복구 및 변화 관리 등 포함) •4.2: 지속적인 개선을 위한 측정 가능한 활동을 AI 시스템 업데이트와 통합(관련 AI 행위자를 포함하여 이해관계자의 정기적 참여 포함) •4.3: 사고 및 오류와 관련, AI 행위자와 의사소통을 진행(영향 받는 커뮤니티 포함)하고, 사고를 추적·대응·복구하는 프로세스를 추진하고 문서화

(3) NIST 생성형 AI 프로파일

ㅇ NIST는 바이든 행정부의 "안전하고 신뢰할 수 있는 인공지능에 관한 행정명령 제14110호"에 따라, NIST AI 리스크 관리 프레임워크(RMF)를 생성형 AI에 적용한 "생성형 AI 프로파일('24.7월)"을 발표하였다.

생성형 AI 기술은 텍스트·이미지·음악·영상 등 다양한 콘텐츠를 자동으로 생성하며, 창작자의 작업을 지원하는 동시에 새로운 콘텐츠 제작 방식과 비즈니스 모델을 가능하게 함으로써 문화콘텐츠 산업에서 빠르게 확산되고 있다. 본 연구는 문화콘텐츠 분야에서의 AI 윤리지침을 연구 대상으로 하므로, NIST 생성형 AI 프로파일을 함께 살펴보고자 한다.

이를 위해 문서 분석(Document Analysis), 내용 분석(Content Analysis)을 기반으로 아래와 같이 프로파일의 목적, 체계, 주요 리스크 항목, 대응 방안, 정책적 함의 등을 체계적으로 살펴보았다.

1) 생성형 AI 주요 리스크 개요

ㅇ NIST는 본 프로파일에 다양한 이해 관계자의 의견을 반영하기 위해 생성형 AI 공공 워킹그룹(GAI PWG)을 통해 산업·학계·공공 기관 피드백을 포함하여 폭넓은 논의를 거쳤고, 여러 차례 AI 리스크에 대한 공공의견을 수렴하여 이를 반영하였다.

ㅇ NIST는 생성형 AI의 주요 리스크를 아래와 같이 12가지 유형으로 구분하고, 각 유형별 AI 리스크 관리 프레임워크(RMF) 상의 AI 신뢰성 특성을 함께 제시함으로써, 조직의 리스크 관리를 지원하고자 하였다.

1. CBRN 정보 또는 역량	• 생화학·방사능·핵 무기나 기타 위험 물질과 관련된 실제 악의적 정보나 설계 역량에 손쉽게 접근하거나 이를 합성 - 생화학적으로 위협이 되는 지식이나 정보는 공개적으로 접근 가능한데, 대규모 언어 모델(LLM)은 과학적 훈련이나 전문성이 없는 개인도 쉽게 이를 분석 또는 합성할 수 있도록 해줌 - 생화학·방사능·핵 무기 설계를 촉진하는 AI 도구의 성능과 생성형 AI 시스템과 관련 데이터·도구와의 연결·접근을 모니터링하면서 지속적으로 리스크 평가를 강화하는 것이 필요
	• AI 신뢰성 특성: 안정성, 설명 가능성, 해석 가능성

2. 혼동 (Confabulation)	• 확신에 찬 상태로 잘못되거나 거짓된 콘텐츠(흔히 "환각" 또는 "조작"으로 알려짐)를 생성하여 사용자를 오도하거나 속임
	• **AI 신뢰성 특성: 유해한 편향이 관리된 공정성, 안전성, 유효성 및 신뢰성, 설명 가능성 및 해석 가능성**
3. 위험하거나, 폭력적이거나, 혐오스러운 콘텐츠	• 폭력적·선동적·급진적·위협적 콘텐츠 및 자해 또는 불법 활동을 추천하는 콘텐츠를 쉽게 생성하고 접근할 수 있도록 함. 혐오적이고, 비하적이거나 고정관념적인 콘텐츠에 대한 대중의 노출을 통제하기 어려움 - 비하적 또는 고정관념적 콘텐츠는 대표성 피해 악화 가능
	• **AI 신뢰성 특성: 안전성, 보안성 및 회복성**
4. 데이터 프라이버시	• 생체측정·건강·위치 및 기타 개인 식별 정보나 민감 데이터의 누출·무단 사용·공개·비익명화로 인한 영향 - 생성형 AI 시스템 학습에는 대규모 데이터가 필요하므로, 일부 개인정보가 포함될 수 있는데, 대부분의 모델 개발사는 훈련 데이터 출처를 공개하지 않기 때문에, 사용자가 개인 식별 정보가 모델 훈련에 쓰였는지 여부나 어떻게 수집되었는 지 여부를 알기 어려움 - 생성형 AI 모델은 훈련 데이터에 없거나 사용자가 공개하지 않은 민감정보나 개인 식별 정보도 이질적 출처들의 정보 결합 등을 통해 정확하게 추론 가능 - 잘못된 또는 부적절한 개인 식별 정보 추론은 후속 또는 2차 피해 초래 가능(부정적 결정, 대표성 피해, 분배적 피해 등)
	• **AI 신뢰성 특성: 책임성 및 투명성, 프라이버시 강화, 안전성, 보안성 및 회복성**
5. 환경 영향	• 생성형 AI 모델 훈련 및 운영에 대규모 컴퓨팅 리소스 사용으로 인한 영향과 이로 인해 생태계에 미치는 부정적 결과 - 현재 추정 값으로는 단일 트랜스포머 기반 대규모 언어모델(LLM)의 훈련의 경우, 탄소배출량이 샌프란시스코에서 뉴욕 간 항공기 왕복 300회 정도에 해당 - 현재는 생성형 AI로 인한 환경적 영향을 측정하는 합의된 방식이 부재
	• **AI 신뢰성 특성: 책임성 및 투명성, 안정성**
6. 유해한 편향 및 균질화	• 역사적·사회적·시스템적 편향의 증폭 및 악화, 대표성이 없는 훈련 데이터 등으로 인한 하위 그룹간 또는 언어간 성능 차이로 차별·편향의 증폭 또는 성능에 대한 잘못된 가정 발생, 원치 않은 균질성이 시스템이나 모델의 출력을 왜곡시켜 오류를 발생시키거나, 잘못된 의사결정을 초래하거나, 유해한 편향을 증폭 - 생성형 AI가 학습에 스스로 만든 합성 결과물을 반복적으로 재활용하는 비율이 높아지면 시스템에 내재된 편향과 균질화가 심화되어 결국 모델 붕괴 등의 문제 초래
	• **AI 신뢰성 특성: 유해한 편향이 관리된 공정성, 유효성 및 신뢰성**

7. 인간-AI 상호작용 (Configuration)	• 인간과 AI 간의 상호작용이나 배치 방식이 부적절하게 생성형 AI 시스템을 의인화하거나, 알고리즘 회피, 자동화 편향, 과도한 의존, 또는 생성형 AI와 정서적 얽힘을 경험하게 하는 상황 초래
	• **AI 신뢰성 특성: 책임성 및 투명성, 설명 가능성 및 해석 가능성, 유해한 편향이 관리된 공정성, 프라이버시 강화, 안전성, 유효성 및 신뢰성**
8. 정보 무결성	• 사실과 의견이나 허구를 구별하지 않거나 불확실성을 인정하지 않는 콘텐츠의 교환과 소비를 생성하거나 지원하는 진입장벽을 낮추거나 허위 정보 및 오정보 캠페인에 활용 가능 - (정보 무결성) 정보 및 그 생성, 교환, 소비와 관련된 패턴의 스펙트럼 - (높은 무결성 정보) 정확성, 신뢰성, 검증 및 인증 가능성, 출처 관리 체계의 명확성, 합리적 유효기간 예측성이 있는 정보 - 생성형 AI 기술은 허위 정보와 잘못된 정보를 대량으로 생산하고 배포하는 과정을 용이하게 하여, 악의적 사용자들이 사실과 유사한 딥페이크 기술 등을 악용해 허위 정보 캠페인, 사기, 사회적 및 경제적 혼란 초래 가능
	• **AI 신뢰성 특성: 책임성 및 투명성, 안전성, 유효성 및 신뢰성, 해석 가능성 및 설명 가능성**
9. 정보 보안	• 자동화된 취약점 발견과 악용 등을 통해 해킹·악성 소프트웨어·피싱·공격적 사이버 운영 또는 기타 사이버 공격을 용이하게 하는 공격적 사이버 역량의 장벽이 낮아지고, 사이버 공격의 목표 범위가 확대되어 시스템의 가용성이나 훈련 데이터·코드 또는 모델 가중치의 기밀성 또는 무결성 손상 가능 - (생성형 AI 기반 시스템의 주요 정보 보안 리스크) ① 사이버 보안 공격 실행 자동화를 용이하게 하고, 실행 장벽을 낮춤으로써, 새로운 사이버 보안 리스크를 발견하거나 가능하도록 함, ② 생성형 AI 가 요청 주입(prompt injection)이나 데이터 중독(data poisoning)과 같은 공격에 취약하여, 공격 가능 범위가 확대됨
	• **AI 신뢰성 특성: 프라이버시 강화, 안전성, 보안성 및 회복성, 유효성 및 신뢰성**
10. 지식재산권	• 저작권이 있거나, 상표등록이 되었거나 라이선스가 부여된 컨텐츠를 승인 없이(공정 사용에 해당하지 않는 경우) 생성 또는 복제가 용이, 영업비밀 노출이나 표절 또는 불법 복제 용이 - 생성형 AI의 훈련 데이터에 저작권이 있는 자료가 포함된 경우, 그 데이터 암기 결과물이 저작권 침해 가능 - 생성 콘텐츠가 유사하지만 엄격히 복제한 것은 아닌 경우 등 생성형 AI와 저작권의 관계에 대한 법적 논의 진행 중 - 허락 없는 개인 신상, 유사성, 음성 등과 관련하여서도 유사한 논의 진행 중
	• **AI 신뢰성 특성: 책임성 및 투명성, 유해한 편향이 관리된 공정성, 프라이버시 강화**

11. 외설적·비하적·학대적 콘텐츠	• 아동 성학대 자료(CSAM) 및 비동의 성인 친밀 이미지(NCII) 합성을 포함하여, 외설적·비하적·학대적 콘텐츠 생성 및 접근이 용이
	• **AI 신뢰성 특성: 유해한 편향이 관리된 공정성, 안전성, 프라이버시 강화**
12. 가치 사슬 및 구성 요소 통합	• 투명하지 않거나 추적 불가능한 방식으로 초기 단계(upstreaming) 제3자 구성요소(생성형 AI 자동화 증가로 인해 부적절하게 획득 또는 미처리 및 미정제된 데이터 포함) 통합, AI 라이프사이클 전반에 공급업체 검증 부적절, 후속 사용자에 대한 투명성 또는 책임성 약화 문제
	• **AI 신뢰성 특성: 책임성 및 투명성, 설명 가능성 및 해석 가능성, 유해한 편향이 관리된 공정성, 프라이버시 강화, 안전성, 보안성 및 회복성, 유효성 및 신뢰성**

2) 생성형 AI 리스크 관리 조치 제안

ㅇ 본 프로파일은 AI 리스크 관리 프레임워크(RMF)의 4가지 핵심기능(①거버넌스, ②맵핑, ③측정, ④관리)을 기반으로 생성형 AI 리스크를 관리하는 조치를 제안한다.

이는 아래 각 기능별 조치 예시에서 볼수 있듯이, 생성형 AI의 확산과 함께 증가하는 윤리적·법적 문제 해결을 위한 실무적 가이드라인으로 활용할 수 있도록 설계되었다.

• 거버넌스(Govern) 1.1: AI 관련 법적 및 규제적 요건을 이해·관리·문서화		
조치 ID	**제안 조치**	**생성형 AI 리스크**
거버넌스 1.1-001	• 데이터 프라이버시, 저작권, 지식재산법을 포함하여 적용가능한 관련 법 및 규정에 합치하여 생성형 AI 개발 및 사용	데이터 프라이버시; 유해한 편향 및 균질화; 지식재산권
AI 행위자 과업: 거버넌스 및 감독		
• 맵핑(MAP) 1.1: 의도된 목적, 잠재적으로 유익한 사용 사례, 특정 맥락의 법률·규범·기대사항 및 AI 시스템이 배포될 가능성이 있는 환경을 이해하고 문서화, 고려사항: 특정 사용자 집단 또는 유형 및 이들의 기대사항, 개인·커뮤니티·조직·사회·세계에 미치는 시스템 사용의 잠재적 긍정·부정적 영향, 개발 전반 및 AI 제품 수명주기 전반에 걸친 AI 시스템 목적·사용·리스크에 대한 가정과 관련 한계, 관련 TEVV(테스트, 평가, 검증, 유효성 확인) 및 시스템 지표		
조치 ID	**제안 조치**	**생성형 AI 리스크**
맵핑 1.1-001	• 의도된 목적을 정의할 때, 내부 및 외부 사용, 적용 범위의 좁고 넓음, 미세 조정, 데이터 소스의 다양성(예: 그라운딩, 검색-증강 생성) 등을 고려	데이터 프라이버시; 지식재산권

맵핑 1.1-002	• 사회 문화적 및 기타 세부 분야 전문가들과 협업하여, 생성형 AI 시스템의 예상되는 그리고 허용 가능한 사용 맥락을 정의하고 문서화할 때 다음 항목을 평가: 가정 및 한계; 조직이 갖게 되는 직접적 가치; 의도된 운영 환경 및 관찰된 사용 패턴; 개인·공공 안전·그룹·커뮤니티·조직·민주적 기관 및 물리적 환경에 미치는 긍정·부정적 영향; 사회적 기준 및 기대	유해한 편향 및 균질화
맵핑 1.1-003	• 확인된 리스크를 해결하기 위한 리스크 측정 계획 문서화. 계획에는 가능한 다음 항목을 포함 : 생성형 AI 시스템의 설계·구현·사용과 관련한 AI 행위자들의 개인적 및 집단적 인지적 편향 (예: 확인 편향, 펀딩 편향, 집단 사고), 알려진 과거의 생성형 AI 시스템의 사고 및 오류 유형; 맥락 내 사용 및 예측 가능한 오용·남용 및 비허가 사용; 사용 맥락에서 한계에 대한 충분한 인지 없이 정량적 지표와 방법론에 과도하게 의존; 표준화된 측정 및 구조화된 인간 피드백 접근법; 예상되는 인간-AI 상호작용(Configuration)	인간-AI 상호작용 (Configuration); 유해한 편향 및 균질화; 위험하고, 폭력적이며, 혐오스러운 콘텐츠
맵핑 1.1-004	• 조직의 리스크 수용 한계를 넘어서는 생성형 AI 시스템의 예측 가능한 불법적 사용 또는 적용을 식별하고 문서화	CBRN 정보 또는 역량; 위험하거나, 폭력적이거나, 혐오스러운 콘텐츠; 외설적·비하적·학대적 콘텐츠
AI 행위자 과업: AI 배포		
• 측정(Measure) 1.1: MAP 기능에서 열거된 AI 리스크의 측정을 위한 접근법과 지표를 선정하고, 가장 중요한 AI 리스크부터 시행, 측정하지 않거나 측정이 불가능한 리스크와 신뢰성 특성은 적절히 문서화		
조치 ID	**제안 조치**	**생성형 AI 리스크**
측정 1.1-001	• 디지털 콘텐츠의 출처 및 수정 추적 가능 방법 채택	정보 무결성
측정 1.1-002	• 콘텐츠 출처를 분석, 이상 데이터 감지, 디지털 서명의 진위성 검증, 잘못된 정보나 조작 관련 패턴을 식별하기 위해 설계된 도구들을 통합	정보 무결성
측정 1.1-003	• 인구집단에 따라 콘텐츠 출처 확인 메커니즘이 작동하는 방식에 차이 여부를 식별하기 위해 평가 지표를 인구 통계학적 요소에 따라 세분화	정보 무결성; 유해한 편견 및 균질화
측정 1.1-004	• 대표성이 있는 AI 행위자들이 제공한 정보에 기반하여 구조화된 공개 피드백 활동을 평가하는 지표를 개발	인간-AI 상호작용 (Configuration); 유해한 편향 및 균질화; CBRN 정보 또는 역량

측정 1.1-005	• 모델이 유효하고, 신뢰성 있고, 사실적으로 정확한 결과물을 산출해 내는 능력을 유지하면서, 콘텐츠 출처·사이버 공격·CBRN을 포함한 생성형 AI 관련 리스크를 측정하는 새로운 방법과 기술을 평가	정보 무결성; CBRN 정보 또는 역량; 외설적·비하적·학대적 콘텐츠
측정 1.1-006	• 생성형 AI 산출물이 다양한 하위 인구집단에 형평성이 있는 지 여부를 식별하기 위해 생성형 AI 시스템 영향에 대한 지속적 모니터링 구현, 구조화된 피드백 메커니즘이나 결과물을 모니터링하고 개선하기 위한 레드팀 작업을 통해 영향받는 커뮤니로부터 능동적·직접적 피드백 수렴	유해한 편향 및 균질화
측정 1.1-007	• 훈련 사용 데이터의 품질·무결성과 AI 생성 콘텐츠의 출처를 평가(예: 카오스 엔지니어링과 같은 기술을 사용, 이해관계자 피드백 수렴 방식으로 진행)	정보 무결성
측정 1.1-008	• 사용 맥락을 기준으로 생성형 AI 리스크 측정 및 관리를 위해 구조화된 인간 피드백 활동(예: 생성형 AI 레드팀 작업 등)이 가장 이로운 경우, 사용 사례·사용 맥락·성능·부정적 영향을 정의	유해한 편향 및 균질화; CBRN 정보 또는 역량
측정 1.1-009	• 정량적으로 측정이 불가능한 모든 생성형 AI 리스크와 관련된 리스크 및 기회를 추적하고 문서화(일부 리스크를 측정 불가능한 이유(예: 기술적 한계, 자원 제한 또는 신뢰성 고려사항 등)에 대한 설명 포함), 측정불가 리스크는 한계 리스크(Marginal Risks)에 포함	정보 무결성
AI 행위자 과업: AI 개발, 특정 분야 전문가, 테스트·평가·검증·유효성 확인(TEVV)		
• 관리(Manage) 1.3: 맴핑 기능에서 확인된 높은 우선순위로 여겨지는 AI 리스크에 대한 대응을 개발·기획·문서화, 리스크 반응 선택사항에는 완화·이전·회피 또는 수용이 포함될 수 있음.		
조치 ID	**제안 조치**	**생성형 AI 리스크**
관리 1.3-001	• 조직의 리스크 허용 한계를 넘지 않는 리스크를 대상으로 절충, 의사 결정 과정·관련 측정 결과 및 피드백 결과를 문서화, 모델 출시 경우, 모델 출시의 다른 접근법을 고려(예: 단계적 출시 접근 활용), 모델의 사용 및 예상되는 사용 사례 맥락에서의 출시 접근법을 검토, 조직의 리스크 허용 한계를 초과하는 리스크를 완화·이전·회피	정보 보안
관리 1.3-002	• 리스크 통제와 완화 계획의 견고성 및 효과성을 모니터링(예: 레드팀 활동, 현장 테스트, 참여형 활동, 성능 평가, 사용자 피드백 메커니즘)	인간-AI 상호작용 (Configuration)
AI 행위자 과업: AI 개발·배포·영향평가, 운영 및 모니터링		

(4) 행정명령 제14110호 "안전하고 신뢰할 수 있는 인공지능 개발 및 사용"(Executive Order 14110 of October 30, 2023 "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence")

ㅇ 행정명령 제14110호(Executive Order 14110, "Safe, Secure, and Trustworthy Artificial Intelligence")는 AI 개발 및 활용과 관련된 법적·윤리적 원칙을 규정하고, 연방 정부 기관이 AI를 어떻게 사용하고 관리해야 하는지에 대한 지침을 제시한다. 하지만, 일부 조항에서는 **민간 기업에 대한 직접적인 의무도 포함**하고 있어 미국 정부의 AI 규제 정책의 방향을 살펴볼 수 있다.

본 항(Subsection)에서는 문화 콘텐츠 분야와 관련된 주요 조항을 중심으로, 문서 분석(Document Analysis), 내용 분석(Content Analysis), 문헌 연구(Literature Review)를 통해 행정명령의 전체 구조 및 관련 조항의 세부 내용, 주요 쟁점을 함께 검토하였다.

1) 구조 및 개요

ㅇ 본 행정명령은 전체 총 13개 조항으로 구성되어 있고, AI의 안전성, 윤리성, 경제적 영향, 소비자 보호, 국가 안보, 국제 협력 등의 요소를 규정한다.

구분	주요 내용
제1조 (목적, Purpose)	안전하고 책임감 있는 AI 개발과 활용을 위한 연방 정부의 목표 및 방향
제2조 (정책 및 원칙, Policy and principles)	행정부가 AI 개발과 활용을 발전시키고 관리하기 위한 8가지 원칙 (①안전성 및 보안성 확보, ②책임 있는 혁신·경쟁·협업 촉진, ③근로자 보호 및 지원, ④평등과 시민권 보호, ⑤소비자 보호 강화, ⑥개인정보 및 시민 자유 보호, ⑦연방정부의 AI 활용 및 책임 강화, ⑧연방정부의 AI 국제 리더십 강화 등)
제3조 (정의, Definitions)	행정명령에 사용된 용어 정의 (인공지능(AI), AI 모델, AI 시스템, 이중 용도 기초 모델(dual-use foundation model), 생성형 AI, 합성 콘텐츠, 워터마킹 등)
제4조 (AI 기술의 안전 및 보안 보장, Ensuring the Safety and Security of AI Technology)	AI 안전성 및 보안 표준 촉진을 위한 지침 확립, NIST 생성형 AI 프로파일 개발, AI 기업 및 클라우드 서비스 제공업체의 미국 정부 보고 의무 등 포함

구분	주요 내용
제5조 (혁신 및 경쟁의 촉진, Promoting Innovation and Competition)	미국 내 AI 혁신과 경쟁을 촉진하기 위한 조치, AI 인재 유치, 지식 재산 보호, AI 연구 지원 등
제6조 (근로자에 대한 지원, Supporting Workers)	AI가 노동 시장에 미치는 영향 관련 보고서 제출, AI로 인한 직업 변화를 대비하기 위한 지원 조치 등 규정
제7조 (형평성 및 시민권의 발전, Advancing Equity and Civil Rights)	AI가 차별과 불공정한 결과를 초래하지 않도록 하는 조치 등 규정
제8조 (소비자·환자·승객 및 학생에 대한 보호, Protecting Consumers, Patents, Passengers, and Students)	AI를 활용한 의료·금융·교육·교통 등의 분야에서 안전성과 공정성을 보장하기 위한 연방 차원의 조치 등 규정
제9조 (개인정보 보호, Protecting Privacy)	AI 기반 데이터 수집 및 활용 과정에서 발생할 수 있는 위험을 평가하고, 보호조치 강화 등 규정
제10조 (연방정부의 AI 사용의 발전, Advancing Federal Governmnet Use of AI)	연방정부의 AI 개발 및 사용 체계 구축, AI 사용에 대한 지침 제시, AI 혁신 지원 환경 조성 등 규정
제11조 (미국의 국제 리더십 강화, Strengthening American Leadership)	AI의 리스크를 관리하고 이익을 활용하기 위한 강력한 국제 프레임워크 구축, 글로벌 AI 정책 및 기술 표준 주도, 국제 협력 확대 등
제12조 (이행, Implementation)	행정명령 이행을 총괄하고, 관련 정책의 효과적 실행을 위한 연방 정부 차원의 매커니즘 구축 - 대통령실 내 백악관 인공지능 위원회(White House AI Council) 설립 및 연방 기관 간 조정 기능 강화 등
제13조 (일반 조항, General Provisions)	행정명령의 해석 및 시행에 관한 일반 규정

2) 주요 조항 및 관련 쟁점 – 제4조를 중심으로

○ 본 행정명령 제4조는 미국 국립표준기술원(NIST)이 AI 리스크 관리 프레임워크(RMF)를 기반으로, 생성형 AI 프로파일을 마련하도록 규정(제4.1항)하고 있으며, 민간 AI 기업의 연방정부에 대한 정보제공 의무(제4.2항)도 포함하고 있다.

특히, AI 기업 및 클라우드 서비스 제공업체에 대한 정보 제출 의무는 글로벌 AI 기업 뿐 아니라 해외 콘텐츠 제작자에게도 영향을 미칠 수 있으므로, 관련 문헌을 통해 주요 쟁점을 구체적으로 분석하고자 한다.

(i) AI 기업의 정보 제출 의무 관련(제4조제4.2항(a)목(i))

ㅇ 이중 용도 파운데이션 모델(dual-use foundation models)을 개발 중이거나, 개발할 의도를 가진 기업은 연방정부에 아래 사항과 관련하여, 지속적으로 정보·보고서·기록을 제출해야 한다.

- 모델의 훈련·개발·생산과 관련된 모든 활동(복잡한 위협에 대한 훈련 과정의 무결성 보장 보안조치 포함)
- 모델의 모델 가중치에 대한 소유권 및 보유 현황(모델 가중치 보호를 위한 보안 조치 포함)
- 레드팀의 모델 성능 테스트 결과 및 안전 목표 달성 조치 사항 설명 등

❑ **주요쟁점**[55]

본 조항은 기업이 모델 개발, 가중치, 보안 조치 등 기업의 민감한 정보를 정부에 공개하도록 강제하고 있어, 영업 비밀 침해 및 경쟁력 저하에 대한 우려를 초래할 수 있다. 특히, 중소기업 및 스타트업의 경우, 정보 제출 및 AI 레드팀 테스트 수행에 필요한 인적·기술적 자원이 부족할 가능성이 높아, 정보 제출 의무 자체가 상당한 부담이 될 수 있다

(ii) 美 클라우드 업체 외국 기업 정보 제출 의무(제4조제4.2항(c), (d)목)

① 해외 고객 AI 모델 훈련 보고 의무(제4.2항(c)목)

• 미국 클라우드(IaaS) 제공업체는 외국 고객이 "악의적인 사어비 행위에 사용될 가능성이 있는 AI 모델"을 훈련하는 경우, 다음 사항을 보고하여야 한다.

  - AI 훈련을 수행하는 외국인의 신원 정보

55) O'Brien, J., Ee, S., & Kraprayoon, J. (2024). Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI. arXiv. Retrieved from https://arxiv.org/pdf/2407.01420; Mehra, A. (2025, January 21). Executive Orders on AI: How to (Lawfully) Apply the Defense Production Act. Mercatus Center. Retrieved from https://www.mercatus.org/research/policy-briefs/executive-orders-ai-how-lawfully-apply-defense-production-act

- AI 모델 훈련 여부
- 기타 상무부가 요구하는 추가 정보

② 해외 리셀러 및 외국 고객 신원 보고 의무(제4.2항(d)목)

• 미국 클라우드(IaaS) 제공업체는 아래와 같이, 해외 리셀러(재판매자) 및 이를 이용하는 외국 고객에 대한 신원확인·보고 의무를 부담한다.
  - 해외 리셀러에게 요구해야 하는 최소 기준
    · 클라우드(IaaS) 제품·서비스 임대 외국인의 신원 확인에 필요한 문서 및 절차
    · 외국인이 계정을 생성·유지할 경우, 리셀러가 보관해야 할 기록
      > 외국인의 신원(이름, 주소 포함)
      > 결제 수단 및 출처(연관 금융 기관, 신용카드 번호, 계좌번호, 가상 화폐 지갑 주소 등)
      > 외국인 신원 확인을 위한 이메일 및 전화번호
      > 계정 액세스에 사용된 IP 주소, 로그인 날짜 및 시간

❑ 주요쟁점[56)]

ㅇ 제4.2항(c)목의 경우, 정보 제출 의무 대상의 기준이 되는 "악의적인 사이버 행위에 사용될 가능성이 있는 AI 모델"의 명확한 기준이 제시되지 않았다. 따라서, 일반적인 AI 연구 및 훈련이나 특정 국가의 AI 연구가 의심만으로 보고 대상이 될 가능성이 있다.

또한, 제4.2항(d)목에 따라, 해외 AI 기업이 미국 클라우드 서비스를 이용할 경우, 해외 리셀러조차 고객의 신원을 보고해야 한다.

이같은 규정은 미국 정부가 AI 연구기관 및 기업을 광범위하게 감시할 수 있다는 우려를 초래하고 있으며, 이에 따라, AI 연구기관·기업들이 미국 클라우드 사용을 기피할 수 있다는 지적이 제기되고 있다.

ㅇ 특히, AI 기업들의 동의 없이 신원과 훈련 정보를 제출하도록 요구하는 것은 개인 및 기업 데이터 보호에 중점을 두고 있는 유럽연합의 GDPR(일반데이터보호규정) 등 글로벌 데이터 보호 규정과 충돌할 가능성이 있다.

---

56) Heim, L., Fist, T., Egan, J., Huang, S., & Zekany, S. (2024). Governing through the cloud: The intermediary role of compute providers in AI regulation. arXiv. Retrieved from https://arxiv.org/pdf/2403.08501; Congressional Research Service. (2024, April 3). Highlights of the 2023 executive order on artificial intelligence for Congress. Retrieved from https://crsreports.congress.gov/product/pdf/R/R47843

3) 기타 문화콘텐츠 분야 관련 주요 조항

○ 이 외에도 본 행정명령에서 콘텐츠 제작·배포·보호에 영향을 미치는 문화 콘텐츠 분야 관련 주요 조항은 다음과 같이 살펴볼 수 있다.

구분	주요 내용
• 정책 및 원칙 (제2조)	- AI가 생성한 콘텐츠임을 알 수 있도록 라벨 지정·콘텐츠 출처 매커니즘 개발 지원, 발명가·창작자 보호를 위한 지식재산 보호 및 소비자 보호 원칙, 개인정보 및 시민 자유 보호 규정
• 합성 콘텐츠 식별 및 출처 이력 확인 등 (제4조제4.5항)	- AI 시스템 활용 합성 콘텐츠를 식별하고 라벨링, 진위성 및 출처 이력 추적 기능 정비 등을 위해, 표준·기술 개발 보고서 제출(콘텐츠 인증 및 출처 추적, 워터마킹 등을 이용한 합성 콘텐츠 라벨링, 합성 콘텐츠 발견, 아동성학대 자료 생성 및 미합의 은밀한 이미지 생성 방지, 합성 콘텐츠 감사 및 유지 등 포함) 및 관련 지침 개발
• AI 활용 콘텐츠의 지식재산 보호 (제5조)	- AI(생성형 AI 포함)의 발명 및 사용에 관한 지침, AI와 지식재산권 관련 추가 지침, 저작권 및 AI 관련 잠재적 행정조치에 대한 권고사항 발표(AI를 활용한 저작물의 보호범위와 AI 훈련 시 저작물의 처리 등) 관련 사항 규정 - AI 관련 지식재산권 위험 완화를 위한 교육, 분석 및 평가 프로그램 개발(지식재산 도용에 관한 보고서 수집·분석 등)
• 생성형 AI 출력 워터마크 표시 (제10조)	- 생성형 AI를 이용한 차별적·오해유발·선동적·안전하지 않거나 기만적 출력물 및 아동성학대 제작 및 미동의 개인적 은밀한 이미지 제작 방지, 생성형 AI 출력에 워터마크 또는 라벨 표시

3. 유럽연합과 미국의 인공지능(AI) 규제 비교·분석

ㅇ 앞서 살펴본 바와 같이, 유럽연합(EU)은 인공지능법(AI Act)을 통해 AI 시스템의 위험도에 기반한 구속력이 있는 규제를 채택하고 있고, 미국은 AI 권리장전, NIST AI 리스크 관리 프레임워크, NIST 생성형 AI 프로파일, 행정명령 제14110호 등을 통해 기업 중심의 자율 규제 및 지침에 따른 접근 방식을 도입하고 있다.

