유럽연합 인공지능법

범용인공지능모델(GPAI) 실천강령 안내서

인공지능안전연구소 AI안전정책및대외협력실

ETRI 한국전자통신연구원 www.etri.re.kr

인공지능안전연구소 AI Safety Institute

유럽연합 인공지능법
범용인공지능모델(GPAI) 실천강령 안내서

최민석 실장(cooldenny@etri.re.kr)
송경호 선임연구원(songkyungho@etri.re.kr)
주지연 석사후연수연구원(jiyeon.joo@etri.re.kr)
신민규 석사후연수연구원(mgsin@etri.re.kr)
한국전자통신연구원 인공지능안전연구소 AI안전정책및대외협력실

Guidebook to the Code of Practice for General-Purpose AI Models under the EU AI Act

Minn-seok Chio(Assistant Director for AI Safety Policy and Strategic Cooperation)
Kyungho Song(Senior Researcher)
Jiyeon Joo(Research associate)
Min-gyu Sin(Research associate)
AI Safety Policy and Collaboration Section, Korea AI Safety Institute(AISI),
Electronics and Telecommunications Research Institute(ETRI)

Contents

Ⅰ 개요

1 안내서의 목적

- 유럽연합(European Union, 이하 EU)의 『인공지능법(Artificial Intelligence Act, 이하 AI법)』[1]은 범세계적인 영향력을 갖는 포괄적 인공지능 규범 체계다. EU 집행위원회(European Commission, EC, 이하 위원회)는 AI법 제56조에 근거하여 「범용인공지능모델 실천강령(Code of Practice for General-Purpose AI Models, 이하 CoP)」[2]을 마련하고, AI법 이행에 필요한 세부 실행 지침과 운영 기준을 구체화하였다.

- CoP는 범용인공지능(General-purpose AI, 이하 GPAI)모델[3]에 관한 규율 체계를 제시하는 문서로서, GPAI모델의 정의와 책임 범위, 시장 배치[4] 기준, 위험 분류, 조치 의무 등을 포괄적으로 규정하고 있다. CoP는 법적 구속력을 지닌 규범은 아니지만, 자율적 실행기준으로서 기능하며 EU AI법 준수를 입증하는 데 있어 핵심 준거가 된다. 나아가 감독기관의 평가, 민간 인증 절차, 시장 진입 요건 등 다양한 상황에서 실질적인 기준선으로 활용될 수 있다. 이에 따라 한국의 기업 및 기관은 EU 시장에서 직면할 수 있는 규제 불확실성을 최소화하고, 서비스 출시 지연, 제재 부과와 같은 리스크를 예방하기 위하여 CoP의 기술적 개념, 법적 구성요건 및 절차적 의무를 체계적으로 이해할 필요가 있다.

- 본 안내서는 EU AI법 및 CoP에 따른 GPAI모델의 규율체계와 주요 준수사항에 대한 이해를 돕기 위해 작성되었으며, 이를 위해 CoP의 체계와 구조를 충실히 반영하는 것을 기본 방향으로 하였다. 구체적으로, CoP의 가이드라인 및 3개 장(Chapter)의 순서와 조문 구조를 원문과 동일하게 유지하는 방식으로 번역을 진행하였다. 용어는 EU AI법 및 CoP의 정의와 용례를 기준으로 정리하였으며, 주요 개념과 번역어는 별도의 용어집으로 수록하였다. 아울러 해석에 영향을 줄 수 있는 사항은 각주로 보완하였으며, 이때 원주는 기울임체로, 역주는 별도 표시 없이 구분하여 기술하였다. 이와 같은 구성 방식과 번역 원칙을 통해 CoP 원문의 의미를 충실히 전달하는 동시에, 문맥에 따라 필요한 해석상 참고사항도 함께 확인할 수 있도록 하였다.

---

1) PE/24/2024/REV/1. 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). ELI: http://data.europa.eu/eli/reg/2024/1689/oj

2) European Commission. The General-Purpose AI Code of Practice. Published 10 July 2025. European Commission Digital Strategy. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

3) GPAI(General-purpose AI), 특정 목적이 아닌 다양한 용도에 사용할 수 있도록 설계된 인공지능 모델을 의미하며, 대표적으로 거대언어모델(LLM)이 포함됨

4) 시장배치, 개발이 완료된 AI모델을 실제 환경에 설치하거나 서비스에 통합하여 작동시키는 과정

- 이를 바탕으로 본 안내서는 다음과 같은 부분에 기여하고자 한다.:

(1) EU 규제에 대한 한국 기업의 대응 역량 제고

EU AI법 및 CoP는 GPAI모델 제공자, 다운스트림 수정자[5] 등 AI제품 및 서비스를 개발·운영하는 다양한 행위자에 대한 규율을 포함하며, 시스템적 위험[6]에 대한 평가 의무와 오픈소스 예외 요건 등 다층적인 규제 체계를 형성하고 있다. 본 안내서는 기술·법·운영 요소를 모두 포함한 CoP의 절차적 요구 사항을 명확하게 제시함으로써, 한국 기업이 EU 진출 과정에서 직면할 수 있는 규제 해석의 불확실성과 리스크를 최소화하는 것을 목표로 한다.

(2) 한국 법·정책 체계와의 정합성 확보

EU AI법 및 CoP는 안전 평가, 투명성 의무, 데이터 거버넌스, 저작권 준수, 위험의 식별·평가·완화 등 다양한 요소를 포함하는 인공지능 규제 체계로서, 한국의 인공지능기본법[7]과 규범적 연관성을 가진다. 다만, 양 규범은 개념 정의와 적용 방식, 의무의 범위 등에서 차이를 보이며, 이러한 차이는 국제 규범과 국내 제도 간 해석 및 적용의 간극을 초래할 수 있어 정합성 확보가 필수적이다. 이에 본 안내서는 CoP의 개념틀과 준수 기준을 제시함으로써, 한국의 법·제도·가이드라인이 국제 규범과 조화된 방향으로 발전할 수 있는 기준선을 제공한다.

(3) EU AI사무국 및 유관 기관과의 협력 기반 구축

EU AI법 및 CoP에 대한 공통된 이해와 용어 정합성은 EU AI사무국(AI Office), 위원회 및 유관 기관과 정책·기술 협력을 추진하기 위한 중요한 기반이 된다. 특히 GPAI모델의 위험 평가, 안전 조치 및 준수체계와 관련하여 동일한 개념과 기준을 공유하는 것은 국제 공조 및 제도 협력 과정에서 필수적이다. 본 안내서는 CoP의 주요 개념과 절차적 요구사항을 국내 기준에 맞추어 체계적으로 정리함으로써, 국내 AI 유관 기관·연구자·기업 등이 EU 측 기관과 보다 일관된 기준 아래 소통하고 협력할 수 있는 기반을 제공한다.

(4) 국내 산업·공공부문의 안전 거버넌스 고도화

CoP는 AI모델 수명주기[8]에 기반한 위험관리와 시스템적 위험의 식별·평가·완화 등에 관한 프레임워크를 제시한다. 이는 AI 안전 관리체계 구축 과정에서 참고할 수 있는 선제적·예방적 기준으로 활용될 수 있다. 본 안내서는 국내 기업·공공기관·연구자 및 개발자 등이 CoP의 주요 개념과 요구사항을 보다 쉽게 이해하고 실무적으로 참고할 수 있도록 지원하고자 한다.

---

5) Downstream Modifier, 기본 모델을 활용해 애플리케이션, 제품, 서비스 등을 개발·운영하는 행위자
6) Systemic Risk, AI시스템이 사회 전체에 영향을 줄 수 있는 구조적 위험으로, 단일 사고가 아닌 누적 효과나 상호작용을 통해 발생하는 위험
7) 『인공지능 발전과 신뢰 기반 조성 등에 관한 기본법』. 법률 제20676호. 공포 2025년 1월 21일; 시행 2026년 1월 22일. https://www.law.go.kr/lsInfoP.do?lsiSeq=268543
8) AI Model Lifecycle, AI모델의 설계, 학습, 배포, 운영, 업데이트, 폐기까지 포함한 전 주기

2 실천강령(CoP)의 개요 및 구성

- CoP는 독립 전문가 주도의 다자 이해관계자 절차를 통해 개발되었으며, GPAI모델 제공자, 다운스트림 제공자[9], 산업계, 학계 및 시민사회 등 여러 주체의 참여를 바탕으로 마련되었다. 이러한 참여 기반 구조는 AI법 제56조에서 AI사무국이 CoP 마련 과정에서 이해관계자의 참여를 촉진하고 그 필요와 이익을 고려하도록 규정한 데에 근거한다.

법률 EU AI법 제56조(Code of Practice)

1. AI사무국은 국제적 접근을 고려하여, 본 규정의 적절한 적용에 기여하기 위하여 EU 차원의 실천강령 수립을 장려하고 촉진하여야 한다.
2. AI사무국과 AI위원회(Board)는 실천강령이 최소한 제53조 및 제55조에 규정된 의무를 포괄하도록 보장하는 것을 목표로 하며, 다음 사항을 포함하여야 한다.:
   (a) 제53조(1) (a) 및 (b)에 언급된 정보가 시장 및 기술 발전에 비추어 최신 상태로 유지되도록 하는 수단;
   (b) 학습에 사용된 콘텐츠에 관한 요약의 적정한 상세 수준;
   (c) 필요한 경우, 그 출처를 포함하여 EU 차원에서의 시스템적 위험의 유형 및 성격, 그 발생 원인의 식별;
   (d) EU 차원에서의 시스템적 위험의 평가 및 관리에 관한 조치, 절차 및 방식, 그 문서화. 해당 조치는 위험의 심각성과 발생 가능성을 고려하여 비례적으로 설계되어야 하며, AI 가치사슬 전반에서 위험이 발생하고 실현될 수 있는 가능한 방식에 따른 대응상의 특수성을 반영하여야 한다.

… (생략) …

- CoP는 AI법 제56조에 근거하여 마련된 자율적 준수 도구로서, GPAI모델 제공자의 효과적인 의무 이행을 지원하기 위해 도입되었다. CoP는 GPAI모델 제공자에 대한 의무가 적용되는 시점(2025년 8월)과 표준 채택 시점(2027년 8월 이후) 사이의 공백을 보완하는 과도기적 준수 수단으로, 조화표준이 마련되기 전까지 제53조(GPAI 제공자의 의무) 및 제55조(시스템적 위험이 있는 GPAI모델 제공자의 의무)에 따른 의무 준수 사실을 입증하는 수단으로 활용된다. CoP는 법적 구속력이 없으나, AI법상 의무 준수 여부를 입증하는 실질적인 기준으로 기능하며, CoP를 활용하지 않는 GPAI모델 제공자는 향후 의무 이행 사실을 증명하는 과정에서 보다 높은 수준의 입증 부담을 질 수 있다.

9) Downstream provider, AI모델을 통합하여 애플리케이션, 제품, 서비스 형태로 개발·제공하는 행위자

- CoP는 적용 대상, 구조 및 운영 원칙 등에 관한 개괄적 내용을 담은 가이드라인(Guidelines)과 투명성(Transparency), 저작권(Copyright), 안전 및 보안(Safety and Security)의 3개 장(Chapter)으로 구성되어 있다. 이 중 투명성 및 저작권 장은 모든 GPAI모델 제공자에게 적용되며, 안전 및 보안 장은 시스템적 위험을 수반하는 GPAI모델 제공자(General-purpose AI model with systemic risk, 이하 GPAI-SR)에게만 적용된다. 각 장은 주요 이행사항을 약정(Commitment)과 이를 구체화하는 이행조치(Measure)[10]의 구조로 제시함으로써, 의무 준수를 위한 구체적 기준을 체계적으로 제시하고 있다.

- [투명성(Transparency)] 장은 GPAI모델 제공자가 AI법 제53조에 따른 문서화 및 정보 제공 의무를 이행할 수 있도록 모델설명서양식(Model Documentation Form)을 중심으로 한 체계를 제시한다. 특히 모델의 기능, 역량, 학습 및 테스트 방식, 배포 및 활용 정보 등 주요 사항을 체계적으로 문서화하고 최신 상태로 유지하도록 요구하며, 이를 통해 다운스트림 제공자, AI사무국 및 관할 당국이 모델의 특성과 활용 환경을 적절히 이해하고 감독할 수 있도록 지원한다.

- [저작권(Copyright)] 장은 GPAI모델 제공자가 EU 저작권 및 저작인접권 관련 법령을 준수하기 위한 정책 및 내부 관리체계를 제시한다. 특히 권리유보(opt-out) 신호의 식별 및 준수, 적법한 데이터 수집 및 활용, 저작권 관련 위험관리, 권리자 대응 및 불만처리 절차 등을 포함함으로써 학습데이터의 수집·이용 과정 전반에 걸친 저작권 컴플라이언스 체계 구축을 지원한다.

- [안전 및 보안(Safety and Security)] 장은 GPAI-SR모델 제공자를 대상으로 시스템적 위험(systemic risk)의 식별, 분석, 평가(ass), 수용 가능 여부 판단, 완화 및 사후 모니터링에 이르는 종합적인 위험관리 체계를 제시한다. 특히 시스템적 위험 평가(ass) 및 완화 절차, Safety and Security Framework의 구축·운영, 모델 평가 및 보고, 사고 보고, 보안 통제 등을 요구함으로써 GPAI-SR모델이 초래할 수 있는 시스템적 위험을 체계적으로 관리하고 AI법상 안전·보안 의무의 이행을 지원한다.

- 이와 같이 CoP는 투명성, 저작권 및 안전·보안 분야별 기준과 절차를 체계화함으로써 GPAI모델 제공자의 AI법상 의무 이행을 지원하는 실무적 기준을 제공한다. CoP는 GPAI모델 제공자의 자발적 참여를 전제로 운영되며, 서명기관은 CoP 준수를 통해 AI법상 관련 의무를 이행하고 있음을 입증하는 중요한 수단으로 활용할 수 있다. 또한 CoP의 운영과 참여기관 간 협력을 지원하기 위하여 AI사무국 주도의 협의체가 구성·운영되고 있으며, 참여기관 목록도 지속적으로 업데이트되고 있다. 이하에서는 CoP 서명기관 현황, GPAI 관련 EU AI법의 주요 추진 일정, 본 안내서 작성에 활용된 주요 용어(EU AI법 제3조) 및 CoP 관련 참고자료를 함께 제시한다.

10) 'Commitment'와 'Measure'는 문맥에 따라 각각 '책무'와 '조치사항' 또는 '서약'과 '조치' 등으로 번역될 수 있으나, 본 안내서에서는 CoP의 구조상 Commitment가 상위 수준의 이행 약속을, Measure가 이를 구체화하는 하위 실행 항목을 의미한다는 점을 고려하여 각각 '약정'과 '이행조치'로 번역함.

참고 CoP 서명기관 목록

- Accexible
- AI Alignment Solutions
- Aleph Alpha
- Almawave
- Amazon
- Anthropic
- Bria AI
- Cohere
- Cyber Institute
- Domyn
- Dweve
- Euc Inovação Portugal
- Fastweb
- Google
- Humane Technology
- IBM
- Lawise
- LINAGORA
- Microsoft
- Mistral AI
- Open Hippo
- OpenAI
- Pleias
- re-inventa
- ServiceNow
- Virtuo Turing
- WRITER
- xAI signed up to the Safety and Security Chapter

https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

참고 GPAI 관련 EU AI법 주요 일정

일정	추진 경과
2024년 7월 12일	·AI법이 EU 관보에 게재됨. 새로운 법률의 공식적인 공표(제113조)
2024년 8월 1일	·AI법 시행 ·이 시점에서는 AI법의 어떠한 요구사항도 적용되지 않으며, 향후 단계적으로 시행될 예정임(제113조)
2024년 11월 2일	·회원국은 기본권 보호 담당 기관·기구를 지정·공개해야함 ·해당 정보를 위원회 및 다른 회원국에 통보해야 함(제77조(2))
2025년 2월 2일	·특정 AI시스템에 대한 금지 규정이 시행됨. AI 리터러시에 관한 요구사항이 적용되기 시작함(제1장, 제2장, 제113조(a), 전문 179)
2025년 5월 2일	·위원회는 이날까지 CoP를 마련하여야 함(제56조(9), 전문 179).
2025년 8월 2일	·AI법의 주요 규정이 본격적으로 적용되기 시작함 –지정기관(Notified bodies) 관련 규정(제3장 제4절) –GPAI모델 관련 규정(제5장) –거버넌스 체계(제7장) –비밀유지 의무(제78조) –벌칙 규정(제99조 및 제100조) · (근거: 제113조(b))
2025년 8월 2일	·해당 날짜 이전에 출시하거나 서비스를 개시한 GPAI모델의 제공자는 2027년 8월 2일까지 법 준수 의무를 이행해야 함(제111조(3))
2025년 8월 2일	· 회원국은 국가 관할 당국(통지 기관 및 시장감독 기관 포함)을 지정해야 함 · 지정 결과를 위원회에 통보하고, 관련 기관의 연락처를 공개해야 함(제70조(2))
2025년 8월 2일	(이후 2년마다 반복) · 회원국은 국가 관할 당국의 재정 및 인적 자원 현황을 위원회에 보고해야 함(제70조(6))
2025년 8월 2일	· 회원국은 벌칙 및 과징금에 관한 규정을 마련하고 이를 위원회에 통보해야 함 · 또한 해당 규정이 실효성 있게 집행되도록 해야 함(전문 179)
2025년 8월 2일	· CoP가 확정되지 않거나 AI사무국이 CoP가 적정하지 않다고 판단하는 경우, 위원회는 GPAI모델 제공자의 의무 이행을 위한 공통 규칙을 이행행위를 통해 마련할 수 있음.(제56조(9))
2025년 8월 2일	(이후 매년 반복) · 위원회는 금지된 AI 관행에 대해 연례 검토를 실시함 · 필요 시 관련 규정을 개정할 수 있음(제112조(1))
2026년 2월 2일	· 위원회는 제6조의 실질적 이행을 구체화하는 지침을 마련해야 함 · 해당 지침에는 사후 시장 모니터링 계획이 포함됨(제6조(5), 제72조(3))

3 주요 용어(EU AI Act Art.3) 및 참고자료

용어	정의
AI system AI시스템	다양한 수준의 자율성을 가지고 작동하도록 설계되었으며 배포 후 적응성을 보일 수 있고, 명시적 또는 묵시적 목적을 위하여, 수신한 입력으로부터 예측, 콘텐츠, 추천 또는 의사결정과 같이 물리적 또는 가상 환경에 영향을 미칠 수 있는 출력을 생성하는 방법을 추론하는 기계 기반 시스템
risk 위험	위해(harm) 발생 확률과 해당 위해의 심각성의 결합
provider 제공자	AI시스템 또는 GPAI모델을 개발하거나 AI시스템 또는 GPAI모델을 개발하도록 하여 자신의 명의나 상표로 이를 출시하거나 AI시스템을 서비스를 개시하는 자연인, 법인, 공공기관, 대행기관 또는 기타 기관을 의미하며, 유상 또는 무상 여부를 불문함
deployer 운용자	AI시스템이 개인적인 비전문가적 활동의 과정에서 사용되는 경우를 제외하고, 자체 권한 하에 AI시스템을 사용하는 자연인, 법인, 공공기관, 대행기관 또는 기타 기구
authorised representative 권한을 위임받은 대리인	AI시스템 또는 GPAI모델의 제공자로부터 본 규정에 의해 설립된 의무와 절차를 대신 이행하고 준수할 권한을 서면으로 위임 받고 이를 수락한 자로써 EU 내에 소재하거나 설립된 자연인이나 법인
importer 수입업자	EU에 소재하거나 설립된 자연인이나 법인으로서 제3국에서 설립된 자연인이나 법인의 이름 또는 상표를 지닌 AI시스템을 출시하는 자
distributor 유통자	공급망에 속하는 자연인이나 법인 중 제공자 또는 수입업자가 아닌 자로서 EU 시장에 AI시스템을 제공하는 자
operator 운영자	제공자, 제품 제조업자, 운용자, 권한을 위임받은 대리인, 수입업자 또는 유통자
placing on the market 시장 출시	AI시스템 또는 GPAI모델을 EU 시장에 최초로 공급하는 행위
making available on the market 시장 공급	상업적 활동의 과정에서 EU 시장 내 유통 또는 사용을 위해 AI시스템 또는 GPAI모델을 공급하는 것을 의미하며, 유상 또는 무상 여부를 불문함
putting into service 서비스 개시	그 의도된 목적을 위하여 EU 시장에서 처음 사용하기 위해 제공자가 운용자에게 AI시스템을 직접 공급하거나 또는 자체적으로 사용하기 위해 공급하는 것