또한, 최근에는 AI 규제의 글로벌 조율을 위한 국제적 협력이 활발히 진행되면서, 미국과 유럽연합도 "무역기술위원회(TTC)"를 통한 협력 등 글로벌 AI 규제 표준을 정립하기 위한 노력을 강화하고 있다.

ㅇ 이에 따라, 본 절(Section)에서는 내용 분석(Content Analysis), 문서 분석(Document Analysis), 문헌 연구(Literature Review)를 통해 유럽연합과 미국의 인공지능(AI) 규제를 비교·분석하고, 주요 개념과 원칙 등에 있어 공통점과 차이점을 도출하였다. 또한, 양측의 최근 국제 표준 정비 노력과 협력 동향을 검토하여, 그 의미와 전망을 살펴보았다.

(1) 비교·분석 주요 내용(범용 AI 및 생성형 AI 중심)[57]

항목	**유럽연합(EU)** 인공지능법(AI Act)	**미국** AI 권리장전, NIST AI 리스크 관리 프레임워크·생성형 AI 프로파일, 행정명령 제14110호
법적성격	• 구속력 있는 법률(강제 규제)	• 비구속적 지침 및 행정명령(일부 의무 규정 채택)

57) Haie, A.-G., Golodny, A., Shenk, M., Avramidou, M., Goodwin, E., & Arethusa, V. (2024, April 30). A comparative analysis of the EU, US, and UK approaches to AI regulation. StepTechToe. Retrieved from https://www.steptoe.com/en/news-publications/steptechtoe-blog/a-comparative-analysis-of-the-eu-us-and-uk-approaches-to-ai-regulation.html; Engler, A. (2023, April 25). The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Brookings Institution. Retrieved from https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/; Lumenova AI. (2024, June 20). AI policy analysis: European Union vs. United States. Retrieved from https://www.lumenova.ai/blog/ai-policy-eu-vs-us-comparison/?utm_source=chatgpt.com

58) 유럽연합 인공지능법(EU AI Act) 제3조(1) 참조 (경제협력개발기구(OECD) 인공지능(AI)

항목	**유럽연합(EU)** 인공지능법(AI Act)	**미국** AI 권리장전, NIST AI 리스크 관리 프레임워크·생성형 AI 프로파일, 행정명령 제14110호
AI관련 정의	• **인공지능 시스템(AI System)** : 다양한 수준의 자율성을 갖고 작동하도록 고안되었고, 배포 이후 적응성(변화)을 보여줄 수 있으며, 명시적 또는 암시적 목적을 위해 입력된 정보로부터, 물리적 또는 가상 환경에 영향을 줄 수 있는 예측·콘텐츠·추천 또는 결정과 같은 결과물을 어떻게 생성해 낼 지 추론해 내는 기계 기반 시스템[58]	• **인공지능(AI)** : 인간이 정의한 일련의 주어진 목표에 대해 실제 또는 가상 환경에 영향을 미치는 예측, 권고 또는 결정을 내릴 수 있는 기계 기반 시스템[59]
접근 방식	• **수평적 접근 방식(Horizonal Approach)** - AI를 포괄적으로 규제하는 통합적 법률로서 모든 산업과 영역에 걸쳐 적용 - AI의 안전성과 신뢰성 확보, 법적 일관성 유지, AI 규제 표준화를 통한 EU 시장 내 공통 규범 확립 도모 • **위험 기반 규제(Risk-Based Regulation)** - 위험 수준을 4단계로 구분하고, 위험 수준별 규제적용	• **특정 영역별 규제 접근 방식 (Sectorally Specific Approach)** - 기존 산업별 법률과 표준을 활용하면서, AI 관련 지침 및 행정명령 추가 적용 방식 • **원칙에 기반한 위험관리 (Principle-Based Risk Management)** - AI 권리장전, NIST AI 리스크 관리 프레임워크, 생성형 AI 프로파일 등 자율적 적용이 가능한 원칙적 지침 제시 - 행정명령 제14110호를 통해 AI의 안전성 및 국가 안보 보호를 목적으로 대규모 AI 모델 및 AI 훈련 보고 의무를 부과, 특정 AI 모델 및 인프라에 대한 감시 및 보고 체계 구축
강조하는 핵심 가치	• 인공지능법 목적[60] - 내부 시장기능을 개선하여 인간중심적으로 신뢰할 수 있는 AI 활용 촉진, 기본권·안전·건강 보호, 혁신 지원	• 시민권·시민자유, 프라이버시, 공정성, 접근성, 안전성, 유효성·신뢰성, 보안성·회복성, 책임성·투명성, 설명 가능성·해석 가능성

원칙(권고안)의 'AI 시스템' 정의 채택
59) 행정명령 제14110호 제3조(b) 참조
60) 유럽연합 인공지능법(EU AI Act) 제1조제1항

항목	**유럽연합(EU)** 인공지능법(AI Act)	**미국** AI 권리장전, NIST AI 리스크 관리 프레임워크·생성형 AI 프로파일, 행정명령 제14110호
주요 규제·의무사항	• AI로 금지되는 행위 규정 - 개인의 의사결정 능력을 저하시켜 해를 끼치는 결과를 초래하는 조작·기만적 방식의 AI, 특정 개인이나 단체에 대한 차별적 행위, 감정 추론, 정치·종교·신념 등을 추론하기 위해 생체 데이터에 기반해 개인 분류 등 • 투명성 의무(AI 시스템이 작동중임을 고지, AI 생성 또는 조작물임을 표시-합성 콘텐츠, 딥페이크, 공익 텍스트 등) • 저작권 및 관련 권리 보호 • 유럽연합 내 공인 대리인 지정	• 알고리즘 차별, 개인정보 보호 • AI 작동에 대한 고지 및 설명 • 인간 대안 선택 및 인간 검토 • AI 생성 콘텐츠 표시, AI 시스템 활용 합성 콘텐츠 식별·라벨링·진위여부·출처 이력 추적을 위한 표준·기술 개발 (콘텐츠 인증 및 출처 추적, 워터마킹 등을 이용한 합성 콘텐츠 라벨링, 합성 콘텐츠 발견, 아동성학대 자료 생성 및 미합의 은밀한 이미지 생성 방지 등 포함) 및 관련 지침 개발 • 콘텐츠 출처 매커니즘 개발 • 저작권 및 지색재산 보호 • 소비자 보호
기업의 정보공개 의무	• 범용 AI 모델 훈련 콘텐츠에 대한 상세한 요약을 대중에게 공개 • AI 사무국이 API 또는 소스코드를 포함한 기술적 수단을 통해 범용 AI 모델에 접근 요청 가능	• 이중 용도 파운데이션 모델 개발 기업의 정보 제출 의무(모델 훈련·개발·생산 관련 활동, 모델 가중치 정보 등) • 美클라우드 업체 외국 고객 정보 제출 의무(해외 고객 AI 모델 훈련 보고, 리셀러 외국 고객 신원 등)

○ AI 규제에 있어 유럽연합과 미국의 가장 큰 차이는 규제 접근 방식에서 나타난다. 유럽연합은 수평적 접근 방식(Horizontal Approach)을 적용하여 단일 법률을 통한 포괄적 규제를 시행하는 반면, 미국은 특정 영역별 접근 방식(Sectorally Specific Approach)을 적용하여, 기존 산업별 법률과 표준을 활용하면서 AI 리스크 관리 원칙을 반영한 지침과 행정명령을 도입하고 있다.

○ 이처럼 접근 방식에는 차이가 있지만, 양측 모두 신뢰할 수 있는 AI(Trustworthy AI)라는 공통의 목표를 바탕으로, 리스크 관리와 규제의 기본 원칙을 마련하고 있다. AI 생성 콘텐츠에 대한 워터마킹·라벨링 등의 표시 및 출처 추적을 규정하고, 책임 있는 AI 개발을 촉진하며, 차별 방지 및 프라이버시 보호를 핵심 원칙으로 반영하고 있다.

ㅇ 이러한 공통의 목표를 바탕으로, 양측은 지속적으로 글로벌 AI 규제 협력을 추진 중이며, 특히 미국-유럽연합 무역기술위원회(TTC)와 같은 협의체를 통해 AI 기술의 국제 표준화를 논의하고, AI의 안전성과 책임성 강화를 위한 국제적 협력을 확대하고 있다. 아래에서는 이 같은 양측 협력의 최근 동향을 구체적으로 살펴보고자 한다.

(2) 미국-EU AI 규제 협력 및 표준화 논의

ㅇ 디지털 경제 및 기술혁신이 급속히 발전함에 따라, 공정한 기술 시장 구축과 AI 및 데이터 거버넌스 규제 협력의 필요성이 증대되면서, 미국과 유럽연합은 2021년 9월 무역기술위원회(TTC, Trade and Technology Council)[61]를 출범하고, AI, 데이터 거버넌스, 반도체 공급망, 디지털 인프라 등의 분야에서 협력을 추진하고 있다.

ㅇ 제1차('21.9월)부터 제6차('24.4월)까지 진행된 총 6차례의 회의를 통해 추진된 인공지능(AI) 관련 협의 경과는 아래와 같다.

회차	AI관련 주요 협의 진행 경과
1차[62], '21.9.29. (미국, 피츠버그)	• AI의 미래와 공동의 가치 (잠재력, 위험성, 인간 중심) • 신뢰할 수 있는 AI 개발 및 표준화 협력 (기술, 윤리, 안전) • AI의 경제적, 사회적 영향 공동 연구 및 포용적 성장 추구
2차[63], '22.5.16. (프랑스, 파리)	• AI 윤리 및 안전 표준 협력 강화(유럽연합 AI법, 미국 AI 권리 장전, NIST AI 위험 관리 프레임워크 등 양측 AI 이니셔티브 공유를 통한 OECD 권고 발전 및 협력) • 공동 허브 개발을 통한 AI 표준 및 평가 체계 구축 • AI 경제·사회적 영향 공동 연구, 미래 노동시장 공동 대응 모색

61) 美-EU 무역기술위원회(TTC): 미국-EU 간 장관급(미 국무장관, 상무부 장관, 무역대표부(USTR) 대표 및 EU 집행위원회 부위원장 등 공동 주재) 정기 협의체로 운영 (출처- European Commission. (2021). Joint EU-US statement on the Trade and Technology Council (TTC).; 대한무역투자진흥공사(KOTRA). (2024). 제5차 미국-EU 무역기술위원회(TTC) 주요 내용 및 현지 반응 (US24-4))

62) European Commission. (2021, September 29). EU-US Trade and Technology Council Inaugural Joint Statement. Retrieved from https://ec.europa.eu/commission/presscorner/detail/el/STATEMENT_21_4951

63) U.S. Department of Commerce & European Commission. (2022, May 16). U.S.-EU Joint Statement of the Trade and Technology Council. Retrieved from https://www.commerce.gov/sites/default/files/2022-05/US-EU-Joint-Statement-Trade-Technology-Council.pdf

회차	AI관련 주요 협의 진행 경과
3차[64], '22.12.4. (미국 워싱턴 DC)	• AI 로드맵 발표 및 AI 신뢰성 및 위험 측정 기준·방법론 공유 저장소 구축을 통한 AI 위험 관리·신뢰성 평가 및 국제 표준 기구 협력 강화 • 책임감 있는 AI 개발과 개인 정보 보호 균형 모색(개인정보 강화 기술·합성 데이터 활용 파일럿 프로젝트 추진) • 극한 기상 예측, 보건, 전력망, 농업, 응급 대응 등 공익을 위한 AI 및 컴퓨팅 공동 연구 프로젝트 추진
4차[65], '23.5.31 (스웨덴, 롤레오)	• 생성형 AI를 포함하여 신뢰할 수 있는 AI 표준 및 위험 관리 협력 강화(AI 용어 및 분류체계, 신뢰할 수 있는 AI 및 위험 관리를 위한 AI 표준 및 도구 협력, 기존 미 신규 AI 위험 모니터링 및 측정을 위한 각각의 전담 전문가 그룹을 출범하여 AI 로드맵 실행 추진) • 글로벌 과제(극한 기상 및 기후 예측, 응급 대응 관리, 보건 및 의학 개선, 에너지망 최적화, 농업 최적화) 해결을 위한 AI 협력
5차[66], '23.12.6. (미국, 워싱턴 DC)	• G7 AI 지침 원칙 및 행동 강령 지지, 국제 AI 거버넌스 협력 지속
6차[67], '24.4.4. (벨기에, 브뤼셀)	• G7, OECD, G20, UN 등 다자간 협력 지속 및 국제 행동 강령 적용 장려 • 미국-유럽연합 간 AI 대화 협의체(Dialogue) 구축(벤치마크, 위험, 기술 동향 등 과학 정보 교환, 신뢰할 수 있는 AI 표준 및 도구 개발 협력) • 미국-유럽간 워킹 그룹을 통한 글로벌 과제(극한 기상, 에너지, 응급 대응, 재건) 협력 및 영국, 캐나다, 독일 등 파트너와 협력하여 아프리카 AI 개발 지원

ㅇ 위와 같이 미국과 유럽연합은 무역기술위원회와 같은 협의체를 통해 개별 국가의 법적·윤리적 접근을 넘어, AI 규제와 국제 기술의 표준을 정립하기 위한 협력을 강화하고 있다.

64) European Commission. (2022, December 5). Joint Statement following the third EU-U.S. Trade and Technology Council. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/statement_22_7516
65) European Commission. (2023, June 15). Joint Statement following the fourth EU-U.S. Trade and Technology Council. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/statement_23_2992
66) European Commission. (2024, April 5). Fifth meeting of the EU-U.S. Trade and Technology Council. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/ip_24_575
67) European Commission. (2024, April 5). Joint Statement EU-US Trade and Technology Council. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_24_1828

유럽연합 인공지능법(AI Act)과 미국 AI 윤리지침 및 행정명령은 일부 정책적 차이를 보이면서도, 책임성·투명성·안전성 확보 등 공통된 가치와 목표를 공유하고 있다. 따라서, 양측 협력의 핵심은 글로벌 기술 기업들이 준수해야 할 일관된 국제 컴플라이언스 기준을 마련하고 이를 선도하는 데 있다.

ㅇ 美-EU 무역기술위원회는 AI 기술 표준화 및 규제 조율을 위한 핵심 협의체로 자리 잡고 있으며, AI 개발 및 배포와 관련한 지침 마련, 리스크 평가 기준 정립, 기술 투명성 강화 등의 논의를 지속하고 있다. 특히, 각국의 AI 규제 체계를 조율하여 책임성을 확보하는 방향으로 나아가고 있으며, 글로벌 AI 기업들은 이를 기반으로 기술적·윤리적 대응 전략을 구축하고 있다. 이는 단순한 규제 대응을 넘어, AI 시스템을 설계·운영하는 방식 전반에 걸친 새로운 기준 확립을 의미한다.

ㅇ 다음 장에서는 미국 주요 AI 기업들이 미국과 유럽연합의 AI 규제 및 윤리지침 준수를 위해 채택한 기술적·법적 접근 방식과 대응 전략을 검토하고, 이를 바탕으로 시사점을 도출하고자 한다.

Ⅲ. 미국 글로벌 AI 기업 규제 준수 현황 및 대응 전략

1. 주요 AI 기업의 투명성 현황

ㅇ 지금까지 미국과 유럽연합의 AI 윤리 원칙 및 규제에 대해 살펴보았다. 이제 글로벌 AI 기업들이 이 같은 윤리원칙 및 규제 요구사항에 어느 정도 부합하고 있는 지 살펴보기 위해, 본 절(Section)에서는 "The Foundation Model Transparency Index(FMTI)"를 활용한 스탠포드 대학 보고서[68]를 통해 주요 파운데이션 모델의 투명성 및 각 기업의 규제 준수 현황을 살펴보고자 한다.

(1) The Foundation Model Transparency Index(FMTI) 개요

ㅇ 유럽연합 인공지능법(AI Act)과 미국의 AI 권리장전·NIST AI 리스크 관리 프레임워크·NIST 생성형 AI 프로파일·행정명령 제14110호 등 AI 윤리 관련 규제 및 가이드라인이 마련되면서, 실제 AI 기업들이 이를 어느 정도 수준으로 준수하고 있는지를 평가하는 지표가 필요하게 되었다.

이에 따라, 스탠포드 대학의 파운데이션 모델 연구센터(Center for Research on Foundation Models, CRFM)에서는 스탠포드, MIT, 프린스턴 대학 출신의 다학적 연구팀을 구성하여, AI 모델의 투명성을 정량적으로 평가할 수 있는 "The Foundation Model Transparency Index(FMTI)"를 개발하였다[69].

ㅇ 파운데이션 모델 투명성 지수(FMTI)의 평가 지표는 유럽연합 인공지능법과 미국의 AI 관련 윤리지침 및 행정명령에서 강조하는 투명성, 설명 가능성, 안전성 등의 요소를 반영하여 설계되어, 이를 통해 AI 기업들이 규제 요구 사항을 어느 수준으로 충족하는지 간접적으로 평가할 수 있다. IBM은 자사의 Granite 모델에 대해 FMTI의 평가 기준을 적용하여, 학습 데이터 출처·모델 구성·위험 완화 조치 등에 반영함으로써, AI

68) Stanford Center for Research on Foundation Models. (2024, May). The Foundation Model Transparency Index v1.1. Retrieved from https://crfm.stanford.edu/fmti/paper.pdf.
69) Miller, K. (2023, October 18). Introducing the Foundation Model Transparency Index. Stanford Institute for Human-Centered AI. Retrieved from https://hai.stanford.edu/news/introducing-foundation-model-transparency-index.

윤리 기준에 부합하기 위해 노력하고 있다[70]. 이 같이 FMTI 지표는 현재 AI 기업을 비롯해 스탠포드 인간중심 AI 연구소(HAI)의 AI Index 보고서 등 연구와 정책 논의에도 중요한 참고 지표로 활용되고 있다[71].

(2) FMTI 구성 및 평가방식

1) FMTI 구성

○ 파운데이션 모델 투명성 지수(FMTI)의 지표는 투명성을 포괄적으로 평가하기 위해 3개 주요 영역의 100개 지표로 구성된다.

① 업스트림 지표(Upstream Indicators)
• 파운데이션 모델 구축과 관련한 구성 요소 및 절차 확인 • 6개 하위 영역 총 32개 지표로 구성 ❶데이터(10개), ❷데이터 노동(7개), ❸데이터 접근(2개), ❹컴퓨팅(7개), ❺방법(4개), ❻데이터 완화(2개)
② 모델 지표(Model Indicators)
• 파운데이션 모델의 속성과 기능을 확인 • 8개 하위 영역 총 33개 지표로 구성 ❶모델 기본정보(6개), ❷모델 접근성(3개), ❸성능(5개), ❹한계(3개), ❺리스크(7개), ❻모델 완화(5개), ❼신뢰성(2개), ❽추론(2개)
③ 다운스트림 지표(Downstream Indicators)
• 파운데이션 모델의 사용 및 배포 관련 세부사항 확인 • 9개 하위 영역 총 35개 지표로 구성 ❶배포(7개), ❷사용 정책(5개), ❸모델 행동 정책(3개), ❹사용자 인터페이스(2개), ❺사용자 데이터 보호(3개), ❻모델 업데이트(3개), ❼피드백(3개), ❽사회적 영향(7개), ❾다운스트림 문서화(2개)

(i) 업스트림 지표(Upstream Indicators)

❶데이터(Data): 10개 지표 - 모델 구축에 사용된 데이터의 규모 및 구성과 관련한 투명성 평가 - 데이터에 나타난 콘텐츠의 창작자와 관련한 투명성 평가 - 데이터의 큐레이션(curate) 및 증강(augment)과정과 관련한 투명성 평가 - 개인정보·저작권보호·라이선스 데이터 포함 여부와 관련한 투명성 평가

70) IBM Research. (2024, May 21). IBM Granite and the Foundation Model Transparency Index: Advancing transparency in AI. IBM Research Blog. Retrieved from https://research.ibm.com/blog/ibm-granite-transparency-fmti

71) Jackson, K. (2024, April 18). Stanford HAI AI Index report highlights responsible AI challenges. AI Wire. Retrieved from https://www.aiwire.net/2024/04/18/stanford-hai-ai-index-report-responsible-ai/

• **데이터 규모(Data size):** 모델 구축에 사용된 데이터 규모가 공개되는가?
• **데이터 출처(Data source):** 모델 구축에 사용된 데이터 출처가 공개되는가?
• **데이터 창작자(Data creator):** 모델 구축 데이터를 창작한 사람들에 대한 속성이 제시되는가?
• **데이터 출처 선택(Data source selection):** 데이터 출처를 포함하거나 제외하는 선정 프로토콜이 공개되는가?
• **데이터 큐레이션(Data curation):** 모든 데이터 출처에 대한 큐레이션 프로토콜이 공개 되는가?
• **데이터 증강(Data augmentation):** 데이터 출처를 증강(보강)하기 위해 개발자가 취하는 절차가 공개되는가?
• **유해 데이터 필터링(Harmful data filtration):** 유해한 콘텐츠를 제거하기 위해 데이터가 필터링되는 경우, 관련 필터에 대해 설명이 되는가?
• **저작권 데이터(Copyrighted data):** 모델 구축에 사용된 모든 데이터에 대해, 관련 저작권 현황이 공개되는가?
• **데이터 라이선스(Data license):** 모델 구축에 사용된 모든 데이터에 대해, 관련 라이선스 현황이 공개되는가?
• **데이터 내 개인정보(Personal information in data):** 모델 구축에 사용된 데이터에 개인정보의 포함 또는 제외 여부가 공개되는가?
❷데이터 노동(Data Labor): 7개 지표 - 모델 구축에 사용된 데이터 생성 과정에 인간 노동 사용과 관련한 투명성 평가 (데이터 주석(annotation) 및 큐레이션 작업에 참여한 노동자의 급여, 노동 보호 조치, 고용주 및 지리적 분포 포함) - 모델 구축을 위해 파운데이션 모델 개발사와 파트너 협력을 맺은 제3자 관련 투명성 평가
• **인간 노동 사용(Use of human labor):** 인간 노동이 관여된 데이터 파이프라인의 단계가 공개되는가?
• **데이터 노동자 고용(Employment of data laboreres):** 데이터 노동에 관여된 사람들을 직접 채용한 조직이 데이터 파이프라인 각 단계별로 공개되는가?
• **데이터 노동자 지리적 분포(Geographic distribution of data laborers):** 데이터 노동에 관여된 사람들에 대한 지리적 정보가 데이터 파이프라인 각 단계별로 공개되는가?
• **급여(wages):** 데이터 노동을 수행하는 사람들의 급여가 공개되는가?
• **데이터 생성 지침(Instructions for creating data):** 데이터 노동을 수행하는 사람들에게 제공된 지침이 공개되는가?
• **노동 보호(Labor protections):** 데이터 노동을 수행하는 사람들을 위한 노동 보호 조치가 공개되는가?
• **제3자 파트너(Third party partners):** 모델 개발에 참여했거나 참여하고 있는 제3자가 공개되는가?

❸데이터 접근(Data Access): 2개 지표 - 외부 관계자들을 고려해 데이터 접근 범위를 평가
• **쿼리를 통한 외부의 데이터 접근 가능 여부(Queryable external data access):** 외부 기관이 쿼리를 통해 데이터 구축에 사용된 데이터에 접근이 가능한가?
• **직접적인 외부의 데이터 접근 가능 여부(Direct external data acess):** 외부 기관이 모델 구축에 사용된 데이터에 직접적으로 접근이 가능한가?
❹컴퓨팅(Compute): 7개 지표 - 에너지 사용 결과 및 환경적 영향을 포함해 모델 구축에 사용된 하드웨어 및 연산(computation)과 관련한 투명성 평가
• **컴퓨팅 리소스 사용량(Compute usage):** 모델 구축에 필요한 컴퓨팅 리소스(compute)가 공개되는가?
• **개발 기간(Development duration):** 모델 구축에 필요한 시간이 공개되는가?
• **컴퓨팅 하드웨어(Compute hardware):** 모델 구축에 사용된 주요 하드웨어의 유형 및 수량이 공개되는가?
• **하드웨어 소유자(Hardware owner):** 모델 구축에 사용된 주요 하드웨어의 소유자가 공개되는가?
• **에너지 사용(Energy usage):** 모델 구축에 사용된 에너지 양이 공개되는가?
• **탄소 배출(Carbon emissions):** 모델 구축 과정에 (에너지 사용과 연관된) 탄소 배출량이 공개되는가?
• **광범위한 환경 영향(Broader environmental impact):** 탄소 배출량 외에 모델 구축으로 인한 광범위한 환경 영향이 공개되는가?
❺방법(Methods): 4개 지표 - 모델 훈련 과정 및 목표에 대한 기본적인 기술적 사양 평가(사용된 소프트웨어 프레임워크 및 의존성 포함)
• **모델 단계(Model stage):** 모델 개발 과정의 모든 단계가 공개되는가?
• **모델 목표(Model objectives):** 모든 단계를 설명할 때, 관련 학습 목표와 모델 업데이트의 본질적 특성이 명확하게 제시되어 있는가?
• **핵심 프레임워크(Core frameworks):** 모델 개발에 사용된 핵심 프레임워크가 공개되는가?
• **추가 의존성(Additional dependencies):** 데이터, 컴퓨팅, 코드 외에 모델 구축에 필요한 다른 의존성이 공개되는가?
❻데이터 완화(Data Mitigations): 2개 지표 - 데이터 프라이버시 및 저작권 문제 완화 조치와 관련한 투명성 평가
• **프라이버시 완화(Mitigations for privacy):** 데이터 내에 있는 개인정보를 보호하기 위해 개발자가 취한 조치를 공개하는가?
• **저작권보호 완화(Mitigations for copyright):** 개발자가 데이터에 저작권이 있는 정보를 줄이기 위해 취한 조치를 공개하는가?

(ii) 모델 지표(Model Indicators)

❶ **모델 기본정보(Model Basics): 6개 지표** - 입력 방식(modalities), 크기(size), 아키텍처(architecture), 중앙화된 문서화의 존재 등 모델과 관련한 기본정보에 대한 투명성 평가
• **입력 방식(Input modality):** 모델의 입력 방식(input modalities)이 공개되는가?
• **출력 방식(Output modality):** 모델의 출력 방식(output modalities)이 공개되는가?
• **모델 구성요소(Model components):** 모델의 모든 구성 요소가 공개되는가?
• **모델 크기(Model size):** 모델의 모든 구성 요소에 대한 모델 크기가 공개되는가?
• **모델 아키텍처(Model architecture):** 모델 아키텍처가 공개되는가?
• **중앙화된 모델 문서화(Centralized model documentation):** 모델에 대한 주요 정보가 모델 카드와 같은 중앙화된 문서에 포함되는가?
❷ **모델 접근성(Model Access): 3개 지표** - 외부 기관에 허용된 모델 접근 범위 평가
• **외부 모델 접근 프로토콜(External model access protocol):** 외부 기관이 모델에 접근할 수 있는 프로토콜이 공개되는가?
• **블랙박스 외부 모델 접근(Blackbox external model access):** 외부 기관에 모델 접근 허용이 블랙박스 방식으로 이루어지는가?
• **완전한 외부 모델 접근(Full external model access):** 외부 기관에 모든 부분에 대한 모델 접근이 허용되는가?
❸ **성능(Capabilities): 5개 지표** - 모델의 성능(평가(evaluation) 포함)과 관련한 투명성 평가
• **성능 설명(Capabilities description):** 모델 성능이 설명되는가?
• **성능 시연(Capabilities demonstration):** 모델 성능이 시연되는가?
• **성능 평가(Evaluation of capabilities):** 모델 성능이 엄격히 평가되고, 모델 초기 배포 이전에 또는 동시에 이 같은 평가 결과가 보고되는가?
• **성능 평가의 외부 재현성(External reproductibility of capabilities evaluation):** 모델의 성능 평가를 외부 기관이 재현할 수 있는가?
• **제3자 성능 평가(Third party capabilities evaluation):** 제3자가 모델 성능을 평가하는가?
❹ **한계(Limitations): 3개 지표** - 모델의 한계(평가(evaluation) 포함)와 관련한 투명성 평가
• **한계 설명(Limitation description):** 모델의 한계가 공개되는가?
• **한계 시연(Limitation demonstration):** 모델의 한계가 시연되는가?
• **제3자 한계 평가(Third party evaluation of limitations):** 제3자가 모델의 한계를 평가할 수 있는가?

❺ 리스크(Risks): 7개 지표 - 모델의 리스크(평가(evaluation) 포함)와 관련한 투명성 평가, 특히 의도하지 않은 피해(예: 편향)와 의도적인 악용(예: 사기)에 초점
• **리스크 설명(Risks description):** 모델의 리스크가 공개되는가?
• **리스크 시연(Risks demonstration):** 모델의 리스크가 시연되는가?
• **비의도적인 유해성 평가(Unintentional harm evaluation):** 비의도적 유해성 관련 모델 리스크가 엄격히 평가되고, 그 평가 결과가 모델 초기 배포 전 또는 동시에 보고되는가?
• **비의도적 유해성 평가의 외부 재현성(External reproducibility of unintentional harm evaluation):** 외부 기관이 비의도적 유해성과 관련한 모델 리스크 평가를 재현할 수 있는가?
• **의도적 유해성 평가(Intentional harm evaluation):** 의도적 유해성과 관련한 모델 리스크가 엄격히 평가되고, 그 평가 결과가 모델의 초기 배포 전 또는 동시에 보고되는가?
• **의도적 유해성 평가의 외부 재현성(External reproducibility of intentional harm evaluation):** 외부 기관이 의도적 유해성과 관련한 모델 리스크 평가를 재현할 수 있는가?
• **제3자 리스크 평가(Third party risks evaluation):** 제3자가 모델 리스크를 평가하는가?
❻ 모델 완화(Model Mitigations): 5개 지표 - 모델에 적용된 완화조치(model-level mitigations) 및 그 효과성 평가에 대한 투명성 평가
• **완화 설명(Mitigations description):** 모델의 완화 조치가 공개되는가?
• **완화 시연(Mitigations demonstration):** 모델의 완화 조치가 시연되는가?
• **완화 평가(Mitigations evaluation):** 모델의 완화 조치가 엄격히 평가되고, 그 평가 결과가 보고되는가?
• **완화 평가의 외부 재현성(External reproducibility of mitigations evaluation):** 외부 기관이 모델의 완화조치 평가를 재현할 수 있는가?
• **제3자 완화 평가(Third party mitigations evaluation):** 제3자가 모델의 완화조치를 평가할 수 있는가?
❼ 신뢰성(Trustworthiness): 2개 지표 - 모델의 신뢰성(평가(evaluation) 포함)에 관한 투명성 평가
• **신뢰성 평가(Trustworthiness evaluation):** 모델의 신뢰성이 엄격히 평가되고, 그 평가 결과가 공개되는가?
• **신뢰성 평가의 외부 재현성(External reproducibility of trustworthiness evaluation):** 외부 기관이 신뢰성 평가를 재현할 수 있는가?
❽ 추론(Inference): 2개 지표 - 모델 추론의 표준화에 대한 투명성 평가
• **추론 시간 평가(Inference duration evaluation):** 특정 하드웨어 환경에서 특정 작업에 대해 모델 추론에 필요한 시간이 공개되는가?
• **추론 컴퓨팅 평가(Interence compute evaluation):** 특정 하드웨어 환경에서 특정 작업에 대해 모델 추론을 위한 컴퓨팅 리소스 사용량(compute usage)이 공개되는가?