용어	정의
intended purpose 의도된 목적	기술 문서뿐만 아니라 사용 설명서, 홍보 또는 판매 자료 및 진술에서 제공자가 제공한 정보에 명시된 바와 같이, 구체적인 사용 맥락 및 조건을 포함하여 제공자가 AI시스템에 대하여 의도한 용도
reasonably foreseeable misuse 합리적으로 예측 가능한 오용	AI시스템을 의도된 목적에 맞게 사용하지 않는 행위로서, 합리적으로 예측 가능한 인간의 행동이나 다른 AI시스템 등 다른 시스템과의 상호작용의 결과로서 일어나는 것
safety component 안전 요소	제품이나 AI시스템의 요소 중에서 안전 기능을 담당하는 요소, 또는 장애나 오작동을 일으키면 사람의 건강이나 안전 또는 재산을 위험에 빠뜨리게 되는 요소
instructions for use 사용 지침	AI시스템의 의도된 목적과 올바른 사용법을 운용자에게 특별히 알리기 위해 제공자가 제공하는 정보
recall of an AI system AI시스템 리콜	운용자에게 공급한 AI시스템을 제공자에게 반품하거나 서비스를 중단하거나 사용을 비활성화하는 조치
withdrawal of an AI system AI시스템 회수	공급망 내 AI시스템을 시장에 공급하는 것을 방지하기 위한 조치
performance of an AI system AI시스템의 성능	AI시스템이 의도된 목적을 달성하는 능력
notifying authority 통보 기관	적합성 평가 기구의 평가, 지정 및 통보와 모니터링을 위하여 필요한 절차를 설정하고 수행할 책임이 있는 국가 기관
conformity assessment 적합성 평가	고위험 AI시스템에 관한 제3장 제2절에 규정된 요건이 충족되었는지 여부를 입증하는 절차
conformity assessment body 적합성 평가기관	시험, 인증 및 검사를 포함하여 제3자 적합성 평가 활동을 수행하는 기관
notified body 지정기관	본 규정 및 기타 관련 EU 조화 법령(Union harmonisation legislation)에 따라 통지된 적합성 평가기관

용어	정의
substantial modification 실질적 변경	AI시스템이 시장에 출시되거나 서비스가 개시된 후, 제공자가 수행한 최초의 적합성 평가에서 예견되거나 계획되지 않은 수정으로서, 그 결과로 제3장 제2절에 규정된 요구사항에 대한 AI시스템의 적합성이 영향을 받거나 해당 AI시스템이 평가받았던 의도된 목적에 수정이 발생하는 변경
CE marking CE 마크	제공자가 AI시스템이 제3장 제2절에 규정된 요구사항 및 그 부착을 규정하는 기타 적용 가능한 EU 조화 법령을 준수함을 나타내는 마크
post-market monitoring system 사후 시장 모니터링 시스템	AI시스템 제공자가 시장에 출시하거나 서비스를 개시한 AI시스템의 사용으로부터 얻은 경험을 수집하고 검토하기 위해 수행하는 모든 활동을 의미하며, 이는 필요한 시정조치 또는 예방 조치를 즉시 적용할 필요성을 식별하기 위한 목적을 가짐
market surveillance authority 시장 감독 기관	Regulation (EU) 2019/1020에 따라 활동을 수행하고 조치를 취하는 국가 기관
harmonised standard 조화 표준	Regulation (EU) 제1025/2012호 제2조(1)(c)에 정의된 조화된 표준
common specification 공통 기준	Regulation (EU) 제1025/2012호 제2조(4)에 정의된 일련의 기술 사양을 의미하며, 본 규정 하에 확립된 특정 요구사항을 준수하기 위한 수단을 제공함
training data 학습 데이터	학습 가능한 파라미터를 조정함으로써 AI시스템을 학습시키는 데 사용되는 데이터
validation data 검증 데이터	학습된 AI시스템에 대한 평가를 제공하고, 특히 과소적합(underfitting) 또는 과대적합(overfitting)을 방지하는 등을 목적으로 해당 시스템의 학습 불가능한 파라미터 및 학습 과정을 조정하는 데 사용되는 데이터
validation data set 검증 데이터셋	고정되거나 가변적인 분할 방식에 따라, 별도로 분리된 데이터셋 또는 학습 데이터셋의 일부
testing data 테스트 데이터	AI시스템이 시장에 출시되거나 서비스가 개시되기 전에 해당 시스템의 예상 성능을 확인하기 위해 독립적인 평가를 제공하는 데 사용되는 데이터
input data 입력 데이터	AI시스템이 산출물을 생성하는 근거로서 해당 시스템에 제공되거나 시스템에 의해 직접 획득되는 데이터
biometric data 생체 데이터	얼굴 이미지나 지문 데이터와 같이 자연인의 신체적, 생리적 또는 행동적 특성에 관한 특정 기술적 처리를 통해 생성된 개인정보를 의미

용어	정의
biometric identification 생체 인식	자연인의 신원을 확인하기 위하여 해당 개인의 생체 데이터와 데이터베이스에 저장된 개인들의 생체 데이터를 비교함으로써 인간의 신체적, 생리적, 행동적 또는 심리적 특징을 자동적으로 인식하는 것
biometric verification 생체 확인	자연인의 생체 데이터를 이전에 제공된 생체 데이터와 비교함으로써 해당 자연인의 신원을 인증과 같은 자동화된 방식으로 일대일 확인하는 것
special categories of personal data 개인정보의 특정 범주	Regulation (EU) 2016/679 제9조(1), Directive (EU) 2016/680 제10조 및 Regulation (EU) 2018/1725 제10조(1)에서 언급된 개인정보 범주
sensitive operational data 민감 운영 데이터	범죄의 예방, 탐지, 수사 또는 기소 활동과 관련된 운영 데이터를 의미하며, 그 공개가 형사 절차의 무결성을 위태롭게 할 수 있는 운영 데이터
emotion recognition system 감정 인식 시스템	생체 데이터를 기반으로 자연인의 감정 또는 의도를 식별하거나 추론할 목적으로 사용되는 AI시스템
biometric categorisation system 생체 분류 시스템	자연인의 생체 데이터를 기반으로 특정 범주로 분류할 목적으로 사용되는 AI시스템을 의미함. 단, 다른 상업적 서비스에 부수적이고 객관적인 기술적 이유로 엄격히 필요한 경우는 제외
remote biometric identification system 원격 생체 인식 시스템	자연인의 생체 데이터를 레퍼런스 데이터베이스에 포함된 생체 데이터와 비교함으로써, 주로 먼 거리에서 해당 자연인의 적극적인 관여 없이 그 신원을 식별할 목적으로 사용되는 AI시스템을 의미
real-time remote biometric identification system 실시간 원격 생체 인식 시스템	생체 데이터의 캡처, 비교 및 식별이 상당한 지체 없이 이루어지는 원격 생체 인식 시스템을 의미하며, 이는 즉각적인 식별뿐만 아니라 회피를 방지하기 위한 제한적인 짧은 지연 시간도 포함
post remote biometric identification system 사후 원격 생체 인식 시스템	실시간 원격 생체 인식 시스템이 아닌 원격 생체 인식 시스템

용어	정의
publicly accessible space 공개적으로 접근 가능한 공간	접근에 대한 특정 조건이 적용될 수 있는지 여부나 잠재적인 수용 인원 제한 여부와는 관계 없이, 불특정 다수의 자연인들이 접근할 수 있는 공공 또는 사적 소유의 물리적 장소
law enforcement authority 법 집행 기관	(a) 공공 안전에 대한 위협을 방지하고 이에 대비하는 기능을 포함하여, 범죄의 예방, 탐지, 수사 또는 기소, 혹은 형사 처벌의 집행을 담당하는 모든 공공 기관 (b) 공공 안전에 대한 위협으로부터 보호하고 이를 방지하는 기능을 포함하여, 범죄의 예방, 탐지, 수사 또는 기소, 혹은 형사 처벌의 집행을 목적으로, 회원국 법률에 따라 공권력 및 공공 권한을 행사하도록 위임받은 기타 모든 기구 또는 실체;
law enforcement 법 집행	범죄의 예방, 탐지, 수사 또는 기소, 혹은 형사 처벌의 집행을 위하여 법 집행 기관이 직접 수행하거나 이를 대리하여 수행하는 활동을 의미하며, 공공 안전에 대한 위협으로부터 보호하고 이를 방지하는 기능을 포함함
AI Office AI사무국	2024년 1월 24일 자 위원회 결정에 따라 규정된 AI시스템 및 GPAI모델의 이행, 모니터링, 감독과 AI 거버넌스에 기여하는 위원회의 기능을 의미함. 본 규정에서 AI사무국에 대한 언급은 위원회에 대한 언급으로 해석됨
national competent authority 국가 관할 당국	통보 기관 또는 시장 감독 당국을 의미함. 유럽연합의 기관, 대행기관, 사무소 및 기구에 의해 서비스에 개시되거나 사용되는 AI시스템과 관련하여, 본 규정에서 국가 관할 당국 또는 시장 감독 당국에 대한 언급은 유럽개인정보보호감독관(European Data Protection Supervisor)에 대한 언급으로 해석함
serious incident 중대 사고	직접적 또는 간접적으로 다음 중 하나를 초래하는 AI시스템과 관련된 사건 또는 오작동을 의미함. (a) 사람의 사망 또는 사람의 건강에 대한 중대한 위해; (b) 중요 기반시설의 관리 또는 운영에 대한 중대하고 돌이킬 수 없는 중단; (c) 기본권 보호를 목적으로 하는 EU 법령상 의무의 침해; (d) 재산 또는 환경에 대한 중대한 위해;
personal data 개인정보	Regulation (EU) 2016/679 제4조(1)에 정의된 개인정보
non-personal data 비개인정보	Regulation (EU) 2016/679 제4조(1)에 정의된 개인정보가 아닌 데이터
profiling 프로파일링	Regulation (EU) 2016/679 제4조 제4호에 정의된 프로파일링

용어	정의
real-world testing plan 실제 시험 계획	실제 환경 조건에서 시험의 목적, 방법론, 지리적·인구통계적·시간적 범위, 모니터링, 조직 및 수행 과정을 기술한 문서
sandbox plan 샌드박스 계획	참여하는 제공자와 관할 당국 간에 합의된 문서로서, 샌드박스 내에서 수행되는 활동의 목적, 조건, 기간, 방법론 및 요구사항을 기술한 것
AI regulatory sandbox AI 규제 샌드박스	관할 당국이 구축한 통제된 프레임워크를 의미하며, 이는 AI시스템의 제공자 또는 예비 제공자에게 규제 감독 하에 제한된 기간 동안, 필요한 경우 실제 환경 조건에서, 샌드박스 계획에 따라 혁신적인 AI시스템을 개발, 학습, 검증 및 시험할 수 있는 기회를 제공하는 것
AI literacy AI 리터러시	제공자, 운용자 및 영향을 받는 자들이 본 규정의 맥락에서 각자의 권리와 의무를 고려하여, AI시스템을 충분한 정보에 입각하여 활용할 수 있도록 하고, AI의 기회와 위험 및 AI가 초래할 수 있는 잠재적 위해에 대한 인식을 제고할 수 있도록 하는 기술, 지식 및 이해
testing in real-world conditions 실제 환경 조건에서의 시험	신뢰할 수 있고 견고한 데이터를 수집하고 본 규정의 요구사항에 대한 AI시스템의 적합성을 평가하고 검증하기 위해, 실험실 또는 기타 모의 환경 외부의 실제 환경 조건에서 AI시스템을 의도된 목적에 따라 일시적으로 시험하는 것을 의미함. 또한 이는 제57조 또는 제60조에 명시된 모든 조건이 충족되는 경우, 본 규정의 의미 내에서 AI시스템을 시장에 출시하거나 서비스를 개시하는 것으로 간주하지 않음
subject 대상자	실제 환경에서의 시험 목적으로, 실제 환경 조건에서의 시험에 참여하는 자연인
informed consent 사전 고지 동의	시험 참여 결정과 관련된 시험의 모든 측면에 대해 고지를 받은 후, 특정 실제 환경 조건에서의 시험에 참여하겠다는 의사를 대상자가 자유롭게, 구체적으로, 명확하게, 자발적으로 표현한 것
deep fake 딥페이크	실존하는 인물, 사물, 장소, 실체 또는 사건과 유사하며 사람에게 진실하거나 사실인 것처럼 허위로 보일 수 있는, AI에 의해 생성되거나 조작된 이미지, 오디오 또는 비디오 콘텐츠

용어	정의
widespread infringement 광범위한 침해	개인의 이익을 보호하는 유럽연합 법령에 위배되는 모든 작위 또는 부작위로서 다음 중 하나에 해당하는 것을 의미함 (a) 다음의 회원국을 제외한 최소 2개 이상의 회원국에 거주하는 개인들의 집단적 이익에 위해를 가하였거나 가할 가능성이 있는 경우: (i) 해당 작위 또는 부작위가 시작되거나 발생한 회원국; (ii) 관련 제공자 또는 해당하는 경우 그 권한을 위임받은 대리인이 소재하거나 설립된 회원국; (iii) 운용자에 의해 침해가 저질러진 경우, 해당 운용자가 설립된 회원국; (b) 동일한 운영자에 의해 최소 3개 이상의 회원국에서 동시에 발생하고, 동일한 불법 관행이나 침해된 동일한 이익을 포함하여 공통된 특징을 가지며, 개인들의 집단적 이익에 위해를 가하였거나, 가하고 있거나, 가할 가능성이 있는 경우;
critical infrastructure 핵심 인프라	Directive (EU) 2022/2557 제2조(4)에 정의된 핵심 인프라
general-purpose AI model 범용 AI(GPAI)모델	대규모 자기 지도 학습을 통해 대량의 데이터로 학습된 경우를 포함하여, 상당한 범용성을 나타내고 모델이 시장에 출시되는 방식과 관계없이 다양한 유형의 별개 과업을 능숙하게 수행할 수 있으며 다양한 다운스트림 시스템 또는 애플리케이션에 통합될 수 있는 AI모델. 단 시장에 출시되기 전 연구, 개발 또는 시제품 제작 활동에 사용되는 AI모델은 제외
high-impact capabilities 고영향 역량	가장 발전된 GPAI모델들에서 기록된 역량과 일치하거나 이를 상회하는 역량
systemic risk 시스템적 위험	GPAI모델의 고영향 역량에 특유한 위험으로서, 그 도달 범위로 인하여, 혹은 공중 보건, 안전, 공공의 안전, 기본권 또는 사회 전체에 대한 실제적이거나 합리적으로 예견 가능한 부정적 영향으로 인하여 EU 시장에 중대한 영향을 미치며, 가치 사슬 전반에 걸쳐 대규모로 확산될 수 있는 위험
general-purpose AI system 범용 AI(GPAI)시스템	GPAI모델을 기반으로 하며, 직접적인 사용뿐만 아니라 다른 AI시스템으로의 통합을 포함하여 다양한 목적에 부합하는 역량을 갖춘 AI시스템
floating-point operation(FLOP) 부동소수점 연산	부동소수점 수와 관련된 모든 수학적 연산 또는 할당을 의미하며, 부동 소수점 수는 실수의 부분집합으로서 일반적으로 컴퓨터에서 고정된 기수(base)의 정수 지수에 의해 크기가 조정된 고정 정밀도 정수로 표현됨
downstream provider 다운스트림 제공자	해당 AI모델이 스스로에 의해 제공되어 수직적으로 통합되었는지 또는 계약 관계에 기반하여 다른 실체에 의해 제공되었는지 여부와 관계없이, GPAI시스템을 포함하여 AI모델을 통합하는 AI시스템의 제공자

참고자료 1 The General-Purpose AI Code of Practice

2025년 7월 10일 공개된 「General-Purpose AI Code of Practice(GPAI CoP)」는 EU AI법의 GPAI 모델 규율체계를 실효적으로 집행하기 위한 핵심 자율규범이다. 본 문서는 단순한 업계 가이드라인을 넘어 GPAI모델 제공자가 AI법상의 의무를 보다 구체적이고 체계적으로 이행할 수 있도록 지원하는 준규범적(soft law) 성격의 이행 수단으로 평가된다. 특히 OpenAI, Google, Microsoft, Anthropic, Amazon, Mistral AI 등 주요 글로벌 AI 사업자들이 서명에 참여함에 따라, CoP는 향후 국제적인 AI 컴플라이언스 기준으로 발전할 가능성이 높은 규범으로 주목받고 있다.

EU 위원회는 GPAI모델 제공자에게 부과되는 AI법상 의무의 이행을 지원하기 위해 본 CoP를 마련하였다. CoP는 독립 전문가와 산업계, 시민사회 등이 참여한 다중 이해관계자(multi-stakeholder) 방식으로 작성되었으며, EU AI사무국과 AI위원회는 이를 "AI법 준수를 입증하기 위한 적절한 자율적 수단"으로 공식 인정하였다. 이에 따라 GPAI모델 제공자는 CoP 준수를 통해 법적 의무 이행을 보다 효율적으로 입증할 수 있으며, 규제 불확실성과 행정 부담 역시 상당 부분 완화할 수 있게 되었다.

본 사이트는 CoP의 세 개의 장인 Transparency, Copyright 및 Safety and Security의 내용을 제공하고 있다. 각 장은 AI법이 GPAI모델 제공자에게 요구하는 의무를 구체화하고, 이를 실무적으로 이행할 수 있는 방법을 제시한다는 점에서 중요한 의미를 가진다.

먼저 투명성(Transparency) 장은 AI법 제53조에 따른 투명성 의무를 구체화한 내용으로, 특히 모델설명서양식(Model Documentation Form)을 통해 모델의 기능, 학습 방식, 사용 제한, 성능 특성 등을 체계적으로 문서화할 수 있도록 지원한다. 이는 단순한 정보 공개를 넘어 AI시스템의 설명가능성과 책임성을 확보하기 위한 핵심 규제수단으로 평가된다.

저작권(Copyright) 장은 생성형 AI의 학습 데이터 활용과 관련하여 급속히 부상하고 있는 저작권 문제에 대응하기 위한 운영체계를 제시한다. 여기에는 데이터 수집 과정에서의 권리자 대응, 권리유보(opt-out) 메커니즘의 준수, 내부 기록관리 및 위험 완화 절차 등이 포함되어 있으며, 단순한 선언적 준법 수준을 넘어 조직 차원의 저작권 컴플라이언스 거버넌스 체계 구축을 요구한다는 점에서 의미가 있다.

안전 및 보안(Safety and Security Chapter) 장은 GPAI-SR 제공자를 대상으로 한다. 이는 AI법 제55조에 따라 시스템적 위험이 있는 GPAI모델 제공자에게 적용되는 내용으로, 모델 평가, 위험 식별 및 완화, 사이버보안, 사고 보고 체계 등 고도화된 AI 안전관리 체계를 제시하고 있다.