(iii) 다운스트림 지표(Downstream Indicators)

❶배포(Distribution): 7개 지표 - 출시 과정, 모델 배포 채널, 내부 사용을 통해 발생한 상품 및 서비스와 관련한 투명성 평가 - 모델 라이선스, 약관 및 모델 생성 콘텐츠 감지 매커니즘 존재 평가
• **출시 의사결정(Release decision-making):** 모델 출시 여부를 결정하는 개발자의 프로토콜이 공개되는가?
• **출시 과정(Release process):** 모델 출시 과정에 대한 설명이 공개되는가?
• **배포 채널(Distribution channels):** 모든 배포 채널이 공개되는가?
• **상품 및 서비스(Products and services):** 개발자가 제공하는 상품 및 서비스가 해당 모델에 의존적인지 여부를 공개하는가?
• **기계 생성 콘텐츠의 감지(Detection of machine-generated content):** 모델이 생성한 콘텐츠를 감지하는 매커니즘이 공개되는가?
• **모델 라이선스(Model License):** 모델 라이선스가 공개되는가?
• **이용약관(Terms of service):** 각 배포 채널의 이용 약관이 공개되는가?
❷사용 정책(Usage Policy): 5개 지표 - 특정 사용이나 사용자에 대한 제한과 같은 개발자의 허용 가능한 사용 정책과 해당 정책 시행 방식에 대한 투명성 평가
• **허용 및 금지된 사용자(Permitte and prohibited users):** 모델을 누가 사용할 수 있고 사용할 수 없는 지에 대한 설명이 공개되는가?
• **허용, 제한 및 금지된 사용(Permitted, restricted, and prohibited uses):** 모델의 허용·제한·금지된 사용이 공개되는가?
• **사용 정책 시행(Usage policy enforcement):** 사용 정책에 대한 시행 프로토콜이 공개되는가?
• **시행 조치 정당성(Justification for enforcement action):** 사용자가 사용 정책 위반으로 시행 조치 대상이 된 경우, 이에 대한 정당한 이유를 제공받는가?
• **사용 정책 위반 이의 제기 매커니즘(Usage policy violation appeals mechanism):** 사용 정책 위반에 대한 이의제기 매커니즘이 공개되는가?
❸모델 행동 정책(Model Behavior Policy): 3개 지표 - 허용 가능한 모델 행동 및 허용 불가능한 모델 행동에 대한 개발자 정책과 관련한 투명성 평가 - 정책 시행 및 위반 시 기대되는 조치에 대한 투명성 평가
• **허용·제한·금지된 모델 행동(Permitted, restricted, and prohibited model behaviors):** 허용·제한·금지된 모델 행동이 공개되는가?
• **모델 행동 정책 시행(Model bahavior policy enforcement):** 모델 행동 정책의 시행 프로토콜이 공개되는가?
• **사용 정책과 모델 행동 정책의 연계(Interoperability of usage and model behavior policies):** 사용 정책과 모델 행동 정책의 연계 방식이 공개되는가?

❹사용자 인터페이스(User Interface): 2개 지표
- 개발자의 주요 배포 채널에 사용자 인터페이스가 포함되는 경우, 사용자 인터페이스에 대한 투명성 평가

- **사용자와 AI 시스템 상호작용(User interaction with AI system):** 사용자가 직접 접하는 인터페이스가 있는 배포 채널에서, 사용자가(i) AI 시스템과 상호작용하고 있음을 고지받는가?, (ii) 사용자가 상호작용하는 특정 파운데이션 모델을 고지받는가?, (iii) 결과물이 기계가 생성한 것임을 고지받는가?
- **사용 면책 조항(Usage disclaimers):** 사용자가 직접 접하는 인터페이스가 있는 배포 채널의 경우, 사용자에게 모델 사용에 대한 면책 조항이 제공되는가?

❺사용자 데이터 보호(User Data Protection): 3개 지표
- 데이터가 저장·공유·접근되는 방식 등 사용자 데이터 보호와 관련한 개발자 정책 투명성 평가

- **사용자 데이터 보호 정책(User data protection policy):** 개발자가 사용자 데이터를 저장·접근·공유하는 방식에 대한 포로토콜이 공개되는가?
- **사용자 데이터의 허용 및 금지된 사용(Permitted and prohibited use of user data):** 사용자 데이터의 허용된 사용과 금지된 사용이 공개되는가?
- **사용 데이터 접근 프로토콜(Usage data access protocol):** 외부 기관이 사용 데이터에 접근할 수 있는 권한 부여 프로토콜이 공개되는가?

❻모델 업데이트(Model Update): 3개 지표
- 개발자의 버전 관리 프로토콜, 변경 기록 및 중단 정책에 대한 투명성 평가

- **버전 관리 프로토콜(Versioning protocol):** 모델 버전 및 버전 관리 프로토콜이 공개되어 있는가?
- **변경 기록(Change log):** 모델의 변경 기록이 공개되어 있는가?
- **중단 정책(Deprecation policy):** 개발자가 적용하는 중단 정책이 공개되어 있는가?

❼피드백(Feedback): 3개 지표
- 모델에 대한 피드백, 접수된 피드백 요약 및 관련 정부 요청 보고 매커니즘에 대한 투명성 평가

- **피드백 메커니즘(Feedback mechanism):** 피드백 매커니즘이 공개되는가?
- **피드백 요약(Feedback summary):** 개발자가 받은 피드백 또는 개발자가 피드백에 대응한 방식과 관련한 보고나 요약이 공개되는가?
- **정부 요청(Government inquiries):** 모델과 관련해서 개발자가 받은 정부 요청에 대한 요약이 공개되는가?

❽영향(Impact): 7개 지표
- 모델이 사회에 미치는 다운스트림 영향(영향받는 시장 부문, 개인 및 지역 등)에 대한 투명성 평가
- 다운스트림 애플리케이션, 사용 통계, 사용 모니터링 매커니즘, 사용자에 해를 끼친 경우에 시정 조치에 대한 투명성 평가

• **모니터링 매커니즘(Monitoring mechanism):** 각 배포 채널에 대해 모델 사용을 추적하기 위한 모니터링 매커니즘이 공개되는가?
• **다운스트림 어플리케이션(Downstream applications):** 모든 형태의 다운스트림 어플리케이션 사용 전반에 걸쳐, 파운데이션 모델에 의존하는 어플리케이션의 수가 공개되는가?
• **영향받는 시장 부문(Affected market sectors):** 모든 다운스트림 어플리케이션에 걸쳐, 각 마켓 부문에 상응하는 어플리케이션의 비율이 공개되는가?
• **영향받는 개인(Affected individuals):** 모든 형태의 다운스트림 사용에 걸쳐 파운데이션 모델에 영향받는 개인의 수가 공개되는가?
• **사용 보고(Usage reports):** 사용자에 대한 모델의 영향을 설명하는 사용 통계가 제시된 사용 보고서가 공개되는가?
• **지리적 통계(Geographic statistics):** 모든 형태의 다운스트림 사용에 걸쳐, 지역별 모델 사용 통계가 공개되는가?
• **구제 매커니즘(Redress mechanism):** 피해에 대해 사용자에게 구제 조치를 제공하는 매커니즘이 공개되는가?
❾다운스트림 문서화(Downstream Documentation): 2개 지표 - 다운스트림 사용에 대한 중앙화된 문서화 및 책임있는 다운스트림 사용을 위한 문서화의 존재 평가
• **다운스트림 사용에 대한 중앙 문서화(Centralized documentation for downstream use):** 다운스트림 사용 관련 문서가 중앙에서 관리되는 문서로 제공되는가?
• **책임있는 다운스트림 사용의 문서화(Documentation for responsible downstream use):** 책임있는 다운스트림 사용을 위한 문서가 공개되는가?

2) 평가방식

ㅇ 파운데이션 모델 투명성 지수(FMTI)는 2023년 Ver.1.0을 최초로 발표한 이후, 2024년 Ver.1.1.로 개정되면서, 아래와 같이 개발사들의 참여 방식과 평가 방법이 변경되었다.

항목	FMTI Ver.1.0 (2023)	FMTI Ver.1.1 (2024)
평가지표	• 100개 지표	• 동일(100개 지표 유지)
평가방식	• 연구팀이 공개된 정보만을 수집하여 평가	• 개발사들이 100개 지표에 대해 직접 제출한 보고서를 평가
점수산정	• 이진 점수(0(비공개), 1(공개))	• Ver.1.0과 동일(이진 점수 유지)
개발사 피드백	• 2주간 이의제기 가능	• 이메일·화상회의 등을 통해 반복적 피드백 과정 추가
대상기업[72]	• 10개 개발사	• 14개 개발사
대표모델	• 연구팀이 선정	• 개발사가 지정

항목	FMTI Ver.1.0 (2023)	FMTI Ver.1.1 (2024)
평균점수	• 37/100	• 58/100 (21점 상승)
공개방식	• 점수 공개	• 기업별 투명성 보고서 함께 공개

(3) 주요 AI 기업 투명성 현황 분석

○ 본 연구는 최근 보고서인 FMTI Ver.1.1을 기반으로 기업별 투명성 현황을 분석[73]하였다. 특히, 문화콘텐츠 분야의 미국 사례 중심 연구임을 반영하여, 평가 대상 14개 기업 중, 미국 기반 인공지능(AI) 선도 기업[74]으로, 콘텐츠 생성 및 유통과 밀접한 관련이 있는 아래 5개 기업의 AI 모델 투명성 현황을 검토하였다.

기업명	FMTI Ver.1.1. 평가 대상 모델	문화콘텐츠 분야 주요 기능
OpenAI	GPT-4	• 텍스트 및 이미지 생성
Google	Gemini 1.0 Ultra	• AI 번역, 음악 및 영상 데이터 처리
Meta	Llama 2	• 텍스트 기반 콘텐츠 제작 등 창작 지원
Microsoft	Phi-2	• 텍스트 기반 콘텐츠 창작 지원
Anthropic	Claude 3	• 대화형 AI 기반 콘텐츠 창작 지원

1) 평가결과

○ 파운데이션 모델 투명성 지수(FMTI)는 앞서 살펴본 바와 같이, 업스트림(Upstream), 모델(Model), 다운스트림(Downstream)의 세 가지 영역에 속하는 100개 지표를 평가한다. FMTI Ver.1.1에서 평가한 14개 대상기업 중, 분석 대상 5개 기업의 영역별 결과 및 순위는 다음과 같다.

---

72) 기업 유형, 파운데이션 모델 유형 등 기업의 다양성을 고려 선정(출처: Stanford Center for Research on Foundation Models. (2024, May). The Foundation Model Transparency Index v1.1)

73) Stanford Center for Research on Foundation Models. (2024, May). Foundation Model Transparency Index (FMTI) May 2024 dataset. GitHub. Retrieved from https://github.com/stanford-crfm/fmti/tree/main/May2024

74) Geneva Internet Platform. (2023, October 30). Google DeepMind, Meta, OpenAI among top AI firms sharing safety policies ahead of UK summit. Geneva Internet Platform. Retrieved from https://dig.watch/updates/google-deepmind-meta-openai-among-top-ai-firms-sharing-safety-policies-ahead-of-uk-summit; Antonio, V. (2025, January 3). What is the difference between OpenAI, Google DeepMind, Anthropic, and Microsoft AI? Foreign Affairs Forum. Retrieved from https://www.faf.ae/home/2025/1/3/what-is-difference-between-open-ai-and-google-deepmind-anthropic-and-microsoft-ai

기업명	평가모델	업스트림 (Upstream)	모델 (Model)	다운스트림 (Downstream)	총점	순위
Microsoft	Phi-2	20/32 (62.5%)	18/33 (54.55%)	24/35 (68.57%)	62	5위
Meta	Liama 2	15/32 (46.88%)	25/33 (75.76%)	20/35 (57.14%)	60	6위
Anthropic	Claude 3	7/32 (21.88%)	21/33 (63.64%)	23/35 (65.71%)	51	10위
OpenAI	GPT-4	7/32 (21.88%)	20/33 (60.61%)	22/35 (62.86%)	49	11위
Google	Gemini 1.0 Ultra	6/32 (18.75%)	18/33 (54.55%)	23/35 (65.71%)	47	12위

ㅇ 위 표에서 알 수 있듯이, Microsoft(5위)와 Meta(6위)는 총 14개 기업 중 상위 50%내에 해당하는 비교적 높은 투명성을 기록하였고, Anthropic(10위), OpenAI(11위)와 Google(12위)은 하위 50%에 속하는 낮은 투명성을 보였다.

Microsoft를 제외한 4개 기업은 모두 3가지 영역 중 업스트림(upstream) 영역에서 다른 2개 영역 대비 상대적으로 낮은 점수를 기록하고 있어, 기업들이 해당 영역의 정보를 공개하지 않는 경향을 보였다.

2) 공통 미공개 지표 분석

ㅇ AI 윤리 및 규제 프레임워크에서 핵심은 기업이 AI 모델이 생성하고 활용하는 데이터·훈련 과정·성능 평가·사회적 영향 등을 공개하는지 여부이다. 따라서, 주요 AI 기업의 투명성 현황 분석 시, 평가 대상 5개 기업이 공통적으로 공개하지 않은 16개 지표를 중점적으로 검토하였다.

기업이 특정 지표를 공통적으로 공개하지 않고 있다는 것은 구조적·전략적 이유가 존재할 가능성을 시사한다. 따라서, 비공개 16개 지표를 평가 대상 3개 영역인 업스트림, 모델, 다운스트림의 3개 영역으로 나누어 분석하고, 5개 기업이 공통적으로 불투명성을 보이는 부분을 구체적으로 살펴보았다.

(i) 업스트림 미공개 지표

ㅇ 파운데이션 모델 구축과 관련한 구성 요소 및 절차의 투명성을 평가하는 업스트림 영역은 기업들이 가장 많이 공개하고 있지 않은 부분이다. 5개 기업이 공통적으로 공개하고 있지 않은 지표는 아래 8개 지표로, 미공개 사유를 함께 검토하였다.

하위영역	미공개 지표	공개 시 우려사항
데이터 (Data)	데이터 창작자 (Data creator)	• 공개 시, 데이터 출처와 저작권 문제 등 법적 논란 발생 가능성
	데이터 저작권 (Data copyright status)	• 데이터 저작권 및 라이선스 정보 공개 시, 법적 분쟁 발생 가능성
	데이터 라이선스 (Data license status)	
	데이터 내 개인정보 (Personal information in data)	• 다양한 개인정보 보호 규제 위반 제기 가능성
데이터 접근 (Data Access)	쿼리를 통한 외부의 데이터 접근 가능 여부 (Queryable external data access)	• 개인정보 보호 및 데이터 보안 문제에 대한 추가 규제 제기 가능성 • 한정된 기관에만 데이터 접근 권한을 부여할 경우, 공정한 데이터 접근 관련 논란 발생 가능성
	직접적인 외부의 데이터 접근 가능 여부 (Direct external data access)	
컴퓨팅 (Compute)	광범위한 환경 영향 (Broader environmental impact)	• 환경적 책임 문제 제기 가능성
데이터 완화 (Data Mitigation)	저작권보호 완화 (Mitigations for copyright)	• 저작권 보호조치 미흡 등 법적 책임 문제 제기 가능성

(ii) 모델 미공개 지표

ㅇ 모델 영역은 파운데이션 모델의 속성과 기능에 대한 투명성을 평가하며, 5개 기업 모두 공개하고 있지 않은 지표는 아래 4개 지표이다.

하위영역	미공개 지표	공개 시 우려사항
한계 (Limitation)	한계 시연(Limitations demonstration)	• 모델의 한계가 공개될 경우, 신뢰도 하락 및 규제 강화 가능성

하위영역	미공개 지표	공개 시 우려사항
모델완화 (Model Mitigations)	완화 평가의 외부 재현성(External reproducibility of mitigations evaluation)	• 완화 조치 불충분 등 문제 제기 및 법적·규제적 책임 강화 가능성 • 산업 표준 대비 부족 또는 문제 발견 시, 기업 신뢰도·시장 경쟁력 저하 가능성
신뢰성 (Trustworthiness)	신뢰성 평가의 외부 재현성(External reproducibility of trustworthiness evaluation)	• 모델의 평가 및 검증 절차에 대해 외부 기관의 감사를 받게 될 가능성 • 외부 기관 재현을 위해 데이터 접근 허용 시, 기업 내부 운영 방식이나 알고리즘 세부 내용 누출 가능성
추론 (Inference)	추론 컴퓨팅 평가 (Inference compute evaluation)	• 모델 최적화 방식 및 기술적 노하우·효율성·비용 구조 노출 위험

(iii) 다운스트림 미공개 지표

ㅇ 다운스트림 영역은 파운데이션 모델의 사용 및 배포와 관련한 투명성을 평가하며, 5개 기업 공통 미공개 지표는 "영향"에 속하는 아래 4개 지표이다.

하위영역	미공개 지표	공개 시 우려사항
영향 (Impact)	영향받는 시장 부문 (Affected market sectors)	• 시장 점유율이 독점적일 경우, 공정경쟁 저해 및 타 기업 접근 제한 문제 등 제기 가능성 • 규제 당국의 기술 독점 활용 등에 대한 조사, 반독점 규제 및 경쟁법 위반 여부 검토 가능성 • 서비스 가격 및 시장 지배력 문제 제기 가능성
	영향받는 개인 (Affected individuals)	• 특정 사용자 그룹의 영향 분석이 평향성과 차별 문제 논란 야기 가능 • 차별적 영향에 대한 법적·윤리적 책임 문제, 공공 신뢰도 저하 및 규제 조치 추가 가능성
	사용 보고 (Usage reports)	• 특정 사용자 그룹에 미치는 영향이 편향성 논란으로 이어질 가능성, 이로 인한 사회적 논란 및 규제 당국의 조사 초래 가능성 • 기업의 타겟층 비즈니스 전략 노출 위험
	지리적 통계 (Geographic statistics)	• 지역별 차별로 보일 경우, 디지털 격차, 공정한 기술 배분에 대한 문제 제기 가능성

3) 미공개 요인 및 시사점

ㅇ 위와 같이 기업들의 공통 미공개 지표를 구체적으로 분석해 본 결과, 특정 지표를 공통적으로 공개하지 않는 사유는 아래와 같이 크게 세 가지로 볼 수 있다.

① 기술 보호 및 경쟁 우위 유지

영역	하위영역	미공개 지표
업스트림 (Upstream)	데이터 접근 (Data Access)	• 쿼리를 통한 외부의 데이터 접근 가능 여부(Queryable external data access) • 직접적인 외부의 데이터 접근 가능 여부(Direct external data acess)
모델 (Model)	한계(Limitation)	• 한계 시연(Limitations demonstration)
	모델완화 (Model Mitigations)	• 완화 평가의 외부 재현성(External reproducibility of mitigations evaluation)
	신뢰성 (Trustworthiness)	• 신뢰성 평가의 외부 재현성(External reproducibility of trustworthiness evaluation)
	추론(Inference)	• 추론 컴퓨팅 평가(Inference compute evaluation)

- 외부 기관의 데이터 접근 허용, 모델 자체의 한계점, 외부 검증 가능성, 추론 연산 정보 공개와 관련된 지표들은 기술 전략 및 최적화 방식 등 기업의 핵심 기술력 공개 여부를 평가함. 따라서, 기업에 생존과 직결되는 정보 노출이라는 부담으로 작용할 수 있음.
- 기업들은 기술 보호 및 경쟁 우위 유지 측면에서 공통적으로 이 같은 지표들을 비공개로 유지하는 경향을 보임

② 비즈니스 및 시장논리

영역	하위영역	미공개 지표
다운스트림 (Downstream)	영향(Impact)	• 영향받는 시장 부문(Affected market sectors)

- AI 모델이 특정 산업에 미치는 영향을 공개할 경우, 시장 진입 전략, 점유율, 타겟층 등 시장 경쟁력이 노출될 우려가 있음. 이 때문에, 기업들이 공통적으로 해당 지표를 비공개로 유지하는 경향을 보임

③ 규제 대응 및 사회적 논란 최소화

영역	하위영역	미공개 지표
업스트림 (Upstream)	데이터 (Data)	• 데이터 창작자(Data creator), • 데이터 저작권(Data Copyright status) • 데이터 라이선스(Data License status) • 데이터 내 개인정보(Personal information in data)

영역	하위영역	미공개 지표
업스트림 (Upstream)	컴퓨팅 (Compute)	• 광범위한 환경 영향(Broader environmental impact)
	데이터 완화 (Data Mitigation)	• 저작권보호 완화(Mitigations for copyright)
다운스트림 (Downstream)	영향 (Impact)	• 영향받는 개인(Affected individuals) • 사용 보고(Usage reports) • 지리적 통계(Geographic statistics)

- 개인정보 보호, 저작권, 환경적 영향, 사용자 편향성 및 차별 등의 문제가 제기될 수 있는 지표로 공개 시, 법적 책임 증가와 사회적 신뢰도 저하의 우려가 있음. 따라서 기업들이 공통적으로 해당 지표를 비공개하는 전략을 취하는 경향을 보임

ㅇ 이와 같이, AI 모델의 투명성은 정보 공개만의 문제만이 아니라, 기술 보호, 경제적 이해관계, 사회적 책임이 복합적으로 엮여있다.

따라서, AI 규제 및 윤리지침 수립 시에는 아래와 같이 공공윤리·기업윤리·시장 논리가 상충하는 부분을 함께 고려하여, 공공성과 기업의 기술 보호 전략 간의 균형을 고려한 정책적 접근이 필요하다.

• 공공 윤리	• 인공지능(AI) 기술이 사회 전체에 미치는 영향을 고려 - AI 모델이 학습하는 데이터 출처, 알고리즘 설계 과정, 사회적 영향에 대한 투명성 기준 마련 필요
• 기업 윤리	• 기업이 사회적 책임을 다하면서 동시에 경쟁력 유지를 고려 - 기술적 경쟁력을 유지하면서, 사회적 책임을 수행할 수 있도록 윤리적 의사결정 체계 수립 필요
• 시장 논리	• 비즈니스 전략과 경쟁력 보호 고려 - AI 기술이 독점되지 않으면서, 기업 간 공정한 경쟁과 협력이 이루어질 수 있는 시장 환경 조성 및 기업간 투명한 정보 공유를 유도할 수 있는 가이드라인 마련 필요

즉, 단순한 규제 도입이 아니라, 기업들의 정보 공개 부담을 현실적으로 경감할 수 있는 제도적 지원을 마련하고, 지속적인 투명성 확보가 가능한 환경을 조성하는 방안*이 함께 마련될 필요가 있다.

***예시)[75]**

- **정보 공개 표준화 및 가이드라인 개발:** 기업이 정보 공개 대상 및 방식을 명확히 이해할 수 있도록 표준화된 기준을 제공하여, 일관성을 확보할 수 있도록 지원
- **기업 기밀 보호와 공공 투명성 간 균형을 위한 정책 설계:** 기술 보호와 시장 경쟁력을 고려하여 정보 공개 범위를 설정하고, 이에 대한 법적·제도적 장치 마련
- **자율규제 및 인센티브 제공**: 투명성 향상을 위해 정보 공개를 적극적으로 수행하는 기업에 대해 정부 조달 우대, 연구개발 지원, 세제 혜택 등 인센티브를 제공하여 정보 공개를 유도
- **데이터 공개 및 리스크 부담 완화를 위한 기술적 지원:** 데이터 마스킹(masking), 차등적 프라이버시(Differential Privacy) 등 민감 정보를 보호하면서 동시에 AI 투명성을 확보할 수 있는 기술적 해결책을 지원하고, AI 모델 학습 데이터의 접근성을 개선할 수 있는 안전한 데이터 공유 기술과 표준 마련

2. 기업별 규제 대응 전략 및 정책 방향

ㅇ 지금까지 파운데이션 모델 투명성 지수(FMTI) 분석을 통해, 주요 AI 기업들의 전반적인 투명성 현황을 검토하였다. 이어서, 본 절(Section)에서는 각 기업이 공개한 공식 자료(시스템 카드(System Cards), 모델 카드(Model Cards), 기술 보고서(Technical Report) 등)를 기반으로, 개별 기업들의 구체적인 AI 규제 대응 전략을 분석하고자 한다.

분석 대상은 파운데이션 모델 투명성 지수(FMTI) 평가 분석 대상 5개 기업 중, 생성형 AI 분야를 주도하는 OpenAI, 기술적·학문적 AI 연구를 선도하는 구글(DeepMind), 소셜 플랫폼 분야에서 AI 활용을 선도하고 있는 메타(Meta)를 대상[76]으로 선정하였다.

---

75) Bertomeu, J., Lin, Y., Liu, Y., & Ni, Z. (2024, November 27). Voluntary disclosure, misinformation, and AI information processing: Theory and evidence. SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5026360; Kilpala, M., & Kärkkäinen, T. (2024, April 17). Artificial intelligence and differential privacy: Review of protection estimate models. In A. Khezerlou, R. Schmidt, & K. Hu (Eds.), Advances in artificial intelligence and security: Volume 1 (pp. XX-XX). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-031-57452-8_3; Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October 24). Deep learning with differential privacy. arXiv. Retrieved from https://arxiv.org/abs/1607.00133; IBM. (n.d.). IBM InfoSphere Optim Data Privacy. Retrieved from https://www.ibm.com/products/infosphere-optim-data-privacy

76) Reuters. (2025, February 19). Google develops AI 'co-scientist' to aid researchers. Retrieved from https://www.reuters.com/technology/artificial-intelligence/google-develops-ai-co-scientist-aid-researchers-2025-02-19/; Moreno, J. (2022, December 29). OpenAI positioned itself as the AI leader in 2022. But could Google supersede it in '23? Forbes. Retrieved from

ㅇ 이들 기업은 글로벌 AI 산업을 이끌어 가는 핵심 기업들로서, 기술 발전뿐 아니라 윤리적 규제 대응 측면에서도 선도적 역할을 수행하고 있다. 이에 본 연구에서는 이들이 시행하고 있는 규제 대응 전략과 정책 방향성을 분석하고, AI 윤리 및 규제에 미치는 실질적인 시사점을 도출하고자 한다.

(1) OpenAI

1) GPT-4o 시스템 카드 분석

ㅇ 시스템 카드는 OpenAI가 AI 모델 개발 시 적용한 윤리적 고려사항과 안전 조치를 구체적으로 기술한 문서이다. OpenAI가 공개한 최근 모델인 GPT-4o 시스템 카드[77]는 모델의 기능과 한계, 안전성 평가와 관련해 다양한 측면에서 분석한 상세한 내용이 공개되어 있으며, 제3자 평가 결과 및 사회에 미칠 잠재적 영향 등을 포함하고 있다.

본 보고서에서는 최근 모델인 GPT-4o의 시스템 카드를 기반으로 OpenAI의 규제 대응 전략 및 방향을 분석하고자 한다. 이를 위해, 먼저 시스템 카드에 공개된 주요 내용을 항목별로 살펴보고, 각 항목이 대응하고 있는 규제 및 추가 대응 과제를 검토하였다.

특히, 본 연구는 문화콘텐츠 분야 AI 윤리지침 관점에서, 앞서 범용 AI 및 생성형 AI를 중심으로 규제 및 윤리지침을 검토하였는바, 이를 바탕으로 분석을 진행하였다.

❶ 모델 데이터 및 학습
- **'23년 10월까지 데이터를 사용하여 사전 훈련**
  - 업계 표준 데이터셋 및 웹 크롤링 등 공개적으로 이용 가능한 데이터
  - 데이터 파트너십을 통한 독점 데이터
- **GPT-4o 성능에 기여하는 주요 데이터셋 구성요소**
  - 웹 데이터: 공개 웹 페이지 데이터를 통해 다양한 관점과 주제 학습
  - 코드 및 수학: 추론 능력 개발 지원
  - 멀티모달 데이터: 비텍스트 입·출력 해석 및 생성법 학습

---

https://www.forbes.com/sites/johanmoreno/2022/12/29/openai-positioned-itself-as-the-ai--leader-in-2022-but-could-google-supersede-it-in-23/

77) OpenAI. (2024, August 8). GPT-4o system card. Retrieved from https://cdn.openai.com/gpt-4o-system-card.pdf

<table>
<tr><td colspan="2">• 배포 전, 사전 훈련·사후 훈련·제품 개발·정책 전반의 모든 개발 단계에 복합적 방식을 사용하여 정보 유해성·편향·차별 및 기타 사용 정책 위반 컨텐츠 등 잠재적 리스크 평가 및 완화
- 사후 훈련 단계 평가 및 완화 조치
· 모델을 인간 선호도에 맞춰 조정
· 레드팀 테스트 진행
· 모니터링 및 시행 조치 등 제품 단계의 완화 조치 추가
· 사용자에게 중재 도구 및 투명성 보고서 제공
- 사전 훈련 단계 필터링 및 안전 완화 조치
· CSAM, 혐오 콘텐츠, 폭력, CBRN 등 유해한 데이터 필터링(Moderation API, safety classifiers 사용)
· 이미지 생성 데이터셋에서 성적 자료 및 CSAM 필터링
· 개인정보 보호를 위해 고급 데이터 필터링 적용
· DALL-E 3에서 이미지 학습 제외(opt-out) 기능 도입(핑거프린팅 기술 활용)</td></tr>
<tr><td>규제 대응</td><td>• (미국 AI 권리장전) 데이터 프라이버시 원칙
- 사용자 개인정보 보호 및 최소한의 데이터 수집
• (NIST AI RMF) 안정성, 보안성 및 회복성, 책임성 및 투명성, 프라이버시 강화, 유해한 편향이 관리된 공정성, 유효성 및 신뢰성
• (NIST GAI 프로파일) 위험하거나, 폭력적이거나, 혐오스러운 콘텐츠, 데이터 프라이버시, 유해한 편향 및 균질화 리스크 유형 대응
• (행정명령 제14110호) AI 기업의 정보 제출 의무(제4조제4.2항(a)목(i))
- 모델의 훈련·개발·생산과 관련된 모든 활동(복잡한 위협에 대한 훈련 과정의 무결성 보장 보안조치 포함)
• (EU AI Act) 범용 AI 모델 공급자의 의무(제53조)
- 훈련·테스트 과정·평가 결과를 포함한 기술문서 작성·최신 유지
- 범용 AI 모델 훈련 콘텐츠에 대한 충분히 상세한 요약 공개</td></tr>
<tr><td>대응 과제</td><td>• 훈련 데이터 상세 비공개 문제
- 기술문서 작성 규제 요구 등을 반영하고 있으나, 내부 데이터셋의 세부적인 정보 등은 비공개
- 데이터 출처·구성 요건 공개 규정 준수 미흡, 투명성·검증 가능성 부족 등 문제 제기 가능</td></tr>
<tr><td colspan="2">❷ 위험 식별, 평가 및 완화
❷-1. 외부 레드 팀 활동
• 외부 레드팀(29개국 출신, 45개 언어 사용, 100명 이상 참여) 평가를 거쳐, 보안·안전성 검증, 모델의 유해 콘텐츠 생성 가능성·허위 정보 생성 여부 등 점검</td></tr>
<tr><td>규제 대응</td><td>• (미국 AI 권리장전) 안전하고 효과적인 시스템 원칙
• (NIST AI RMF) 프라이버시 강화, 안전성, 보안성 및 회복성, 유효성 및 신뢰성
• (NIST GAI 프로파일) 정보 보안, 위험하거나, 폭력적이거나, 혐오스러운 콘텐츠 리스크 유형 대응
• (행정명령 제14110호) 시스템의 안전성·신뢰성 확보를 위한 레드 팀 테스트 수행(제4조제4.1항(a)목(ii))
• (EU AI Act) AI 시스템의 위험 평가 및 완화 조치 필수 조항(고위험 시스템 관련 조항 제9조)</td></tr>
</table>

❷-2. 관찰된 안전 문제, 평가 및 완화책

❷-2-1. 무단 음성 생성(Unauthorized voice generation)

- 사기 및 음성 사칭, 허위 정보 확산에 악용될 가능성 및 모델이 의도치 않게 사용자의 목소리를 모방할 가능성
- 사전 승인 음성만 사용하도록 제한, 사후 훈련 과정에 선택된 음성 인식 학습 (음성 출력 분류기 활용 이상 음성 감지)

❷-2-2. 화자 식별(Speaker Indentification)

- 음성 입력을 통해 특정인 식별 위험 가능성(개인의 프라이버시 침해 위험)
- 특정인 목소리를 기반으로 신원 파악 요청을 거부하도록 모델을 사후 훈련. 단, 유명한 명언(quote)의 출처 식별 요청 등 예외적인 경우 응답 허용.

❷-2-3. 저작권 보호 콘텐츠 생성

- 저작권 있는 콘텐츠 생성 위험 가능성
- 저작권 콘텐츠 요청 거부 훈련, 기존 텍스트 필터를 오디오 대화에도 적용
- 음악 포함 출력을 감지하고 차단하는 필터 구축

❷-2-4. 음성 입력에서 성능 차이

- 사용자의 억양(accent)에 따라 상이하게 작동 가능(공정성 문제 초래)
- GPT-4o가 다양한 사용자 음성을 균일하게 처리하도록 사후 훈련
- 다양한 음성 데이터셋을 학습하여, 사용자 억양이나 목소리에 관계없이 모델의 성능과 동작이 일관되게 유지되도록 조정

❷-2-5. 근거 없는 추론/ 민감한 특성 귀속

- 근거없이 모델이 특정 화자의 능력, 성향, 특성 등을 추론할 가능성
- 근거 없는 추론 요청 거부 훈련, 억양·인종 등 민감 특성 질문에는 신중한 표현을 사용하도록 훈련

❷-2-6. 오디오 출력에서 금지된 콘텐츠 생성

- 폭력 및 혐오 발언 등 금지된 콘텐츠 출력 차단
- 기존의 콘텐츠 필터링 시스템 활용, 오디오 입력을 텍스트로 변환 후 분석하고 차단

❷-2-6. 에로틱·폭력적 발언 출력

- 음성 입력에 에로틱하거나 폭력적 내용 포함 시, 부적절한 응답 생성 가능성
- 입력 프롬프트에 에로틱하거나 폭력적인 언어가 포함되었을 경우 출력 차단
- 기존의 콘텐츠 필터링 시스템 활용, 오디오 입력을 텍스트로 변환 후 분석하고 차단

규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템, 데이터 프라이버시, 알고리즘 차별 보호, 고지 및 설명 원칙 • **(NIST AI RMF)** 유해한 편향이 관리된 공정성, 안전성, 유효성 및 신뢰성, 설명 가능성 및 해석 가능성, 보안성 및 회복성, 책임성 및 투명성, 프라이버시 강화 • **(NIST GAI 프로파일)** 정보 보안, 데이터 프라이버시, 지식재산권, 유해한 편향 및 균질화, 혼동(Confabulation), 위험하거나, 폭력적이거나, 혐오스러운 콘텐츠 리스크 유형 대응 • **(행정명령 제14110호)** 데이터 보호 및 프라이버시 강화 지침 수립 의무(제4조제4.1항(a)목(ii)), AI 생성 콘텐츠의 식별과 워터마킹 적용, 저작권 보호 콘텐츠의 무단 생성 방지 조치, 허위 정보 생성 및 민감 개인 정보 연결 방지, 폭력적이거나 혐오적 콘텐츠 생성 방지를 위한 기술적 조치 마련 등 (제4조제4.5항)

	• **(EU AI Act)** 인간의 행동을 조작하거나, 심리적·신체적 위해를 초래할 수 있는 AI 시스템의 사용을 금지(제5조), 범용 AI 모델 제공자의 저작권 및 관련 권리를 준수 의무(제53조)
❸ 제3자 평가 • 모델의 성능과 안전성을 검증하기 위해 독립적인 외부 연구기관인 METR 및 Apollo Research와 협력하여 제3자 평가 진행, AI 모델의 자율적 행동 가능성, 위협 평가, 모델 오용 가능성을 객관적으로 점검	
규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템, 알고리즘 차별 보호 원칙 • **(NIST AI RMF)** 유효성 및 신뢰성, 안전성, 보안성 및 회복성, 책임성 및 투명성, 설명 가능성 및 해석 가능성, 프라이버시 강화, 유해한 편향이 관리된 공정성 • **(EU AI Act)** 고위험 AI 시스템 독립적 평가 및 투명성 요구(제43조)
❹ 사회적 영향 **❹-1. 의인화 및 감정적 의존** • 초기 테스트 및 내부 테스트 결과, AI 모델과 정서적 유대감을 형성하는 경향이 일부 나타남 • AI 모델과 상호작용이 사용자의 사회적 관계에 영향을 미칠 가능성(고립된 개인에게 긍정적 영향이 될 수도 있지만, 인간과의 사회적 관계를 악화시킬 위험도 있음)	
규제 대응	• **(미국 AI 권리장전)** 알고리즘 차별 보호 원칙 • **(NIST AI RMF)** 책임성 및 투명성, 설명 가능성 및 해석 가능성 • **(NIST GAI 프로파일)** 인간-AI 상호작용(Configuration) 리스크 유형 대응 • **(EU AI Act)** 인간 행동 조작 및 심각한 심리적 영향을 미치는 행위 금지(제5조), AI모델과 상호작용 시, 사람이 아님을 인식하도록 설명할 의무(제52조)
❹-2. 건강 • 의료정보 제공 가능성 및 음성 기반 상호작용을 통해 환자와 의료진 커뮤니케이션 지원 가능성 • 모델의 의료 지식 평가 결과, 임상 문서 처리, 임상 시험 모집, 진료 지원 등의 분야에서 이전 모델보다 향상 • 신뢰성에는 위험이 있음. 모델이 부정확한 의료 정보 제공 시, 사용자 건강에 부정적 영향을 미칠 가능성이 있음	
규제 대응	• **(미국 AI 권리장전)** 데이터 프라이버시 원칙 • **(NIST AI RMF)** 유효성 및 신뢰성, 안전성, 보안성 및 회복성 • **(NIST GAI 프로파일)** 혼동(Confabulation), 데이터 프라이버시, 유해한 편향 및 균질화, 정보 보안 • **(EU AI Act)** 의료 진단 및 치료 지원 목적 AI 시스템 고위험 AI로 분류
❹-3. 과학적 성능 • GPT-4o는 과학적 연구 및 분석 지원에 탁월한 성능, 복잡한 과학적 개념 탐구에 있어 연구 지원, 새로운 가설 도출 및 복잡한 시뮬레이션 분석 가능성 • 양자물리학 및 생명과학 같은 전문 분야에 효과적으로 작동 사례 보고	
규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템 원칙 • **(NIST AI RMF)** 유효성 및 신뢰성, 설명 가능성 및 해석 가능성 • **(NIST GAI 프로파일)** 혼동(Confabulation), 유해한 편향 및 균질화, 정보 무결성, 지식재산권 리스크 유형 대응 • **(EU AI Act)** 혁신을 위한 AI 규제 샌드박스(제57조 내지 제63조)

❹-4. 소외된 언어 • 아프리카 언어를 포함 다수의 소외 언어에서 이해 및 추론 성능 개선 • 영어와 성능 격차 존재, 특정 언어에서 부정확하거나 제한적 답변 제공 가능성	
규제 대응	• **(미국 AI 권리장전)** 알고리즘 차별 보호 원칙 • **(NIST AI RMF)** 유해한 편향이 관리된 공정성, 설명 가능성 및 해석 가능성 • **(NIST GAI 프로파일)** 유해한 편향 및 균질화 리스크 유형 대응

2) OpenAI AI 규제 대응 현황

ㅇ 위에서 살펴본 바와 같이 OpenAI는 GPT-4o 시스템 카드를 통해 AI 모델의 안전성·투명성·공정성·프라이버시 보호·사회적 영향을 고려한 규제 대응 전략을 마련하고 있다.

특히, EU 인공지능법을 포함하여, 미국의 AI 권리장전, NIST AI RMF 및 GAI 프로파일과 행정명령 제14110호의 주요 원칙을 반영하는 방향으로 대응하고 있다.

- **(데이터 관리 및 투명성 강화)**
  - 미국 행정명령 제14110호 및 EU 인공지능법 규정에 따라 훈련 데이터(공개 웹 데이터 및 독점 데이터) 및 주요 데이터셋 구성 요소를 명시
  - NIST AI RMF 및 GAI 프로파일의 데이터 프라이버시 요건 준수를 위해 사용자 데이터 처리 방식을 설명
- **(위험 평가 및 완화 조치)**
  - 외부 레드팀 평가 및 내부 안전성 테스트를 통해 모델의 위험 요소를 사전 평가하고, 모델 배포 후 지속적인 모니터링 수행
  - 미국 행정명령 제14110호의 레드팀 테스트 수행 요건 및 EU 인공지능법의 고위험 AI 시스템 규정에 따라, 외부 레드팀 평가 시행
- **(제3자 평가 및 신뢰성 확보)**
  - 외부 연구기관(METR 및 Apollo Research)과 협력하여, 독립적인 제3자 평가를 수행함으로써, AI 모델의 신뢰성과 안정성을 객관적으로 검증
  - EU 인공지능법의 고위험 AI 시스템 독립적 평가 및 NIST AI RMF 신뢰성·투명성 원칙을 반영
- **(사회적 영향 및 AI 윤리 고려)**
  - AI 모델의 의인화, 감정적 의존, 건강 및 과학적 활용, 언어 다양성 등 사회적 영향 분석

- EU 인공지능법 제5조의 인간 행동 조작·심리적 위해 행위 금지 규정에 따라, AI 모델에 감정적으로 의존하지 않도록 설계
- NIST AI RMF 및 GAI 프로파일의 공정성 및 프라이버시 보호 원칙에 따라, AI가 특정 인구 집단에 차별적으로 작용하지 않도록 조치

3) OpenAI AI 규제 대응 전략 및 정책 방향

ㅇ GPT-4o 시스템 카드 분석 결과, OpenAI는 AI 규제 환경 변화에 대응하기 위해 자율적인 AI 거버넌스를 강화하면서 동시에 주요 규제 요구사항을 반영하는 방향으로 정책을 추진하고 있음을 확인할 수 있다.

- **(자율적 AI 거버넌스 강화 및 규제요구 선제적 반영)**
  - 규제 당국의 요구를 수동적으로 따르는 것이 아니라, 시스템 카드에 공개한 바와 같이 독립적 평가 및 리스크 완화 조치 등의 방식으로 자율적 AI 거버넌스를 구축하는 전략을 취하고 있음.
  - 이는 AI 정책 형성 과정에도 적극적으로 개입하는 방식으로 대응하고 있음을 보여줌
- **(데이터 투명성과 AI 모델 신뢰성 확보를 위한 조치 강화)**
  - 공개 웹 데이터 및 독점 데이터 등 훈련 데이터와 멀티모달 데이터 등 주요 데이터셋 구성 요소를 공개하면서 출처 등 세부 사항은 비공개
  - 공개 의무를 준수하면서, 기술적 경쟁력을 보호하기 위해 훈련 데이터의 세부 공개는 신중한 태도를 유지하는 전략을 채택
- **(AI 안전성 및 리스크 평가 체계 구축)**
  - AI 모델 신뢰성 향상을 위한 외부 레드팀 평가, 내부 안전성 테스트, 제3자 연구기관(METR, Apollo Research)과의 지속적 협력 추진,
  - EU 인공지능법의 고위험 AI 시스템 평가 및 NIST AI RMF 신뢰성 원칙에 부합
- **(AI 윤리적 설계 및 사회적 영향 고려 확대)**
  - AI 모델이 인간과 상호 작용에서 의인화되지 않도록 설계하고, AI 생성 콘텐츠 출처를 명확히 할 수 있는 기술적 조치를 연구하는 등 사회적 영향을 고려하는 방식으로 대응
  - EU 인공지능법 제5조의 인간 행동 조작 및 심리적 위해 금지행위 및 GAI 프로파일의 공정성 원칙을 반영

- **(벤치마크 및 평가 방법의 국제 표준화 선도)**
  - 레드팀 테스트·제3자 검증·사회적 영향 분석 등 AI 모델 성능 및 위험 평가 방법을 선제적으로 채택하여, 향후 AI 평가 및 안전성 검증의 국제 기준을 선도
  - 현재 AI 평가와 관련한 국제 표준이 부재한 상황에서, OpenAI가 선례로 보여주는 신뢰성 평가방식은 글로벌 AI 정책에 중요 참고 사례가 될 수 있음
  - 규제 당국은 기 도입되고 있는 평가방식과 벤치마크를 기반으로 AI 규제 정책을 구체화할 가능성이 높고, 이는 OpenAI가 AI 거버넌스 및 정책 표준화를 주도할 가능성을 시사

ㅇ 이와 같은 OpenAI의 규제 대응 정책 방향은 단순한 규제 준수를 넘어, 글로벌 AI 거버넌스 형성과 규제 환경에서 기업이 주도적인 역할을 수행하는 방향으로 발전하고 있음을 보여준다.

AI 규제 환경이 점차 구체화됨에 따라, OpenAI가 개발한 평가 기준과 리스크 완화 조치가 향후 AI 규제의 국제적 표준으로 자리 잡을 가능성이 있고, 이는 AI 기업과 규제 당국 간 협력 방식에도 중요한 변화를 가져올 것으로 예상된다.

(2) 구글(Google DeepMind)

1) Gemini 1.0 Ultra 모델 카드 분석

ㅇ Gemini 1.0 Ultra 모델은 생성형 AI 기술이 적용된 멀티모달 모델로, 구글(Google DeepMind)은 파운데이션 모델 투명성 지수(FMTI) 보고서에서 해당 모델을 자사의 대표 파운데이션 모델로 지정하였다.

ㅇ Gemini 1.0 Ultra 모델 카드[78]는 아키텍처, 입·출력, 활용 사례, 데이터 구성, 평가 방법 및 제한 사항 등을 기술한 문서로, 이를 통해 모델의 주요 특징과 활용 범위, 안전성 및 윤리적 고려사항 등을 파악할 수 있다.

78) Google DeepMind. (2024). *Gemini: A family of highly capable multimodal models.* Retrieved from https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

ㅇ 본 연구에서는 Gemini 1.0 Ultra 모델카드를 분석하여 구글(Google DeepMind)의 AI 규제 대응 전략 및 정책 방향을 도출하였다. 특히, 앞서 범용 AI와 생성형 AI를 중심으로 미국과 유럽연합의 규제 및 윤리지침을 검토하였으므로, 해당 규제를 중심으로 분석을 진행하였다.

<table>
<tr><td colspan="2">❶ 모델 개요(Model summary)
❶-1. 모델 아키텍처
• 디코더 전용(Decoder-only) 트랜스포머 구조 채택, 멀티 쿼리 어텐션(Multi-Query Attention) 적용
• 파라미터 수: Gemini Ultra > Pro > Nano, 구체적 수치 비공개
• 컨텍스트 길이: 최대 32K 토큰 지원
❶-2. 입·출력
• 입력(Input): 텍스트, 이미지, 동영상, 오디오 파일
• 출력(Output): 생성된 텍스트(질문에 대한 답변·문서 요약·비교 분석 등)</td></tr>
<tr><td>규제 대응</td><td>• (미국 AI 권리장전) 안전하고 효과적인 시스템 원칙
• (NIST AI RMF) 안정성, 보안성 및 회복성, 유효성 및 신뢰성
• (NIST GAI 프로파일) 정보 무결성, 유해한 편향 및 균질화, 데이터 프라이버시 리스크 유형 대응
• (행정명령 제14110호) AI 모델 개발·배포 투명성 의무(제4조제4.2항(a)목)
• (EU AI Act) 범용 AI 모델 공급자의 기술문서 제공 및 투명성 의무(제53조)</td></tr>
<tr><td>대응 과제</td><td>• 모델 투명성 미흡
- 모델의 파라미터 수치, 학습 데이터 크기 및 연산량 정보 비공개
- 시스템의 입·출력 처리 방식에 대한 상세 설명 부족</td></tr>
<tr><td colspan="2">❷ 사용(Usage)
❷-1. 애플리케이션
• 언어 모델 연구 가속화, Google 제품 및 특정 애플리케이션(Gemini App, Search Generative Experience)에서 활용되도록 설계, 외부 개발자에게 Google Cloud Vertex API, Google Labs를 통해 제공하며, 안전 정책 준수를 위한 추가 절차 및 기술적 보호조치 적용(상세 비공개)
❷-2. 주의사항(Known Caveats)
• 범용 AI 서비스로 직접 제공 불가
• 특정 애플리케이션에서 안전 및 공정성 평가(상세 비공개) 없이 사용 금지</td></tr>
<tr><td>규제 대응</td><td>• (미국 AI 권리장전) 안전하고 효과적인 시스템 원칙
• (NIST AI RMF) 안정성, 보안성 및 회복성, 유효성 및 신뢰성
• (NIST GAI 프로파일) 정보 무결성, 유해한 편향 및 균질화, 데이터 프라이버시, 정보 보안 리스크 유형 대응
• (행정명령 제14110호) AI 모델 개발·배포 투명성 의무(제4조제4.2항(a)목)
• (EU AI Act) 범용 AI 모델 공급자의 기술문서 제공 및 투명성 의무(제53조)</td></tr>
<tr><td>대응 과제</td><td>• 위험 완화 조치 및 안전성 보장 메커니즘 투명성 미흡
- 안전성 및 공정성 평가 방법에 대한 세부 사항 비공개
- 안전 정책 준수를 위한 추가 절차 및 기술적 보호조치 상세 비공개</td></tr>
</table>

❸ 모델 구현 프레임워크(Implementation Frameworks)

❸-1. 하드웨어 & 소프트웨어

- 하드웨어: TPUv4 및 TPUv5e에서 훈련 수행
- 소프트웨어: JAX 및 ML Pathways 사용
  - JAX: 최신 하드웨어(TPU 포함)를 활용한 고속 및 효율적 대규모 모델 훈련 지원
  - ML Pathways: 다양한 태스크에 일반화할 수 있는 인공지능 시스템 구축을 위한 인프라 소프트웨어, 특히, Gemini V1.0 모델과 같은 대규모 언어모델 등 파운데이션 모델에 적합
  - JAX, ML Pathways는 단일 컨트롤러(single controller) 프로그래밍 모델을 적용하여 하나의 Python 프로세스로 전체 훈련을 조정하여 개발 워크플로우를 최적화함

❸-2. 연산 요구사항: 비공개

규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템 원칙 • **(NIST AI RMF)** 안정성, 보안성 및 회복성, 유효성 및 신뢰성 • **(NIST GAI 프로파일)** 정보 무결성, 유해한 편향 및 균질화, 데이터 프라이버시, 정보 보안 리스크 유형 대응 • **(행정명령 제14110호)** AI 모델 개발·배포 투명성 의무(제4조제4.2항(a)목) • **(EU AI Act)** 범용 AI 모델 공급자의 의무(제53조) - 범용 AI 모델 제공자의 기술문서 제공 및 투명성 의무
대응 과제	• 사용된 하드웨어 및 소프트웨어 구성에 대한 구체적인 성능 및 안전성 평가 자료 비공개 • 연산 요구사항(예: TPU/GPU 종류 및 개수, 연산 성능(FLOPs). 메모리 요구사항, 전력 소모량 등) 비공개→에너지 소비 및 환경적 영향 평가 등에 대한 투명성 미흡

❹ 모델 특성(Model Characteristics)

❹-1. 모델 초기화(Model initialization)

- 초기 사전 훈련 시, 랜덤으로 초기화되며, 이후 사전 훈련 후반 단계에서 저장된 체크포인트를 기반으로 모델 재초기화, 체크포인트를 감독을 통한 파인 튜닝으로 조정한 후, 보상 모델 훈련과 인간 피드백을 통한 강화학습(RLHF) 재초기화를 진행하면서 모델을 점진적으로 개선

❹-2. 모델 상태(Model Status)

- 오프라인 데이터셋으로 훈련된 정적 모델, 실시간 업데이트 없음

❹-3. 모델 통계(Model Stats)

- 구체적인 파라미터 수 및 연산량 정보 비공개

규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템 원칙 • **(NIST AI RMF)** 책임성 및 투명성, 유효성 및 신뢰성, 안정성 • **(NIST GAI 프로파일)** 정보 무결성, 유해한 편향 및 균질화, 데이터 프라이버시 리스크 유형 대응 • **(행정명령 제14110호)** AI 모델 개발·배포 투명성 의무(제4조제4.2항(a)목) • **(EU AI Act)** 범용 AI 모델 공급자의 기술문서 제공 및 투명성 의무(제53조)
대응 과제	• **모델 특성에 대한 투명성 미흡** - 통계(연산량 정보 등) 비공개, AI 모델의 학습 과정 및 검증 과정에 대한 상세 정보 비공개

❺ 데이터 개요(Data Overview)

❺-1. 훈련 데이터셋(Training Dataset)

- 웹 문서·코드·이미지·오디오·비디오 데이터를 포함한 멀티 모달 데이터 및 다국어(multilingual) 데이터셋으로 훈련

❺-2. 평가 데이터셋(Evaluation Dataset)

- 사전 훈련 및 사후 훈련 버전을 기존 외부 대형 언어 모델 및 Google PaLM과 비교하여 평가, 문자 기반 학술 벤치마크(추론, 독해력, 과학·기술·공학·수학(STEM), 코딩 등)를 통해 성능 분석
- 다중 모달 성능 평가를 위해 4가지 주요 능력 테스트

(1) 캡셔닝 및 질문-응답 태스크를 활용한 고수준 객체 인식(VQAv2 등)
(2) 세밀한 문자 인식 및 스크립트(TextVQA, DocVQA 등 태스크 이용)
(3) 차트 이해 및 공간적 레이아웃 분석(ChartQA, InfographicVQA 태스크 이용)
(4) 다중 모달 추론 능력(AI2D, MathVista, MMMU 등 태스크 활용)

❺-3. 사후-훈련 데이터셋(Post-training Dataset)

- 실사례를 반영할 수 있도록 다양한 프롬프트 수집, 이후, 감독을 통한 파인튜닝을 위해 특정 프롬프트에 대해 모델이 생성해야 할 이상적인 출력 등 시연 데이터를 수집, 또한, 동일한 프롬프트에 대한 다양한 가능한 응답을 수집하고 이에 대한 피드백 데이터를 활용하여 보상 모델을 훈련

규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템 원칙 • **(NIST AI RMF)** 책임성 및 투명성, 유효성 및 신뢰성, 안정성 • **(NIST GAI 프로파일)** 정보 무결성, 유해한 편향 및 균질화, 데이터 프라이버시 리스크 유형 대응 • **(행정명령 제14110호)** AI 기업의 정보제출 의무 및 데이터 관련 투명성 의무(제4조제4.2항(a)목) • **(EU AI Act)** 범용 AI 모델 공급자의 기술문서 제공 및 투명성 의무(제53조)
대응 과제	• 데이터 출처 및 처리 방식에 대한 세부사항 비공개

❻ 평가 결과(Evaluation Results)

- **벤치마크 정보 및 평가 결과**
  - **(학술적 벤치마크 성능평가)**
    - MMLU(Massive Multitask Language Understanding, 지식 및 추론능력 평가) 벤치마크에서 90%의 정확도를 기록, 최초로 인간 전문가 수준(81.8%) 초과
    - HumanEval(코딩) 벤치마크에서 78.9% 정확도를 기록하며, Python 코드 생성 성능에서 최고 수준 달성
  - **(다중 모달 성능 평가)**
    - VQAv2(77.8% 정확도 기록, 기존 모델 대비 높은 성능 달성), TextVQA(81.8% 정확도 기록, Google PaLI-17 모델 대비 향상), DocVQA(90.1% 정확도 기록, GPT-4V와 동등한 수준)를 활용한 고수준 객체 인식 및 문자 인식 성능 평가 수행
    - ChartQA(80.8% 정확도, Google DePlot(78.5%)보다 향상) 및 InfographicVQA(80.3% 정확도, GPT-4V(75.1%) 대비 개선)를 통한 차트 해석 및 공간적 레이아웃 이해 능력 측정
    - AI2D(71.5% 정확도 기록, 기존 모델 대비 높은 성능), MathVista(63.0% 정확도 기록, 수학적 추론 분야에서 개선) 등의 태스크를 활용한 다중 모달 추론 능력 평가 수행

・ MMMU(다학제적 대학 수준 문제) 벤치마크에서 61.8% 정확도를 기록하며, GPT-4V(56.8%)를 초과하는 성능 달성	
규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템 원칙 • **(NIST AI RMF)** 책임성 및 투명성, 유효성 및 신뢰성 • **(NIST GAI 프로파일)** 정보 무결성, 유해한 편향 및 균질화 리스크 유형 대응 • **(행정명령 제14110호)** AI 기업의 정보제출 의무 및 데이터 관련 투명성 의무(제4조제4.2항(a)목) • **(EU AI Act)** AI 모델의 투명한 성능평가(제43조)
대응 과제	• AI 모델의 실사례에서의 성능과 벤치마크 성능 간의 차이 및 실제 환경에서의 성능 검증 평가 자료 비공개

❼ 모델 사용 및 한계(Model Usage & Limitation)

❼-1. 민감한 사용(Sensitive Use)

- Google은 Gemini 모델을 개발하고 배포하는 과정에서 구조화된 책임감 있는 접근 방식을 채택하여, 사회적 영향을 평가하고 지속적으로 관리
- 모델 영향 평가(Model Impact Assessment): 다양한 문헌, 외부 전문가, 내부 윤리 및 안전 연구를 반영하여 진행
  - (모델 수준 평가) Gemini 모델의 다양한 기능(예: 텍스트-텍스트, 이미지-텍스트, 비디오-텍스트)에 걸쳐 식별된 다운스트림 이점과 위험을 평가
  - (리스크 평가) 생성형 멀티모달 모델이 사회적으로 미칠 위험을 고려하여, 사용자가 유해한 콘텐츠(폭력적, 혐오적, 성적으로 노골적인 표현 등)에 노출될 가능성, 아동 안전 문제, 감시 기술에 악용될 가능성 등을 분석하여 평가
  - (제품 수준 평가) Gemini 모델을 기반으로 구축된 특정 제품(예: Gemini Advanced)과 관련된 추가적인 위험 및 영향평가
- Gemini 모델의 다중 모달 특성을 활용하여 텍스트, 이미지, 비디오를 포함한 정보 처리를 효율화할 수 있으며, 정보 요약, 영상 캡션 생성, 분석 분야 등 효율성 향상, AI 모델의 환경적, 경제적 영향을 고려하며, 사이버 보안 위협과 같은 위험 요소를 지속적으로 연구하고 평가

❼-2. 한계(Known Limitations)

- Gemini 모델은 특정 도메인에서 제한된 성능을 보일 가능성이 있으며, 안전성 및 공정성 평가 필요
- 일부 학습 데이터의 한계로 인해 특정 언어 및 문화적 맥락에서 성능 저하 가능, 고도로 전문화된 분야(예: 법률, 의학, 과학 연구)에서 정확도가 떨어질 수 있으며, 추가적인 검증 절차 필요
- 모델이 생성하는 응답이 과장되거나 부정확한 정보를 포함할 가능성이 있어, 신뢰성과 검증 가능한 정보 제공 방안 필요
- 하위 애플리케이션에서 사용될 경우, 잠재적 위험 및 부작용을 고려해야 함
- AI 모델 악용 가능성 방지를 위해 추가적인 안전 조치·필터링 시스템 필요

❼-3. 윤리적 고려 사항 및 위험(Ethical Considerations & Risks)

- 개발 단계에서 모델의 책임 기준을 향상시키기 위한 내부 평가 및 외부 학술 벤치마크 테스트 수행
- 모델 개발 팀 외부에서 모델의 안전성을 검토하며, 안전 정책 위반 여부를 지속적으로 점검하는 위험성 보증 평가(Assurance Evaluations) 수행
- 독립적인 외부 전문가 그룹이 모델을 평가하며, 잠재적 위험을 식별하기 위해 스트레스 테스트 등 외부 평가(External Evaluations) 진행

• Gemini 모델에 대한 적대적 테스트를 수행하여 보안 취약점을 발견하고 보완하는 레드 팀 테스트(Red Teaming) 수행	
규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템, 알고리즘 차별 보호 원칙 • **(NIST AI RMF)** 책임성 및 투명성, 유효성 및 신뢰성, 안정성 • **(NIST GAI 프로파일)** 정보 무결성, 유해한 편향 및 균질화, 데이터 프라이버시 리스크 유형 대응 • **(행정명령 제14110호)** AI 기업의 정보제출 의무 및 데이터 관련 투명성 의무(제4조제4.2항(a)목) • **(EU AI Act)** 범용 AI 모델 공급자의 기술문서 제공 및 투명성 의무(제53조)
대응 과제	• AI 모델의 하위 애플리케이션에서의 안전성 및 공정성 평가 절차 부족 • Gemini 모델이 다양한 환경에서 실제 적용 시, 발생할 수 있는 윤리 문제 및 책임성 명확화 필요

2) 구글(Google DeepMind) AI 규제 대응 현황

ㅇ 구글(Google DeepMind)은 Gemini 1.0 Ultra 모델 카드를 통해 AI 모델의 성능과 신뢰성을 보장하는 동시에, 모델의 윤리적·사회적 영향을 최소화하기 위한 규제 대응 전략을 보여주고 있다. 모델 카드에는 AI 안전성 확보를 위한 위험 완화 조치, 데이터 사용 정책, 모델의 활용 제한 및 공정성 유지 방안 등이 포함되어 있다.