정책·법률적 관점에서 CoP의 가장 중요한 의의는 AI 규제가 단순한 원칙 선언 중심의 접근에서 벗어나 실제 기업 내부의 문서화, 위험관리, 거버넌스 및 감사체계를 요구하는 운영 중심 규제로 발전하고 있음을 보여준다는 데 있다. 형식적으로는 자율규범이지만, 감독기관의 공식 인정과 주요 글로벌 사업자의 참여를 통해 상당한 규범적 영향력을 확보하고 있으며, 향후 글로벌 AI 컴플라이언스의 사실상 국제표준으로 기능할 가능성이 높다. 이러한 점에서 CoP는 EU AI법의 실질적인 이행 메커니즘인 동시에 향후 국제 AI 규범질서 형성에 중요한 영향을 미칠 핵심 거버넌스 모델로 평가할 수 있다.

참고자료: European Commission. "The General-Purpose AI Code of Practice." Shaping Europe's Digital Future, European Commission. Published 10 July 2025; last update 10 December 2025. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

참고자료 2 Questions and Answers on the Code of Practice for General-Purpose AI

「Questions and Answers on the Code of Practice for General-Purpose AI」는 EU GPAI CoP와 관련된 주요 쟁점과 절차를 설명하기 위해 마련된 공식 질의응답(Q&A) 페이지이다. 본 페이지는 CoP의 목적과 법적 성격, 작성 과정, 가입 절차, 기대효과 및 향후 운영계획 등을 종합적으로 설명함으로써 GPAI모델 제공자와 이해관계자의 이해를 지원하는 역할을 수행한다.

EU 위원회에 따르면 CoP는 GPAI모델 제공자가 AI법상 의무를 이행하고 준수 여부를 입증할 수 있도록 지원하기 위한 자율적 수단이다. CoP는 GPAI모델이 유럽 시장에서 안전하고 투명하게 제공될 수 있도록 지원하는 것을 목적으로 하며, 특히 가장 발전된 GPAI모델에 대해서는 시스템적 위험관리에 관한 구체적인 이행방안을 제시한다. CoP는 Transparency, Copyright, Safety and Security의 세 개의 장으로 구성되며, Transparency 장과 Copyright 장은 모든 GPAI모델 제공자에게 적용되고, Safety and Security는장은 GPAI-SR모델 제공자에게 적용된다.

본 페이지는 CoP가 어떠한 과정을 통해 작성되었는지에 대해서도 상세히 설명한다. CoP는 2024년 7월부터 시작된 다중 이해관계자 방식의 논의를 통해 마련되었으며, 산업계, 학계, 시민사회, 권리자 단체, 회원국 대표 등 1,400명 이상이 참여하였다. 또한 EU AI사무국이 임명한 독립 전문가들이 초안을 작성하고, 수차례의 공개 협의와 워크숍을 거쳐 최종안이 마련되었다는 점을 소개하고 있다. 이는 CoP가 특정 이해관계자의 입장이 아니라 폭넓은 의견 수렴 과정을 거쳐 형성된 규범이라는 점을 보여준다.

아울러 본 페이지는 CoP의 법적 성격과 효과를 명확히 설명한다. CoP는 AI법에 새로운 의무를 부과하거나 기존 의무를 확대하는 문서가 아니라, 이미 AI법에 규정된 의무를 이행하기 위한 자율적 준수수단이다. 따라서 CoP 자체는 법적 구속력을 갖지 않지만, GPAI모델 제공자가 CoP를 준수하는 경우 AI법 준수를 보다 간편하고 투명하게 입증할 수 있으며, 행정적 부담과 규제 불확실성을 줄일 수 있다는 점을 강조하고 있다.

또한 CoP 적용 일정과 향후 집행 방향에 관한 정보도 제공한다. AI법에 따른 GPAI 관련 의무는 2025년 8월 2일부터 적용되며, 시스템적 위험이 있는 GPAI모델 제공자는 해당 시점부터 관련 의무를 준수하여야 한다. EU AI사무국은 초기 이행 단계에서 CoP 가입 사업자와 긴밀히 협력할 예정이며, 2026년 8월 이후에는 GPAI모델 제공자의 의무 이행 여부에 대한 본격적인 감독과 집행이 이루어질 예정이다. 또한 기술 발전과 위험 환경 변화에 대응하기 위하여 CoP를 최소 2년마다 검토하고 필요 시 개정할 수 있는 체계를 마련하고 있음을 설명하고 있다.

이와 함께 본 페이지는 EU 위원회가 마련한 GPAI모델 관련 가이드라인의 역할도 소개한다. 해당 가이드라인은 어떤 모델이 GPAI모델에 해당하는지, 어떤 모델이 시스템적 위험을 가진 GPAI모델에 해당하는지, 그리고 누가 GPAI모델 제공자로 간주되는지 등 AI법 적용의 핵심 개념을 명확히 설명한다. 이러한 점에서 본 페이지는 CoP 자체에 대한 설명뿐만 아니라 향후 GPAI 규제체계의 적용과 집행 방향을 이해하기 위한 중요한 참고자료로 활용될 수 있다.

참고자료: European Commission. “Questions and answers on the Code of Practice for General-Purpose AI.” Shaping Europe’s Digital Future, European Commission. Last Updated: 11 July 2025.
https://digital-strategy.ec.europa.eu/en/faqs/questions-and-answers-code-practice-general-purpose-ai

참고자료 3 Guidelines on the Scope of Obligations for Providers of General-Purpose AI Models under the AI Act

「Guidelines on the Scope of Obligations for Providers of General-Purpose AI Models under the AI Act」는 EU 위원회가 「Shaping Europe's Digital Future」 포털을 통해 공개한 공식 가이드라인으로, AI법상 GPAI모델 제공자에게 적용되는 의무의 범위와 적용 기준을 해석·설명하기 위해 마련된 문서이다. 이 가이드라인은 2025년 8월 2일부터 GPAI모델 제공자에 대한 의무가 본격적으로 적용됨에 따라, 어떤 모델과 어떤 사업자가 규제 대상에 해당하는지, 그리고 어떠한 의무를 부담하게 되는지를 명확히 하기 위한 목적을 가진다. 특히 EU 위원회는 본 가이드라인이 AI 가치사슬 전반의 사업자에게 법적 명확성을 제공하고 규제 예측가능성을 높이기 위한 수단임을 강조하고 있으며, GPAI CoP와 함께 AI법 집행체계의 핵심적인 보조 규범으로 기능하도록 설계하고 있다.

이 가이드라인의 가장 중요한 특징은 AI법상 GPAI모델에 해당하는 범위를 보다 구체적으로 제시하고 있다는 점이다. 문서에 따르면 일반적으로 10^{23} FLOP를 초과하는 연산자원을 사용하여 학습된 모델은 GPAI모델에 해당할 가능성이 높은 것으로 간주되며, 특히 텍스트, 이미지, 음성, 비디오 등 다양한 유형의 콘텐츠를 생성하거나 여러 과업을 수행할 수 있는 범용성을 갖춘 경우 GPAI모델로 판단될 수 있다. 다만 이는 절대적인 기준이 아니라 모델의 범용성, 수행 가능한 작업의 범위 및 실제 활용 가능성 등을 종합적으로 고려하기 위한 해석 기준으로 제시되고 있다.

가이드라인은 GPAI모델 제공자에게 적용되는 핵심 의무도 체계적으로 설명하고 있다. 모든 GPAI모델 제공자는 기술문서의 작성 및 유지, 다운스트림 제공자에 대한 정보 제공, EU 저작권법 준수를 위한 정책 수립, 학습데이터에 관한 충분히 상세한 요약본의 공개 등의 의무를 부담한다. 이는 GPAI모델의 투명성과 책임성을 강화하기 위한 기본적인 규제체계를 구성한다.

여기에 더하여 시스템적 위험이 있는 GPAI모델 제공자에게는 보다 강화된 의무가 적용된다. 해당 제공자는 모델 평가, 위험 식별 및 완화, 중대한 사고 보고, 사이버보안 조치, AI사무국에 통지 등의 추가 의무를 부담하게 된다. AI법은 현재 누적 학습 연산량이 10^{25} FLOP를 초과하는 모델을 시스템적 위험이 있는 GPAI모델로 추정하고 있으며, 이는 가장 발전된 AI모델에 대하여 별도의 규율체계를 적용하려는 EU의 위험 기반 접근방식을 보여준다.

또한 이 가이드라인은 오픈소스 모델에 대한 예외 적용 기준도 설명하고 있다. 일정한 투명성 요건을 충족하는 자유·오픈소스 GPAI모델은 일부 의무에서 제외될 수 있으나, 시스템적 위험을 수반하는 경우에는 오픈소스 여부와 관계없이 추가 의무가 적용될 수 있음을 명확히 하고 있다. 이는 혁신 촉진과 위험관리 사이의 균형을 추구하는 EU의 정책 방향을 반영하는 것으로 이해할 수 있다.

정책·법률적 관점에서 볼 때 이 가이드라인의 가장 중요한 의미는 AI법의 추상적인 규정을 실제 집행 가능한 기준으로 구체화하고 있다는 데 있다. 특히 GPAI모델 해당 여부, GPAI모델 제공자 판단 기준, 시스템적 위험 해당 여부 등을 계량적·기능적 요소와 연계하여 설명함으로써, 향후 AI 규제가 기술적 성능과 시장 영향력을 중심으로 한 위험 기반 규율체계로 발전하고 있음을 보여준다. 비록 가이드라인 자체는 법적 구속력을 가지지 않지만, EU 위원회와 AI사무국이 향후 감독 및 집행 과정에서 참고할 공식적인 해석기준이라는 점에서 높은 규범적 중요성을 가진다. 따라서 본 문서는 단순한 참고자료를 넘어 향후 EU GPAI 규제의 실제 적용 방향과 감독 기준을 이해하기 위한 핵심 해석자료로 평가할 수 있다.

참고자료: European Commission. "Guidelines on the scope of the obligations for providers of general-purpose AI models under the AI Act." Shaping Europe's Digital Future, European Commission. Published 18 July 2025; last update 31 July 2025.
https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act

참고자료 4 Learn More About the Guidelines for Providers of General-Purpose AI Models

「Learn More about the Guidelines for Providers of General-Purpose AI Models」는 AI법상 GPAI모델 제공자 의무와 관련된 가이드라인의 주요 내용을 소개하는 정책 자료이다. 본 자료는 GPAI모델의 판단 기준, GPAI모델 제공자의 범위, 시스템적 위험이 있는 모델에 대한 추가 의무, 오픈소스 모델에 대한 예외 적용, 그리고 향후 집행 일정 등 가이드라인의 핵심 내용을 간략하게 설명하고 있다.

자료에 따르면 일반적으로 10^{23} FLOP를 초과하는 연산자원을 사용하여 학습되고 텍스트·오디오·이미지·비디오 생성 기능을 수행할 수 있는 모델은 GPAI모델에 해당할 가능성이 높은 것으로 제시된다. 또한 GPAI모델 제공자는 다운스트림 제공자에 대한 정보 제공, 기술문서 작성 및 유지, EU 저작권법 준수를 위한 정책 수립 등의 의무를 부담하며, GPAI-SR모델의 경우에는 위험 평가(ass) 및 완화, 안전성 확보 등 추가적인 의무가 적용된다. 아울러 본 자료는 기존 모델을 수정하거나 변경한 경우, 어떠한 상황에서 새로운 GPAI모델 제공자로 간주될 수 있는지, 자유·오픈소스 라이선스로 공개된 모델에는 어떠한 예외가 적용될 수 있는지 등에 대해서도 설명하고 있다. 또한 GPAI CoP와의 관계, 2025년 8월 2일부터 시작되는 GPAI모델 관련 의무의 적용, 그리고 2026년 8월 2일부터 예정된 본격적인 집행 일정도 함께 소개하고 있다.

참고자료: European Commission. “Learn more about the guidelines for providers of general-purpose AI models.” Shaping Europe’s Digital Future, European Commission. Publication 18 July 2025; last update 31 July 2025.
https://digital-strategy.ec.europa.eu/en/news/learn-more-about-guidelines-providers-general-purpose-ai-models

참고자료 5 Guidelines for Providers of General-Purpose AI Models

「Guidelines for Providers of General-Purpose AI Models」는 AI법상 GPAI모델 제공자에게 적용되는 의무의 범위와 집행 기준을 설명하는 자료이다. 본 자료는 어떤 AI모델이 GPAI모델에 해당하는지, 누가 GPAI모델 제공자로 간주되는지, 그리고 어떠한 의무가 적용되는지를 중심으로 주요 내용을 소개하고 있다.

특히 일반적으로 10^{23} FLOP를 초과하는 연산량으로 학습되고 텍스트·오디오·이미지·비디오 생성 기능을 수행할 수 있는 모델을 GPAI모델로 설명하고 있으며, 기존 모델에 실질적 변경(significant modifications)이 이루어진 경우 새로운 GPAI모델 제공자로 판단될 수 있다는 점을 제시하고 있다. 또한 자유·오픈소스 라이선스로 공개된 모델에 대해서는 일정한 조건 아래 일부 의무 예외가 적용될 수 있음을 설명한다. 아울러 GPAI모델 제공자에게 적용되는 주요 의무와 집행 일정도 안내하고 있다. GPAI모델 제공자는 정보 제공, 기술문서 작성, 저작권 준수 정책 수립 등의 의무를 부담하며, 시스템적 위험이 있는 GPAI모델 제공자는 위험 평가(ass)·완화, 안전성 확보, AI사무국에 통지 등 추가 의무를 부담한다. 또한 2025년 8월 2일부터 관련 의무가 적용되고, 2026년 8월 2일부터는 EU 위원회의 본격적인 집행이 시작됨을 설명하고 있다.

참고자료: European Commission. “Guidelines for providers of general-purpose AI models.” Shaping Europe’s Digital Future, European Commission. Last updated: 17 October 2025.
https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers

참고자료 6 Code of Practice Overview

「Overview of the Code of Practice」는 Future of Life Institute가 운영하는 EU AI법 해설 웹사이트에 수록된 자료로, GPAI CoP의 전체 구조와 주요 내용을 장(chapter), 약정, 이행조치 단위로 정리하여 설명하고 있다. 특히 Transparency, Copyright, Safety and Security 각 장의 주요 의무와 이행 방안을 체계적으로 요약함으로써 CoP의 핵심 내용을 이해하기 쉽게 소개하고 있다는 점이 특징이다.

Transparency 장과 관련해서는 모델설명서양식(Model Documentation Form)을 중심으로 모델의 기술적 사양, 학습데이터, 연산자원 및 에너지 사용량, 활용 목적, 시장 공급 방식 등의 정보를 문서화하고 관리하는 체계를 설명한다. Copyright 장에서는 저작권 정책 수립, 적법한 데이터 수집, 권리유보(opt-out) 신호 준수, 저작권 침해 위험 완화, 권리자 불만 처리 체계 등을 주요 내용으로 소개하고 있다. 또한 Safety and Security 장에 대해서는 시스템적 위험이 있는 GPAI모델 제공자에게 적용되는 위험관리 체계를 중심으로 설명한다. 여기에는 Safety and Security Framework 구축, 시스템적 위험 식별·분석·평가(ass), 위험 완화, 사고 보고, 보안 통제 및 내부 책임 배분 등이 포함되며, 모델의 전체 수명주기에 걸친 지속적인 위험관리 체계를 요구하고 있음을 정리하고 있다.

참고자료: Future of Life Institute. “Overview of the Code of Practice (General-Purpose AI Code of Practice).” Published 30 July 2025.https://artificialintelligenceact.eu/code-of-practice-overview/

참고자료 7 EU의 GPAI 실천 강령: 시스템적 위험 평가를 중심으로

「EU의 범용 AI 실천 강령: 시스템적 위험 평가를 중심으로」는 EU GPAI CoP의 주요 내용 가운데 Safety and Security 장에 규정된 시스템적 위험 평가 체계를 중심으로 분석한 자료이다. 본 자료는 EU가 2025년 7월 발표한 GPAI CoP가 AI법 제53조 및 제55조에 따른 GPAI모델 제공자의 의무 이행을 지원하기 위한 자율적 이행 수단이라는 점을 설명하고, 투명성, 저작권, 안전 및 보안으로 구성된 CoP의 전체 구조를 함께 소개하고 있다.

특히 시스템적 위험이 있는 GPAI모델 제공자에게 적용되는 Safety and Security 장에 주목하여, 안전 및 보안 프레임워크 구축, 시스템적 위험 식별·분석·평가(ass), 위험 수용가능성 판단, 위험 완화, 모델 보고서 작성 및 사고 보고 체계 등 AI법상 의무 이행을 위한 주요 절차를 검토한다. 또한 시스템적 위험 평가(ass)가 AI모델의 전 생애주기에 걸쳐 수행되는 위험관리 체계임을 강조하면서, 모델 평가, 적대적 테스트, 외부 평가, 사후 모니터링 등 구체적인 평가 및 관리 수단도 함께 다루고 있다.

아울러 본 자료는 EU가 CoP를 통해 기업의 자율적인 참여와 혁신을 장려하는 동시에, 고도화된 AI모델이 사회 전반과 기본권, 공공안전에 미칠 수 있는 위험에 대해서는 체계적인 관리 의무를 부과하고 있음을 분석한다. 이를 바탕으로 한국 역시 산업 진흥과 신뢰 확보 간 균형을 고려한 AI 위험관리 체계를 마련할 필요가 있으며, GPAI CoP는 향후 인공지능기본법과 하위법령 및 관련 정책 설계 과정에서 참고할 수 있는 주요 사례임을 시사하고 있다.

참고자료: 정보통신정책연구원(KISDI). “KISDI Perspectives, No. 1 | [동향] EU의 GPAI 실천 강령: 시스템적 위험 평가를 중심으로”. 2025년 8월 18일.
https://www.kisdi.re.kr/report/view.do?key=m2102058837181&masterId=4334696&arrMasterId=4334696&artId=1869856

참고자료 8 EU, 'GPAI모델 제공자 실천 강령' 제정

「EU, '범용 AI(GPAI)모델 제공자 실천 강령' 제정」은 한국지능정보사회진흥원(NIA)이 발간한 디지털 법제 Brief 자료로, 2025년 7월 최종 공개된 EU GPAI CoP의 구조와 주요 내용을 체계적으로 분석하고 있다. 특히 AI법 제53조 및 제55조에 따른 GPAI모델 제공자와 시스템적 위험이 있는 GPAI모델 제공자의 의무를 중심으로, 투명성·저작권·안전 및 보안 각 장의 주요 내용과 정책적 의미를 정리하고 있다는 점이 특징이다.

Transparency 장과 관련해서는 모델설명서양식(Model Documentation Form)을 중심으로 GPAI모델 제공자가 제공해야 하는 정보의 범위와 절차를 검토한다. 모델의 일반 정보, 아키텍처, 학습데이터, 컴퓨팅 자원, 에너지 사용량, 시장 공급 방식 및 사용 목적 등을 체계적으로 문서화하고 지속적으로 최신 상태로 유지하도록 요구하고 있으며, 다운스트림 제공자, AI사무국(AI Office), 국가 관할 당국에 대한 정보 제공 체계도 함께 소개하고 있다. 또한 정보의 품질·무결성·보안 확보와 영업비밀 및 지식재산권 보호 원칙이 어떻게 적용되는지도 설명한다.

Copyright 장에서는 EU 저작권법 준수를 위한 정책 수립 의무와 함께 웹 크롤링 및 텍스트·데이터 마이닝(TDM) 과정에서 준수해야 할 기준을 다룬다. 특히 robots.txt 및 기계판독형 권리유보 표시(machine-readable rights reservations)의 식별·준수, 불법 복제 사이트 크롤링 배제, 저작권 침해 결과물 생성 위험 완화, 권리자 불만 처리 절차 구축 등 구체적인 이행조치를 정리하고 있다. 이를 통해 GPAI모델 학습 과정에서 데이터의 적법한 접근과 권리유보 준수가 핵심적인 컴플라이언스 요소로 자리잡고 있음을 보여준다.