특히, EU 인공지능법과 미국 AI 권리장전, NIST AI RMF 및 GAI 프로파일, 행정명령 제14110호 등 주요 윤리지침과 규제 요건을 준수하기 위해 일정 수준의 정보를 공개하는 방향으로 대응하면서, 핵심적인 모델 학습 데이터, 안전성 평가 등의 세부 절차 및 상세 내역은 비공개로 유지하는 전략을 취하고 있다. 이는 규제를 준수하는 동시에, 기업 정보 보호·경쟁력 유지를 위한 기술적 비밀 유지·데이터 보안·법적 리스크 관리 등을 고려한 조치로 보인다.

- **(데이터 관리 및 투명성 강화)**
  - 학습 데이터가 웹 문서, 코드, 이미지, 오디오, 비디오 등으로 구성되어 있음을 공개하면서, 세부적인 데이터 출처, 상세 내역 및 처리 방식은 비공개로 유지
  - 미국 행정명령 제14110호 및 EU 인공지능법의 데이터 프라이버시 및 투명성 요건 준수를 위해 일정 수준의 정보 공개, 훈련 데이터의 구성과 관련하여 정보 무결성 및 프라이버시 보호 조치 적용

- **(위험 평가 및 완화 조치)**
  - 다양한 벤치마크 테스트를 활용하여 성능과 신뢰성을 평가하고, 다중 모달 환경에서 강력한 이해 및 추론 성능을 제공하도록 설계하여, 미국 NIST AI RMF의 안전성 및 신뢰성 원칙과 EU 인공지능법의 고위험 AI 시스템 평가 요건 반영
  - AI 모델이 특정 위험 요소를 방지하도록 설계하고, 다중 모달 평가(VQAv2, TextVQA, ChartQA 등)를 통해 안전성을 검증하여, 미국 AI 권리장전의 안전하고 효과적인 시스템 원칙과 EU 인공지능법의 투명성 및 신뢰성 요건 대응
  - 모델의 사회적 영향을 최소화하기 위해 사용 제한 및 감시 시스템 적용, 특정 용도의 직접적 활용 제한, AI 모델이 공정성을 유지하고 민감한 사용 사례에서 안전하게 활용될 수 있도록 설계하여, 미국 AI 정책의 프라이버시 보호 원칙과 EU 인공지능법의 인간 중심 설계 원칙 반영
- **(윤리적 사항 고려 및 책임감 있는 배포)**
  - AI 모델이 사회에 예상치 못한 부정적 영향을 미치지 않도록 모델 및 제품 평가 체계를 구축, 모델 개발 수명 주기 전반에 걸쳐 책임감 있는 배포 프레임워크를 적용, 미국 AI 권리장전의 인간 대안 및 검토·대책 원칙과 EU의 인간 중심 설계 원칙 반영을 준수하는 방식으로 운영
  - AI 모델의 안전성 정책을 수립하여 허위 정보 및 유해 콘텐츠 생성 방지, 모델 훈련 단계에서 데이터 필터링 및 안전 조치를 적용하고, 사후 학습 과정에서 안전 평가를 수행하여 AI 모델의 윤리적 문제를 최소화, 미국 행정명령 제14110호 및 EU 인공지능법의 투명성 원칙 반영

3) 구글(Google DeepMind) AI 규제 대응 전략 및 정책 방향

ㅇ Gemini 1.0 Ultra 모델 카드 분석 결과, 구글(Google DeepMind)은 AI 규제 환경에서 윤리 위원회 운영·AI 안전성 평가·내부 정책 조정 등 자율적인 거버넌스를 강조하면서, 당국의 규제 요건을 일정 수준에서 충족하는 전략을 취하고 있다.

특히, 미국과 EU의 AI 규제에 정보 공개 정도에 대한 세부 기준이 아직 명확히 확정되지 않았다는 점을 고려하여, 일정 부분 정보를 공개하는

방식으로 대응하되 핵심 기술과 데이터는 보호하는 방식을 유지하고 있다. 이는 대외적으로 규제 준수 입장을 표명하는 동시에 대내적으로는 기업의 경쟁력을 유지하고자 하는 접근법으로 해석될 수 있다.

- **(데이터 투명성과 모델 신뢰성 확보)**
  - 훈련 데이터 출처 및 구성 요소를 일정 수준 공개하면서, 모델의 신뢰성 확보를 위해 학술적 벤치마크를 활용한 성능 검증을 지속
  - 기술적 경쟁력 유지를 위해 훈련 데이터의 상세 공개는 제한하는 전략을 채택
- **(AI 모델의 안전성 및 위험 평가 체계 구축)**
  - AI 시스템 안전성 확보 및 사용자 신뢰 구축을 위해 모델 개발 수명 주기 전반에 걸쳐 외부 검증 진행
    - ❶**구조화된 평가**: 사전 정의된 방법론으로 모델의 유해성, 편향성, 견고성 등을 평가
    - ❷**질적 조사**: 모델 행동 방식과 잠재적 위험성을 심층 분석·평가
    - ❸**비구조화된 레드팀 테스트**: 모델의 취약점을 찾기 위해 다양한 공격 시나리오를 시뮬레이션하여 테스트
  - 모델의 잠재적 위험(예: 허위 정보 생성, 혐오 발언, 음성 사칭 등)을 사전에 식별하고, AI 모델의 안전성 및 윤리성에 대한 규제 기관의 요구사항을 충족하는 방향으로 리스크 완화 조치 시행
- **(윤리적 설계 및 사회적 영향 고려 확대)**
  - AI 모델과 인간이 상호작용할 때, 발생할 수 있는 윤리적 문제를 최소화하기 위한 조치 적용
    - ❶**유해 콘텐츠 생성 방지**: 학습 데이터 필터링(폭력·혐오·차별·선정성 등 유해 콘텐츠 포함 데이터 사전 필터링), 안전 필터링(모델 출력 시 유해 콘텐츠 탐지 및 차단하는 안전 필터 적용), 사용자 피드백(유해 콘텐츠 생성 시 사용자 신고 기능 제공 및 신고 내용 신속 처리)
    - ❷**사생활 침해 방지**: 개인정보 최소화(모델 학습 및 운영에 필요한 최소한의 개인정보만 수집, 익명화 또는 가명화하여 사용), 데이터 보안 강화(개인정보를 안전하게 저장하고 관리하기 위한 기술적, 관리적 조치 적용), 사용자 동의(개인정보 수집 및 이용 시 사용자에게 명확하게 고지하고 동의 확인)

❸**사회적 편견 및 고정관념 강화 방지**: 다양한 데이터셋 구축(다양한 인구 집단의 데이터를 포함하는 학습 데이터셋 구축, 데이터 증강 기법 등을 활용하여 데이터셋의 다양성 확보), 편향성 완화 기술 적용(모델 학습 시 편향성을 완화하기 위한 적대적 학습 등 여러 가지 기술적 기법 적용), 다양한 평가 지표 활용(모델 평가 시 인구집단 영향을 평가할 수 있는 지표 활용)

- 다양한 인구 집단에 대한 공정성을 보장하기 위한 노력 강화

❶**학습 데이터의 다양성 확보 및 편향성 완화**: 다양한 데이터셋 구축(각기 다른 연령·성별·인종·문화적 배경을 가진 사람들의 데이터를 수집하고, 데이터 증강 기법 등을 활용하여 데이터셋의 다양성 확보), 데이터 편향성 분석 및 개선(학습 데이터에 존재하는 편향성을 분석하고, 데이터 증강, 편향 제거 알고리즘 등을 사용하여 편향성 완화)

❷**모델 평가 시 인구집단에 대한 영향평가 및 개선**: 다양한 평가 지표 활용(모델 성능 평가 시, 정확도·재현율 등 외에도 형평성·포용성·대표성 등 다수의 지표 활용), 인구집단 평가(모델이 다양한 인구집단에 대해 차별적인 응답을 생성하는지 평가하고, 문제 발견 시 모델 수정), 외부 전문가 검토(모델 평가 과정에 외부 전문가를 참여시켜 다양한 관점에서 모델의 공정성을 검토)

- 모델 개발·배포 시 사회적 영향에 대한 지속적인 모니터링 및 평가

- **(국제 표준화 대응 및 AI 규제 환경 선도)**
  - AI 모델의 신뢰성과 위험 평가를 위한 글로벌 벤치마크 및 평가 방법을 선제적으로 적용하여 AI 규제 표준을 주도하는 전략 추진
  - AI 안전성 및 투명성 요건을 반영한 기술문서를 선제적으로 공개하고, 관련 정책을 수립하여 국제 AI 규제 환경을 주도

ㅇ 이와 같은 구글(Google DeepMind)의 AI 규제 대응 정책의 방향은 자율적인 거버넌스를 강조하는 동시에 주요 규제 요건을 선제적으로 반영하여, 규제 준수의 기준을 제시하고자 하는 것으로 보인다.

특히, 데이터 관리·위험 평가·신뢰성 검증·윤리적 설계 등 AI 안전성 확보에 대한 규제 준수 선례를 제시하여 이를 국제 규제 표준 논의에 반영하고자 하는 전략을 취하고 있다.

따라서, 이와 같은 구글의 AI 모델 투명성 및 리스크 완화 조치가 향후 글로벌 AI 규제 국제 표준 마련 과정에 핵심적인 참고 사례 중 하나로 활용될 가능성이 있다.

(3) 메타(Meta)

1) Liama 2 모델 카드 분석

ㅇ Llama 2는 '23년 공개된 연구 및 상업적 사용을 지원하는 대규모 언어 모델(LLM)로, 메타(Meta)는 파운데이션 모델 투명성 지수(FMTI) 보고서에서 Llama 2를 자사의 대표 파운데이션 모델로 지정하였다.

ㅇ Llama 2 모델카드[79]는 해당 모델의 구조·파라미터·학습률 등 세부 사항, 환경적 영향, 훈련 데이터, 평가 결과, 윤리적 고려사항 등이 기술된 문서로, 이를 통해 모델의 성능과 안전성, 책임성과 관련한 기업 정책과 미국 및 유럽연합의 AI 윤리지침·규제에 대한 대응 전략을 파악해 볼 수 있다.

ㅇ 본 Llama 2 모델카드 분석은 OpenAI GPT-4o 시스템 카드, 구글(Google DeepMind) Gemini 1.0 Ultra 모델 카드 분석과 마찬가지로, 문화콘텐츠 분야 AI 윤리지침 관점에서, 범용 AI 및 생성형 AI의 주요 규제 요건을 중심으로 진행하였다.

❶ 모델 상세(Model Details)
- **유형(Variation):** 7B(70억), 13B(130억), 70B(700억) 매개변수 유형, 사전 훈련 및 미세 조정(fine-tuned) 유형
- **입·출력:** 텍스트만 가능
- **모델 아키텍처:** 자기회귀 언어모델(최적화된 트랜스포머 사용), 미세조정은 감독을 통해 미세조정 및 인간 피드백을 활용한 강화 학습(RLHF)을 사용해 유용성과 안전성 향상
- **훈련 데이터(Training Data)** - 메타 사용자 데이터는 불포함
  - 2조 개 토큰을 학습(공개적으로 이용 가능한 데이터)한 대규모 언어 모델
  - 파인튜닝 데이터는 공개적으로 이용 가능한 인스트럭션 데이터셋(instruction datasets)과 인간 어노테이션 데이터(human-annotated examples) 포함
- **Llama 2 모델군:** 토큰 개수는 사전 훈련 데이터만 기준으로 함, 모든 모델이 글로벌 배치 크기로 400만개 토큰을 사용하여 훈련, 70B 모델은 GQA를 적용하여 추론 확장성 향상
- **모델 훈련기간:** '23.1월~7월

79) Meta. (2024). LLaMA model card. GitHub. Retrieved from https://github.com/meta-llama/llama/blob/main/MODEL_CARD.md

• **상태:** 오프라인 데이터셋으로 훈련된 정적 모델
• **라이선스:** 맞춤형 상업용 라이선스 제공

	훈련 데이터	매개 변수	컨텍스트 길이	GQA (Grouped -Query Attention)	토큰	학습률
Llama2	공개적으로 이용 가능한 온라인 데이터의 새로운 조합	7B	4k	x	2.0T	$3.0x10^{-4}$
		13B	4k	x	2.0T	$3.0x10^{-4}$
		70B	4k	✔	2.0T	$1.5x10^{-4}$

규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템, 고지 및 설명, 알고리즘 차별 보호 원칙 • **(NIST AI RMF)** 안정성, 보안성 및 회복성, 유효성 및 신뢰성 • **(NIST GAI 프로파일)** 유해한 편향 및 균질화, 데이터 프라이버시 리스크 유형 대응 • **(행정명령 제14110호)** AI 모델 개발·배포 투명성 의무(제4조제4.2항(a)목) • **(EU AI Act)** 범용 AI 모델 공급자의 기술문서 제공 및 투명성 의무(제53조)
대응 과제	• **훈련 데이터 출처 세부 정보 비공개** - 모델 세부 정보(아키텍처, 훈련 데이터 매개변수, 훈련 방식 등)를 공개하고 있지만, 훈련 데이터의 출처는 공개적으로 이용 가능한 온라인 데이터로만 명시하고, 세부 정보 비공개

❷ 의도된 사용(Intended Use)
• 영어 기반 연구 및 상업적 용도로 개발
- 세부 조정 모델: 어시스턴트 역할의 챗봇에 적합
- 사전 훈련 모델: 다양한 자연어 생성 과제에 맞춰 조정 가능
• 사용 범위를 벗어난 사용
- 관련 법률 및 규정을 위반하는 방식의 사용(무역 규정 포함)
- 허용 가능한 사용 정책 및 Llama 2 커뮤니티 라이선스에 금지된 기타 모든 방식의 사용

규제 대응	• **(미국 AI 권리장전)** 알고리즘 차별 보호 • **(NIST AI RMF)** 안정성, 보안성 및 회복성, 유해한 편향이 관리된 공정성, 유효성 및 신뢰성, 책임성 및 투명성, 설명 가능성 및 해석 가능성, 프라이버시 강화 • **(NIST GAI 프로파일)** 위험하거나, 폭력적이거나, 혐오스러운 콘텐츠, 유해한 편향 및 균질화, 인간-AI 상호작용(Configuration), 정보 무결성, 지식재산권, 외설적·비하적·학대적 콘텐츠 리스크 유형 대응 • **(EU AI Act)** AI로 금지되는 행위(제5조)
대응 과제	• **비영어권 사용자에 대한 접근성 및 공정성 문제** - 영어 기반 설계로 비영어권 사용자에게 불공정한 영향을 미칠 가능성

❸ 하드웨어·소프트웨어 및 환경적 영향
• **훈련 요소(Training Factors)**
- **사전훈련:** 맞춤형 훈련 라이브러리, 메타의 연구 수퍼 클러스터(Meta's Research Super Cluster) 및 프로덕션 클러스터 사용,

- **파인튜닝(fine-tuning), 어노테이션(annotaion), 평가(evaluation):** 제3자 클라우드 연산으로 수행

• **탄소 배출량(Carbon Footpring) - 사전훈련**

- 소요시간: 총 330만 GPU 시간
- 사용 하드웨어: A100-80GB(TDP 350-400W 전력소모)
- 예상 총 탄소 배출량: 539 tCO2eq(메타의 지속가능 프로그램으로 100% 상쇄)

	시간(GPU 시간)*	전력소비(W)**	탄소 배출량(tCO_2eq)***
Llma 2 7B	184320	400	**31.22**
Llma 2 13B	368640	400	**62.44**
Llma 2 70B	1720320	400	291.42
Total	3311616		539

(*시간(GPU시간): 각 모델을 훈련하는 데 사용된 총 GPU 시간, **전력 소비: 사용된 GPU의 개별 장치당 최대 전력 용량을 전력 사용 효율(PUE)에 맞춰 조정한 값, ***탄소 배출량: 사전 훈련 중 발생한 CO_2 배출량)

규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템 원칙 • **(NIST AI RMF)** 유효성 및 신뢰성, 보안성 및 회복성, • **(NIST GAI 프로파일)** 환경 영향, 가치 사슬 및 구성요소 통합 리스크 유형 대응 • **(행정명령 제14110호)** 환경적 영향 및 지속 가능성 고려(제5조제5.2항(g)목), AI 모델 개발·배포 투명성 의무(제4조제4.2항(a)목), • **(EU AI Act)** 범용 AI 모델 공급자의 기술문서 제공 및 투명성 의무(제53조)
대응 과제	• **제3자 클라우드 서비스 관련 정보 비공개** - 파인튜닝, 어노테이션, 평가 과정이 제3자 클라우드 컴퓨팅으로 수행되었음을 공개했지만, 해당 업체 정보나 이와 관련한 보안 및 프라이버시 정책에 대한 세부 정보 비공개 • **탄소 배출 상쇄의 구체적 정보 및 외부 검증 여부 비공개** - 사전 훈련 과정에 발생한 탄소 배출량이 100% 상쇄되었다고 공개했지만, 상쇄 내역에 대한 상세 정보와 이에 대한 외부 검증 여부는 비공개

❹ 평가 결과(Evaluation Results)

• 평가목적: Llama 1과 비교하여 Llama 2 모델의 발전 정도 및 성능 평가

• 평가항목: 표준화된 학술 벤치마크 총 10개 평가(코딩, 상식적 추론, 세계 지식, 독해, 수학, MMLU*, BBH**, AGI Eval***, TruthfulQA****, Toxigen*****)

벤치마크	평가내용[80]
***MMLU(Massive Multitask Language Understanding)**	• 다양한 주제(과학, 역사, 법률 등)에서 AI 모델의 일반 지식 및 이해력을 평가하는 벤치마크, 모델의 광범위한 지식과 문제 해결 능력을 측정하는데 사용
****BBH (Big-Bench Hard)**	• LLM의 복잡한 추론 능력을 평가하기 위한 벤치마크, 논리적 추론·상식·수학적 문제 해결 등 고난이도 테스트셋으로, 인간과 유사한 추론 능력이 필요한 문제를 포함
*****AGI Eval**	• 인공일반지능(AGI) 수준의 문제 해결 능력을 평가하는 벤치마크, 모델의 포괄적인 지능을 평가

벤치마크	평가내용[80]
****TruthfulQA	• AI 모델이 거짓 없이 사실적인 응답을 생성하는 능력 측정(점수가 높을수록 더 사실적인 답변), 잘못된 정보나 오해의 소지가 있는 답변 생성 경향을 측정
*****ToxiGen	• AI가 생성하는 응답 중 유해하거나 편향된 콘텐츠의 비율을 평가하는 벤치마크. 모델의 안전성 평가에 사용(점수가 낮을수록 AI가 더 안전함)

- 평가방식: 내부 평가 라이브러리 사용
- 평가 결과: ToxiGen(유해 콘텐츠 생성) 평가를 제외한 모든 평가에서 Llama 1보다 성능 개선(특히 코딩, 수학, BBH(논리적 추론), AGI Eval(인공 일반 지능)에서 큰 폭의 성능 향상)

모델 (Model)	크기	코딩 (Code)	상식적 추론	세계 지식	독해	수학	MMLU	BBH	AGI Eval
Llama 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
Llama 1	13B	18.9	66.1	52.6	62.3	10.9	46.9	37	33.9
Llama 1	33B	26	70	58.4	67.6	21.4	57.8	39.8	41.7
Llama 1	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
Llama 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
Llama 2	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
Llama 2	70B	37.5	71.9	63.6	69.4	35.2	68.9	51.2	54.2

※ **▲(코딩(Code))** HumanEval 및 MBPP에서 모델의 pass@1(AI가 첫 번째 시도에서 정답을 맞출 확률) 평균 점수, **▲(상식적 추론)** PIQA, SIQA, HellaSwag, WinoGrande, ARC (easy & challenge), OpenBookQA, CommonsenseQA의 평균 점수, CommonSenseQA는 7-shot(제시 예제 7개) 평가, 다른 벤치마크는 0-shot(예제 없음) 평가 사용, **▲(세계 지식)** NaturalQuestions 및 TriviaQA에서 5-shot(제시 예제 5개) 평균 점수, **▲(독해)** SQuAD, QuAC, BoolQ에서 0-shot(예제 없음) 평균 점수, **▲(수학)** GSM8K (8-shot, 제시 예제 8개) 및 MATH (4-shot, 제시 예제 4개) 벤치마크의 평균 top-1(AI가 가장 높은 확률로 예측한 최상위 답변의 정확도)

모델(Model)	크기	TruthfulQA	Toxigen
lama 1	7B	27.42	23.00
Llama 1	13B	41.74	23.08
Llama 1	33B	44.19	22.57
Llama 1	65B	48.71	21.77
Llama 2	7B	33.29	21.25
Llama 2	13B	41.86	26.10
Llama 2	70B	50.18	24.60
Llama-2-Chat*	7B	57.04	0.00
Llama-2-Chat	13B	62.18	0.00
Llama-2-Chat	70B	64.14	0.01

(*파인튜닝이 진행된 Llama-2-Chat은 다른 안전 데이터셋으로 평가)

규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템 원칙 • **(NIST AI RMF)** 안정성, 보안성 및 회복성, 유효성 및 신뢰성 • **(NIST GAI 프로파일)** 정보 무결성, 유해한 편향 및 균질화 • **(행정명령 제14110호)** AI 모델 개발·배포 투명성 의무(제4조제4.2항(a)목) • **(EU AI Act)** 범용 AI 모델 공급자의 기술문서 제공 및 투명성 의무(제53조)

대응 과제	• **외부 기관 독립적 검증 여부 비공개** - 모델 평가가 내부 평가 라이브러리를 통해 수행 - 외부 기관의 독립적 검증에 대한 언급이 없음 - 독립적 외부 검증 기관 AI 모델 성능 및 안전성 평가 공개 필요 • **평가 데이터셋 상세 정보 비공개** - 평가에 사용된 데이터셋 상세 정보 비공개 • **안전성 성능 향상을 위한 추가 조치 필요** - Toxigen 평가 결과, 유해 콘텐츠 생성 관련 성능 개선 필요
❺ 윤리적 고려 및 한계(Ethical Considerations and Limitations) • 테스트가 영어 중심으로 수행, 모든 시나리오 포괄에는 한계가 있음 • 잠재적 출력 사전 예측 불가, 일부 부정확하거나 편향되거나 부적절한 응답 생성 가능 • Llama 2 활용 애플리케이션 배포 전, 모델의 특정 애플리케이션 상황에 맞춰 안전성 테스트 및 튜닝 필요	
규제 대응	• **(미국 AI 권리장전)** 안전하고 효과적인 시스템, 알고리즘 차별 보호 원칙 • **(NIST AI RMF)** 안정성, 보안성 및 회복성, 유효성 및 신뢰성 • **(NIST GAI 프로파일)** 정보 무결성, 유해한 편향 및 균질화 • **(행정명령 제14110호)** AI 모델의 안전성·투명성 요건(제4조제4.1항(a)목) • **(EU AI Act)** 범용 AI 모델 공급자의 모델 성능·한계 기술문서 제공 의무(제53조)
대응 과제	• **모델의 편향성 및 공정성 개선 조치 필요** - 영어 기반 테스트로 다른 언어 및 문화적 맥락에서 편향된 결과 생성 가능성 및 사회적·윤리적 문제에 대한 부정확하거나 편향된 응답 생성 가능성 완화를 위해 다양한 맥락에서 모델의 안전성과 신뢰성 검증 필요

2) 메타(Meta) AI 규제 대응 현황

ㅇ Llama 2 모델카드 분석을 통해, 메타(Meta)가 모델의 투명성, 안전성, 공정성을 확보하기 위한 정보 공개 정책을 수립하고 있으며, 동시에 미국과 EU의 AI 윤리지침 및 규제 요건을 반영하는 전략을 마련하고 있음을 확인할 수 있다.

---

80) Ip, Jeffrey. (2025, March 5). Benchmarks: MMLU. DeepEval Cloud. Retrieved from https://docs.confident-ai.com/docs/benchmarks-mmlu; Vongthongsri, K. (2025, January 19). LLM benchmarks: MMLU, HellaSwag, and beyond. Confident AI. Retrieved from https://www.confident-ai.com/blog/llm-benchmarks-mmlu-hellaswag-and-beyond; Klu.ai. (n.d.). TruthfulQA eval. Retrieved from https://klu.ai/glossary/truthfulqa-eval; Lin, B., Hilton, J., & Evans, R. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3214–3235. Retrieved from https://aclanthology.org/2022.acl-long.234/

ㅇ 메타(Meta)는 OpenAI 및 구글(Google DeepMind)과 비교했을 때, 훈련 데이터 구성·모델 매개변수·학습률 등과 같은 모델 상세 정보와 에너지 소비량·탄소 배출량 등의 환경적 영향 정보를 상대적으로 상세히 공개하여 투명성을 강조하는 방향으로 접근하고 있음을 보여준다.

하지만, OpenAI 및 구글(Google DeepMind)이 일부 모델 평가 과정에 독립적 연구기관과 협력한 것과 달리, 메타는 내부 평가 라이브러리를 활용하여 모델을 검증하는 방식을 채택하고 있고, 외부 검증과 관련된 정보는 모델카드에 포함하고 있지 않다. 또한, 훈련 데이터의 구체적인 출처 및 세부 구성, 탄소 배출 상쇄 과정 등에 대한 정보도 제공하고 있지 않아, 일부 영역에서 정보 공개가 다소 제한되는 경향을 보인다.

- **(데이터 관리 및 투명성 확보)**
  - 모델 매개변수, 학습 토큰, 컨텍스트 길이, 학습률 등 모델 정보와 구체적인 훈련 데이터 정보 공개
  - 훈련 데이터 출처는 공개적으로 이용 가능한 온라인 데이터라고만 명시하고 세부 정보는 비공개
  - 미국 행정명령 제14110호 및 EU 인공지능법의 AI 투명성과 데이터 공개 의무를 준수하는 방향으로 정보를 제공하되, 기업의 데이터 보호 전략도 고려하여 대응
- **(환경적 영향 고려 및 지속 가능성 전략)**
  - AI 모델 훈련 과정에서 발생한 총 탄소 배출량, 모델별 세부 배출량, 소비 전력, GPU 사용량 등 환경적 영향과 관련한 상세 정보를 정량적으로 평가·공개하면서 투명성을 강조하는 전략을 취함
  - 행정명령 제14110호 환경적 지속 가능성 요건 및 EU 인공지능법의 투명성 요건을 반영하는 방식으로 대응, 단, 탄소 배출량 100% 상쇄라고 명시하면서도, 구체적인 내역 및 외부 검증 여부는 모델카드에 제공하지 않음
- **(위험 식별 및 안전성 강화)**
  - 총 10개의 다양한 벤치마크를 활용하여 모델의 성능 및 안전성 검증
  - 미국 행정명령 제14110호 및 EU 인공지능법의 투명성 요건을 준수

하는 전략을 취하면서, OpenAI와 구글(Google DeepMind)이 독립적 제3자 평가 및 외부 연구기관과의 협력을 밝히고 있는 것과 달리, 내부 평가 라이브러리를 기반으로 검증하는 방식을 유지

- **(사회적 영향 및 윤리적 고려사항)**
  - Llama 2 테스트가 영어 기반으로 수행되었고, 모든 사용 사례를 완전히 포괄할 수 없는 한계를 인정하고, 부정확하거나 편향된 응답 생성 가능성을 명시
  - Llama 2 활용 애플리케이션 배포 이전에 안전성 테스트 및 사용 사례에 맞는 조정 수행을 권장

3) 메타(Meta) AI 규제 대응 전략 및 정책 방향

ㅇ Llama 2 모델카드에서는 모델 매개변수·훈련 데이터·학습률 등 모델 상세 내역 및 환경 영향적 정보도 상당 부분 세부 정보를 공개하고 있다. 이 같은 선제적 공개는 향후 AI 규제 논의에서 메타(Meta)의 공개 기준을 표준 정립에 반영하고자 하는 전략적 접근으로 해석될 수 있다.

또한, 메타(Meta)는 안전성 및 성능 평가 과정에서도 외부 검증에 대한 언급 없이 내부 평가 라이브러리를 활용하였음을 명시하고, 상세한 평가 결과를 공개하는 방식을 채택하고 있다. 이는 내부 라이브러리를 활용한 검증도 투명성을 제공할 수 있으며, 충분히 규제 준수의 대안이 될 수 있음을 시사하는 접근법으로 볼 수 있다.

- **(AI 정보 공개 기준 및 규제 논의 선도)**
  - 모델 아키텍처·매개변수·학습률·데이터 규모 등 훈련 데이터 및 모델 세부 정보를 공개하여 투명성을 강조
  - 모델카드를 통해 선제적으로 훈련 데이터 및 모델 정보를 공개함으로써, 해당 공개 수준이 업계 표준으로 자리 잡고, 향후 AI 규제 논의에 반영되도록 하는 정책 방향을 추진
  - AI 규제 환경에서 모델 정보의 투명성이 강조되는 만큼, 이러한 세부 정보 공개 방식이 규제 당국에 의해 하나의 선례로 활용될 가능성이 높음

- **(AI 모델의 환경적 책임 및 지속 가능성 강조)**
  - 모델 훈련에 사용된 GPU 시간·소비 전력·총 탄소 배출량 등 환경적 영향을 투명하게 공개하고, 지속 가능성 강조
  - 탄소 배출량 100% 상쇄를 명시하여, AI 모델 개발 과정에서 환경적 영향을 최소화하려는 정책 방향을 표명
  - AI 지속 가능성 관련 규제 강화 시, 메타의 환경 정보 공개 및 탄소 배출 상쇄 접근 노력이 글로벌 AI 기업들의 환경 책임 기준으로 자리 잡을 가능성이 있음
- **(내부 검증 방식의 규제 준수 가능성 모색)**
  - 총 10개의 벤치마크 평가 결과를 상세히 공개하면서, 내부 평가 라이브러리를 사용했음을 명시하고, 외부 검증 여부에 대한 정보는 제공하지 않음
  - 이러한 방식의 평가 결과를 상세히 공개한 것은 향후 AI 규제 논의에서 외부 검증 없이도 충분히 투명성과 신뢰성을 확보할 수 있는지 여부를 검토하는 선례가 될 수 있음.

ㅇ 이와같이, 메타(Meta)는 AI 모델 및 훈련 데이터 상세 정보 공개, 환경적 영향 및 지속 가능성, 내부 평가 라이브러리를 활용한 다양한 벤치마크 평가방식을 중심으로 AI 규제에 대응하는 전략을 추진하고 있다. 이같은 선제적 모델카드 공개는 자사 기준을 규제 표준에 반영하고자 하는 정책 방향으로 해석된다.

이에 따라, 향후 AI 거버넌스 논의에서 메타(Meta)가 기술적 선도 기업으로서 정보 공개 수준 및 AI 평가방식의 표준 마련 등에 정책적 영향력을 강화할 것으로 예상된다.

(4) 기업별 AI 규제 대응 시사점

ㅇ AI 규제 환경이 점차 구체화 되면서, 글로벌 AI 선도기업인 OpenAI, 구글(Google DeepMind), 메타(Meta)의 규제 대응 방식이 향후 업계 표준 형성에 영향을 미칠 중요한 참고 사례로 주목받고 있다.

ㅇ 각 기업이 공식적으로 공개한 대표 AI 모델의 시스템 카드 및 모델 카드 분석 결과, 세 기업 모두 규제 준수를 위해 선제적으로 정보 공개 노력을 보이면서, 기업 경쟁력 유지를 위해 핵심 기술 정보(훈련 데이터 출처 및 상세내역, 평가 데이터셋, 평가 방식, 실제 환경에서 성능 검증 등)는 제한적으로 공개하고 있다. 이는 기업들이 AI 규제 흐름을 준수하면서 기술적 경쟁력을 유지하기 위한 전략적 대응으로 볼 수 있다.

OpenAI, 구글(Google DeepMind), 메타(Meta)는 정보 공개 수준[81]과 검증 방식에서 차이를 보이지만, 공통적으로 규제 번화에 대해 신중하게 접근하고 있다. 따라서, 규제 표준이 확립되기 전까지는 기업 정보 보호를 우선하여 대응 범위를 가급적 최소화하려는 경향을 보일 것이다.

또한, AI 규제 표준이 아직 확립되는 과정에 있으므로, 기업들은 선제적인 대응을 통해 업계 표준을 정립하고, 규제 표준 마련 시, 가능한 자사의 공개 수준이 반영될 수 있도록 정책적 참여에 적극적으로 기여할 것으로 예상된다.

81) 파운데이션 모델 투명성 지수(FMTI) 평가 순위: 메타(6위)>OpenAI(11위)>구글(12위)

Ⅳ. 연구 결과 및 시사점

1. 연구 결과 및 향후 연구 과제

(1) 연구 결과 주요 내용

ㅇ 인공지능(AI)은 문화콘텐츠 산업에서 콘텐츠 생성과 유통 방식에 혁신을 가져오고 있으며, 특히 범용 AI와 생성형 AI의 발전은 콘텐츠 제작·홍보·배포 과정에 상당한 변화를 일으키고 있다.

이에 따라, 미국과 EU를 비롯한 각국은 자국의 AI 기술 발전 촉진과 함께 AI 기술이 초래할 수 있는 윤리적·법적 문제 해결을 위한 규제 정책을 강화하고 있으며, 이같은 환경 변화는 AI 기업들의 기술 개발과 운영 방식에 직접적인 영향을 미치고 있다.

ㅇ 이와 같이 AI 규제 환경이 빠르게 진화하고 있는 가운데, 각국은 자국 산업 보호와 국익 관점에 기반해 서로 다른 규제 접근법을 채택하고 있다. 이러한 차이로 인해 AI 기업들은 서로 다른 국가별 규제 환경에 맞춰 사업 운영 방식과 전략을 조정해야 하는 상황에 직면해 있다.

이에 본 연구는 문화 콘텐츠 분야에서 가장 광범위하게 활용되고 있는 범용 AI 및 생성형 AI를 중심으로, 세계 최초의 포괄적 AI 규제법인 EU의 인공지능법과 글로벌 AI 산업을 주도하고 있는 미국의 AI 윤리지침 및 행정명령을 비교·분석하고, 이에 대한 글로벌 AI 기업들의 대응 전략 및 정책 방향을 검토하여, 그 시사점을 도출하는 것을 목표로 하였다.

ㅇ 본 연구의 분석 결과, AI 규제는 AI 기술이 초래할 수 있는 기술적·윤리적 문제 완화 조치와 더불어 국가별 산업 경쟁력과 기술 주도권 확보 전략과 밀접하게 연관되어 있는 것으로 나타났다.

- **(EU 인공지능법)** 위험 기반 접근법(Risk-based Approach)을 통해 AI 시스템의 위험 수준을 4단계로 구분하고, 위험 수준별 규제를 적용하고 있으며, 위반 시 연 매출액 7% 과징금 부과 등 강력한 법적 의무를 부여하고 있다.

특히, 국제 무역 규범과 충돌 가능성 등 여러 논란에도 불구하고, AI 시스템이 생성하는 결과물이 EU 내에서 사용되기만 해도 법 적용을 받도록 하는 역외적용 규정, 제3국 범용 AI 모델 공급자의 제품 출시 전 EU내 공인 대리인 지정 의무 규정, 소스 코드를 비롯한 기술적 수단을 통해 범용 AI 모델 내부 접근 허용 요구 규정 등을 명시하고 있는데, 이는 국내 시장 보호와 함께 EU의 AI 규제를 글로벌 표준으로 자리 잡도록 유도하고 있는 것으로 보인다.

즉, 법 규제를 통한 기술적·윤리적 통제를 넘어, EU의 인공지능(AI) 산업 경쟁력을 강화하고, 글로벌 AI 거버넌스에서 주도권을 확보하려는 전략적 접근과 맞닿아 있는 것으로 해석된다.

- **(미국 AI 윤리지침 및 행정명령)** 법적 강제성을 띠는 규제보다는, AI 윤리지침과 행정명령을 통해 AI 기술의 안전성과 투명성을 확보하는 방향을 채택하고 있다. 특히, 기업의 자율성을 보다 강조하는 한편, AI 규제에 있어서는 특정 영역별 규제 접근 방식(Sectorally Specific Approach)을 취하고 있다.

  즉, 일괄적인 AI 규제법을 제정하지 않고, 기존 산업별 법률과 표준을 활용하면서, 윤리지침과 행정명령을 추가 적용하는 방식으로 규율하고 있다. 이를 통해, 전 세계 AI 기술을 선도하고 있는 미국 기업들의 기술혁신을 촉진하면서, 동시에 필요한 경우, 정부가 개입할 수 있는 정책적 유연성을 확보하고자 하는 전략으로 해석된다.

ㅇ 이에 따라, 글로벌 AI 기업들은 각국 규제의 공통 요건인 투명성 요구에 대응한 기술문서 공개 체계 구축, 법적 규제 준수를 위한 기술적 조치 마련, 자율적 규율 마련을 위한 내부 관리 시스템 구축 및 거버넌스 강화 등 국가별 공통 규제 요건과 상이한 규제 접근에 균형 있게 대응함으로써, 진화하는 AI 규제 환경에 따라 운영 전략을 조정할 필요가 있다.

- **(법적 규제 준수 중심 접근)** 데이터 공개 및 투명성과 관련한 규제 요건 준수를 위해, 공식적인 정보 공개 문서(AI 모델 시스템 카드, 모델 카드,

기술 보고서 등) 체계 도입, 생성형 AI 콘텐츠 출처 명시 등 기술 개발

- **(자율적 내부 정책 강화 중심 접근)** 데이터 관리 시스템 구축 및 감사 등 내부 AI 거버넌스 체계 강화

(2) 연구 결과 의미 및 향후 연구 과제

ㅇ 본 연구는 문화콘텐츠 산업에 점차 그 영향력이 확대되고 있는 범용 AI 및 생성형 AI를 중심으로 세계 최초의 포괄적 AI 규제법인 EU 인공지능법과 글로벌 AI 기술을 선도하고 있는 미국의 AI 윤리지침 및 행정명령을 비교·분석하고, 이에 대응하는 글로벌 AI 기업들의 전략을 검토하여 그 시사점을 도출했다는 점에 의의가 있다.

- **(미국-EU 규제 비교·분석)** EU와 미국의 AI 윤리지침 및 규제 내용을 범용 AI 및 생성형 AI를 중심으로 구체적으로 분석하고, 미국-EU 기술위원회 등 AI 기술 발전과 규제 환경의 조화를 도모하려는 노력을 검토하여, 향후 AI 규제 환경의 변화를 전망
- **(글로벌 AI 기업 규제 대응 전략 분석 및 시사점 도출)** 스탠포드 대학 파운데이션 모델 투명성 지수(FMTI) 보고서와 글로벌 AI 기업들이 자발적으로 공개하고 있는 AI 모델 시스템 카드·모델 카드 등 공식 문서를 실증적으로 분석하여, 미국과 EU의 AI 규제가 글로벌 AI 선도 기업들의 기술 개발과 운영 방식에 미친 영향과 기업 대응 전략 및 정책 방향을 검토하고 그 시사점을 도출

ㅇ 그러나, 실제 기업의 대응 사례가 현재 진행 중으로 계속 변화하는 과정에 있다는 점에 있어 본 연구 분석에 한계가 있다. EU 인공지능법과 미국의 행정명령 제141110호도 최근 발표된 규제로, 아직 규정의 세부 기준이나 표준이 마련 중인 단계이므로, 기업들의 장기적인 대응 전략은 지속적인 관찰이 필요하다.

따라서, 향후 연구에서는 글로벌 AI 규제 환경의 변화를 지속적으로 추적하고, 각국의 AI 규제에 대응하기 위한 기업의 정책적·기술적 전략을 보다 구체적으로 분석하는 것이 필요할 것이다.

2. AI 관련 규제 환경 변화 및 시사점

(1) AI 관련 최신 규제 환경 동향

○ 글로벌 AI 규제 환경은 각국의 데이터 보호 및 기술 산업 보호 전략과 밀접하게 연관되어 있다. 특히 미국·EU·중국은 AI 규제와 함께 데이터 규제를 통해 국가 산업 경쟁력과 안보·미래 전략을 강화해 나가고 있다.

❶ (미국) 국가 안보를 근거로 데이터 보호 및 타국 기술 기업 규제 강화
- 틱톡(TikTok) 강제매각법 사례

- 미국 내 가입자 수가 1억 7천만명이 넘는 틱톡(TikTok)이 미국 사용자 데이터를 수집하여, 이를 감시나 조작 등에 악용할 우려 제기[82)]
- 美정부가 틱톡(TikTok)이 미국 사용자 데이터를 중국 정부와 공유할 가능성과 관련해, 국가 안보를 근거로 중국 모기업 바이트댄스(ByteDance)에 미국 내 사업권 매각을 강제[83)]

내용[84)]	• **(금지 조치)** 외국 적대국이 통제하는 애플리케이션의 배포·유지·업데이트 금지 • **(강제 매각)** 바이트댄스는 270일 내에 틱톡의 미국 사업권을 매각, 매각 후 해당 앱은 외국 적대국과의 모든 운영적 관계를 단절해야 함 • **(매각 조건)** 매수자는 외국 적대국과 무관해야 하며, 대통령이 이를 승인해야 함 • **(위반 시 제재)** 앱스토어(애플, 구글)와 인터넷 호스팅 서비스 제공자가 틱톡 배포 및 업데이트를 지원하지 못하도록 제재함으로써 앱 기능을 점진적으로 저하시킴
경과[85)]	• ('20.8.6.) 트럼프 대통령, 틱톡 금지 행정명령 서명 • ('24.3.13.) 미 하원, 틱톡 강제매각법 통과 • ('24.4.24.) 바이든 대통령, 틱톡 강제매각 조항 포함 법안 서명 • ('25.1.17.) 美대법원, 틱톡 강제매각법 합헌 판결 • ('25.1.18.) 틱톡, 미국 내 서비스 중단 • ('25.1.19.) 트럼프 대통령, 강제매각 기한 75일 연장 행정명령 서명 • ('25.2.13.) 틱톡, 앱스토어에서 다시 다운로드 가능 • ('25.3.6.) 트럼프 대통령, 틱톡 금지 기한 추가 연장 가능성 시사, ※ '25년 3월 기준 바이트댄스는 4.4.까지 틱톡을 매각해야 하는 상황이며, 일론 머스크, 미국 내 비영리 단체 등이 잠재적 구매자로 거론되고 있음.

82) Holland & Knight. (2025, January 17). U.S. Supreme Court upholds TikTok sale or ban law. Retrieved from https://www.hklaw.com/en/insights/publications/2025/01/us-supreme-court-upholds-tiktok-sale-or-ban-law

83) Kiely, E. (2025, February 3). TikTok and U.S. national security. Retrieved from https://www.factcheck.org/2025/02/tiktok-and-u-s-national-security/

84) Holland & Knight. (2025, January 17). U.S. Supreme Court upholds TikTok sale or ban law. Retrieved from

ㅇ 틱톡 사례는 미국이 데이터 주권을 국가 안보의 핵심 요소로 간주하고 있음을 보여준다. 중국 정부가 요청하면 중국 기업은 데이터를 제공해야 한다는 중국 국가정보법(제7조, 제14조, 제16조) 규정[86]은 미국이 외국 기술 기업의 데이터 처리 방식을 강하게 규제하는 계기가 되었다.

이는 AI 윤리 및 규제논의에서도 데이터 주권이 중요한 고려 요소가 될 것임을 시사한다. 각국은 AI 관련 규제를 설계할 때, AI 모델이 학습하고 의사결정을 내리는 과정에 데이터의 주권적 통제를 강화하는 방향으로 나아갈 것으로 예상된다.

ㅇ 또한, 틱톡 강제매각법은 앱 매각의 문제가 아니라, 미중 간 기술 패권 경쟁의 일환으로도 해석될 수 있다. 중국은 틱톡 추천 알고리즘 매각을 금지하며 맞대응했고, 이는 양국 간 긴장을 고조시키고 있다[87]. 이는 AI 및 데이터 규제가 국가 간 경제·안보 전략과 밀접하게 연결되어 있으며, 각국의 정책적 이해관계에 따라 경쟁적으로 전개되고 있음을 보여준다.

---

https://www.hklaw.com/en/insights/publications/2025/01/us-supreme-court-upholds-tiktok-sale-or-ban-law; Restrict Act. (n.d.). Retrieved from https://www.restrict-act.com/; Vattikonda, N., & Kelley, B. (2025, January 9). The D.C. Circuit Court's TikTok ban decision—Explained. Retrieved from https://www.lawfaremedia.org/article/the-d.c.-circuit-court's-tiktok-ban-decision--explained

85) Lutkevich, B. (2025, February 18). TikTok bans explained: Everything you need to know. Retrieved from https://www.techtarget.com/whatis/feature/TikTok-bans-explained-Everything-you-need-to-know; Basso-Davis, F. (2024, February 11). TikTok ban timeline and updates. Retrieved from https://wwcecho.news/news/tiktok-ban-timeline-and-updates; Nazzaro, M. (2025, March 6). Trump considers extending TikTok ban. Retrieved from https://thehill.com/policy/technology/5180939-trump-considers-extending-tiktok-ban/; Thomson Reuters. (2025, January 17). Supreme Court upholds TikTok ban. Retrieved from https://www.cbc.ca/news/world/supreme-court-tiktok-ban-1.7434082; Bachman, J. (2025, January 17). Supreme Court upholds U.S. law forcing TikTok ban or sale. Retrieved from https://www.legaldive.com/news/supreme-court-upholds-us-law-forcing-tiktok-ban-or-sale/737704/

86) 김청중. (2025, February 2). 중국, '국가정보법'으로 데이터 수집 우려. 세계일보. Retrieved from https://www.segye.com/newsView/20250202508426

87) 정혜진. (2024, May 8). "강제매각법은 표현의 자유 침해"…틱톡, 美 정부와 소송전. 서울경제. Retrieved from https://www.sedaily.com/NewsView/2D93YFDUK3

이러한 사례는 다른 외국 기업에도 언제든 적용될 가능성이 열려 있기 때문에, 글로벌 테크 기업들은 각국의 규제 환경에 적응하기 위해 지역별 데이터 분리 및 운영 전략을 수립할 필요가 있다.

❷ (EU) 강력한 AI 및 데이터 규제로 기업 활동 제한

- 애플 과징금 사례

- 애플 iOS 인앱 결제 시스템 강제, 최대 30%의 수수료 부과, 외부 결제 링크 금지 정책으로 경쟁 제한 및 소비자 선택권 침해 논란
- '19.3월 스웨덴 음악 스트리밍 서비스 업체 스포티파이(Spotify)가 애플이 자사 앱스토어 정책을 통해 경쟁사 서비스보다 애플 뮤직을 유리하게 만든다고 주장하며 경쟁법(반독점법) 위반 행위로 EU에 제소[88]

내용[89]	• **EU 경쟁법(반독점법, Competition Law)** : 시장에서의 공정 경쟁을 촉진하고, 소비자 이익을 보호하기 위해 발전된 규정 ※ 1993년 마스트리히트 조약 발효로 EC(European Community) 경쟁법이 EU 반독점법으로 발전, 2009년 리스본 조약 발효로 현재의 유럽연합 기능에 관한 조약(TFEU)에 반독점 규정이 포함되면서 현재의 형태를 갖추게 됨 - **(주요 내용)** 독점적 행위, 시장 지배적 지위 남용, 반경쟁적 행위 규제 - **(위반 시 처벌)** 글로벌 연 매출의 최대 10% 과징금 부과(DMA와 달리 반복 위반 시 상향 규정은 없으나, 위반 행위 지속성·고의성에 따라 제재 강화 가능) • **EU 디지털시장법(Digital Markets Act, DMA)** : 빅테크 기업*의 독점적 행위를 규제하고, 공정경쟁과 소비자 보호를 촉진하기 위해 제정('23.5월 발효) - **(주요 내용)** 앱스토어 수수료 제한, 대체 결제 옵션 허용, 제3자 앱스토어 지원 등 - **(위반 시 처벌)** 글로벌 연 매출의 최대 10% 과징금 부과 가능, 반복 위반 시 20%까지 상향 가능, 상습적 위반 시 인수합병(M&A) 일시 금지 * 시가총액 750억유로(약 100조원), 연매출 75억유로(10조원), 월간 사용자 4500만명 이상인 IT 기업을 "게이트키퍼(gatekeeper)로 지정하고 규제(아마존, 페이스북 모기업인 메타, 구글 모기업 알파벳, 마이크로소프트(MS), 애플 등 미국 빅테크 기업들이 이에 해당)
경과[90]	• ('19.3.13.) 스포티아, 애플 상대로 EU 집행위원회에 경쟁법(반독점법) 위반 제소 • ('24.3.4.) EU는 애플이 음악 스트리밍 시장에서 경쟁사를 불리하게 만들었다고 판단하고, 경쟁법(반독점법) 위반에 대해 €18억(약 $19.7억) 과징금 부과. • ('24.6.24.) EU는 애플이 DMA를 위반했다고 발표하며 추가 조사 시작 • ('24.11.6.) EU는 애플에 대해 DMA 위반으로 추가 과징금 부과 예정 발표. • ('25.1-2월) 애플이 DMA 준수를 위해 앱스토어 정책을 일부 변경, EU는 여전히 불충분하다고 판단

88) Bostoen, F. (2019, April 24). Spotify vs. Apple: The Core of the Dispute. Retrieved from https://www.lexxion.eu/en/coreblogpost/spotify-apple/

	※ EU는 애플에 "소규모" 과징금을 부과할 예정이라고 발표하였으나, 과징금 규모는 아직 확정되지 않음, EU는 애플의 DMA 준수 여부를 지속적으로 평가 중이며, 이번 과징금은 미국 트럼프 정부와의 무역 갈등을 고려해 최대치(글로벌 연 매출의 10%)에는 미치지 않는, 상대적으로 적은 금액으로 책정될 것으로 예상 ※ EU는 벌금보다는 규정 준수를 유도하는 데 초점을 맞추고 있으며, 애플이 향후 정책을 어떻게 변경할지 주목하고 있음.

ㅇ EU의 경쟁법과 디지털시장법(DMA)은 AI 기술이 디지털 플랫폼 전반에 광범위하게 사용되는 상황에서, AI의 윤리적 활용과 시장 공정성을 확보하는 핵심 규제 도구로 자리 잡고 있다. 특히, 디지털 플롯폼에서 AI 기술을 활용한 독점적 관행을 억제하고, AI 기술이 시장 지배력 강화에 악용되는 것을 방지하는 역할을 한다.

- AI 기반 추천 알고리즘이 특정 서비스를 우대하거나, 경쟁사를 불리하게 만드는 경우, 공정경쟁법과 DMA의 규제 대상
- DMA의 "게이트키퍼(gatekeeper)"로 지정된 기업(알파벳, 아마존, 애플, 메타, 마이크로소프트 등)은 자사 서비스 우대를 금지하고, 데이터 결합을 제한하며, 상호운용성을 보장하도록 요구

89) Spotify Newsroom. (2024, August 14). The European Commission confirms Apple's anti-competitive behavior is illegal and harms consumers. Retrieved from https://newsroom.spotify.com/2024-03-04/the-european-commission-confirms-apples-anti-competitive-behavior-is-illegal-and-harms-consumers/; 박효재. (2022, March 25). EU, 미 IT공룡 독점 금지 '디지털 시장법' 합의. 경향신문. Retrieved from https://www.khan.co.kr/article/202203251053001; Usercentrics. (2023, October 12). Digital Markets Act Timeline. Retrieved from https://usercentrics.com/knowledge-hub/digital-markets-act-timeline/; European Parliament. (2022, November 30). Digital Markets Act: What's Next? Retrieved from https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2022)739226

90) McCallum, S. (2024, March 4). Apple Faces EU Scrutiny Under Digital Markets Act. Retrieved from https://www.bbc.com/news/technology-68467752; PYMNTS. (2024, November 5). Apple Faces Fine in First Charge Under EU's Digital Markets Act. Retrieved from https://www.pymnts.com/apple/2024/apple-faces-fine-in-first-charge-under-eus-digital-markets-act/; Cooban, A. (2024, June 24). Apple and EU Competition Rules. Retrieved from https://www.cnn.com/2024/06/24/business/apple-eu-competition-rules/index.html; Clover, J. (2025, March 10). Apple Receives 'Modest' Fine Under EU Digital Markets Act. Retrieved from https://www.macrumors.com/2025/03/10/apple-modest-fine-eu-digital-markets-act/; Christoffel, R. (2025, March 10). Report: Apple Will Be Fined by EU for Alleged Violation of DMA. Retrieved from https://9to5mac.com/2025/03/10/report-apple-will-be-fined-by-eu-for-alleged-violation-of-dma/; Owen, M. (2025, March 10). Third Time's the Charm? EU Considers Fining Apple Over DMA Yet Again. Retrieved from https://appleinsider.com/articles/25/03/10/third-times-the-charm-eu-considers-fining-apple-over-dma-yet-again

ㅇ 이같은 디지털 규제는 EU의 인공지능법과 함께 AI 기술이 소비자 선택권을 제한하거나 시장 경쟁을 저해하지 않도록 투명성과 책임성을 요구하고, 높은 수준의 과징금을 규정하고 있다. 이는 기업에 강력한 억제력을 제공하여 규제 준수를 유도하고, 공정경쟁과 책임 있는 AI 기술 사용을 보장하기 위한 EU의 일관된 정책 방향을 반영한 것으로 해석될 수 있다.

ㅇ 하지만, 실제 규제 대상이 대부분 미국의 글로벌 빅테크 기업이라는 점에서, 글로벌 기술 패권을 둘러싼 미국과 EU간 규제 갈등으로 확산되는 양상을 보이고 있다.

- **(미국)** ▲EU의 규제가 자국 기업에 과도한 부담을 주고 있고, 이는 불공정한 "과세"라고 주장[91], ▲'25.2월 트럼프 대통령 디지털 서비스세(DST) 및 기타 규제에 대해 무역 보복 검토 발표(관세 부과 가능성을 포함하며, 미국 기업 보호를 위한 조치로 해석됨)[92]
- **(EU)** 특정 국가나 기업을 겨냥한 것이 아니라 공정한 경쟁 환경을 조성하기 위한 조치임을 강조[93]

ㅇ EU는 세계 최초의 포괄적 인공지능(AI) 규제인 인공지능법(AI Act)을 비롯해 디지털시장법(DMA) 등을 제정하여, 글로벌 기술 규제의 기준을 정립하고 이를 국제 표준으로 확산시키려는 전략을 취하고 있다. 현재 한국, 일본, 캐나다, 호주 등 여러 국가들이 EU의 선례를 참고하여 유사한 법안을 도입하고 있어, 이 같은 전략이 실제 국제적 영향력을 발휘하고 있음을 보여준다.

---

91) Kirkwood, M. (2025, March 6). Digital Markets Act roundup: February 2025. TechPolicy.Press. Retrieved from https://www.techpolicy.press/digital-markets-act-roundup-february-2025/; Parry, J., & Micheletti, F. (2025, March 7). EU tech chiefs say they don't target US tech. POLITICO. Retrieved from https://www.politico.eu/article/eu-tech-chiefs-say-they-dont-target-us-tech/

92) Allen, B. E., Leiter, M. E., Werry, S., & Bell, J. F. (2025, February 28). Trump revives and expands the battle over digital services taxes. Skadden Publication / Executive Briefing: Latest Updates on the Trump Administration. Retrieved from https://www.skadden.com/insights/publications/2025/02/trump-revives-and-expands-the-battle-over-digital-services-taxes

93) European Commission. (2024, March 25). The Digital Markets Act: Fair and competitive digital markets. Retrieved from https://data.europa.eu/en/news-events/news/digital-markets-act-fair-and-competitive-digital-markets

ㅇ 이러한 상황은 AI 및 디지털 규제가 국가 간 경제적 이해관계와 기술 패권 경쟁의 연장선상에서 작동하고 있음을 보여준다.

이제 글로벌 테크 기업들은 국가마다 상이한 강제 규정과 처벌수위를 고려하여, AI 기술 활용 방식을 조정하고, 이에 따른 추가 비용과 운영 복잡성을 부담해야 하는 과제에 직면해 있다. 따라서, 글로벌 규제 흐름을 면밀히 분석하고, 각국의 법적 요건을 반영하여 AI 개발 및 운영 전략을 포함한 체계적인 내부 정책을 마련해야 할 필요가 있다.

❸ (중국) 데이터를 국가 전략 자산으로 간주하여 데이터 해외 반출 금지 - 테슬라의 중국 내 데이터 저장 정책 사례

- 중국, 데이터를 국가 전략 자산으로 간주하고 해외 반출 엄격히 제한[94)]
  - 중국 정부의 국가 안보 및 기술 주도권 확보를 위한 정책의 일환
  - 외국 기업이 중국 시장 내 운영을 위하여, 현지화된 데이터 관리 시스템을 구축하도록 압박
- 테슬라(Tesla), 중국 시장의 중요성을 인식하고 데이터 규제 수용[95)]
  - 중국 내 판매 차량 바이두 지도를 사용하고 중국 내에서만 데이터 저장
  - 외국 기업 최초로 데이터 안전 검사에서 "적합" 판정을 받음
  - 외국 기업이 중국 시장에서 활동하려면 규제를 준수해야 한다는 중요한 선례가 됨

내용[96)]	• **데이터 3법(데이터 안전법, 네트워크 안전법, 개인정보 보호법)** : 데이터를 국가의 핵심 자산으로 규정하여 강력히 관리, 외국 기업들은 중국 내 서비스와 글로벌 서비스를 물리적으로 분리해야 함 - 데이터를 수집·저장·전송 등 처리 활동에 대해 국가의 역할을 명시하며, 국가 보안, 국민경제, 공공 이익 관련 데이터를 국가 핵심 데이터로 규정(데이터 안전법 제3조)

94) Economic Times. (2024, April 29). Musk's big win in China: Tesla clears data security, full self-driving hurdles for locally-made cars. The Economic Times. Retrieved from https://economictimes.indiatimes.com/industry/renewables/musks-big-win-in-china-tesla-clears-data-security-full-self-driving-hurdles-for-locally-made-cars/articleshow/109678866.cms
95) Cheng, E. (2024, April 29). Musk visits Beijing as Tesla's China-made cars pass security rules. CNBC. Retrieved from https://www.cnbc.com/2024/04/29/musk-visits-beijing-as-teslas-china-made-cars-pass-security-rules.html; Yang, Z. (2023, August 14). Chinese Tesla data to be stored on mainland. China Daily. Retrieved from https://www.chinadailyhk.com/hk/article/345833

	- 핵심 데이터는 해외 반출 금지 조항을 통해 해외로 전송 불가(데이터 안전법 제31조 및 네트워크 안전법 제37조)
경과[97)]	• ('21.4월) 중국 정부, 테슬라에 중국 내 수집된 모든 데이터를 현지에 저장하도록 요구 • ('21.4.21.) 테슬라, 상하이에 데이터 센터를 구축하여 중국에서 수집된 데이터를 처리하겠다고 발표 • ('21.10.25.) 테슬라, 상하이 데이터 센터 완공 및 운영 시작 공식 발표 • ('24.4.28.) 테슬라 CEO 일론 머스크, 중국 방문하여 총리와 회담, 중국 자동차 제조업 협회에서 테슬라 차량이 데이터 보안 요구사항을 충족했다고 발표 • ('24.4.29.) 중국자동차공업협회(CAAM)와 국가컴퓨터네트워크응급기술처리협조센터(CNCERT/CC)가 공동으로 테슬라의 중국산 모델(Model 3 및 Model Y)이 중국의 데이터 보안 규정을 준수한다고 확인 • ('24.11.8.) 테슬라, "자동차 개인정보 보호 인증"을 획득, 이는 중국에서 생산된 모든 차량이 자동차 데이터 보안 규정을 준수하고 있음을 의미하며, 외국 기업으로서는 최초로 받은 인증으로 평가됨 • ('25.2.24.) 테슬라, 중국에서 완전 자율 주행(FSD) 시스템 출시를 승인받음. 이에 따라 북미·유럽에 이어 중국에서도 완전 자율 주행(FSD) 기능이 도입될 예정 ※ '25년 3월 기준 테슬라는 완전 자율 주행(FSD) 시스템의 중국 내 출시를 준비 중이며, 이를 통해 자율 주행 기술 경쟁에서 우위를 점하려 하고 있음. 또한, 데이터 로컬라이제이션 및 현지 파트너십 전략을 통해 규제를 철저히 준수하고 있음

96) 중화인민공화국 네트워크 안전법. (2024, February 2). 중화인민공화국 네트워크 안전법. 세계법제정보센터. Retrieved from https://world.moleg.go.kr/web/wli/lgslInfoReadPage.do?CTS_SEQ=41114&AST_SEQ=53; 중화인민공화국 데이터 안전법. (2024, February 2). 중화인민공화국 데이터 안전법. 세계법제정보센터. Retrieved from https://world.moleg.go.kr/web/wli/lgslInfoReadPage.do?CTS_SEQ=49497&AST_SEQ=53

97) Pandaily. (2024, April 28). Tesla meets China's data security standards, restrictions to be lifted. Retrieved from https://pandaily.com/tesla-meets-chinas-data-security-standards-restrictions-to-be-lifted/; Reuters. (2024, April 29). Baidu, Tesla said to have agreed on mapping deal for FSD in China. The Economic Times. Retrieved from https://economictimes.indiatimes.com/news/international/business/baidu-tesla-said-to-have-agreed-on-mapping-deal-for-fsd-in-china/articleshow/109678784.cms; Alvarez, S. (2024, April 29). Tesla China partners with Baidu for maps to clear FSD hurdle. Teslarati. Retrieved from https://www.teslarati.com/tesla-china-partners-baidu-maps-clear-fsd-hurdle/; Gubina, E. (2024, April 29). Tesla cars have been tested for compliance with data security requirements in China. Mezha.Media. Retrieved from https://mezha.media/en/2024/04/29/tesla-cars-have-been-tested-for-compliance-with-data-security-requirements-in-china/; Alvarez, S. (2024, November 8). Tesla China earns "Automotive Privacy" certification in China. Teslarati. Retrieved from https://www.teslarati.com/tesla-china-first-company-automotive-privacy-certification/; Park, W. (2025, February 24). Tesla's full self-driving: A game changer in China. AInvest. Retrieved from https://www.ainvest.com/news/tesla-full-driving-game-changer-china-2502/; Fox, E. (2021, April 20). Tesla to build new data center in China in response to law on local storage of EV data. Tesmanian. Retrieved from https://www.tesmanian.com/blogs/tesmanian-blog/tesla-will-build-a-new-data-center-in-china-due-to-the-new-law-on-the-local-storage-of-ev-data; Borak, M. (2021, April 21). Tesla

ㅇ 중국의 데이터 3법과 같은 강력한 디지털 규제 정책은 데이터 주권, 국가 보안 등을 이유로 데이터의 이동성과 접근성을 제한하고 있다. AI 모델의 학습과 발전은 방대한 데이터의 수집과 분석을 전제로 하기 때문에 강화된 디지털 규제는 AI 기반 서비스의 운영 방식과 신뢰성·공정성을 보장하기 위한 AI 윤리지침에 직접적인 영향을 미치고 있다. 테슬라가 중국의 데이터 규제를 준수하면서 자율 주행 같은 AI 기반 서비스를 유지한 사례는 AI 기술이 각국의 디지털 규제 환경 속에서 어떻게 적응해나가고 있는지를 보여주는 중요한 사례로 볼 수 있다.