Safety and Security 장에서는 시스템적 위험이 있는 GPAI모델 제공자에게 요구되는 위험관리 체계를 중점적으로 분석한다. 시스템적 위험의 식별·분석·평가(ass)·수용 가능 여부 판단·완화·사후 모니터링·사고 보고 등 모델 수명주기 전반에 걸친 위험관리 절차를 설명하고, Safety and Security Framework 구축, 모델 보고서(Model Report) 제출, 독립적인 외부평가, 사이버보안 조치, 내부 책임 배분 체계 등 구체적인 요구사항을 소개한다. 또한 화학·생물·방사능·핵(CBRN) 위험, 통제력 상실, 사이버 공격, 유해한 조작 등 CoP가 제시하는 주요 시스템적 위험 유형을 정리하면서, EU가 GPAI 규제를 공공안전과 기본권 보호를 위한 위험관리 체계로 접근하고 있음을 보여준다.

아울러 본 자료는 Transparency, Copyright, Safety and Security 각 장에 대해 1차 초안부터 최종안까지의 변화 과정을 비교·분석하고 있다는 점에서도 의미가 있다. 정보 제공 범위 확대, robots.txt 준수 의무 명확화, 독립 외부평가 강화, 중대 사고 보고 체계 구체화 등 최종안에서 강화된 주요 내용을 정리함으로써 CoP의 발전 과정을 한눈에 파악할 수 있도록 한다.

또한 GPAI CoP가 향후 글로벌 AI 규범의 중요한 기준으로 발전할 가능성이 높다고 평가하면서, 한국 역시 GPAI모델에 특화된 위험관리 체계와 자율규범 기반의 정책 수단을 마련할 필요가 있다는 시사점을 제시한다. 특히 국내 스타트업과 중소기업의 부담을 고려한 비례성 원칙 및 단계적 준수 방안은 향후 국내 제도 설계 과정에서 참고할 수 있는 요소로 제시되고 있다.

참고자료 : 한국지능정보사회진흥원(NIA). "지능사회 이슈분석 | EU, 'GPAI (GPAI) 모델 제공자 실천 강령' 제정." 2025년 9월 9일.
https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=82618&bcIdx=28577&parentSeq=28577

II CoP 가이드라인

EU 집행위원회

Brussels, 2025. 7. 18.
C(2025) 5045 최종본
부속서

위원회에 대한 통신문(Communication)의 부속서(Annex)

– Regulation (EU) 2024/1689 (AI법)에서 정한 GPAI모델 제공자의 의무 범위에 관한 가이드라인」에 관한 위원회 통신문 초안 내용의 승인[11)]

〈목 차〉

11) European Commission. Communication to the Commission — Approval of the content of the draft Communication from the Commission – Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act). C(2025) 5045 final. Brussels, 18 July 2025. Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act: https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act

1 위원회 가이드라인의 배경 및 목적

- (1) 2024년 6월 13일 유럽의회 및 이사회(European Parliament and the Council)가 채택한 인공지능에 관한 조화된 규칙을 제정하고 일부 규정을 개정하는 법률((EU) 2024/1689, 이하 "AI법")은 2024년 8월 1일 발효되었다. 이 법은 EU 내에서 AI의 혁신과 활용 확산을 촉진하는 한편, 민주주의와 법치주의를 포함한 건강·안전 및 기본권에 대한 높은 수준의 보호를 보장하는 것을 목적으로 한다.

- (2) GPAI모델은 EU 내 AI의 혁신과 활용 확산에 중요한 역할을 한다. 이는 GPAI모델이 다양한 작업에 활용될 수 있을 뿐만 아니라 여러 다운스트림 AI시스템에 통합될 수 있기 때문이다. 따라서 GPAI모델 제공자는 AI 가치사슬(AI value chain) 전반에서 중요한 역할과 책임을 부담하며, 특히 다운스트림 제공자(downstream providers)가 해당 모델과 그 역량(capabilities)을 충분히 이해할 수 있도록 하는 데 중요한 역할과 책임을 진다. 이러한 이해는 다운스트림 제공자가 해당 모델을 자신의 제품에 통합하고 AI법에 따른 의무를 이행하는 데 중요하다.

- (3) 이러한 이유로 AI법은 비례적인 투명성 조치를 규정하고 있다. 여기에는 GPAI모델 제공자는 다운스트림 제공자 및 요청 시 AI사무국·국가 관할 당국에 제공되는 문서를 작성하고 이를 최신 상태로 유지하할 의무가 포함된다(AI법 제53조(1) (a) 및 (b) 참조). 한편, 자유 및 오픈소스 라이선스(free and open-source licence)[12] 하에 공개되는 GPAI모델의 제공자는 일정한 조건 하에서 이러한 투명성 관련 요구사항이 면제된다. 다만, 해당 모델이 GPAI-SR에 해당하는 경우에는 면제되지 않는다.

- (4) GPAI모델은 대량의 텍스트, 이미지, 비디오 및 기타 데이터를 기반으로 학습되는 경우가 많으며, 이러한 학습 데이터에는 저작권 보호 콘텐츠가 포함될 수 있다. 이에 따라 AI법은 GPAI모델 제공자에게 EU 저작권법 준수 정책을 마련하도록 요구하고 있으며(AI법 제53조(1)(c) 참조), 학습에 사용된 콘텐츠의 요약을 공개하도록 규정하고 있다(AI법 제53조(1)(d) 참조).

- (5) GPAI모델은 시스템적 위험을 수반할 수 있으며, 이는 가장 발전된 GPAI모델에 특유한 위험으로서 EU 시장에 중대한 영향을 미칠 수 있는 위험을 의미한다(AI법 제3조(65) 참조). AI법에 따르면, 누적 연산 자원량(cumulative amount of computational resources)이 10^{25} 부동 소수점 연산("FLOP[13]", AI법 제3조(67) 참조)을 초과하여 학습된 GPAI모델은 GPAI-SR모델로 추정된다(AI법 제51조(2) 참조). 또한 위원회는 AI법 부속서 XIII의 기준에 따라 GPAI모델을 GPAI-SR모델로 지정할 수 있다. GPAI-SR모델 제공자는 GPAI모델 제공자에게 적용되는 의무 외에도

12) free and open-source licence, 자유 및 오픈소스 라이선스. 소프트웨어의 소스 코드를 공개하여 누구나 자유롭게(Liberties) 사용, 연구, 수정, 배포할 수 있는 권리를 부여하는 법적 계약

13) FLOP(Floating Point Operation)은 부동소수점 연산 1회를 의미하고, FLOPs(Floating-Point Operations)는 누적 부동소수점 연산 횟수를 의미함. 본 안내서는 EU AI Act 및 CoP 가이드라인 원문의 용례를 따라 모델의 누적 학습 연산량(training compute)을 'FLOP'으로 표기하였으며, 이는 FLOPs와 동일한 의미로 사용됨

해당 위험의 평가(ass) 및 완화를 위한 추가적인 의무를 부담하며, 여기에는 모델 평가 수행, 중대 사고(serious incident) 보고 및 적정 수준의 사이버보안 보호조치 확보 등이 포함된다.

- (6) AI법에 따라 위원회는 시스템적 위험의 유무와 관계없이 GPAI모델 제공자가 AI법 제V장에 규정된 의무를 준수하도록 집행할 독점적 권한을 가진다. 해당 집행 권한은 유럽 AI사무국(AI법 제3조(47) 참조)에 위임되어 있다.

- (7) AI법 제96조(1)에 따라 위원회는 AI법의 실질적 이행에 관한 가이드라인을 마련하여야 한다. 2025년 8월 2일 GPAI모델 제공자에 대한 의무의 적용이 임박한 상황을 고려하여, AI법상 GPAI모델 제공자의 의무 범위를에 중점을 두고 있다. 또한 본 가이드라인은 AI법에 한정하여 적용되며, 책임에 관한 규정을 포함한 다른 EU 법령에는 적용되지 않는다.

- (8) 구체적으로, 본 가이드라인은 다음 네 가지 핵심 주제를 다룬다.: (i) GPAI모델(제2절), (ii) GPAI모델을 출시하는 제공자(제3절), (iii) 오픈소스(open-source)로 공개된 GPAI모델 제공자에 대한 일부 의무 면제(제4절), (iv) GPAI모델 제공자의 의무 준수(제5절).

  이를 통해 본 가이드라인은 AI 가치사슬(value chain)에 참여하는 행위자가 다음 사항을 판단하는 데 도움을 주고자 한다.: (i) 자신의 모델이 GPAI모델에 해당하는지 여부, (ii) 자신이 해당 GPAI모델을 출시하는 제공자인지 여부, (iii) GPAI모델 제공자에게 적용되는 일부 의무가 면제되는지 여부 및 (iv) 특히 2025년 8월 2일 의무 적용 직후 예상되는 위원회의 의무 준수 집행 방식.

- (9) 본 가이드라인은 GPAI모델 제공자에 대하여 구속력을 가지지 않으며, AI법에 대한 유권해석은 EU 사법재판소(Court of Justice of the European Union, “CJEU”)만이 제시할 수 있다. 그럼에도 불구하고 본 가이드라인은 AI법에 대한 위원회의 해석 및 적용 방식을 제시하는 것으로서, 이를 바탕으로 관련 집행 행위를 수행하게 된다. 이는 제공자의 의무 준수를 용이하게 하고 AI법의 효과적인 이행에 기여한다. 다만, 개별 사안의 특수성을 고려하기 위하여 사안별 평가(ass)는 항상 필요하다. 또한 본 가이드라인은 기술적·사회적·시장적 발전에 비추어 필요한 경우 검토 및 업데이트될 예정이다(제6절).

- (10) 본 가이드라인은 AI법에 따른 GPAI모델 제공자의 적절한 의무 이행을 지원하기 위한 위원회의 다른 이니셔티브와 상호보완적인 관계에 있으며, 특히 AI법 제56조에 따라 마련된 CoP와 상호보완적인 관계에 있다. 본 가이드라인은 AI법상 GPAI모델 제공자 의무의 범위와 그 이행에 관한 사항(준수 입증에 있어 CoP의 역할 포함)을 다룬다. 한편, CoP는 AI법 제53조(1) 및 제55조(1)에따른 의무 준수를 위하여 GPAI모델 제공자가 이행할 수 있는 구체적인 조치를 규정한다. 또한 본 가이드라인은 AI법 제53조(1)(d)에 따른 GPAI모델 학습에 사용된 콘텐츠에 대한 공개 요약을 위한 위원회가 제공하는 템플릿과도 상호보완적인 관계에 있다.

- (11) 본 가이드라인은 위원회가 2025년 4월 22일부터 2025년 5월 22일 실시한 공개 다중 이해관계자

협의(public multi-stakeholder consultation)를 통해 수집한 의견을 반영하였다. 또한 본 가이드라인의 이전 버전은 2025년 6월 30일 유럽 인공지능위원회(European Artificial Intelligence Board, 이하 "AI위원회")를 통해 회원국에 제시되었다. 아울러 본 가이드라인은 AI모델을 GPAI모델 및 GPAI-SR모델로 분류하는 것과 관련하여 AI사무국에 자문을 제공하기 위해 위원회의 공동연구센터(Joint Research Centre, 이하 "JRC")가 구성한 전문가 풀과 그 밖의 전문가들로부터 수렴한 의견도 반영하였다.

2 GPAI 모델

- (12) 본 절은 'GPAI모델'의 개념에 중점을 둔다. 우선, 위원회가 어떠한 경우에 모델을 GPAI모델로 판단하는지에 관한 지표적 기준을 제시한다(제2.1절). 또한 GPAI모델의 '수명주기(lifecycle)' 개념과 해당 개념이 GPAI모델 제공자의 의무에 미치는 영향에 대하여 설명한다(제2.2절). 마지막으로 AI법 제51조 및 제52조에 따라 GPAI모델이 GPAI-SR모델에 해당하는 경우를 설명한다(제2.3절).

2.1. GPAI모델 해당 여부 판단 기준

- (13) AI법 제3조(63)은 'GPAI모델'을 다음과 같이 정의한다.: '대규모 자기 지도 학습(self-supervision)[14]을 통해 대량의 데이터로 학습된 경우를 포함하여, 출시 방식과 관계없이 상당한 범용성을 갖추고 다양한 개별 작업을 적절한 수준으로 수행할 수 있으며, 다양한 다운스트림 시스템 또는 애플리케이션(application)에 통합될 수 있는 AI모델을 의미한다. 다만, 시장에 출시되기 전에 연구·개발 또는 프로토타이핑(prototyping) 활동에 사용되는 AI모델은 제외된다.'

  이 정의는 모델이 GPAI모델에 해당하는지를 판단하기 위한 요소를 개괄적으로 제시하고 있다. 그러나 잠재적 제공자가 자신의 모델이 GPAI모델에 해당하는지 여부를 평가(ass)하는 데 활용할 수 있는 구체적인 기준까지 제시하고 있지는 않다. 따라서 GPAI모델 제공자에 해당하는지 여부를 평가(ass)하여야 하는 다수의 행위자의 부담을 줄이기 위해서는 확인이 용이한 기준을 제시하는 것이 중요하다.

- (14) GPAI모델의 역량과 사용 사례가 매우 다양하다는 점을 고려할 때, 모델이 GPAI모델에 해당하는지를 판단하기 위한 구체적인 역량이나 수행 가능한 작업을 일률적으로 제시하는 것은 현실적으로 어렵다.

- (15) 이에 따라 위원회는 모델이 GPAI모델에 해당하는지를 판단하기 위한 지표적 기준을 마련하기 위하여, FLOP 단위로 측정되는 모델 학습에 사용된 연산 자원량(학습 연산량(training compute), 본 가이드라인에서의 의미는 제115항 참조)과 모델의 모달리티를 활용하는 접근을 취한다. 이러한 접근은 AI법 전문 98에 근거한다. AI법 전문 98은 "최소 10억 개 이상의 파라미터를 가지며, 대량의 데이터를 활용한 대규모 자기 지도 학습의 방식으로 학습된 모델은 상당한 범용성을 나타내고 다양한 개별 작업을 유능하게 수행할 수 있는 것으로 보아야 한다."고 규정하고 있다. 또한 AI법 전문 99는 "대규모 생성형 AI모델은 텍스트, 오디오, 이미지 또는 비디오 등의 형태로 콘텐츠를 유연하게 생성할 수 있어 다양한 개별 작업에 쉽게 활용될 수 있다는 점에서 GPAI모델의 전형적인 사례에 해당한다."고 설명하고 있다.

14) Self-supervision, 레이블(label, 정답)이 없는 대규모 데이터로부터 데이터 자체의 구조나 관계를 활용하여 학습 신호를 생성하고, 이를 통해 모델이 학습하는 머신러닝 기법

- (16) 학습 연산량은 제공자가 파라미터 수와 학습 예시 수를 비교적 쉽게 하나의 수치로 결합하여 산정할 수 있다는 장점이 있다. 학습 연산량은 일반적으로 파라미터 수와 학습 예시 수를 곱한 값에 비례하므로, 모델 규모와 학습 데이터 규모에 대하여 각각 별도의 임계값을 설정하는 대신 단일 임계값을 설정할 수 있다. 학습 연산량은 범용성과 역량을 완벽하게 대변하는 지표(proxy)는 아니지만, 위원회는 현 시점에서는 학습 연산량 임계값을 포함한 지표적 기준(indicative criterion)을 설정하는 것이 가장 적절한 접근이라고 본다. 다만 기술 및 시장의 발전에 따라 위원회의 접근방식은 향후 변경될 수 있다. 또한 위원회는 특히 소규모 행위자를 위하여, 범용성과 역량을 비교적 용이하게 평가(ass)할 수 있는 다른 기준의 활용 가능성에 대한 검토를 지속할 예정이다.[15)]

- (17) 위와 같은 고려사항에 따라 모델이 GPAI모델에 해당하는지 판단하기 위한 지표적 기준은 해당 모델의 학습 연산량이 10^{23} FLOP을 초과하고, 언어(텍스트[16)] 또는 오디오[17)] 형태 여부와 관계없음), 텍스트-이미지 생성(text-to-image) 또는 텍스트-비디오 생성(text-to-video)이 가능하다는 것이다.

- (18) 이러한 임계값, 즉 10^{23} FLOP은 대량의 데이터를 활용하여 약 10억 개의 파라미터를 갖는 모델을 학습시키는데 통상적으로 요구되는 연산량 수준에 해당한다. 다만 동일한 모델이라도 사용되는 데이터의 양에 따라 학습에 필요한 연산량은 달라질 수 있다. 그럼에도 불구하고, 작성 시점을 기준으로 대량의 데이터를 활용하여 학습된 모델의 경우 10^{23} FLOP이 일반적인 수준이다(부속서 A.3 예시 참조)[18)].

- (19) 모달리티는 모델이 텍스트 또는 음성(오디오의 한 유형) 형태에 관계없이 언어를 생성하도록 학습된 경우, 언어를 사용하여 의사소통하고 지식을 저장하며 추론할 수 있다는 점을 고려하여 선정되었다. 어떠한 모달리티도 이처럼 광범위한 역량을 제공하지 않는다. 따라서 언어 생성 모델은 일반적으로 다른 모델보다 더 다양한 작업을 수행할 수 있는 역량을 가진다. 한편, 이미지 또는 비디오를 생성하는 모델은 통상적으로 언어 생성 모델에 비해 역량과 사용 사례의 범위가 더 제한적이지만, 그럼에도 GPAI모델로 간주될 수 있다. 특히 텍스트-이미지 생성 및 텍스트-비디오 생성 모델은 다양한 시각적 결과물을 생성할 수 있으며, 이를 통해 다양한 개별 작업에 쉽게 활용될 수 있는 유연한 콘텐츠 생성이 가능하다.

- (20) GPAI모델이 위 (17)의 기준을 충족하더라도, 예외적으로 상당한 범용성을 나타내지 않거나 다양한 개별 작업을 유능하게 수행할 수 없는 경우에는 GPAI모델에 해당하지 않는다. 반대로, GPAI모델이 (17)의 기준을 충족하지 않더라도 예외적으로 상당한 범용성을 나타내고 다양한 개별 작업을 유능하게 수행할 수 있는 경우에는 GPAI모델로 해당한다.

---

15) 향후 위원회는 필요하다고 판단되는 경우, 모델이 GPAI모델에 해당하는지 여부를 판단하기 위하여 벤치마크를 고려할 수 있다.
16) 위원회는 '텍스트(text)' 모달리티(modality)에 '코드(code)'가 포함되는 것으로 이해한다.
17) 위원회는 '오디오(audio)' 모달리티(modality)에 '음성(speech)'이 포함되는 것으로 이해한다.
18) 본 가이드라인에 대한 공개 의견수렴(public consultation) 당시 제안되었던 기준값에서 해당 임계값이 업데이트되었다. 이전에는 부속서 A.3의 참고 사례에 10^{23} FLOP 미만의 학습 연산량으로 학습된 모델의 사례가 여러 개 포함되어 있었다. 그러나 이러한 모델은 모두 비상업적 행위자(예: 대학 연구그룹 또는 비영리 연구기관)가 연구 목적으로만 학습한 것이므로, AI법의 적용 대상에 포함되어야 하는 모델을 대표하는 사례로 보기 어렵다고 판단되었다.

〈 적용 대상 모델의 예시 〉

• 인터넷 및 기타 출처에서 수집·정제된 광범위한 자연어 데이터(즉, 텍스트)를 활용하여, 10^{24} FLOP를 사용해 학습된 모델(현재 언어 모델에서 일반적인 방식)
  - (17)의 기준에 따르면, 해당 모델은 텍스트를 생성할 수 있고 학습 연산량이 10^{23} FLOP를 초과하므로 GPAI모델에 해당하는것으로 보아야 한다. 또한 광범위한 자연어 데이터를 기반으로 학습된 점은 해당 모델이 상당한 범용성을 나타내며 다양한 개별 작업을 적절히 수행할 수 있음을 시사한다. 따라서 해당 모델은 GPAI모델에 해당할 가능성이 높다.