- **(데이터 현지화(localization) 요구)** 미국·중국을 비롯한 각국의 데이터 주권 강화 움직임은 AI 기업들에게 데이터 현지화(localization)를 필수적으로 요구하는 방향으로 변화하고 있다.

글로벌 테크 기업들은 시장 진출을 원하는 국가의 데이터 보호 규제 준수를 위해 현지에 데이터 저장 인프라를 구축해야 하는 부담을 안게 된 것이다. 특히, AI 모델 학습 및 데이터 분석 시, 특정 국가의 데이터 활용이 해당 국가 내에서만 이루어져야 하는 경우, AI 기업들의 기술 개발과 운영 방식은 근본적으로 변화해야 할 가능성이 크다. 테슬라가 중국 내 데이터 저장 정책을 도입한 사례는 이러한 흐름을 단적으로 보여준다.

---

promises new China data centre as Beijing lays down the law on local storage of EV data. South China Morning Post. Retrieved from https://www.scmp.com/tech/big-tech/article/3130339/tesla-promises-new-china-data-centre-beijing-lays-down-law-local; Cheng, E. (2024, April 29). Musk makes surprise visit to Beijing as Tesla's China-made cars pass data security rules. CNBC. Retrieved from https://www.cnbc.com/2024/04/29/musk-visits-beijing-as-teslas-china-made-cars-pass-security-rules.html; Geopolitechs. (2024, April 28). China lifted ban on Tesla vehicles during Musk's China visit. Retrieved from https://www.geopolitechs.org/p/china-lifted-ban-on-tesla-vehicles; Fox, E. (2021, October 25). Tesla launches Shanghai R&D center & Gigafactory data center in China. Tesmanian. Retrieved from https://www.tesmanian.com/blogs/tesmanian-blog/tesla-launches-shanghai-r-d-center-and-gigafactory-data-center-in-china; 심두보. (2024, April 30). '중국 데이터 안전 검사 통과' 급등한 테슬라 주가. 딜사이트. Retrieved from https://dealsite.co.kr/articles/121887; 신화통신 한국어 뉴스 서비스. (2024, May 1). 중 차량 보안 테스트 통과 ICV 모델 발표...BYD·테슬라 등 76종. 신화망 한국어판. Retrieved from https://kr.news.cn/20240501/7fb69de887b44ac187549fa2473e304f/c.html

이는 AI 기업들이 해당 국가에서 사업을 지속적으로 운영하기 위해서 반드시 준수해야 하는 현실이 되고 있다. 따라서, 유연한 데이터 거버넌스 모델과 현지 규제 준수를 반영한 기술 아키텍처를 구축해 이에 대응해 나가야 할 필요가 있다.

- **(데이터 보호와 AI 윤리의 조화)** AI 기술의 신뢰성과 책임성을 확보하기 위해서는 이러한 데이터 보호 규제를 AI 윤리에 반영해야 한다. AI 학습 데이터는 법적·윤리적으로 적절한 방식에 따라 수집·저장·활용되어야 하는데, 이는 해당 국가가 보호하려는 데이터 규제 기준을 준수하는 방향으로 설계되어야 하기 때문이다.

데이터 보호 규제는 단순히 기업에 법적 의무를 부과하는 것이 아니라, 신뢰받는 AI 기술을 위해 필수적인 윤리적 장치로 기능한다. 테슬라가 중국의 데이터 규제를 준수하면서 AI 기술을 활용한 서비스를 지속한 사례는, 법적 요건 충족을 넘어서, AI 시스템이 해당 국가에서 신뢰받는 방식으로 운영됨을 승인받았다는 점에서 의미가 있다.

향후 데이터 규제와 AI의 윤리 준수 여부를 함께 검증하는 인증 제도나 규제 체계가 강화될 것으로 예상되며, 이는 현지에서의 기업 신뢰성 확보뿐 아니라 글로벌 시장에서의 경쟁력과도 직결될 것이다.

이에 따라, 기업들은 각국의 데이터 규제 요건 준수와 AI의 신뢰성을 확보할 수 있도록 AI 개발·운영방식 및 데이터 관리 전략을 수립해야 할 것이다.

- (AI 및 데이터 규제 국제 표준화) 점차 데이터 규제가 강화되는 방향으로 변화하면서, AI 및 데이터 보호 규제의 국제 표준화 논의가 점차 가속화될 것으로 예상된다. 국가마다 다른 규제 차이는 기업에 부담이 될 수밖에 없고, 데이터 이동의 제한은 글로벌 협력을 저해하는 요소로 작용할 수 있다.

따라서 장기적으로 AI 및 데이터 보호 규제의 국제적 조율이 필요한 시점이 도래할 것이다. 따라서, 단기적으로는 각국 규제에 따라 데이터 운영 전략을 최적화하면서, 장기적으로는 글로벌 데이터 보호 표준화에 대비한 전략을 마련해야 한다.

(2) AI 규제 환경 변화와 국내 대응 방향

ㅇ AI 기술 발전과 더불어 각국의 AI 규제 환경도 빠르게 변화하고 있다. AI의 활용은 AI만의 기술적 문제를 넘어 데이터 보호, 국가 안보, 산업 보호와 같은 다양한 법적·정책적 사항과 밀접히 연결되어 있으므로, AI 규제는 AI 자체의 윤리적 문제 뿐만 아니라 국가 경제 및 기술 경쟁력 확보를 위한 전략적 수단으로 활용되고 있다.

때문에 AI 규제는 단순히 기술 규제로만 인식하는 게 아니라, 데이터 보호 규제와 디지털 시장 규제 등 다양한 법적 요구를 함께 고려해야 한다. 이에 따라, AI 기업과 정부는 정책 마련 시, 글로벌 AI 규제 환경의 흐름을 종합적으로 분석하여 대응 전략을 마련할 필요가 있다.

1) AI 규제 환경 변화의 시사점

ㅇ 미국, EU, 중국의 사례에서 살펴본 바와 같이, 각국은 AI 및 데이터 규제를 국가 경제 및 기술 경쟁력 강화를 위한 핵심 정책 수단으로 활용하고 있다.

미국	• 국가 안보를 근거로 틱톡(TikTok)의 강제 매각 추진, 데이터 주권과 AI 규제를 연계하는 전략 추진
EU	• 인공지능법(AI Act), 경쟁법(Competition Law), 디지털시장법(DMA)을 통해 AI 기술이 플랫폼 시장에서 독점적으로 활용되는 것을 막고, 공정 경쟁을 촉진하는 방향으로 규제를 강화
중국	• 데이터 3법(데이터 안전법, 네트워크 안전법, 개인정보 보호법)을 통해 데이터를 국가 전략 자산으로 간주하고, AI 기업들이 중국 내에서 데이터를 저장·처리하도록 의무화

ㅇ 이처럼, 이제 AI 및 데이터 규제는 국제 경제와 무역 환경 전반의 영향을 미치고 있고, 이에 따라, 글로벌 테크 기업들은 AI 규제 대응에 있어 AI 관련 지침과 법령뿐 아니라, 데이터 보호 및 디지털 시장 규제까지 통합적으로 고려하여, 법적·정책적 대응 방안을 수립해야 한다.

테슬라(Tesla)를 비롯한 미국의 글로벌 AI 선도 기업들은 이미 데이터 주권 강화 움직임에 대응해 데이터 현지화 전략을 채택하고, 미국과 EU 정부의 AI 규제 및 윤리지침 준수를 위해 내부 기준을 정립하고, 모델 카드·시스템 카드·기술 보고서 등을 통해 자사 AI 모델의 기술적 정보를 공식적으로 공개하여 신뢰성과 투명성을 확보하는 체계를 마련해 나가고 있다.

이러한 대응 방식은 향후 글로벌 AI 규제 준수 및 시장 접근 전략의 표준이 될 것으로 예상되며, 국내에서도 이같은 흐름을 반영한 대응 전략 마련이 필요하다.

2) 국내 대응 방향 제언

ㅇ 글로벌 AI 규제 환경의 변화 흐름에 따라 우리 정부도 AI 기술 발전을 촉진하면서 신뢰성을 강화하는 방향으로 정책 대응을 추진해 왔다. 특히, 국제 사회에서 논의되고 있는 AI 규제 프레임워크에 대응을 강화하며 글로벌 AI 거버넌스에 적극 참여하고 있다.

(i) 정부 대응 주요 동향 및 법제화 노력

뉴욕구상[98] ('22.9.21.)	- 디지털 심화 시대의 새로운 질서 정립과 국제 사회의 연대 필요성을 처음으로 제시
파리 이니셔티브[99] ('23.6.21.)	- 프랑스 파리 소르본대에서 열린 '디지털 비전 포럼'에서 새로운 디지털 질서 정립을 구체화한 '파리 이니셔티브(Paris Initiative)' 선언, 디지털 규범 제정을 위한 국제기구 설치 제안 (인공지능(AI)에 국한하지 않고 데이터와 컴퓨터 역량, 디지털의 기초부터 심화까지 모든 영역을 포함, 디지털 어느 단계에 있든 모든 국가에 적용할 수 있는 포괄성을 지닌다는 점을 강조)
디지털 권리장전[100] ('23.9.25.)	- 과학기술정보통신부가 발표한 보편적 디지털 질서 규명을 위한 헌장, 세계 시민이 함께 지향해야 할 디지털 공동번영사회의 가치와 5대 원칙* 제시 (*▲디지털 환경에서의 자유와 권리 보장, ▲디지털에 대한 공정한 접근과 기회의 균등, ▲안전하고 신뢰할 수 있는 디지털 사회, ▲자율과 창의 기반의 디지털 혁신의 촉진, ▲인류 후생의 증진)

AI 서울 정상회의[101] ('24.5.21.)	- '23.11월 "AI 안전성 정상회의"(영국 블레츨리 파크)에 이어 개최, AI 안전성 강화 및 혁신을 촉진하고 포용과 상생을 도모하는 AI 발전방안을 포괄적으로 논의 (▲정상급: 11개국(대한민국, 영국, 미국, EU, 호주, 캐나다, 프랑스, 독일, 이탈리아, 일본, 싱가포르) 참여, ▲장관급: 28개국 참여)
AI 기본법 제정[102] ('25.1.21.)	- '25.1.21. 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 (약칭: 인공지능기본법) 제정(시행 '26.1.22.) - (목적) 인공지능의 건전한 발전과 신뢰 기반 조성에 필요한 기본적인 사항을 규정함으로써 국민의 권익과 존엄성을 보호하고 국민의 삶의 질 향상과 국가경쟁력을 강화하는 데 이바지

(ii) 대응 방향 제언

❶ 기업 전략 방향 제언

○ 글로벌 AI 규제 환경에 적절히 대응하면서 국내외 시장에서 경쟁력을 유지하기 위해 AI 기업들은 내부적으로 AI 모델의 신뢰성과 투명성을 확보하기 위한 운영 체계를 정립하고, 대외적으로 국제 규제 준수를 위한 대응 전략을 마련할 필요가 있다.

❶-1. 내부 AI 개발·운영 체계 정립

○ **(AI 신뢰성 및 투명성 확보)** AI 모델 개발·운영 과정에 설명 가능성을 높이고, 투명성과 신뢰성을 확보하기 위해 정보 공개 체계 마련

- 글로벌 AI 기업에서 시행 중인 AI 모델의 시스템 카드·모델 카드·기술 보고서 공개 사례를 참고하여 공식적인 정보 공개 문서화 체계 구축

---

98) 대한민국 대통령실 과학기술비서관실. (2023, September 21). 尹 대통령, 뉴욕구상 1주년을 맞아 새로운 디지털 질서에 대한 원칙을 제시하다. Retrieved from https://www.president.go.kr/newsroom/press/RxNd3Vhr

99) 정아란, 안용수. (2023, June 21). 尹, 세계 '디지털 질서' 정립 제안...'파리 이니셔티브' 공개. 연합뉴스. Retrieved from https://www.yna.co.kr/view/AKR20230621153600001; 박수형. (2023, June 21). 尹, '파리 이니셔티브' 선언...디지털 규범 국제기구 설치 제안. ZDNet Korea. Retrieved from https://zdnet.co.kr/view/?no=20230621181937

100) 과학기술정보통신부. (2023, September 25). 대한민국이 새로운 디지털 규범질서를 전 세계에 제시합니다! 보도자료. Retrieved from https://www.msit.go.kr/bbs/view.do?sCode=user&mPid=238&mId=113&bbsSeqNo=94&nttSeqNo=3183520

101) AI 서울 정상회의. (2024). AI 서울 정상회의 공식 웹사이트. Retrieved from https://aiseoulsummit.kr/kor/; 연합뉴스. (2024, May 22). AI 서울 정상회의 참가국 장관들 '서울 성명'...28개국 참여(종합). Retrieved from https://www.yna.co.kr/view/AKR20240522129000017

102) 법제처. (2025). 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법. Retrieved from https://www.law.go.kr/법령/인공지능기본법

ㅇ **(내부 데이터 거버넌스 체계 정비)** AI 학습 데이터의 출처 명확화 등 데이터 관리 체계를 강화하고, 데이터 보호 및 윤리적 AI 활용 기준을 정비한 내부 가이드라인을 마련하여 지속적으로 개선

❶-2. 대외 규제 환경 대응

ㅇ **(글로벌 AI 규제 준수 적합성 확보)** 미국 및 EU의 AI 윤리지침과 규제 요건을 반영하여 AI 서비스 운영 방식 조정

- 각국의 AI 규제 요건을 반영하여 AI 운영 프로세스 및 정책을 조정하고, 내부 관리 체계 정비
- 글로벌 테크 기업들의 규제 대응 전략을 벤치마킹하여 법적 리스크를 최소화하고, AI 기반 서비스의 지역별 규제 적합성 확보

ㅇ **(지역별 AI 및 데이터 보호 규제 대응)** 각국의 데이터 주권 강화 동향에 따른 AI 및 데이터 규제 변화에 대응하고, 글로벌 시장 경쟁력 확보를 위한 전략 수립

- 미국·중국 등 주요 국가의 데이터 보호 규제 준수를 위해 데이터 로컬라이제이션(국가별로 현지에 데이터를 저장 및 처리) 전략 도입
– EU 경쟁법 및 디지털시장법(DMA) 등 AI 관련 규제 요건을 반영하여 공정 경쟁 원칙을 고려한 대응 방안 수립

❷ 정부 정책 방향 제언

ㅇ AI 기술 발전과 규제 환경 변화에 대응하기 위해 정부는 국내적으로 AI 기술 발전을 지원하면서 글로벌 AI 규제 환경과 조화를 이루는 방향으로 규제를 정비하고, 대외적으로 AI 및 데이터 규제와 관련한 국제 논의에서 적극적인 역할을 수행하여 정부 입장을 명확히 설정하고 협상력을 강화해 나가야 할 것이다.

❷-1. 국내 정책 방향

ㅇ 국내 정책은 AI 산업의 지속적인 성장과 기술혁신을 지원하면서, 글로벌 규제 환경과 조화를 이루는 방향으로 발전해야 한다.

이를 위해 국내 실정과 글로벌 AI 및 데이터 규제 환경을 조율하여 AI 윤리원칙을 수립하고, 규제 강화 보다는 AI 산업 및 기술 발전을 지원하는 방향으로 정책 균형을 유지해야 할 것이다.

ㅇ **(AI 윤리원칙 및 규제 체계 구축)** ▲AI 신뢰성 평가 기준 마련, ▲인공지능(AI) 기본법에 기반하여 AI 안정성에 대한 기술적 검증 및 윤리적 가이드라인 마련, ▲개인정보 보호 강화를 위해 AI 모델 학습 데이터 대상 윤리 가이드라인 및 데이터 규제 정비

ㅇ **(AI 산업 경쟁력 강화 지원)** ▲AI 스타트업 및 중소기업의 글로벌 시장 경쟁력 강화를 위한 AI 연구개발 및 기업 지원 확대, ▲ AI 기술 인재 확보를 위한 교육 및 연구 지원 강화, ▲AI 규제 샌드박스 운영을 통한 신기술 실험 대상 규제 유예 및 AI 기반 신산업 활성화 지원

ㅇ **(산업별 맞춤형 AI 규제 적용)** ▲미국의 산업별 규제 접근(Sectorally Specific Approach) 사례를 참고하여 산업 특성을 고려한 규제 적용, ▲특정 AI 플랫폼 기업의 시장 독점을 방지하고 공정 경쟁 원칙을 강화

❷-2. 국제 협상에서 한국의 역할 강화

ㅇ 글로벌 AI 규제 환경은 AI 윤리 및 기술 규제만이 아니라 데이터 보호, 디지털 시장 규제, 국가 안보 전략과도 긴밀하게 연계되어 있다. 정부는 이러한 글로벌 규제 환경과 조화를 이루면서 국내 AI 산업을 지원하는 균형 잡힌 정책을 추진하고, 이를 통해 AI 및 데이터 산업의 글로벌 경쟁력을 강화해 나가야 할 것이다. 또한, 국제 협상에서 주도적인 역할을 수행하여, 글로벌 규범 형성에 기여하면서, AI 글로벌 리더십을 확보해 나가는 전략적 대응이 필요하다.

ㅇ **(글로벌 AI 및 데이터 규제 협상에서 협상력 강화)** ▲OECD, UN, EU, 미국 등 주요 국가 및 국제기구와 협력하여 AI 규제 및 표준화 논의에 적극 참여하고, 정부 입장 반영을 위한 협상 전략 수립, ▲AI 서울 정상

회의 등 글로벌 AI 거버넌스 구축에 주도적인 역할을 수행하여 AI 기술 발전과 규제를 균형 있게 조율

ㅇ **(글로벌 표준화 논의 주도)** ▲ISO(국제표준화기구), IEC(국제전기기술위원회), IEEE(국제전기전자기술자협회) 등 국제 기술 표준화 기구와 협력하여 AI 기술 및 데이터 보호 표준화 논의에 적극 참여, ▲국내 AI 기본법과 글로벌 규제의 정합성을 고려하여 법·제도 정비

ㅇ **(AI 규제 환경 변화에 대응한 외교적 전략 수립)** ▲미국의 자유로운 AI 개발 정책과 EU·중국의 강력한 규제 대립 사이에서 중립적이고 협상력을 갖춘 정책 추진 필요, ▲국내 AI 기업이 글로벌 AI 규제 환경을 준수하면서 해외 시장에 진출할 수 있도록 해외 진출 지원 정책 마련.

ㅇ AI 기술 발전과 글로벌 규제 환경 변화는 기업과 정부가 긴밀하게 협력하여 대응해야 할 과제이다. 기업은 AI 신뢰성·투명성을 확보하고, 데이터 보호 전략을 정비하면서, 사업 진출 시장의 규제를 준수해야 한다. 그리고, 정부는 AI 기술 발전을 지원하면서 동시에 글로벌 규제와 균형을 유지하며 산업 경쟁력을 강화하는 방향으로 정책을 추진해야 한다.

ㅇ 특히, AI 및 데이터 규제는 기술 발전, 시장 경쟁, 국가 안보 등 다양한 요소가 복합적으로 얽혀 있어, 단기적인 규제 대응을 넘어 장기적인 시각에서 전략을 수립하고, 국제 협력 지속적으로 강화할 필요가 있다.

앞으로 기업은 글로벌 규제 동향을 모니터링하면서 법적 리스크를 최소화하는 전략을 마련해야 할 것이고, 정부는 국제 협상에 적극적으로 참여하여 글로벌 AI 거버넌스 체계 구축 과정에 확고한 기반을 다질 수 있도록 정책을 추진해야 할 것이다.

기업과 정부가 이러한 노력을 조화롭게 이어간다면, AI 기술 혁신과 규제 환경 변화 속에서 경쟁력을 유지하며, 글로벌 AI 시장에서 주도적인 역할을 수행할 수 있을 것이다.

참고문헌

- AI 서울 정상회의. (2024). AI 서울 정상회의 공식 웹사이트. Retrieved March 12, 2025, from https://aiseoulsummit.kr/kor/
- KDI 경제정보센터. (2024, June). 세계 최초로 통과된 EU「AI법」, 우리 기업의 대응 방향은?, 작성자: 정재욱 (주벨기에EU대사관 겸 주NATO대표부 과학관). Retrieved January 20, 2025, from https://eiec.kdi.re.kr/publish/naraView.do?fcode=000020000400001000 10&cidx=14782&sel_year=20
- LG AI 윤리 책무성 보고서. (2023). Retrieved January 18, 2025, from https://www.lgresearch.ai/about/mission
- YTN. (2021, November 26). 유네스코, 세계 첫 'AI 윤리적 사용 지침' 마련. Retrieved March 15, 2025, from https://www.ytn.co.kr/_ln/0104_202111261132522639.
- 경제협력개발기구(OECD). (2023, November 8). Explanatory Memorandum on the Updated OECD Definition of an AI System. Retrieved July 15, 2024, from https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html
- 경제협력개발기구(OECD). (2023, November 8). OECD AI Principles Revised. Retrieved July 15, 2024, from https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html
- 고준성. (2024, July). 국제 차원에서의 인공지능(AI) 거버넌스 논의 동향과 평가 및 시사점. 한국외교협회. Retrieved March 15, 2025, from https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11849732.
- 과학기술정보통신부. (2023, September 25). 대한민국이 새로운 디지털 규범질서를 전 세계에 제시합니다! 보도자료. Retrieved March 12, 2025, from https://www.msit.go.kr/bbs/view.do?sCode=user&mPid=238&mId=113&bbsSeqNo=94&nttSeqNo=3183520
- 김청중. (2025, February 2). 중국, '국가정보법'으로 데이터 수집 우려. 세계일보. Retrieved March 10, 2025, from https://www.segye.com/newsView/20250202508426
- 대외경제정책연구원(KIEP). (2018, December 31). 국제 통상 환경 변화와 한국의 대응 전략. Retrieved January 29, 2025, from https://www.kiep.go.kr/gallery.es?mid=a10101040000&bid=0001&list_no=2328&act=view

- 대한무역투자진흥공사(KOTRA). (2024). 제5차 미국-EU 무역기술위원회(TTC) 주요 내용 및 현지 반응 (US24-4). Retrieved February 14, 2025, from https://kocham.org/wp-content/uploads/2024/02/KOTRA-%EC%A0%9C5%EC%B0%A8-%EB%AF%B8%EA%B5%AD-EU-%EB%AC%B4%EC%97%AD%EA%B8%B0%EC%88%A0%EC%9C%84%EC%9B%90%ED%9A%8CTTC-%EC%A3%BC%EC%9A%94-%EB%82%B4%EC%9A%A9-%EB%B0%8F-%ED%98%84%EC%A7%80-%EB%B0%98%EC%9D%91US24-4.pdf
- 대한민국 대통령실 과학기술비서관실. (2023, September 21). 尹 대통령, 뉴욕구상 1주년을 맞아 새로운 디지털 질서에 대한 원칙을 제시하다. Retrieved March 12, 2025, from https://www.president.go.kr/newsroom/press/RxNd3Vhr
- 박수형. (2023, June 21). 尹, '파리 이니셔티브' 선언...디지털 규범 국제기구 설치 제안. ZDNet Korea. Retrieved March 12, 2025, from https://zdnet.co.kr/view/?no=20230621181937
- 박효재. (2022, March 25). EU, 미 IT공룡 독점 금지 '디지털 시장법' 합의. 경향신문. Retrieved March 10, 2025, from https://www.khan.co.kr/article/202203251053001
- 백악관 과학기술정책실(OSTP). (2022, October 4). Applying the Blueprint for an AI Bill of Rights. Retrieved July 24, 2024, from https://www.whitehouse.gov/ostp/ai-bill-of-rights/applying-the-blueprint-for-an-ai-bill-of-rights/
- 백악관 과학기술정책실(OSTP). (2022, October 4). Blueprint for an AI Bill of Rights. Retrieved July 24, 2024, from https://www.whitehouse.gov/ostp/ai-bill-of-rights/
- 백악관 과학기술정책실(OSTP). Nelson, A., Friedler, S., & Fields-Meyer, A. (2022, October 4). Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age. Retrieved July 24, 2024, from https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/
- 법률신문. (2024, March 4). 챗GPT와 AI 기술 발전, 데이터 보호법의 새로운 도전. Retrieved January 18, 2025, from https://www.lawtimes.co.kr/news/196474
- 법제처 세계법제정보센터. (n.d.). 중국 데이터 안전법 및 주요 내용. Retrieved January 18, 2025, from https://world.moleg.go.kr/web/wli/lgslInfoReadPage.do?CTS_SEQ=49497&AST_SEQ=53&
- 법제처. (2025). 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법. Retrieved March 13, 2025, from https://www.law.go.kr/법령/인공지능기본법

- 세계무역기구(WTO). (1994). 무역관련 지식재산권에 관한 협정, Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS). Retrieved July 17, 2024, from https://www.wto.org/english/docs_e/legal_e/trips_e.htm#art39
- 세계무역기구(WTO). (1995). 서비스무역에 관한 일반협정, General Agreement on Trade in Services (GATS). Retrieved July 11, 2024, from https://www.wto.org/english/docs_e/legal_e/26-gats_01_e.htm
- 신화통신 한국어 뉴스 서비스. (2024, May 1). 중 차량 보안 테스트 통과 ICV 모델 발표...BYD·테슬라 등 76종. 신화망 한국어판. Retrieved March 11, 2025, from https://kr.news.cn/20240501/7fb69de887b44ac187549fa2473e304f/c.html
- 심두보. (2024, April 30). '중국 데이터 안전 검사 통과' 급등한 테슬라 주가. 딜사이트. Retrieved March 11, 2025, from https://dealsite.co.kr/articles/121887
- 연합뉴스. (2024, May 22). AI 서울 정상회의 참가국 장관들 '서울 성명'...28개국 참여(종합). Retrieved March 12, 2025, from https://www.yna.co.kr/view/AKR20240522129000017
- 유럽연합 관보. (2024, July 12). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Retrieved July 22, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ
- 유럽연합 집행위원회(European Commission). (2024). Regulatory framework proposal on artificial intelligence. Retrieved June 25, 2024, from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
- 정아란, 안용수. (2023, June 21). 尹, 세계 '디지털 질서' 정립 제안...'파리 이니셔티브' 공개. 연합뉴스. Retrieved March 12, 2025, from https://www.yna.co.kr/view/AKR20230621153600001
- 정혜진. (2024, May 8). "강제매각법은 표현의 자유 침해"...틱톡, 美 정부와 소송전. 서울경제. Retrieved March 10, 2025 from https://www.sedaily.com/NewsView/2D93YFDUK3
- 중화인민공화국 네트워크 안전법. (2024, February 2). 중화인민공화국 네트워크 안전법. 세계법제정보센터. Retrieved March 11, 2025 from https://world.moleg.go.kr/web/wli/lgslInfoReadPage.do?CTS_SEQ=41114&AST_SEQ=53
- 중화인민공화국 데이터 안전법. (2024, February 2). 중화인민공화국 데이터 안전법. 세계법제정보센터. Retrieved March 11, 2025 from https://world.moleg.go.kr/web/wli/lgslInfoReadPage.do?CTS_SEQ=49497&AST_SEQ=53

- 한국경제. (2024, March 4). AI 기술, 산업 지형을 바꾼다: 챗GPT의 영향과 전망. Retrieved January 18, 2025, from https://www.hankyung.com/article/202403046536i

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October 24). Deep learning with differential privacy. arXiv. Retrieved February 26, 2025, from https://arxiv.org/abs/1607.00133
- Ada Lovelace Institute. (2022, April 8). EU AI Act Explainer. Retrieved January 25, 2025, from https://www.adalovelaceinstitute.org/resource/eu-ai-act-explainer/
- Allen, B. E., Leiter, M. E., Werry, S., & Bell, J. F. (2025, February 28). Trump revives and expands the battle over digital services taxes. Skadden Publication / Executive Briefing: Latest Updates on the Trump Administration. Retrieved March 11, 2025 from https://www.skadden.com/insights/publications/2025/02/trump-revives-and-expands-the-battle-over-digital-services-taxes
- Alvarez, S. (2024, April 29). Tesla China partners with Baidu for maps to clear FSD hurdle. Teslarati. Retrieved March 11, 2025, from https://www.teslarati.com/tesla-china-partners-baidu-maps-clear-fsd-hurdle/
- Alvarez, S. (2024, November 8). Tesla China earns "Automotive Privacy" certification in China. Teslarati. Retrieved March 11, 2025, from https://www.teslarati.com/tesla-china-first-company-automotive-privacy-certification/
- Antonio, V. (2025, January 3). What is the difference between OpenAI, Google DeepMind, Anthropic, and Microsoft AI? Foreign Affairs Forum. Retrieved February 23, 2025, from https://www.faf.ae/home/2025/1/3/what-is-difference-between-open-ai-and-google-deepmind-anthropic-and-microsoft-ai
- Arnold & Porter. (2023, July). European Parliament Adopts Its Version of AI Act, Commencing Negotiations with Council and Commission. Retrieved July 15, 2024, from https://www.arnoldporter.com/en/perspectives/advisories/2023/07/european-parliament-adopts-its-version-of-ai-act
- Bachman, J. (2025, January 17). Supreme Court upholds U.S. law forcing TikTok ban or sale. Retrieved March 10, 2025, from https://www.legaldive.com/news/supreme-court-upholds-us-law-forcing-tiktok-ban-or-sale/737704/