〈 비적용 대상 모델의 예시 〉

• 자연어 데이터를 사용하여 10^{22} FLOP을 사용해 학습된 모델(다양한 개별 작업을 적절히 수행할 수 없는 경우)
  - 해당 모델은 텍스트를 생성할 수 있으나 학습 연산량이 10^{23} FLOP을 초과하지 않으므로, (17)의 기준에 따르면 GPAI모델에 해당하지 않는 것으로 보아야 한다. 또한 다양한 개별 작업을 수행할 수 없다는 점 역시 해당 모델이 GPAI모델이 아님을 뒷받침한다.

• 음성을 텍스트로 변환하는 작업에 특화되어 10^{24} FLOP을 사용하여 학습된 모델
  - 해당 모델은 텍스트를 생성할 수 있고 학습 연산량이 10^{23} FLOP을 초과하므로, (17)의 기준에 따르면 GPAI모델에 해당하는 것으로 보아야 한다. 그러나 수행 가능한 작업이 음성을 텍스트로 변환하는 기능에 한정된 경우에는 GPAI모델로 간주되지 않는다.

• 텍스트로부터 음성을 작업에 특화되어 10^{24} FLOP을 사용하여 학습된 모델
  - 해당 모델은 음성을 생성할 수 있고 학습 연산량이 10^{23} FLOP을 초과하므로, (17)의 기준에 따르면 GPAI모델에 해당하는 것으로 보아야 한다. 그러나 수행 가능한 작업이 텍스트로부터 음성을 생성하는 기능에 한정된 경우에는 GPAI모델로 간주되지 않는다.

• 이미지 해상도를 향상시키는 작업에 특화되어 10^{24} FLOP을 사용하여 학습된 모델
  - 해당 모델은 이미지를 생성할 수 있고 학습 연산량이 10^{23} FLOP을 초과하므로, (17)의 기준에 따르면 GPAI모델에 해당하는 것으로 보아야 한다. 그러나 수행 가능한 작업이 이미지 해상도 향상에 한정된 경우에는 GPAI모델로 간주되지 않는다.

• 손상되거나 결손된 이미지의 일부를 복원하는 기능을 목적으로 10^{24} FLOP을 사용하여 학습된 모델

- 해당 모델은 이미지를 생성할 수 있고 학습 연산량이 10^{23} FLOP을 초과하므로, (17)의 기준에 따르면 GPAI모델에 해당하는 것으로 보아야 한다. 그러나 수행 가능한 작업이 이미지 복원(inpainting)에 한정된 경우에는 GPAI모델로 간주되지 않는다.

• 체스 또는 비디오 게임 수행을 목적으로 10^{24} FLOP을 사용하여 학습된 모델
- 해당 모델의 학습 연산량은 10^{23} FLOP을 초과하나, 텍스트 생성, 음성 생성, 텍스트-이미지 생성(text-to-image) 및 텍스트-비디오(text-to-video) 생성 기능을 갖추지 못한 경우 (17)의 기준에 따르면 GPAI모델에 해당하지 않는 것으로 보아야 한다. 또한 수행 가능한 작업이 체스 또는 비디오 게임 수행에 한정된다는 점 역시 해당 모델이 GPAI모델이 아님을 뒷받침한다.

• 기상 패턴 또는 물리 시스템 모델링을 목적으로 10^{24} FLOP을 사용하여 학습된 모델
- 해당 모델의 학습 연산량은 10^{23} FLOP을 초과하나, 텍스트 생성, 음성 생성, 텍스트-이미지 생성(text-to-image) 및 텍스트-비디오 생성 기능(text-to-video)을 갖추지 못한 경우, (17)의 기준에 따르면 GPAI모델에 해당하지 않는 것으로 보아야 한다. 또한 수행 가능한 작업이 기상 패턴 또는 물리 시스템 모델링에 한정된다는 점 역시 해당 모델이 GPAI모델이 아님을 뒷받침한다.

• 가사를 포함한 텍스트 프롬프트를 기반으로 음악 생성 기능을 목적으로 10^{24} FLOP을 사용하여 학습된 모델
(21) 해당 모델은 음성(노래 가사 형태)을 생성할 수 있고 학습 연산량이 10^{23} FLOP를 초과하므로, (17)의 기준에 따르면 GPAI모델에 해당하는 것으로 보아야 한다. 그러나 수행 가능한 작업이 가사를 포함한 텍스트 프롬프트를 기반으로 음악을 생성하는 기능에 한정된 경우에는 GPAI모델로 간주되지 않는다.
• 음향 효과 생성 기능을 목적으로 10^{24} FLOP을 사용하여 학습된 모델. 해당 모델은 언어(음성)를 생성할 수 있으나, 생성되는 음성 콘텐츠의 품질이 낮은 경우(예: 짧은 대화 동안에도 일관된 내용을 유지하지 못하는 경우)
- 해당 모델은 언어를 생성할 수 있고 학습 연산량이 10^{23} FLOP를 초과하므로, (17)의 기준에 따르면 GPAI모델에 해당하는 것으로 보아야 한다. 그러나 수행 가능한 작업이 음향 효과 생성에 한정된 경우에는 GPAI모델로 간주되지 않는다.

2.2. GPAI모델의 수명주기

- (22) GPAI-SR모델 제공자는 '모델의 전체 수명주기에 걸쳐 적절한 조치를 취하는 것'을 포함하여 '시스템적 위험을 지속적으로 평가(ass)하고 완화할 것'이 요구된다(AI법 전문 114). 또한 '적절한 경우 모델의 전체 수명주기에 걸쳐 모델 및 그 물리적 인프라에 대한 적정 수준의 사이버보안 보호조치를 확보하여야 한다'(AI법 전문 115). 따라서 모델의 '수명주기' 개념은 GPAI-SR모델 제공자의 의무와 관련하여 중요한 의미를 가진다.

- (23) GPAI모델 제공자는 증류(distillation), 양자화(quantisation), 모델 가중치 병합(merging of

model weights) 등의 기법을 통해 반복적이고 상호 연계된 방식으로 개발될 수 있다. 이처럼 모델 개발이 여러 단계에 걸쳐 연속적으로 이루어질 수 있기 때문에, 하나의 모델과 그 수명주기를 명확히 구분하기는 어렵다. 이러한 점을 고려하여 위원회는 '모델'과 그 '수명주기'를 넓은 의미로 이해한다. 실무적으로 위원회는 GPAI모델의 수명주기가 대규모 사전학습 실행이 시작되는 시점부터 개시된다고 본다.[19] 대규모 사전학습 실행(large pre-training run)이란 모델의 일반적 역량을 구축하기 위하여 대량의 데이터를 활용하여 수행되는 기초 학습 실행을 의미한다. 이후 해당 사전학습 실행을 기반으로 이루어지는 후속 개발은 모델의 출시 전후를 불문하고 GPAI모델 제공자 또는 그를 대신하여 수행되는 경우 동일한 모델의 수명주기에 포함되며 새로운 모델을 형성하는 것으로 보지 않는다. 다만 다른 행위자가 해당 모델을 수정하는 경우에는 달리 판단한다(제3.2절 참조). 따라서 모델은 개발·시장 출시·사용에 이르는 전체 수명주기에 걸쳐 동일한 모델로 간주된다. 특히 모델 개발 과정의 서로 다른 단계가 각각 별개의 모델을 구성하는 것으로 보지는 않는다.

- (24) GPAI모델의 수명주기에 관한 이러한 접근방식에 따르면, GPAI모델 제공자의 의무에는 아래와 같은 사항이 적용된다.
  - AI법 제53조(1)(a) 및 (b)에 따라 요구되는 문서는 출시된 각 모델별로 작성되어야 하며, 해당 모델의 전체 수명주기에 걸쳐 최신 상태로 유지되어야 한다.
  - AI법 제53조(1)(c)에 따라 요구되는 저작권 정책은 GPAI모델 제공자의 각 관련 모델의 전체 수명주기에 걸쳐 적용되어야 한다. GPAI모델 제공자는 하나의 정책을 마련하여 자신의 모든 관련 모델에 적용할 수 있다.
  - AI법 제53조(1)(d)에 따라 요구되는 학습에 사용된 콘텐츠의 요약은 GPAI모델 제공자가 출시한 모든 모델에 대하여 작성·공개되어야 한다. 또한 현재 AI사무국이 마련 중인 템플릿은 해당 공개 요약을 어떠한 경우에 업데이트하여야 하는지를 보다 명시할 예정이다.
  - AI법 제55조(1)에 따라 요구되는 시스템적 위험의 평가(ass) 및 완화는 각 모델의 전 수명주기에 걸쳐 지속적으로 수행되어야 한다. 이는 상시적으로 이루어지는 조치뿐만 아니라, 정기적으로 또는 모델 수명주기의 주요 의사결정 이전에 실시되는 보다 포괄적이고 면밀한 조치를 포함할 수 있다. 이러한 조치는 시스템적 위험에 비추어 효과적이고, 적정하며, 비례적이어야 한다. 또한 AI법 제55조(1)에 따라 요구되는 거버넌스 차원의 위험 완화조치는 GPAI-SR모델 제공자의 각 관련 모델의 전 수명주기에 걸쳐 적용되어야 한다. 일부 위험 완화조치의 경우, GPAI-SR모델 제공자는 이를 한 번 마련한 후 자신의 모든 관련 모델에 적용할 수 있다.

19) 대규모 사전학습 실행(large pre-training run)이란 모델의 일반적인 역량을 구축하기 위하여 대량의 데이터를 사용하여 수행되는 모델의 기반이 되는 학습 실행을 의미한다. 이러한 학습 실행은 더 작은 규모의 실험적 학습 실행 이후에 수행될 수 있으며, 이후 특정 목적에 맞게 특화하기 위한 미세조정 또는 그 밖의 사후학습을 통한 성능 개선이 뒤따를 수 있다.

2.3. GPAI모델이 GPAI-SR모델에 해당하는 경우

- (25) GPAI-SR모델은 GPAI모델의 특별한 유형에 해당한다. 이러한 모델의 제공자는 AI법 제52조 및 제55조에 따라 해당 모델이 초래하는 '시스템적 위험'의 평가(ass) 및 완화와 관련된 추가적인 의무를 부담한다. AI법은 '시스템적 위험'을 다음과 같이 정의한다.: 'GPAI모델의 고영향 역량에 특유한 위험으로서, 그 도달 범위로 인하여 또는 공중보건, 안전, 공공안보, 기본권 또는 사회 전체에 대하여 발생하였거나 합리적으로 예견 가능한 부정적 영향으로 인하여 EU 시장에 중대한 영향을 미치며, 가치사슬 전반에 걸쳐 대규모로 확산될 수 있는 위험'(AI법 제3조(65)).

2.3.1. 분류

- (26) AI법 제51조(1)에 따라 GPAI모델은 다음 두 가지 조건 중 어느 하나를 충족하는 경우 GPAI-SR모델로 분류된다.:
  - '고영향 역량'을 보유한 경우, 즉 '가장 발전된 모델에서 기록된 역량에 상응하거나 이를 초과하는 역량'을 보유하고 있는 경우(AI법제3조(64));
  - 위원회가 직권으로 또는 과학자 패널의 적격 경고(qualified alert)에 따라 내린 결정에 근거하여, 부속서 XIII에 규정된 기준을 고려할 때 전 호에서 정한 역량, 즉 가장 발전된 모델에서 기록된 역량에 상응하거나 이를 초과하는 역량과 동등한 수준의 역량 또는 영향을 보유하고 있는 경우.

- (27) GPAI모델이 위 두 가지 조건 중 어느 하나를 충족하는 시점부터 해당 모델은 GPAI-SR모델로 분류되며, GPAI-SR모델 제공자는 관련 의무를 준수하여야 한다.

- (28) 특정 GPAI모델이 고영향 역량을 보유하는지 여부는 AI법 제51조(1)(a)에 따라 적절한 기술적 도구와 방법론, 지표(indicator) 및 벤치마크(benchmark)를 기초로 평가되어야 한다. 이러한 도구와 방법론은 위원회가 위임법령(delegated act)을 채택하여 추가적으로 구체화할 예정이다((28) 참조). 다만 향후 이러한 위임법령이 채택되는지 여부와 관계없이, AI법 제51조(2)는 '부동 소수점 연산(floating-point operations)으로 측정한 누적 학습 연산량이 10^{25}를 초과하는 경우, 해당 GPAI모델은 제1항(a)에 따른 고영향 역량을 보유한 것으로 추정된다'고 규정하고 있다. 이는 GPAI모델의 학습에 사용된 누적 학습 연산량('cumulative training compute', 본 가이드라인에서 해당 개념의 의미는 (116) 참조))을 FLOP으로 측정한 값이 고영향 역량을 식별하기 위한 관련 지표(relevant metric)로 간주되기 때문이다(AI법 전문 111).

- (29) AI법 제51조(3)에 따라 위원회는 AI법 제97조에 따른 위임법령을 채택하여 AI법 제51조(1) 및 (2)에 규정된 임계값을 조정할 권한을 가진다. 또한 위원회는 알고리즘의 개선 또는 하드웨어 효율성

향상과 같은 기술 발전을 반영하여 해당 임계값이 지속적으로 최첨단 기술 수준을 반영하도록 필요한 경우 추가적인 벤치마크 및 지표를 도입할 권한도 가진다.

2.3.2. 통지

- (30) GPAI모델이 고영향 역량[20]을 보유한 것으로 추정되는 요건을 충족하였거나 해당 GPAI모델이 그 요건을 충족하게 될 것임이 알려진 경우, GPAI모델 제공자는 AI법 제52조(1)에 따라 위원회에 이를 통지하여야 한다(AI법 전문 112 참조). 이러한 통지는 '그 요건이 충족된 후 또는 충족될 것임이 알려진 후 지체 없이, 어떠한 경우에도 2주 이내'에 이루어져야 한다(AI법 제52조(1)).

- (31) 특히, GPAI모델 제공자가 해당 모델이 고영향 역량을 보유한 것으로 추정되는 요건을 충족할 가능성이 높다고 합리적으로 예견할 수 있는 경우에는 학습이 완료되기 전이라도 통지가 요구될 수 있다. 이와 관련하여 AI법 전문 112는 'GPAI모델의 학습에는 연산 자원의 사전 할당을 포함한 상당한 수준의 계획이 수반되므로, GPAI모델 제공자는 학습이 완료되기 전에도 해당 모델이 관련 임계값을 충족하게 될지를 알 수 있다'고 설명하고 있다. '계획' 및 '연산 자원의 사전 할당'은 대규모 사전학습 실행이 시작되기 전에 이루어지므로, GPAI모델 제공자는 해당 실행을 시작하기 전에 자신이 사용할 누적 학습 연산량을 추정하여야 한다(학습 연산량의 추정 방법은 본 가이드라인 부록 A.1 및 A.2 참조). 이러한 추정을 바탕으로 GPAI모델 제공자는 다음과 같이 조치하여야 한다.
  - 추정값이 AI법 제51조(2)에 규정된 임계값을 충족하는 경우, GPAI모델 제공자는 AI법 제52조(1)에 따라 '지체 없이, 그리고 어떠한 경우에도 2주 이내에' 위원회에 통지하여야 한다.
  - 추정값이 AI법 제51조(2)에 규정된 임계값을 충족하지 않는 경우에도, GPAI모델 제공자는 학습 과정 전반 및 해당 GPAI모델의 전체 수명주기에 걸쳐 실제 및 예상 연산량 사용 현황을 면밀히 모니터링하여야 한다. 이는 GPAI모델 제공자가 누적 학습 연산량이 해당 임계값을 충족하였거나 충족하게 될 시점을 파악하고, 이에 따라 AI법 제52조(1)에 따라 위원회에 통지할 수 있도록 하기 위한 것이다.

- (32) AI법 제52조(1)에 따라 통지에는 '해당 요건이 충족되었음을 입증하는데 필요한 정보'가 포함되어야 한다. 해당 모델이 AI법 제51조(2)에 규정된 임계값을 충족하였거나 충족할 예정이어서 통지를 하게 된 경우, 또는 해당 모델이 (60)에서 정한 임계값을 충족하였거나 충족하게 될 고영향 역량을 보유한 GPAI모델을 수정한 결과인 경우[21], 해당 정보에는 특히 다음 사항이 포함되어야 한다.

---

20) AI법 제51조(1) 및 (2)에 열거된 임계값을 변경하거나 벤치마크 및 지표를 보완하는 위임법령이 없는 경우, 이에 해당할 수 있는 요건은 해당 모델이 AI법 제52조(1)에서 정한 임계값을 충족하거나 충족할 예정이라는 요건과, 해당 모델이 제60항에서 정한 임계값을 충족하거나 충족할 예정인 고영향 역량을 보유한 GPAI모델을 수정한 결과라는 요건뿐이다.

21) AI법 제51조(1) 및 (2)에 열거된 임계값을 변경하거나 벤치마크 및 지표를 보완하는 위임법령이 없는 경우, AI법 제52조(1)에 따라 위원회에 통지하여야 하는 의무를 발생시키는 요건은 이 두 가지뿐이다.

- GPAI모델 제공자가 추정한 연산량으로서 통지 의무를 발생시킨 연산량. 해당 연산량은 FLOP 단위로, 유효숫자 두 자리(예: 2.3×10^{25} FLOP)로 보고하여야 한다.;
- 해당 연산량을 추정하기 위하여 사용한 접근 방법에 대한 설명. 여기에는 정확한 정보를 이용할 수 없는 경우 근사치를 산정하기 위하여 사용한 방법도 포함되어야 한다.

2.3.3. 분류에 대한 이의 제기 절차

- (33) GPAI모델 제공자가 AI법 제52조(1)에 따라 위원회에 통지하는 경우, 해당 GPAI모델 제공자는 비록 AI법 제51조(2)의 임계값을 충족하더라도, 해당 GPAI모델이 그 구체적인 특성으로 인하여 예외적으로 시스템적 위험을 초래하지 않으며, 따라서 GPAI-SR모델로 분류되어서는 안 된다는 점을 입증할 수 있는 충분한 근거를 통지와 함께 제시할 수 있다(AI법 제52조(2)). 이는 AI법 제51조(2)의 임계값을 충족하거나 고영향 역량을 보유한 GPAI모델이 일반적으로 시스템적 위험을 수반하는 것으로 간주되더라도, 예외적으로 그렇지 않은 경우가 있을 수 있음을 인정하는 것이다.

- (34) GPAI모델 제공자는 자신의 모델이 시스템적 위험을 초래하지 않거나 초래하지 않을 것임을 입증하기 위하여, 해당 모델이 고영향 역량을 보유하지 않거나 보유하지 않을 것이라는 점을 입증하는 주장을 제시할 수 있다. 여기서 고영향 역량이란 '가장 발전된 모델'에서 확인된 역량에 상응하거나 이를 초과하는 역량'을 의미한다(AI법 제3조(64)). 이는 시스템적 위험이 '고영향 역량에 특유한 위험'이기 때문이다(AI법 제3조(65)). 이러한 접근은 모델이 고영향 역량을 보유한 것으로 추정되어 통지가 이루어진 경우, 예를 들어 AI법 제51조(2)의 임계값을 충족하였거나 충족할 것으로 예상되는 경우에 적절하다. 이는 GPAI모델 제공자가 해당 추정이 자신의 모델에 대해서는 타당하지 않다고 판단하는 경우에 적용된다. 또한 위원회는 정량적 임계값의 충족으로부터 발생하는 추정이 해당 모델에 적용되지 않아야 한다는 점에 대한 입증책임은 GPAI모델 제공자에게 있다고 본다.

- (35) GPAI모델 제공자는 자사 모델이 고영향 역량을 보유하지 않거나 보유하지 않을 것임을 입증하기 위하여, 통지 시점에 이용 가능한 해당 모델의 실제 또는 예상 역량에 관한 정보를 제출하여야 한다. 이러한 정보에는 실제 또는 예측된 벤치마크 결과(예: 스케일링 분석(scaling analyses)[22]에 기반한 결과)가 포함될 수 있다. 또한 위원회가 해당 모델이 고영향 역량을 보유하지 않거나 보유하지 않을 것인지, 나아가 시스템적 위험을 초래하지 않는지 여부를 적절히 평가(ass)할 수 있도록 GPAI모델 제공자는 모델의 역량에 영향을 미칠 수 있는 추가적인 정보도 함께 제출하는 것이 권장된다. 이러한 정보에는 모델 아키텍처, 파라미터 수, 학습 예시 수, 데이터 선별 및 처리 기법, 학습 기법, 입력 및 출력 모달리티, 예상되는 도구 활용 방식, 예상 컨텍스트(context) 길이 등이 포함될 수 있다.

22) Scaling analysis, 모델 규모, 학습 데이터, 연산량 같은 확장 가능한 요소가 커질 때 성능·능력·비용·위험이 어떻게 변하는지를 분석하는 방법

- (36) GPAI모델 제공자가 해당 모델이 시스템적 위험을 초래하지 않는다는 점에 관한 주장을 포함한 통지를 제출한 경우, 위원회는 위원회 절차규칙(Rules of Procedure of the European Commission)[23] 및 EU법의 일반 원칙에 따라 해당 주장을 평가(ass)하고 그 수용 여부를 결정한다. 특히 위원회가 해당 주장을 배척하려는 경우에는 EU 기본권 헌장(Charter of Fundamental Rights of the European Union) 제41조에 따라, GPAI모델 제공자는 결정이 내려지기 전에 의견을 진술할 권리를 가지며(결정서 초안을 사전에 송부받는 방식), 기밀 유지 및 직무상·영업상 비밀 보호에 관한 정당한 이익을 존중하는 범위에서 자신의 사건기록을 열람할 권리도 가진다. 또한 위원회의 결정에는 AI법 제52조(2)에 따라 GPAI모델 제공자가 제출한 주장을 수용하거나 배척한 이유가 포함되어야 한다.

- (37) GPAI모델 제공자가 모델이 고영향 역량을 보유하고 있다는 추정을 번복하여 해당 모델이 시스템적 위험을 초래하지 않는다는 점을 입증하기 위한 주장을 제시한 경우, 위원회는 그 주장이 이러한 추정을 명백히 뒤집을 수 있을 정도로 충분한 근거를 제시하고 있는지 여부를 평가(ass)한다.

- (38) 고영향 역량은 정의상 '가장 발전된 모델(the most advanced models)에서 확인된 역량에 상응하거나 이를 초과하는 역량'을 의미한다(AI법 제3조(64)). 또한 이러한 역량은 GPAI모델이 GPAI-SR모델로 분류되는지를 결정하는 기준이 된다. 따라서 모델이 '고영향 역량'을 보유한 것으로 판단되는 수준, 다시 말해 '가장 발전된 모델'에 해당하는 수준은 역량이 향상되고 위험이 변화함에 따라 시간의 경과에 따라 달라질 것으로 예상된다. 특히 위원회는 이를 고정된 역량 수준을 의미하는 것으로 보지 않는다.

- (39) 위원회는 통지 시점에 해당 모델이 '가장 발전된 모델'에 해당하는지를 평가(ass)할 때, 해당 모델의 누적 학습 연산량이 이를 판단하는 데 어느 정도 참고가 되는지를 고려한다. 이와 관련하여, 해당 모델의 누적 학습 연산량이 AI법 제51조(2)에서 정한 임계값을 어느 정도 초과하는지도 함께 고려된다. 또한 위원회는 누적 학습 연산량 외에도 모델의 실제 또는 예상 역량에 영향을 미치는 다른 요소들을 종합적으로 평가(ass)하며, 여기에는 실제 또는 예측된 벤치마크 점수도 포함된다.

- (40) GPAI모델 제공자가 해당 모델이 고영향 역량을 보유하고 있음에도 시스템적 위험을 초래하지 않는다고 주장하는 경우, 위원회는 이미 시행되었거나 향후 시행 예정인 완화조치를 근거로 해당 모델을 GPAI-SR모델 분류 대상에서 제외하는 것은 적절하지 않다고 본다. 이 경우에도 해당 모델은 여전히 시스템적 위험을 수반하는 것으로 보며, 그러한 위험은 지속적으로 평가(ass)되고 완화되어야 한다. 다만 GPAI모델 제공자는 이러한 완화조치를 AI법 제55조(1)에 따른 시스템적 위험 완화를 수행하는 과정에서 활용할 수 있다.

- (41) 위원회가 제공자의 주장을 수용하거나 배척한 결정은 아래와 같은 법적 효과를 가진다.

---

23) 위원회 절차규칙(Rules of Procedure of the European Commission) | EUR-Lex

- 위원회가 해당 주장을 받아들이지 않는 경우, 해당 모델은 GPAI-SR모델로 확정된다. 이 경우 GPAI모델 제공자는 해당 모델이 AI법 제51조(1)(a)의 요건을 충족하는 시점부터 GPAI-SR모델 제공자에 대한 의무를 부담한다. 특히 통지와 함께 주장을 제출하였다는 사실만으로는 GPAI-SR모델 제공자에 대한 의무의 적용이 유예되지 않는다.

- (42) 위원회가 해당 주장을 받아들이는 경우, 해당 GPAI모델은 더 이상 GPAI-SR모델로 분류되지 않는다. GPAI모델 제공자는 수용 결정이 통지된 시점부터 GPAI-SR모델 제공자에 대한 의무를 더 이상 부담하지 않는다. 다만 수용 결정의 기초가 된 사실에 중대한 변경이 발생하거나 그러한 변경이 발생할 것이 확인된 경우, 또는 수용 결정이 중대하게 불완전·부정확·오해의 소지가 있는 정보에 기초한 것으로 밝혀진 경우에는 GPAI모델 제공자는 위원회에 다시 통지하여야 한다. 이 경우 GPAI모델 제공자는 관련 정보를 모두 제출하여야 하며, 위원회는 수용 결정을 배척 결정으로 변경할 수 있다. 이 경우 해당 모델은 GPAI-SR모델로 분류된다. 또한 수용 결정이 이루어졌더라도, 이후 AI법 부속서 XIII의 기준에 따라 위원회가 해당 모델을 별도로 지정하는 경우에는 GPAI-SR모델로 다시 분류될 수 있다(제2.3.4절 참조).

2.3.4. 지정

- (43) 앞서 살펴본 바와 같이, GPAI모델은 고영향 역량을 보유한 경우 GPAI-SR모델로 분류된다. 이와 별도로, AI법은 GPAI모델을 GPAI-SR모델로 지정할 수 있는 메커니즘도 규정하고 있다.

- (44) 구체적으로, AI법 제52조(1), 제51조(1)(b) 및 제52조(4)에 따라 위원회는 직권으로 또는 AI법 제90조(1)(a)에 따른 과학자 패널이 제기한 공식 경고에 근거하여 GPAI모델을 GPAI-SR모델로 지정할 수 있다.

- (45) GPAI모델의 지정은 다음의 경우에 이루어질 수 있다.:
  - AI법 제52조(4)에 따라 위원회가 AI법 부속서 XIII에 규정된 기준에 근거하여 해당 모델이 고영향 역량에 상응하는 역량 또는 영향을 보유한다고 판단하는 경우, 또는
  - AI법 제52조(1)에 따라, AI법 제51조(1)(a)의 요건을 충족하는 GPAI모델의 제공자가 AI법 제52조(1)에 따른 통지 의무를 이행하지 않은 경우. 이 경우 해당 GPAI모델 제공자에게는 AI법 제101조에 따라 과징금이 부과될 수 있다.

- (46) 위 두 경우 모두에서, GPAI모델 제공자는 해당 모델이 GPAI-SR모델로 분류되는 시점부터 GPAI-SR모델 제공자에 대한 의무를 준수하여야 한다. 첫 번째 경우에는 지정 결정이 통지된 시점이 기준이 되며, 두 번째 경우에는 해당 모델이 AI법 제51조(1)(a)의 요건을 충족한 시점이 기준이 된다.

2.3.5. 지정에 대한 이의 제기 절차

- (47) AI법 제52조(4)에 따라 위원회에 의해 GPAI-SR모델로 지정된 GPAI모델 제공자는 AI법 제52조(5)에 근거하여 지정 이후 최소 6개월이 경과한 시점부터 해당 지정에 대한 재평가(reassess)를 요청할 수 있다. 이 경우 GPAI모델 제공자는 해당 모델이 더 이상 시스템적 위험을 초래하지 않는다는 점을 입증하기 위하여, '지정 결정 이후 새롭게 발생한 객관적이고 구체적인 사유'를 제시하여야 한다(AI법 제52조(5)). 위원회가 재평가(reass) 이후에도 기존 지정을 유지하는 경우, GPAI모델 제공자는 해당 결정 이후 최소 6개월이 경과한 시점부터 다시 재평가(reass)를 요청할 수 있다.

3 GPAI 모델을 시장에 출시하는 제공자

- (48) 본 절은 GPAI모델의 '제공자(provider)'와 '시장 출시(placing on the market)' 개념을 중심으로 다룬다. 이 절의 목적은 AI 가치사슬에 참여하는 행위자가 어떠한 경우에 AI법상 GPAI모델 제공자 의무를 부담하게 되는지를 명확히 하는 데 있다. 먼저 사례를 통해 위원회가 누구를 GPAI모델의 '제공자'로 보는지, 그리고 GPAI모델이 언제 '시장에 출시된 것'으로 간주되는지를 설명한다(제3.1절). 이어서 GPAI모델이 AI시스템에 통합되는 경우를 다루고(제3.2절), 마지막으로 GPAI모델을 수정하는 행위자가 어떠한 경우 GPAI모델 제공자로 간주되어 관련 의무를 부담하게 되는지를 설명한다(제3.3절).

3.1. 행위자가 GPAI모델을 출시하는 제공자에 해당하는 경우

- (49) AI법 제3조(3)은 GPAI모델의 '제공자'를 'GPAI모델을 개발하거나 … GPAI모델의 개발을 위탁하고, 이를 자신의 이름 또는 상표로 유상 또는 무상 여부를 불문하고 시장에 출시하는 자연인·법인·공공기관·기관 또는 기타 단체'로 정의한다. 한편, AI법 제3조(9)는 GPAI모델의 '시장 출시'를 'EU 시장에 … GPAI모델을 최초로 공급하는 것'으로 정의하고, AI법 제3조(10)은 '시장 공급'을 '상업적 활동의 일환으로, 유상 또는 무상 여부를 불문하고, … GPAI모델을 EU 시장에서 유통 또는 사용을 위하여 공급하는 것'으로 정의한다. 아울러 AI법 전문 97은 'GPAI모델은 라이브러리, Application Programming Interfaces(APIs), 직접 다운로드 또는 물리적 복제본 등의 다양한 방식으로 시장에 출시될 수 있다'고 설명하고 있다.

- (50) AI법 제2조(1)(a)에 따르면, GPAI모델을 EU 시장에 출시하는 행위자는 해당 GPAI모델의 제공자가 될 수 있으며, 'EU 내에 설립되거나 위치한 경우인지 또는 제3국에 설립되거나 위치한 경우인지와 관계없이' AI법상 GPAI모델 제공자의 의무를 부담한다. 또한 제3국에 설립되거나 위치한 GPAI모델 제공자는 EU 시장에 모델을 출시하기 전에 EU 내에 소재한 권한을 위임받은 대리인(authorised representative)을 지정하여야 한다(AI법 제54조).

3.1.1. GPAI모델 제공자의 예시

- (51) 다음은 다양한 상황에서 GPAI모델의 제공자로 간주되는 행위자에 관한 예시를 제시한다.:
  - 행위자 A가 GPAI모델을 개발하고 이를 출시하는 경우, 행위자 A는 제공자에 해당한다.
  - 행위자 A가 행위자 B로 하여금 자신을 대신하여 GPAI모델을 개발하게 하고, 행위자 A가 해당 모델을 출시하는 경우, 행위자 A가 제공자에 해당한다.
  - 행위자 A가 GPAI모델을 개발하고 이를 행위자 C가 운영하는 온라인 저장소(repository)에

업로드하는 경우, 행위자 A가 제공자에 해당한다.

- 협업체 또는 컨소시엄이 서로 다른 개인이나 조직으로 하여금 자신들을 위하여 GPAI모델을 개발하게 하고, 해당 모델을 출시하는 경우에는 일반적으로 해당 협업체 또는 컨소시엄의 조정자가 제공자가 된다. 다만 협업체 또는 컨소시엄 자체가 제공자가 될 수도 있다. 이러한 사항은 사안별로 평가(ass)되어야 한다.

- (52) 위 사례는 AI법의 관련 규정(특히 AI법 제3조(3))에 따라 해석되어야 하며, 사안별로 평가(ass)되어야 한다.

3.1.2. GPAI모델의 시장 출시 사례

- (53) 다음은 GPAI모델이 언제 시장에 출시된 것으로 간주되는지에 관한 예시로서, AI법 전문 97에서 제시된 사례를 토대로 한다.:
  - GPAI모델이 소프트웨어 라이브러리(library) 또는 패키지를 통해 EU 시장에 최초로 제공되는 경우;
  - GPAI모델이 Application Programming Interface(API)를 통해 EU 시장에 최초로 제공되는 경우;
  - GPAI모델이 공공 카탈로그(catalogue), 허브(hub) 또는 저장소(repository)에 최초로 업로드되어 EU 시장에 직접 다운로드가 가능해지는 경우;
  - GPAI모델이 물리적 복제본의 형태로 EU 시장에 최초로 제공되는 경우;
  - GPAI모델이 클라우드 컴퓨팅 서비스를 통해 EU 시장에 최초로 제공되는 경우;
  - GPAI모델이 고객의 자체 인프라에 복제되는 방식으로 EU 시장에 최초로 제공되는 경우;
  - GPAI모델이 웹 인터페이스를 통해 EU 시장에 최초로 제공되는 챗봇에 통합되는 경우;
  - GPAI모델이 앱 스토어를 통해 EU 시장에 최초로 제공되는 모바일 애플리케이션에 통합되는 경우;
  - GPAI모델이 제3자에게 제품 또는 서비스를 제공하는 데 필수적인 내부 프로세스에 사용되거나, EU 내 자연인의 권리에 영향을 미치는 내부 프로세스에 사용되는 경우.

- (54) 위 예시는 블루 가이드(Blue Guide)[24]에서 제시된 '출시(placing on the market)' 개념에 관한 일반적 해석지침과 AI법의 관련 규정(AI법 제3조(9) 및 (10), 그리고 전문 97)에 따라 해석되어야 하며, 사안별로 평가(ass)되어야 한다.

---

24) 위원회 고시, 「EU 제품 규정의 이행에 관한 '블루 가이드(The Blue Guide)' 2022」(EEA 관련 문서)(2022/C 247/01)

3.1.3. AI시스템에 통합된 GPAI모델 관련 고려사항

- (55) AI법 전문 97에 따르면, 'AI모델은 통상적으로 AI시스템에 통합되거나 그 일부를 구성한다'. 또한 'GPAI모델 및 시스템적 위험을 초래하는 GPAI모델에 관한 규정은 … 해당 모델이 AI시스템에 통합되거나 그 일부를 구성하는 경우에도 적용되어야 한다'. 이에 따라 다음 문단에서는 AI시스템에 통합된 GPAI모델이 시장에 출시된 것으로 간주되는 특수한 사례를 설명한다(제3.1.2절의 예시에 추가되는 내용임).

- (56) 첫째, AI법 전문 97은 'GPAI모델 제공자가 자신의 모델을 시장에 공급하거나 서비스를 개시한 자신의 AI시스템에 통합하는 경우, 해당 모델은 시장에 출시된 것으로 간주되어야 하며, 따라서 이 규정에 따른 모델 관련 의무는 AI시스템 관련 의무와 별도로 계속 적용되어야 한다'고 설명한다.

- (57) 둘째, 업스트림 행위자(upstream actor)가 GPAI모델을 개발하거나 GPAI모델의 개발을 위탁하고, 해당 모델을 EU 시장의 다운스트림 행위자(downstream actor)에게 최초로 공급하는 경우, 해당 모델은 시장에 출시된 것으로 간주되며 업스트림 행위자가 해당 모델의 제공자에 해당한다. 따라서 업스트림 행위자는 GPAI모델 제공자에 대한 의무를 준수해야 한다. 한편, 해당 모델을 AI시스템에 통합하고 그 시스템을 EU 시장에 출시하거나 EU 내에서 서비스를 개시하는 다운스트림 행위자는 AI시스템의 제공자에 해당할 수 있다. 이 경우 다운스트림 행위자는 AI법에 따른 AI시스템 관련 요구사항 및 의무를 준수해야 한다.

- (58) 셋째, 업스트림 행위자(upstream actor)가 GPAI모델을 개발하거나 GPAI모델의 개발을 위탁한 후, 해당 모델을 EU 시장 외의 다운스트림 행위자(downstream actor)에게 최초로 공급하고, 그 다운스트림 행위자가 해당 모델을 AI시스템에 통합하여 그 AI시스템을 EU 시장에 출시하거나 EU 내에서 서비스를 개시하는 경우에는 다음과 같은 사항이 적용된다.
  - EU 내에서 사용되는 모든 GPAI모델이 AI법의 요구사항을 준수하도록 하기 위하여, 해당 모델이 통합된 AI시스템이 EU 시장에 출시되거나 EU 내에서 서비스를 개시하는 시점에 그 GPAI모델도 EU 시장에 출시된 것으로 간주되어야 한다.

- (59) 이 경우 업스트림 행위자는 해당 모델의 제공자로 간주된다. 다만, 업스트림 행위자가 EU 시장[25]에서의 모델의 유통(distribution) 및 사용을 명확하고 분명한 방식으로 배제한 경우에는 예외로 한다. 여기에는 EU 시장에 출시되거나 EU 내에서 서비스가 개시될 예정인 AI시스템에 통합되는 경우도 포함된다. 업스트림 행위자가 이와 같은 배제를 명확히 한 경우에는 GPAI모델을 AI시스템에 통합하여 그 AI시스템을 EU 시장에 출시하거나 EU 내에서 서비스를 개시하는 다운스트림 행위자가 GPAI모델의 제공자로 간주된다.

---

25) AI법 제3조(10)의 '시장 공급(make available on the market)' 정의 참조

3.2. GPAI모델 제공자로서의 다운스트림 수정자

- (60) AI법 전문 97에 따르면, '[GPAI] 모델은 수정되거나 … 미세조정(fine-tuned)되어 새로운 모델로 변형될 수 있다.'[26] 특히, 다운스트림 행위자(해당 모델의 기존 제공자와 구별되며, 그를 대신하여 행위하지 않는 자)는 GPAI모델을 수정[27]할 수 있다(이를 AI시스템에 통합하는 경우와 통합하지 않는 경우를 모두 포함).[28] 다만 AI법은 다운스트림 수정자가 어떠한 경우 수정된 GPAI모델의 제공자로 간주되어야 하는지에 관한 기준을 명시하고 있지 않다.

- (61) 위원회는 GPAI모델에 대한 모든 수정이 곧바로 해당 수정자를 수정된 GPAI모델의 제공자로 간주하지 않는다. 이는 블루 가이드에서 "본래의 성능, 목적 또는 유형을 변경하기 위한 중요한 변경 또는 개조가 이루어진 경우, 해당 제품은 새로운 제품으로 간주될 수 있다"고 설명하는 내용과도 일치한다.

- (62) 이와 달리, 위원회는 수정이 모델의 범용성, 역량[29] 또는 시스템적 위험에 중대한 변화를 초래하는 경우에 한하여 다운스트림 수정자를 수정된 GPAI모델의 제공자로 간주한다.

- (63) 위와 같은 고려사항을 바탕으로, 다운스트림 수정자가 GPAI모델의 제공자로 간주되는지 여부에 관한 참고 기준은 해당 수정에 사용된 학습 연산량이 기존 모델의 학습 연산량의 3분의 1을 초과하는지 여부이다('학습 연산량'의 의미에 대해서는 본 가이드라인 제115항 참조).