- Baker Botts. (2024, April 8). The EU AI Act: Uncharted Territory for General Purpose AI. Retrieved January 29, 2025, from https://www.bakerbotts.com/thought-leadership/publications/2024/april/the-eu-ai-act-uncharted-territory-for-general-purpose-ai.
- Basso-Davis, F. (2024, February 11). TikTok ban timeline and updates. Retrieved March 10, 2025, from https://wwcecho.news/news/tiktok-ban-timeline-and-updates
- Bertomeu, J., Lin, Y., Liu, Y., & Ni, Z. (2024, November 27). Voluntary disclosure, misinformation, and AI information processing: Theory and evidence. SSRN. Retrieved February 26, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5026360
- Bird & Bird. (2024, November 6). European Union Artificial Intelligence Act: A Guide. Retrieved January 25, 2025, from https://www.twobirds.com/-/media/new-website-content/pdfs/capabilities/artificial-intelligence/european-union-artificial-intelligence-act-guide.pdf
- Blank Rome LLP. (2023, July 2). Data Matters: Risks and Best Practices for the Use of Generative AI. Retrieved January 27, 2025, from https://www.blankrome.com/publications/data-matters-risks-and-best-practices-use-generative-ai
- Bloomberg Law. (2024, July 31). European AI Act Training Disclosures Expose US Copyright Risks. Retrieved January 29, 2025, from https://news.bloomberglaw.com/ip-law/european-ai-act-training-disclosures-expose-us-copyright-risks.
- Borak, M. (2021, April 21). Tesla promises new China data centre as Beijing lays down the law on local storage of EV data. South China Morning Post. Retrieved March 11, 2025, from https://www.scmp.com/tech/big-tech/article/3130339/tesla-promises-new-china-data-centre-beijing-lays-down-law-local
- Borghetti, J. (2023). Balancing Innovation and Copyrights: The Legal Framework for AI Training in the EU. SSRN Electronic Journal. Retrieved January 25, 2025, from https://arno.uvt.nl/show.cgi?fid=176939
- Borghetti, J. (2023). Infringing AI: Liability for AI-generated Outputs Under International, EU, and UK Copyright Law. European Journal of Risk Regulation. Retrieved January 25, 2025, from https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/infringing-ai-liability-for-aigenerated-outputs-under-international-eu-and-uk-

copyright-law/C568C6B717E9CFC45FB52E58E54B6BEC

- Borghetti, J. (2023). Prohibited Artificial Intelligence Practices in the Proposed EU Artificial Intelligence Act. European Journal of Risk Regulation. Retrieved January 25, 2025, from https://www.sciencedirect.com/science/article/pii/S0267364923000092
- Bostoen, F. (2019, April 24). Spotify vs. Apple: The Core of the Dispute. Retrieved March 10, 2025 from https://www.lexxion.eu/en/coreblogpost/spotify-apple/
- British Council and Goethe-Institut. (2018). Cultural Relations in the Age of Artificial Intelligence. Retrieved January 25, 2025, from https://www.britishcouncil.org/research
- Buchicchio, E., De Angelis, A., San Marco, L., Moschitta, A., Carbone, P., & Santoni, F. (2024). Design, Validation, and Risk Assessment of LLM-Based Generative AI Systems Operating in the Legal Sector. IEEE International Symposium on Systems Engineering.
- Casson, A. (2023, May 16). Transformer FLOPs. Retrieved from https://www.adamcasson.com/posts/transformer-flops.
- Casson, A. (2023, May 16). Transformer FLOPs. Retrieved January 30, 2025, from https://www.adamcasson.com/posts/transformer-flops.
- Center for Art Law. (2024, May 21). Generative AI and Transparency of Databases and Their Content: From a Copyright Perspective. Retrieved January 29, 2025, from https://itsartlaw.org/2024/05/21/generative-ai-and-transparency-of-databases-and-their-content-from-a-copyright-perspective/.
- Chainoglou, K., & Katsios, S. (2024). Threats to Cultural Heritage: Normative Developments on AI and Cultural Heritage. In Digital Humanities and Cultural Heritage. Routledge. Retrieved January 25, 2025, from https://www.taylorfrancis.com/chapters/edit/10.4324/9781003260127-10/threats-cultural-heritage-kalliopi-chainoglou-stavros-katsios
- Cheng, E. (2024, April 29). Musk makes surprise visit to Beijing as Tesla's China-made cars pass data security rules. CNBC. Retrieved March 11, 2025, from https://www.cnbc.com/2024/04/29/musk-visits-beijing-as-teslas-china-made-cars-pass-security-rules.html
- Cheng, E. (2024, April 29). Musk visits Beijing as Tesla's China-made cars pass security rules. CNBC. Retrieved March 11, 2025 from

https://www.cnbc.com/2024/04/29/musk-visits-beijing-as-teslas-china-made-cars-pass-security-rules.html

- Christoffel, R. (2025, March 10). Report: Apple Will Be Fined by EU for Alleged Violation of DMA. Retrieved March 10, 2025, from https://9to5mac.com/2025/03/10/report-apple-will-be-fined-by-eu-for-alleged-violation-of-dma/
- Clover, J. (2025, March 10). Apple Receives 'Modest' Fine Under EU Digital Markets Act. Retrieved March 10, 2025, from https://www.macrumors.com/2025/03/10/apple-modest-fine-eu-digital-markets-act/
- Congressional Research Service. (2024, April 3). Highlights of the 2023 executive order on artificial intelligence for Congress. Retrieved February 12, 2025, from https://crsreports.congress.gov/product/pdf/R/R47843
- Cooban, A. (2024, June 24). Apple and EU Competition Rules. Retrieved March 10, 2025, from https://www.cnn.com/2024/06/24/business/apple-eu-competition-rules/index.html
- Credo AI. (2023, December). EU AI Act Political Agreement: What You Need to Know for 2024. Retrieved July 15, 2024, from https://www.credo.ai/blog/eu-ai-act-political-agreement-what-you-need-to-know-for-2024
- DLA Piper. (2024, February). Extra-territorial Application of the AI Act: How Will It Impact Australian Organisations? Retrieved January 24, 2025, from https://www.dlapiper.com/en/insights/publications/2024/02/extra-territorial-application-of-the-ai-act-how-will-it-impact-australian-organisations
- Ebers, M., & Navas, S. (2024). The Fundamental Rights Impact Assessment (FRIA) in the AI Act. Computer Law & Security Review. Retrieved January 25, 2025, from https://www.sciencedirect.com/science/article/pii/S0267364924000864
- Economic Times. (2024, April 29). Musk's big win in China: Tesla clears data security, full self-driving hurdles for locally-made cars. The Economic Times. Retrieved March 11, 2025 from https://economictimes.indiatimes.com/industry/renewables/musks-big-win-in-china-tesla-clears-data-security-full-self-driving-hurdles-for-locally-made-cars/articleshow/109678866.cms
- ENCATC. (2023). Artificial Intelligence Embraced: the future of the cultural and creative sector. In Book of Proceedings, 2023. Retrieved January 25, 2025,

from https://encatc.org/media/7193-2023-book-of-proceedings.pdf#page=59

- Engine. (2024, October 24). AI Essentials: What Is Compute and How Is It Measured? Retrieved January 30, 2025, from https://www.engine.is/news/category/ai-essentials-what-is-compute-and-how-is-it-measured.
- Engler, A. (2023, April 25). The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Brookings Institution. Retrieved February 14, 2025, from https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/
- Epoch AI. (2021, November 29). Measuring FLOPs Empirically. Retrieved January 30, 2025, from https://epoch.ai/blog/measure-FLOP-empirically.
- Epoch AI. (2024, April 5). Tracking Large-Scale AI Models. Retrieved January 30, 2025, from https://epoch.ai/blog/tracking-large-scale-ai-models.
- European Commission. (2021). Joint EU-US statement on the Trade and Technology Council (TTC). Retrieved February 14, 2025, from https://ec.europa.eu/commission/presscorner/api/files/document/print/en/statement_21_4951/STATEMENT_21_4951_EN.pdf
- European Commission. (2021, September 29). EU-US Trade and Technology Council Inaugural Joint Statement. Retrieved February 14, 2025, from https://ec.europa.eu/commission/presscorner/detail/el/STATEMENT_21_4951
- European Commission. (2022, December 5). Joint Statement following the third EU-U.S. Trade and Technology Council. Retrieved February 14, 2025, from https://ec.europa.eu/commission/presscorner/detail/en/statement_22_7516
- European Commission. (2023, June 15). Joint Statement following the fourth EU-U.S. Trade and Technology Council. Retrieved February 14, 2025, from https://ec.europa.eu/commission/presscorner/detail/en/statement_23_2992
- European Commission. (2024, April 5). Fifth meeting of the EU-U.S. Trade and Technology Council. Retrieved February 14, 2025, from https://ec.europa.eu/commission/presscorner/detail/en/ip_24_575
- European Commission. (2024, April 5). Joint Statement EU-US Trade and Technology Council. Retrieved February 14, 2025, from https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_24_1828
- European Commission. (n.d.). EU-US Trade and Technology Council. European Commission. Retrieved February 13, 2025, from

https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/stronger-europe-world/eu-us-trade-and-technology-council_en

- European Parliament. (2022, November 30). Digital Markets Act: What's Next? Retrieved March 10, 2025 from https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2022)739226
- Fernandez. (2024, May 14). The Luxembourg Model of AI Governance: Toward Human-Centric AI Regulation. University of Luxembourg. Retrieved January 27, 2025, from https://orbilu.uni.lu/handle/10993/62607
- Fox, E. (2021, April 20). Tesla to build new data center in China in response to law on local storage of EV data. Tesmanian. Retrieved March 11, 2025, from https://www.tesmanian.com/blogs/tesmanian-blog/tesla-will-build-a-new-data-center-in-china-due-to-the-new-law-on-the-local-storage-of-ev-data
- Fox, E. (2021, October 25). Tesla launches Shanghai R&D center & Gigafactory data center in China. Tesmanian. Retrieved March 11, 2025, from https://www.tesmanian.com/blogs/tesmanian-blog/tesla-launches-shanghai-r-d-center-and-gigafactory-data-center-in-china
- Future of Life Institute. (2024). Artificial Intelligence Act; Timeline of Developments. Retrieved June 26, 2024, from https://artificialintelligenceact.eu/developments/
- Future of Privacy Forum (FPF). (2022, November 3). GDPR and the AI Act Interplay: Lessons from FPF's ADM Case Law Report. Retrieved January 25, 2025, from https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report/
- Future of Privacy Forum (FPF). (2023, November). Conformity Assessments under the EU AI Act. Retrieved January 25, 2025, from https://fpf.org/wp-content/uploads/2023/11/OT-FPF-comformity-assessments-ebook_update2.pdf
- Geneva Internet Platform. (2023, October 30). Google DeepMind, Meta, OpenAI among top AI firms sharing safety policies ahead of UK summit. Geneva Internet Platform. Retrieved February 23, 2025, from https://dig.watch/updates/google-deepmind-meta-openai-among-top-ai-firms-sharing-safety-policies-ahead-of-uk-summit
- Geopolitechs. (2024, April 28). China lifted ban on Tesla vehicles during Musk's China visit. Retrieved March 11, 2025, from

https://www.geopolitechs.org/p/china-lifted-ban-on-tesla-vehicles

- Gibson Dunn. (2023, December 9). EU Agrees on a Path Forward for the AI Act. Retrieved July 15, 2024, from https://www.gibsondunn.com/eu-agrees-on-a-path-forward-for-the-ai-act/#_ftn3
- Google AI. (n.d.). About Model Cards. Retrieved January 21, 2025, from https://modelcards.withgoogle.com/about
- Google DeepMind. (2024). Gemini: A family of highly capable multimodal models. Retrieved March 4, 2025, from https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf
- Greenstein, B., Kosar, J., Light, C., & Shelton Rose, M. (2024, June 27). Productivity or pioneering? Your industry's GenAI adoption play. strategy+business. Retrieved March 15, 2025, from https://www.pwc.com/gx/en/issues/technology/genai-adoption-game-changer.html.
- Guadamuz. (2024, November 10). AI Regulation in Global Perspective: Challenges for Intellectual Property Law. The Journal of World Intellectual Property, Wiley Online Library. Retrieved January 27, 2025, from https://onlinelibrary.wiley.com/doi/full/10.1111/jwip.12330
- Gubina, E. (2024, April 29). Tesla cars have been tested for compliance with data security requirements in China. Mezha.Media. Retrieved March 11, 2025, from https://mezha.media/en/2024/04/29/tesla-cars-have-been-tested-for-compliance-with-data-security-requirements-in-china/
- Haie, A.-G., Golodny, A., Shenk, M., Avramidou, M., Goodwin, E., & Arethusa, V. (2024, April 30). A comparative analysis of the EU, US, and UK approaches to AI regulation. StepTechToe. Retrieved February 13, 2025, from https://www.steptoe.com/en/news-publications/steptechtoe-blog/a-comparative-analysis-of-the-eu-us-and-uk-approaches-to-ai-regulation.html
- Heim, L., Fist, T., Egan, J., Huang, S., & Zekany, S. (2024). Governing through the cloud: The intermediary role of compute providers in AI regulation. arXiv. Retrieved February 11, 2025, from https://arxiv.org/pdf/2403.08501
- Holland & Knight. (2025, January 17). U.S. Supreme Court upholds TikTok sale or ban law. Retrieved March 10, 2025, from https://www.hklaw.com/en/insights/publications/2025/01/us-supreme-court-upholds-tiktok-sale-or-ban-law
- IAPP. (2024, October). Top Impacts of the EU AI Act: Post-Market Monitoring,

Sharing, and Enforcement. Retrieved January 25, 2025, from https://iapp.org/resources/article/top-impacts-eu-ai-act-post-market-monitoring-sharing-enforcement/

- IBM Research. (2024, May 21). IBM Granite and the Foundation Model Transparency Index: Advancing transparency in AI. IBM Research Blog. Retrieved February 20, 2025, from https://research.ibm.com/blog/ibm-granite-transparency-fmti
- IBM. (n.d.). IBM InfoSphere Optim Data Privacy. Retrieved February 26, 2025, from https://www.ibm.com/products/infosphere-optim-data-privacy
- iKanagai. (2023, July 14). The Secrets of GPT-4 Leaked. Retrieved January 30, 2025, from https://www.ikangai.com/the-secrets-of-gpt-4-leaked/.
- Internet Policy Review. (2024, July 16). Regulating AI through Technical Standards. Retrieved January 25, 2025, from https://policyreview.info/articles/analysis/regulating-ai-through-technical-standards
- Ip, Jeffrey. (2025, March 5). Benchmarks: MMLU. DeepEval Cloud. Retrieved March 7, 2025, from https://docs.confident-ai.com/docs/benchmarks-mmlu
- Jackson, K. (2024, April 18). Stanford HAI AI Index report highlights responsible AI challenges. AI Wire. Retrieved February 20, 2025, from https://www.aiwire.net/2024/04/18/stanford-hai-ai-index-report-responsible-ai/
- Japan Business Council in Europe (JBCE). (2021, August 6). JBCE Views on the European Commission's AI Act. Retrieved February 4, 2025, from https://www.jbce.org/images/JBCE-comments-on-the-AI-Act-FINAL.pdf.
- Kerr, I. (2024). AI Music Outputs and Copyright Challenges. SSRN Electronic Journal. Retrieved January 25, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4072806
- Kiely, E. (2025, February 3). TikTok and U.S. national security. Retrieved from https://www.factcheck.org/2025/02/tiktok-and-u-s-national-security/
- Kilpala, M., & Kärkkäinen, T. (2024, April 17). Artificial intelligence and differential privacy: Review of protection estimate models. In A. Khezerlou, R. Schmidt, & K. Hu (Eds.), Advances in artificial intelligence and security: Volume 1 (pp. XX-XX). Springer. Retrieved February 26, 2025, from https://link.springer.com/chapter/10.1007/978-3-031-57452-8_3
- Kirkwood, M. (2025, March 6). Digital Markets Act roundup: February 2025. TechPolicy.Press. Retrieved March 11, 2025 from

https://www.techpolicy.press/digital-markets-act-roundup-february-2025/

- Klu.ai. (n.d.). TruthfulQA eval. Retrieved March 7, 2025, from https://klu.ai/glossary/truthfulqa-eval
- Kulesz, O. (2024). Artificial Intelligence and International Cultural Relations: Challenges and Opportunities for Cross-Sectoral Collaboration. ifa – Institut für Auslandsbeziehungen. Retrieved January 25, 2025, from https://www.ifa.de/en/publications
- Lakera AI. (2024, March 19). Risks of AI. Retrieved January 27, 2025, from https://www.lakera.ai/blog/risks-of-ai
- Lal, R., & Smith, T. (2024). Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content. International Review of Law, Computers & Technology. Retrieved January 25, 2025, from https://www.tandfonline.com/doi/pdf/10.1080/13600869.2024.2324540
- Latham & Watkins LLP. (2024, July). EU AI Act: Navigating a Brave New World. Retrieved January 25, 2025, from https://www.lw.com/admin/upload/SiteAttachments/EU-AI-Act-Navigating-a-Brave-New-World.pdf
- Lin, B., Hilton, J., & Evans, R. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3214–3235. Retrieved March 7, 2025, from https://aclanthology.org/2022.acl-long.234/
- Lumenova AI. (2024, June 20). AI policy analysis: European Union vs. United States. Retrieved February 14, 2025, from https://www.lumenova.ai/blog/ai-policy-eu-vs-us-comparison/?utm_source=chatgpt.com
- Lutkevich, B. (2025, February 18). TikTok bans explained: Everything you need to know. Retrieved March 10, 2025, from https://www.techtarget.com/whatis/feature/TikTok-bans-explained-Everything-you-need-to-know
- Makraki, I. (2024). Deepfakes: Legal Challenges and Protection Under the AI Act. University of Macedonia. Retrieved January 25, 2025, from https://dspace.lib.uom.gr/handle/2159/30910
- Maksimovic, A. (2024). NEW EU REGULATION OF AUTONOMOUS ARTIFICIAL INTELLIGENCE. University of Ljubljana. Retrieved January 25, 2025, from http://www.cek.ef.uni-lj.si/magister/maksimovic5269-B.pdf

- Margoni. (2024, May 30). Regulating AI through Global Frameworks: Challenges and Opportunities. SSRN. Retrieved January 27, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5036164
- McCallum, S. (2024, March 4). Apple Faces EU Scrutiny Under Digital Markets Act. Retrieved March 10, 2025, from https://www.bbc.com/news/technology-68467752
- Mehra, A. (2025, January 21). Executive Orders on AI: How to (Lawfully) Apply the Defense Production Act. Mercatus Center. Retrieved February 11, 2025, from https://www.mercatus.org/research/policy-briefs/executive-orders-ai-how-lawfully-apply-defense-production-act
- Meta. (2024). LLaMA model card. GitHub. Retrieved March 7, 2025, from https://github.com/meta-llama/llama/blob/main/MODEL_CARD.md
- Microsoft. (2025, January 15). Innovating in Line with the European Union's AI Act. Retrieved January 21, 2025, from https://blogs.microsoft.com/on-the-issues/2025/01/15/innovating-in-line-with-the-european-unions-ai-act/
- Microsoft. (n.d.). Responsible AI. Retrieved January 21, 2025, from https://www.microsoft.com/en-us/ai/responsible-ai
- Miller, K. (2023, October 18). Introducing the Foundation Model Transparency Index. Stanford Institute for Human-Centered AI. Retrieved February 19, 2025, from https://hai.stanford.edu/news/introducing-foundation-model-transparency-index.
- MIT GenAI. (2024, March 27). Towards Standards for Data Transparency for AI Models. Retrieved January 29, 2025, from https://mit-genai.pubpub.org/pub/uk7op8zs/release/2.
- Moreno, J. (2022, December 29). OpenAI positioned itself as the AI leader in 2022. But could Google supersede it in '23? Forbes. Retrieved February 25, 2025, from https://www.forbes.com/sites/johanmoreno/2022/12/29/openai-positioned-itself-as-the-ai--leader-in-2022-but-could-google-supersede-it-in-23/
- Nazzaro, M. (2025, March 6). Trump considers extending TikTok ban. Retrieved March 10, 2025, from https://thehill.com/policy/technology/5180939-trump-considers-extending-tiktok-ban/
- Neuwirth, R. (2022). How the EU Can Achieve Legally Trustworthy AI. SSRN

Electronic Journal. Retrieved January 25, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991

- Neuwirth, R. (2024). Infringing AI: Liability for AI-generated Outputs Under International, EU, and UK Copyright Law. European Journal of Risk Regulation. Retrieved January 25, 2025, from https://www.cambridge.org/core/services/aop-cambridge-core/content/view/C568C6B717E9CFC45FB52E58E54B6BEC/S1867299X24000722a.pdf
- Neuwirth, R. (2024). The EU Artificial Intelligence Act: Regulating Subliminal AI Systems. Routledge. Retrieved January 25, 2025, from https://www.taylorfrancis.com/books/mono/10.4324/9781003319436/eu-artificial-intelligence-act-rostam-neuwirth
- O'Brien, J., Ee, S., & Kraprayoon, J. (2024). Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI. arXiv. Retrieved February 11, 2025, from https://arxiv.org/pdf/2407.01420
- Odelberg, T. (2024, May). Understanding the future of artificial intelligence governance: Comparing the EU AI Act and U.S. Executive Order on Safe AI. University of Michigan. Retrieved February 13, 2025, from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/195195/stpp-future-of-ai-governance%281%29.pdf?sequence=1
- OECD. (2024, May 3). OECD updates AI principles to stay abreast of rapid technological developments [Press release]. Retrieved March 15, 2025, from https://www.oecd.org/en/about/news/press-releases/2024/05/oecd-updates-ai-principles-to-stay-abreast-of-rapid-technological-developments.html.
- OECD. (n.d.). AI principles. Retrieved March 15, 2025, from https://www.oecd.org/en/topics/sub-issues/ai-principles.html.
- Office of the U.S. Trade Representative. (2023, May). US-EU Joint Statement on Trade and Technology Council. Retrieved February 13, 2025, from https://ustr.gov/about-us/policy-offices/press-office/press-releases/2023/may/us-eu-joint-statement-trade-and-technology-council
- Open Future. (2024, June). Transparency Template Requirements under the AI Act. Retrieved January 27, 2025, from https://openfuture.eu/wp-content/uploads/2024/06/240618AIAtransparency_template_requirements-2.pdf
- OpenAI. (2023). GPT-4 Technical Report. Retrieved January 21, 2025, from https://cdn.openai.com/papers/gpt-4.pdf
- OpenAI. (2024, August 8). GPT-4o system card. Retrieved March 3, 2025,

from https://cdn.openai.com/gpt-4o-system-card.pdf

- OpenAI. (2024, December 11). Terms of Use. Retrieved January 27, 2025, from https://openai.com/policies/terms-of-use/
- Owen, M. (2025, March 10). Third Time's the Charm? EU Considers Fining Apple Over DMA Yet Again. Retrieved March 10, 2025, from https://appleinsider.com/articles/25/03/10/third-times-the-charm-eu-considers-fining-apple-over-dma-yet-again
- Pandaily. (2024, April 28). Tesla meets China's data security standards, restrictions to be lifted. Retrieved March 11, 2025, from https://pandaily.com/tesla-meets-chinas-data-security-standards-restrictions-to-be-lifted/
- Park, W. (2025, February 24). Tesla's full self-driving: A game changer in China. AInvest. Retrieved March 11, 2025, from https://www.ainvest.com/news/tesla-full-driving-game-changer-china-2502/
- Parry, J., & Micheletti, F. (2025, March 7). EU tech chiefs say they don't target US tech. POLITICO. Retrieved March 11, 2025 from https://www.politico.eu/article/eu-tech-chiefs-say-they-dont-target-us-tech/
- Ponce del Castillo, A. (2024). The Role of Collective Agreements in Times of Uncertain AI Governance: Lessons from the Hollywood Scriptwriters' Agreement. AI & Society.
- PricewaterhouseCoopers. (2017). Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise? PwC. Retrieved January 18, 2025, from https://www.pwc.com/
- PwC. (2017). Sizing the prize: What's the real value of AI for your business and how can you capitalise? Retrieved March 15, 2025, from https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf.
- PYMNTS. (2024, November 5). Apple Faces Fine in First Charge Under EU's Digital Markets Act. Retrieved March 10, 2025, from https://www.pymnts.com/apple/2024/apple-faces-fine-in-first-charge-under-eus-digital-markets-act/
- Rehse, D., Valet, S., & Walter, J. (2024). Using Market Design to Improve Red Teaming of Generative AI Models. ZEW Policy Brief No. 06/2024. Retrieved January 18, 2025, from https://hdl.handle.net/10419/294875
- Restrict Act. (n.d.). Retrieved March 10, 2025, from

https://www.restrict-act.com/

- Reuters. (2024, April 29). Baidu, Tesla said to have agreed on mapping deal for FSD in China. The Economic Times. Retrieved March 11, 2025, from https://economictimes.indiatimes.com/news/international/business/baidu-tesla-said-to-have-agreed-on-mapping-deal-for-fsd-in-china/articleshow/109678784.cms
- Reuters. (2024, July 9). Disconnected Rules in a Connected World: Ideas for AI Innovation and Regulation. Retrieved January 24, 2025, from https://www.reuters.com/legal/legalindustry/disconnected-rules-connected-world-ideas-ai-innovation-regulation-2024-07-09/
- Reuters. (2025, February 19). Google develops AI 'co-scientist' to aid researchers. Retrieved February 25, 2025, from https://www.reuters.com/technology/artificial-intelligence/google-develops-ai-co-scientist-aid-researchers-2025-02-19/
- Samuelson, P. (2023). Copyright Implications of the Use of Generative AI. SSRN Electronic Journal. Retrieved January 25, 2025, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4531912
- Schröder, C., & Ashkar, D. (2024, October 22). The Artificial Intelligence Act of the European Union (AI Act). Orrick. Retrieved February 1, 2025, from https://www.orrick.com/en/Insights/2024/10/The-Artificial-Intelligence-Act-of-the-European-Union.
- Skadden. (2023, December). Global AI Legislation Tracker. Retrieved January 21, 2025, from https://www.skadden.com/-/media/files/publications/2023/12/2024-insights/a-list-of-ai-legislation-introduced-around-the-world.pdf
- Spotify Newsroom. (2024, August 14). The European Commission confirms Apple's anti-competitive behavior is illegal and harms consumers. Retrieved March 11, 2025 from https://newsroom.spotify.com/2024-03-04/the-european-commission-confirms-apples-anti-competitive-behavior-is-illegal-and-harms-consumers/
- Stanford Center for Research on Foundation Models (CRFM). (2021). On the Opportunities and Risks of Foundation Models. Retrieved January 21, 2025, from https://crfm.stanford.edu/report.html
- Stanford Center for Research on Foundation Models (CRFM). (2023, June 15). Initial Assessment of the EU AI Act. Retrieved January 21, 2025, from https://crfm.stanford.edu/2023/06/15/eu-ai-act.html
- Stanford Center for Research on Foundation Models (CRFM). (2024).

Foundation Model Transparency Index: Methodology and Results. Retrieved January 21, 2025, from https://crfm.stanford.edu/fmti/paper.pdf

- Stanford Center for Research on Foundation Models (CRFM). (2024, August 1). Foundation Models under the EU AI Act. Retrieved January 21, 2025, from https://crfm.stanford.edu/2024/08/01/eu-ai-act.html
- Stanford Center for Research on Foundation Models (CRFM). (2024, May). Foundation Model Transparency Index: May 2024 Update. Retrieved January 21, 2025, from https://crfm.stanford.edu/fmti/May-2024/index.html
- Stanford Center for Research on Foundation Models. (2024, May). Foundation Model Transparency Index (FMTI) May 2024 dataset. GitHub. Retrieved February 24, 2025, from https://github.com/stanford-crfm/fmti/tree/main/May2024
- Stanford Center for Research on Foundation Models. (2024, May). The Foundation Model Transparency Index v1.1. Retrieved February 19, 2025, from https://crfm.stanford.edu/fmti/paper.pdf.
- Tech Policy Press. (2024, December 9). How the EU AI Act Can Increase Transparency Around AI Training Data. Retrieved January 29, 2025, from https://www.techpolicy.press/how-the-eu-ai-act-can-increase-transparency-around-ai-training-data/.
- Thomson Reuters. (2025, January 17). Supreme Court upholds TikTok ban. Retrieved March 10, 2025, from https://www.cbc.ca/news/world/supreme-court-tiktok-ban-1.7434082
- U.S. AI.gov. (2024, May). Proceedings: Towards Standards for Data Transparency for AI Models. Retrieved January 29, 2025, from https://ai.gov/wp-content/uploads/2024/06/PROCEEDINGS_Towards-Standards-for-Data-Transparency-for-AI-Models.pdf.
- U.S. Department of Commerce & European Commission. (2022, May 16). U.S.-EU Joint Statement of the Trade and Technology Council. Retrieved February 14, 2025, from https://www.commerce.gov/sites/default/files/2022-05/US-EU-Joint-Statement-Trade-Technology-Council.pdf
- UNESCO. (n.d.). Forum on the ethics of AI. Retrieved March 15, 2025, from https://www.unesco.org/en/forum-ethics-ai.
- Usercentrics. (2023, October 12). Digital Markets Act Timeline. Retrieved March 10, 2025 from https://usercentrics.com/knowledge-hub/digital-markets-act-timeline/

- V7 Labs. (2022, July 11). Quality Training Data for Machine Learning: A Guide. Retrieved January 29, 2025, from https://www.v7labs.com/blog/quality-training-data-for-machine-learning-guide.
- Vattikonda, N., & Kelley, B. (2025, January 9). The D.C. Circuit Court's TikTok ban decision—Explained. Retrieved March 10, 2025, from https://www.lawfaremedia.org/article/the-d.c.-circuit-court's-tiktok-ban-decision--explained
- Vongthongsri, K. (2025, January 19). LLM benchmarks: MMLU, HellaSwag, and beyond. Confident AI. Retrieved March 7, 2025, from https://www.confident-ai.com/blog/llm-benchmarks-mmlu-hellaswag-and-beyond
- Voronoi. (2024, April 19). Google's Gemini Ultra Cost $191M to Develop. Retrieved January 30, 2025, from https://www.voronoiapp.com/technology/Googles-Gemini-Ultra-Cost-191M-to-Develop--1088.
- Vujović, D. (2024). Generative AI: Riding the New General Purpose Technology Storm. Ekonomika preduzeća. Retrieved January 18, 2025, from https://www.ses.org.rs/uploads/vujovic__240226_125016_135.pdf
- William Fry. (2024, July 23). A Practical Guide to the Extraterritorial Reach of the AI Act. Retrieved January 24, 2025, from https://www.williamfry.com/knowledge/a-practical-guide-to-the-extraterritorial-reach-of-the-ai-act/
- Yang, Z. (2023, August 14). Chinese Tesla data to be stored on mainland. China Daily. Retrieved March 11, 2025 from https://www.chinadailyhk.com/hk/article/345833