- (64) 다만 다운스트림 수정자가 기존 모델의 학습 연산량 값을 알 수 없고(예: 기존 모델의 제공자가 해당 값을 제공하지 않은 경우), 이를 추정할 수도 없는 경우에는(학습 연산량 추정 방법은 본 가이드라인 부속서 A.1 및 A.2 참조), 해당 기준은 다음과 같이 대체된다. 기존 모델이 GPAI-SR모델인 경우에는 모델이 고영향 역량을 보유한 것으로 추정되는 기준값(즉, 현재 10^{25} FLOP, AI법 제51조(2) 참조)의 3분의 1이 적용된다. 그 밖의 경우에는 모델이 GPAI모델로 추정되는 기준값(즉, 현재 10^{23} FLOP, 제2.1절 참조)의 3분의 1이 적용된다.

- (65) 이 기준값은 해당 수준의 연산량을 사용한 수정이 모델에 중대한 변화를 초래하여, 다운스트림 수정자가 GPAI모델 제공자 또는 경우에 따라 GPAI-SR모델 제공자에 대한 의무를 부담하게 될 것이라는 점을 전제로 한다. 구체적으로는 다음과 같은 사정이 고려된다.:
  - 모델 특성의 변화는 위원회, 국가 관할 당국 또는 다운스트림 제공자에게 중요한 의미를 가질 수 있으며, 이는 다운스트림 수정자가 AI법 제53조(1)(a) 및 (b)에 따른 투명성 의무를 부담하는 근거가 된다.

---

26) 위원회는 '미세조정(fine-tuning)'을 GPAI모델에 대한 '수정(modifying)'의 한 형태로 본다.

27) downstream modification, 타인이 만든 AI모델을 후속 사용자나 개발자가 수정·재학습하거나 기능을 추가하는 행위

28) GPAI모델을 수정한 주체가 최초 제공자인지 또는 다운스트림 행위자인지는 사안별로 판단하여야 한다. 이때 중요한 고려 요소 중 하나는 해당 모델의 가중치에 대한 통제권을 누가 보유하고 있는지 여부이며, 예를 들어 API를 통한 미세조정의 경우 이러한 요소가 중요한 고려 요소가 될 수 있다.

29) model capability, AI가 특정 작업을 수행할 수 있는 기능적 범위를 말하며, 단순 추론 역량부터 복잡한 추론·계획까지 포함함

- 해당 기준값을 충족하는 수정은 상당한 양의 데이터를 사용하였을 것으로 예상되며, 이는 AI법 제53조(1)(c) 및 (d)에 따른 의무, 즉 저작권 정책 및 학습에 사용된 콘텐츠에 관한 공개 요약과 관련된다.
- 기존 모델이 GPAI-SR모델인 경우, 수정된 모델은 기존 모델과 비교하여 현저히 다른 시스템적 위험을 초래할 가능성이 있다. 특히 위원회는 기존 제공자가 자신의 시스템적 위험 평가(ass) 및 완화 과정에서 이러한 수정으로 인한 시스템적 위험의 변화를 합리적으로 예견하기는 어렵다고 본다. 이는 다운스트림 수정자가 AI법 제55조에 따른 의무를 부담하게 되는 근거가 된다.

- (66) 이 임계값은 기존 모델의 학습에 사용된 연산량을 기준으로 하는 상대적 기준이다. 이는 (16)에서 설명된 바와 같이, 모델 학습에 사용되는 연산량이 일반적으로 모델의 파라미터 수와 학습 예시 수를 곱한 값에 비례하기 때문이다. 모델 수정은 일반적으로 일정량의 추가 데이터를 사용한 학습을 수반하므로, 동일한 양의 데이터를 사용하더라도 대형 모델은 소형 모델보다 더 많은 연산량을 필요로 한다. 다시 말해, 특정한 수정을 수행하는 데 필요한 연산량은 일반적으로 모델의 파라미터 수에 비례한다. 따라서 어떤 수정이 다운스트림 수정자를 제공자로 보게 하는지를 판단하기 위한 상대적 임계값을 설정하는 것이 타당하다. 이러한 접근은 모델의 파라미터 수와 관계없이 모든 모델을 동일하게 취급하기 위한 것이다. 특히 파라미터 수가 적은 모델은 임계값도 더 낮지만 수정에 필요한 연산량 역시 더 적으므로, 이러한 기준이 파라미터 수가 적은 모델을 수정하는 것이 상대적으로 불리해지는 결과를 초래해서는 안 된다.
- (67) 현재로서는 제60항에서 제시된 기준을 충족하는 수정 사례가 많지 않을 수 있으나, 모델 수정에 사용되는 연산량이 증가함에 따라 GPAI모델의 제공자로 간주되는 다운스트림 수정자의 수는 시간이 지나면서 증가할 수 있다. 따라서 이 기준은 주로 미래를 염두에 둔 것이며, AI법의 위험 기반 접근과도 부합한다. 이에 따라 기술과 시장의 발전에 따라 위원회의 접근 방식도 향후 변경될 수 있다.

3.2.1. 다운스트림 수정자가 GPAI모델 제공자에 해당하는 시기

- (68) AI법 전문 109에 따르면, “모델의 수정 또는 미세조정(fine-tuning)이 이루어진 경우, GPAI모델 제공자에 대한 의무는 해당 수정 또는 미세조정에 한정되어야 하며, 예를 들어 새로운 학습 데이터 출처를 포함한 수정 사항에 관한 정보를 기존 기술문서에 보완하는 방식으로, 이 규정에서 정한 가치사슬 관련 의무를 준수할 수 있다.” 즉, AI법 제53조(1)(a) 및 (b)에 따른 문서화 의무는 수정에 관한 정보로 한정되며, AI법 제53조(1)(c)에 따른 저작권 정책 및 AI법 제53조(1)(d)에 따른 학습에 사용된 콘텐츠의 요약은 수정 과정에서 사용된 데이터로 한정된다.
- (69) GPAI모델의 제공자가 된 다운스트림 수정자는 AI법 제54조에 따른 의무도 준수하여야 한다.

3.2.2. 다운스트림 수정자가 GPAI-SR모델 제공자에 해당하는 시기

- (70) 다운스트림 행위자가 GPAI-SR모델로 분류된 GPAI모델을 수정하여 수정된 GPAI모델의 제공자가 되는 경우, 해당 수정된 모델 역시 고영향 역량을 보유한 것으로 추정된다. 따라서 해당 모델은 AI법 제51조(1)(a)에 따라 GPAI-SR모델로 간주된다.
- (71) 이 경우 다운스트림 수정자는 GPAI-SR모델 제공자에 대한 의무를 준수하여야 한다. 특히 GPAI-SR모델 제공자는 AI법 제52조(1)에 따라 제31항에서 정한 정보를 포함하여 위원회에 통지하여야 한다.

4 오픈소스로 공개된 일부 모델에 대한 특정 의무의 면제

- (72) 원칙적으로, 모든 GPAI모델 제공자는 AI법 제53조 및 제54조에서 정한 의무를 준수하여야 한다. 다만 AI법 제53조(2) 및 제54조(6)은 '모델에 대한 접근, 사용, 수정 및 배포(distribution)를 허용하는 자유·오픈소스 라이선스로 공개되고, 가중치(weights)를 포함한 파라미터, 모델 아키텍처에 관한 정보 및 모델 사용에 관한 정보가 공개적으로 제공되는 AI모델의 제공자'에 대하여 일부 의무 면제를 규정하고 있다. 다만 해당 모델이 GPAI-SR모델에 해당하지 않는 경우에 한한다. 본 절에서는 먼저 이러한 면제가 적용되는 의무를 설명하고(제4.1절), 이어서 이러한 면제가 적용되기 위하여 충족되어야 하는 요건을 설명한다(제4.2절).

4.1. 면제의 범위

- (73) 제4.2절에서 정한 요건을 충족하는 제공자에 대해서는 해당 모델이 GPAI-SR모델에 해당하지 않는 한 다음 의무가 적용되지 않는다.:
  - AI법 제53조(1)(a): 'AI사무국 및 국가 관할 당국의 요청이 있는 경우 이를 제공하기 위한 목적으로, 최소한 부속서 XI에 규정된 정보를 포함하여 모델의 학습 및 시험 과정과 평가 결과를 포함한 기술문서를 작성하고 최신 상태로 유지할 의무';
  - AI법 제53조(1)(b): 'GPAI모델을 자신의 AI시스템에 통합하려는 AI시스템 제공자에게 제공하기 위한 정보 및 문서를 작성하고 최신 상태로 유지하며 이를 제공할 의무';
  - AI법 제54조: 제3국에 설립된 GPAI모델 제공자가 EU 내에 소재한 권한을 위임받은 대리인을 지정하여야 하는 의무.

- (74) 이러한 면제의 근거는 두 가지이다. 첫째, 자유·오픈소스 라이선스 하에 공개된 모델은 "시장 내 연구 및 혁신에 기여할 수 있으며, EU 경제에 상당한 성장 기회를 제공할 수 있다"는 점이다(AI법 전문 102). 둘째, '가중치를 포함한 파라미터, 모델 아키텍처에 관한 정보 및 모델 사용에 관한 정보가 공개적으로 제공되는 경우, 자유·오픈소스 라이선스 하에 공개된 GPAI모델은 높은 수준의 투명성과 개방성을 보장하는 것으로 볼 수 있다'는 점이다(AI법 전문 102). 이러한 점은 특히 투명성 관련 의무에 대한 면제를 정당화한다. 다만 해당 모델이 GPAI-SR모델에 해당하는 경우에는 자유·오픈소스 라이선스 하에 공개되었다는 사실만으로 AI법에 따른 이러한 의무가 면제되지는 않는다.

- (75) 제4.2절에서 설명한 요건을 충족하는 GPAI모델이 자유·오픈소스 라이선스 하에 공개되더라도, 모델의 학습 또는 수정에 사용된 데이터나 저작권법 준수 여부에 관한 실질적인 정보가 반드시 공개되는 것은 아니다. 이러한 이유로, 제4.2절에서 설명한 요건을 충족하는 GPAI모델 제공자라 하더라도, EU

저작권법 준수를 위한 정책을 마련하여야 하는 AI법 제53조(1)(c)에 따른 의무는 면제되지 않는다. 이 의무에는 유럽의회 및 이사회 Directive (EU) 2019/790 제4조(3)에 따른 권리 유보(reservation of rights)를 식별하고 이를 준수하는 것이 포함된다. 같은 이유로, 학습에 사용된 콘텐츠에 관한 요약을 작성하여야 하는 AI법 제53조(1)(d)에 따른 의무 역시 면제되지 않는다.

4.2. 면제 적용 요건

4.2.1. 라이선스 관련 요건

- (76) 제4.1절에서 언급된 면제가 적용되기 위해서는 GPAI모델이 특히 '모델에 대한 접근, 사용, 수정 및 배포(distribution)를 허용하는 자유·오픈소스 라이선스' 하에 공개되어야 한다(AI법 제53조(2) 및 제54조(6)).
- (77) 이 맥락에서 '라이선스(licence)'란 특정한 조건 및 기준에 따라 모델과 관련된 권한을 부여하는 것을 의미한다. 또한 '자유·오픈소스'는 저작권을 활용하여 모델의 광범위한 배포를 가능하게 하고 추가적인 발전을 촉진하는 라이선스 형태로 이해되어야 한다.
- (78) AI법 전문 102는 해당 라이선스가 모델을 '공개적으로 공유'할 수 있도록 하여야 하며, 이용자가 모델 '또는 그 수정된 버전'에 대하여 '자유롭게 접근, 사용, 수정 및 재배포(redistribute)'할 수 있어야 함을 명확히 한다. 이러한 권리(즉, 접근·사용·수정 및 재배포(redistribute) 권한) 중 하나라도 결여된 경우, 해당 라이선스는 AI법상 자유·오픈소스 라이선스로 간주될 수 없다. 이 경우 제공자는 제4.1절에서 언급된 의무의 적용이 면제되지 않는다.
- (79) 위원회는 '접근(access)'을 관심 있는 누구나 별도의 비용 부담이나 기타 제한 없이 모델을 자유롭게 획득할 수 있는 권리로 이해한다. 다만 이용자 인증 절차와 같은 합리적인 안전 및 보안 조치는 허용될 수 있으며, 이러한 조치가 예를 들어 출신 국가 등을 이유로 특정인을 부당하게 차별하지 않는 경우에 한한다.
- (80) 위원회는 이 맥락에서 '이용(usage)'이란 라이선스가 기존 제공자로 하여금 자신의 지식재산권을 근거로 모델의 이용을 제한하거나 이용료를 요구하지 못하도록 하는 것을 의미하는 것으로 이해한다. 또한 해당 라이선스는 원칙적으로 모델의 자유로운 이용을 허용하여야 한다. 다만 저작자 표시(attribution)를 요구하거나, 모델 또는 그 파생물(derivatives)의 배포(distribution)가 동일하거나 이에 상응하는 라이선스 조건에 따라 이루어지도록 하는 정도의 제한은 허용될 수 있다. 예를 들어 모델이 배포(distributed)되는 경우 동일하거나 호환되는 저작권 라이선스 조건에 따라 배포(distributed)하도록 요구하는 의무를 둘 수 있다.

- (81) 위원회는 '수정(modification)'을 누구든지 별도의 비용 부담이나 기타 제한 없이 모델을 자유롭게 변경할 수 있는 권리로 이해한다.

- (82) 위원회는 '배포(distribution)'를 GPAI모델에 접근·이용·수정한 자가 제한적인 조건에 따라 해당 모델을 다른 자에게 자유롭게 배포(distribute it onwards)할 수 있는 권리로 이해한다. 이러한 제한적인 조건은 일반적으로 저작자를 표시하고 저작권 고지를 유지하는 것, 즉 저작자 표시(attribution)에 한정된다. 따라서 라이선스가 적용된 모델을 제공받은 자는 (80)에서 허용되는 제한을 해하지 않는 범위에서 해당 모델을 수정하고, 그 결과 생성된 파생 저작물(derivative work)을 폐쇄형 독점 라이선스(closed-source proprietary)[30] 조건을 포함한 다른 라이선스 조건에 따라 재배포(redistribute)할 수 있어야 한다.

- (83) 다음과 같은 제한 조항이 포함된 라이선스는 이러한 기준을 충족하지 않는 것으로 본다.:
  - 비상업적 사용 또는 연구 목적의 사용만 허용하는 제한(예: "해당 모델은 오직 비상업적이고 비수익적 연구 목적으로만 사용할 수 있음");
  - 모델 또는 그 구성요소의 배포(distributing)를 금지하는 제한(예: "해당 모델은 사용할 수 있으나, 모델 배포(distribute)는 허용되지 않음");
  - 사용자 규모에 따라 사용을 제한하는 조건(예: 월간 활성 사용자 수가 일정 기준을 초과하는 경우 추가 라이선스를 요구하는 경우);
  - 특정 사용 사례에 대해 별도의 상업용 라이선스 취득을 요구하는 조건.

- (84) 앞서 설명한 권리 외에도, 오픈소스 라이선스에는 다양한 조건과 조항이 포함된다. 이용자는 자유·오픈소스 라이선스의 조건과 조항을 준수하는 범위 내에서 GPAI모델을 사용·수정·배포(distribute)할 수 있다. 이러한 조건에는 원저작자 표시, 배포(distribution) 조건 준수, 수정 또는 개선 사항을 동일하거나 유사한 라이선스 조건에 따라 공개하도록 하는 의무 등이 포함될 수 있다. 한편 해당 라이선스는 원칙적으로 모든 목적의 사용을 허용하여야 하나, 라이선스 제공자는 공공의 안전·보안 또는 기본권에 중대한 위험을 초래할 수 있는 응용 분야 또는 영역에서의 사용에 대하여, 안전을 위한 특정 조건(safety-oriented terms)을 둘 수 있다. 다만 이러한 제한은 완화하려는 위험에 비례해야 하며, 특정 이용자 집단이나 기타 행위자를 부당하게 대상으로 하지 않는 객관적이고 비차별적인 기준에 근거해야 한다.

4.2.2. 수익화의 부재

30) closed-source proprietary, 소프트웨어의 소스코드를 공개하지 않고 저작권자가 해당 소프트웨어에 대한 권리를 독점적으로 보유·통제하는 방식으로, 오픈소스(open source)와 대비되는 개념

- (85) AI법 전문 103에서 명확히 한 바와 같이, 수익화(monetisation) 개념은 GPAI모델에 대한 면제 적용 여부를 판단하는 중요한 요소가 된다. 해당 전문에 따르면, 면제가 적용되기 위해서는 AI모델에 대한 접근, 사용, 수정 및 배포(distribution)의 대가로 어떠한 금전적 보상도 요구되어서는 안 된다. 이와 관련하여, 수익화는 단순히 모델을 유상으로 제공하는 경우뿐만 아니라, 그 밖의 다양한 수익 창출 방식까지 포함하는 개념으로 이해되어야 한다.

- (86) 다음은 위원회가 수익화에 해당하는 것으로 간주하는 사례이다.:
  - 자유로운 학술 목적 사용은 허용하되, 상업적 사용 또는 일정 규모 이상의 사용에 대해서는 대가 지급을 요구하는 이중 라이선스(dual licensing) 방식 또는 이에 준하는 방식으로 모델을 제공하는 경우
  - 모델 자체 또는 그 보안과 불가분하게 연결되어 있고, 해당 지원이나 서비스 없이는 모델이 작동하거나 접근될 수 없는 기술 지원 및 기타 서비스를 유상으로 제공하는 경우;
  - 이용자가 모델에 접근하기 위해 지원·교육·유지보수 서비스를 구매하도록 요구하는 경우;
  - 모델이 제공자의 개별 호스팅 플랫폼 또는 웹사이트에서만 제공되고, 이용자가 접근을 위해 대가를 지급하여야 하는 경우(이용자가 플랫폼 또는 웹사이트 자체에는 자유롭게 접근할 수 있더라도 유료 광고가 제공되는 경우를 포함함).

- (87) 위원회는 개인정보 보호가 기본권에 해당하며, 개인정보가 상품처럼 거래되거나 수익 창출의 수단으로 활용되어서는 안 된다는 점을 전적으로 인정한다. 그럼에도 불구하고, 모델에 대한 접근(access)·사용(use)·수정(modification) 또는 재배포(redistribution)를 위해 개인정보의 수집 또는 그 밖의 개인정보 처리가 요구되는 경우에는 이를 수익화 전략과 동일하게 보아야 한다. 다만 해당 개인정보 처리가 오로지 모델의 보안을 확보하기 위한 목적에 한정되고, 그 과정에서 어떠한 상업적·금전적 이익도 발생하지 않는 경우에는 예외로 한다. 또한 어떠한 경우에도 개인정보 처리는 EU 데이터 보호법에 따른 규정을 준수하여야 한다.

- (88) 이에 반해, 다음과 같은 경우에는 위원회는 면제 적용과 관련하여 이를 수익화의 형태로 보지 않는다.:
  - 모델의 사용 가능성 또는 자유로운 이용에 영향을 미치지 않고 전적으로 선택사항인 유료 서비스와 함께 모델이 제공되는 경우(예: 오픈소스 라이선스와 무관한 상업적 서비스를 제공하는 사업모델);
  - 모델과 함께 유료 서비스 또는 지원이 제공되더라도, 이를 구매할 의무가 없고 모델의 사용 및 자유롭고 개방적인 접근이 보장되는 경우. 이러한 서비스 또는 지원에는 고급 기능이나 추가 도구가 포함된 프리미엄 버전의 모델, 업데이트 시스템, 확장 기능(extension) 또는 플러그인(plug-ins) 등이 포함될 수 있으며, 이는 이용자가 오픈소스 모델을 활용하거나 그 기능을 확장하는 데 도움을 주기

위한 것이다.

- (89) 소기업(microenterprises) 간 거래의 경우에는 수익화 요소가 없는 것으로 본다.

4.2.3. 파라미터(parameter)(가중치 포함), 모델 아키텍처 정보 및 모델 사용 정보의 공개

- (90) 제4.1절에서 언급된 면제가 적용되기 위해서는 모델의 '가중치(weights)를 포함한 파라미터, 모델 아키텍처에 관한 정보 및 모델 사용에 관한 정보'가 '공개적으로 제공'되어야 한다(AI법 제53조(2) 및 제54조(6)).
- (91) 해당 정보는 모델에 대한 접근·사용·수정 및 배포(distribution)가 가능하도록, 적절한 형식과 충분한 수준의 명확성 및 구체성을 갖추어 공개되어야 한다.
- (92) 모델 사용에 관한 정보에는 최소한 모델의 입력 및 출력 모달리티, 역량 및 한계에 관한 정보가 포함되어야 한다. 또한 모델 사용에 관한 정보에는 모델을 AI시스템에 통합하기 위하여 필요한 기술적 수단(예: 사용 지침, 인프라, 도구 등)도 포함되어야 하며, 필요한 경우 예정된 사용 사례에 적합한 설정에 관한 정보도 포함될 수 있다. 이러한 정보는 AI시스템의 다운스트림 제공자, 개발자 및 이용자가 해당 모델을 실제 응용 분야에서 활용할 수 있도록 한다.

5 GPAI 모델 제공자의 의무 이행에 대한 집행

- (93) AI법에 따른 GPAI모델 제공자의 의무는 2025년 8월 2일부터 적용되며(AI법 제113조(b)), 위원회는 AI사무국을 통해 이러한 의무를 감독하고 집행할 권한을 부여받는다(AI법 제88조(1)). 본 가이드라인은 먼저 AI사무국과 AI위원회가 적정하다고 평가(ass)한 CoP에 서명하고 이를 이행하는 경우의 효과를 설명하고(제5.1절), 이어서 AI사무국의 AI법 제V장에 대한 감독·조사·집행·모니터링 권한과 관련된 사항을 설명하며(제5.2절), 마지막으로 AI법 제V장이 2025년 8월 2일부터 적용된 이후 AI법 준수에 관하여 위원회가 기대하는 사항을 설명한다(제5.3절).

5.1. 적정하다고 평가(ass)된 CoP 준수의 효과

- (94) AI법 제56조(6)에 따라 GPAI모델 제공자는 AI사무국과 AI위원회가 적정하다고 평가(ass)한 CoP를 준수함으로써 AI법 제53조(1) 및 제55조(1)에 따른 의무를 준수하고 있음을 입증할 수 있다. CoP의 일부 장(Chapter)에 대한 적용을 배제(opt-out)하는 경우, 해당 부분에 대해서는 CoP를 통한 의무 준수 입증의 이점을 상실하게 된다. GPAI모델 제공자는 그 밖의 적절한 수단을 통해서도 이러한 의무의 준수를 입증할 수 있으나, AI사무국과 AI위원회가 적정하다고 평가(ass)한 CoP를 준수하는 것은 그 준수를 입증하는 가장 용이한 방법이다(AI법 제53조(4) 및 제55조(2)). AI사무국과 AI위원회가 적정하다고 평가(ass)한 CoP를 준수하는 GPAI모델 제공자에 대해서는 위원회가 CoP 준수 여부를 중심으로 집행 활동을 수행할 예정이다. 또한 AI사무국과 AI위원회가 적정하다고 평가(ass)한 CoP에 서명한 GPAI모델 제공자는 AI법을 준수하기 위하여 이행하는 조치에 관하여 투명성을 확보하게 되며, 이에 따라 위원회와 그 밖의 이해관계자로부터 더욱 높은 신뢰를 얻을 수 있다.

- (95) AI사무국과 AI위원회가 적정하다고 평가(ass)한 CoP를 준수하지 않는 GPAI모델 제공자는 AI법 제V장에 따른 의무를 다른 적절한 수단을 통해 어떻게 준수하고 있는지를 입증하여야 하며, 자신이 이행한 조치를 AI사무국에 보고하여야 한다. 또한 이러한 제공자는 자신이 이행한 조치가 AI법상 의무의 준수를 어떻게 보장하는지를 설명하여야 하며, 예를 들어 AI사무국과 AI위원회가 적정하다고 평가(ass)한 CoP에서 정한 조치와 자신이 이행한 조치를 비교하는 갭 분석(gap analysis)을 수행할 수 있다. 아울러 이러한 제공자는 AI사무국이 해당 제공자의 AI법상 의무 준수 여부를 충분히 파악하기 어렵고, 일반적으로 모델의 전체 수명주기에 걸쳐 GPAI모델에 이루어진 수정사항을 포함한 보다 상세한 정보가 필요하기 때문에, 모델의 전체 수명주기에 걸쳐 모델 평가를 수행하기 위한 더 많은 정보 제공 요청 및 접근 요청을 받을 수 있다.

- (96) 위원회는 과징금 금액을 산정할 때, 구체적인 사정을 고려하여 AI사무국과 AI위원회가 적정하다고 평가(ass)한 CoP에 따라 이행된 약정을 감경 사유로 고려할 수 있다(AI법 제101조(1)).

- (97) GPAI-SR이 아닌 GPAI모델 제공자도 GPAI-SR모델 제공자와 관련된 약정에 서명하고 이를 이행할 수 있다(AI법 제56조(7)).

- (98) 적정하다고 평가(ass)된 CoP에 따라 이행된 약정은 AI법 제V장에서 정한 GPAI모델 제공자에 대한 의무가 2025년 8월 2일부터 적용되기 시작한 이후에야 AI법 준수 여부를 평가(ass)하는 데 고려될 수 있다(AI법 제113조(b)).

- (99) 위원회는 이행행위(implementing act)를 통하여 CoP를 승인할 수 있으며, 이 경우 해당 CoP는 EU 전역에서 효력을 갖게 된다(AI법 제56조(6)). 또한 2025년 8월 2일까지 CoP가 최종 확정되지 못하는 경우, 위원회는 이행행위를 통해 공통 규칙(common rules)을 채택할 수 있으며, 해당 규칙은 모든 GPAI모델 제공자 및 GPAI-SR모델 제공자에게 적용될 수 있다(AI법 제56조(9)).

- (100) CoP는 조화표준(harmonised standard)[31]이 개발·승인될 때까지 AI법 준수를 입증하기 위한 임시적인 수단이다(AI법 제53조(4) 및 제55조(2), 제40조). CoP 준수와 달리, 조화표준을 준수하는 경우에는 AI법상 해당 의무에 대한 적합성 추정(presumption of conformity)[32]이 인정된다(AI법 제53조(4) 및 제55조(2)). 표준화 요청 이후에도 조화표준이 적시에 마련되지 않거나 충분한 수준으로 개발되지 않는 경우, 위원회는 이행행위를 통해 공통 기준(common specifications)[33]를 채택할 수 있다(AI법 제41조).

5.2. AI법 제V장에 따른 의무의 감독·조사·집행 및 모니터링

- (101) AI사무국은 AI법 제V장에 따른 GPAI모델 제공자의 의무를 감독·조사·집행 및 모니터링하는 역할을 수행하며(AI법 제88조 및 제89조), 모델 제공자와 AI시스템 제공자가 동일한 경우에는 GPAI모델 기반 AI시스템의 AI법상 의무 준수 여부도 모니터링한다(AI법 제75조(1)). 이하에서는 GPAI모델 제공자와 관련된 AI사무국의 감독·조사·집행 및 모니터링 활동에 대하여 설명한다.

- (102) AI사무국은 협력적이고 단계적이며 비례적인 접근방식을 취할 예정이다. AI사무국은 특히 GPAI-SR모델 제공자의 경우, AI법 준수를 원활하게 하고 모델이 적시에 시장에 출시될 수 있도록 GPAI모델의 학습 단계부터 제공자와 긴밀한 비공식 협력을 추진한다. 또한 AI사무국은 GPAI-SR모델 제공자가 CoP에 따른 약정의 일환으로든, 또는 다른 방식의 준수 입증 수단의 일환으로든 선제적 보고(proactive reporting)를 할 것을 기대한다. 마지막으로 AI사무국은 AI법에 따라 부여된 권한에 근거한 공식 절차가 진행되는 후반 단계에서, 제공자가 AI사무국와 협력하고 적극적으로 관여할 것을

31) harmonised standard, CEN·CENELEC·ETSI 같은 유럽 표준화 기구가 제정한 유럽표준으로, 그 적용은 통상 자발적이지만 이를 따르면 해당 EU 조화법령의 요구사항을 충족한다는 추정의 근거로 활용됨

32) presumption of conformity, 조화표준 등 공식 기술 표준을 준수한 경우, 해당 제품·서비스·시스템이 관련 EU 법규나 지침의 요구사항을 충족한 것으로 추정하는 제도

33) common specification, 조화표준이 존재하지 않거나 충분하지 않은 경우 유럽연합 집행위원회가 채택하는 구체적인 기술적 요구사항, 관련 법적 요구사항의 준수를 입증하기 위한 기준, '공통 사양' 혹은 '공통 명세'로도 번역됨

기대한다 (예 : 체계적인 협의(structured dialogues)를 통한 협력) .

- (103) 이러한 선제적 보고는 AI법 제55조(1)(c)에 따른 '중대 사고'에 관련된 정보의 추적, 문서화 및 보고 의무에 영향을 미치지 않는다. AI사무국은 이러한 의무가 모델 또는 그 물리적 인프라와 관련된 중대한 사이버보안 침해를 포함한다고 본다. 여기에는 모델 파라미터의 (자체) 유출((self-)exfiltration) 및 사이버공격이 포함되며, 이는 AI법 제55조(1)(b) 및 (d)에 따른 의무에 영향을 미칠 수 있기 때문이다. 이와 별도로, AI사무국은 AI법 제V장의 맥락에서 '중대 사고'를 GPAI모델의 사고 또는 오작동으로서, AI법 제3조(49)(a) 내지 (d)에서 AI시스템과 관련하여 규정한 사건 중 어느 하나를 직·간접적으로 초래하는 경우로 본다.

- (104) AI사무국은 AI법 준수 여부를 평가(ass)할 때, 해당 제공자가 관련 기술표준(technical standards)을 이행하였는지를 고려할 수 있다. 여기에는 사이버보안, 책무성 및 거버넌스 절차와 관련된 조치도 포함된다. 특히 유럽연합 관보에 게재된 조화표준이 존재하는 경우에는 해당 제공자가 이러한 조화표준을 이행하였는지도 함께 고려할 수 있다.

- (105) AI법 제89조(1)에 따라, AI사무국은 GPAI모델 제공자의 AI법상 의무 준수 여부를 모니터링하는 역할을 수행한다. 이러한 모니터링에는 CoP 준수 여부에 대한 모니터링도 포함된다. 또한 AI법 제89조(2)에 따라, 다운스트림 제공자는 GPAI모델 제공자의 AI법 위반을 주장하며 진정(complaint)을 제기할 권리를 가진다. 이러한 진정은 충분한 근거를 갖추어야 하며, 최소한 AI법 제89조(2)(a) 내지 (c)에 열거된 정보를 포함하여야 한다.

- (106) GPAI모델 제공자에 대한 의무에 관한 위원회의 집행은 AI법에 따라 부여된 권한에 기반한다. 여기에는 다음과 같은 권한이 포함된다: (i) 정보 제공 요청(AI법 제91조), (ii) GPAI모델에 대한 평가 실시(AI법 제92조), (iii) 완화조치의 이행 및 시장에서의 모델 회수를 포함한 조치를 요구할 권한(AI법 제93조), (iv) 전 세계 연간 총매출액의 최대 3% 또는 1,500만 유로 중 더 높은 금액까지의 과징금 부과(AI법 제101조). 이러한 과징금 부과 권한은 2026년 8월 2일부터 적용된다. 또한 이러한 권한이 어떻게 행사될 것인지를 보다 구체화하기 위한 추가 법령이 마련될 예정이며, 특히 AI법 제92조(6) 및 제101조(6)에 따른 이행행위가 이에 해당한다.

- (107) 위원회는 2025년 8월 2일부터 AI법 제52조에 따른 결정을 내릴 수 있다(AI법 제113조(b)). 여기에는 주장의 수용 또는 거부, GPAI모델을 GPAI-SR모델로 지정하는 결정 등이 포함된다.

- (108) 위원회는 AI법 제78조에 따라 감독·조사·집행 및 모니터링 활동 과정에서 취득한 데이터의 기밀성을 보장하여야 하며, 특히 자연인 또는 법인의 지식재산권, 영업상 비밀정보(confidential business information) 또는 영업비밀(trade secrets), 그리고 공공의 안전 및 보안을 보호하여야 한다.

5.3. 2025년 8월 2일 AI법 제V장 적용 이후의 의무 준수

- (109) AI법 제V장에서 규정된 GPAI모델 제공자의 의무는 2025년 8월 2일부터 적용된다(AI법 제113조(b)). 다만, AI법 제111조(3)에 따라 '이전에 시장에 출시된 GPAI모델의 제공자는 2027년 8월 2일까지 본 규정에서 정한 의무를 준수하기 위하여 필요한 조치를 취해야 한다.' 이러한 의무는 해당 모델의 전체 수명주기에 걸쳐 적용된다(제2.2절 참조).

- (110) 위원회는 2025년 8월 2일부터 GPAI모델 제공자에 대한 의무가 적용되기 시작한 이후 수개월 동안 2025년 8월 2일 이전에 시장에 출시된 GPAI모델의 제공자가 2027년 8월 2일까지 AI법상 의무를 준수하는 과정에서 다양한 어려움에 직면할 수 있음을 인정한다. 이에 따라 AI사무국은 해당 GPAI모델 제공자가 2027년 8월 2일까지 의무 준수를 위하여 필요한 조치를 취할 수 있도록 지원하는 데 전념하고 있다.

- (111) 특히, 2025년 8월 2일 이전에 시장에 출시된 GPAI모델의 제공자는 과거에 수행된 행위에 대하여 모델에 대한 재학습(retraining) 또는 언러닝(unlearning)을 수행하는 것이 불가능한 경우, 학습 데이터에 관한 일부 정보를 이용할 수 없는 경우, 또는 해당 정보를 확보하는 것이 제공자에게 과도한 부담을 초래하는 경우에는 그러한 조치를 수행할 의무를 부담하지 않는다. 이러한 경우에는 저작권 정책 및 학습에 사용된 콘텐츠 요약에 이를 명확하게 공개하고 정당화해야 한다.

- (112) 위원회는 2025년 8월 2일 이후 시장에 출시된 GPAI모델의 제공자도 AI법상 의무를 준수하기 위하여 내부 정책 및 절차를 조정하는데 일정한 시간이 필요할 수 있음을 인정한다. 특히 이러한 필요성은 GPAI모델 제공자에 대한 의무가 적용되기 시작한 이후 초기 수개월 동안 나타날 수 있다. 이에 따라 AI사무국은 해당 GPAI모델 제공자가 의무를 준수하기 위하여 필요한 조치를 취할 수 있도록 지원하고 있다. 또한 AI사무국은 해당 GPAI모델 제공자가 AI법을 준수하기 위하여 올바른 조치를 취하고 있는지를 확인할 수 있도록 즉시 그리고 선제적으로 AI사무국에 연락할 것을 권고한다.

- (113) 특히 2025년 8월 2일 현재 GPAI모델을 이미 학습하였거나 학습 중인 GPAI모델 제공자 또는 2025년 8월 2일 이후 시장에 출시할 목적으로 GPAI모델 학습을 계획하고 있는 GPAI모델 제공자는 GPAI모델 제공자에 대한 의무, 특히 GPAI-SR모델 제공자에 대한 의무를 준수하는 데 어려움이 있을 것으로 예상되는 경우, 의무를 준수하기 위하여 필요한 조치를 언제 어떠한 방식으로 취할 것인지에 관하여 AI사무국에 즉시 선제적으로 알려야 한다. 특히 2025년 8월 2일 이전에 GPAI-SR모델을 시장에 출시한 적이 없는 GPAI모델 제공자의 경우, 위원회는 해당 모델이 적시에 시장에 출시될 수 있도록 해당 제공자가 처한 어려운 상황을 특별히 고려할 예정이다.

- (114) 또한 위원회는 집행 과정에서 외부 모델 평가자 생태계가 아직 발전 단계에 있으며, 초기에는 이러한

평가자들이 사용하는 방법론이 일관되지 않을 수 있다는 점을 집행 과정에서 고려할 예정이다. AI사무국은 이러한 상황을 모니터링할 예정이며, 필요한 경우 평가자 네트워크 및 기타 분야별 전문가를 참여시키는 방식 등을 통해 외부 모델 평가자를 위한 공통 접근방식을 마련하기 위하여 노력할 예정이다.

- (115) 마지막으로 AI법은 2025년 8월 2일 직후에는 GPAI모델 제공자가 적정하다고 평가(ass)된 CoP에서 정한 일부 조치 또는 그 밖의 적정한 대체 조치를 즉시 완전히 이행하지 못할 수 있음을 인정한다. 다만 위원회의 집행 권한은 2026년 8월 2일부터 적용되므로, 2025년 8월 2일부터 1년 동안은 위원회가 집행조치를 취할 수 없다. 그러나 이는 2025년 8월 2일부터 AI법 제V장에 따른 의무가 적용된다는 점에 영향을 미치지 않는다(AI법 제113조(b)). 특히 2025년 8월 2일 현재 GPAI-SR모델을 이미 학습하였거나 학습 중인 GPAI모델 제공자, 또는 2025년 8월 2일 이후 시장 출시를 목적으로 GPAI-SR모델의 학습을 계획하고 있는 GPAI모델 제공자는 AI법 제52조(1)에 따라 지체 없이, 어떠한 경우에도 2025년 8월 2일부터 2주 이내에 위원회에 통지하여야 하며, AI법을 완전히 준수할 수 있도록 AI사무국과 협력할 것이 기대된다. 또한 2026년 8월 2일부터는 위원회가 AI법 제101조에 규정된 모든 관련 요소를 고려하여, 그 시점까지 모든 의무를 완전히 준수하지 않은 GPAI모델 제공자에 대하여 과징금을 부과하는 등 AI법상 의무의 준수를 집행할 예정이다.

6 위원회 가이드라인의 검토 및 업데이트

- (116) 위원회는 GPAI모델 제공자가 의무를 이행하는 과정에서 축적한 실무 경험과 관련 기술적·사회적·시장적 발전에 비추어 필요하다고 판단되는 경우, 본 가이드라인을 검토할 예정이다. 여기에는 GPAI모델 제공자에 대한 의무 및 본 가이드라인에서 다루는 AI법의 다른 조항과 관련하여, 집행조치를 통해 축적된 관련 경험과 EU 사법재판소(CJEU)가 제시한 해석도 포함된다. 위원회는 이러한 검토 과정에서 본 가이드라인을 철회하거나 개정할 수 있다. 또한 위원회는 GPAI모델 제공자, 국가 시장감시당국(AI위원회를 통하여), 자문포럼, 과학패널, 연구 공동체 및 시민사회단체가 향후 공개협의, 워크숍 또는 기타 의견수렴 절차에 참여함으로써 이 검토 과정에 기여할 것을 권장한다.

A 부속서: GPAI 모델의 학습 연산량

- (117) AI법 제51조(2)는 GPAI모델의 학습에 사용된 연산량에 관한 임계값을 규정하고 있으며, 이 임계값을 충족하는 경우 해당 GPAI모델은 GPAI-SR모델로 분류될 수 있다. 또한 제2.1절 및 제3.2절에서도 학습 연산량과 관련된 임계값을 정하고 있다. 따라서 GPAI모델이 이러한 임계값 중 어느 하나를 충족하는지를 판단하기 위해, 잠재적 제공자는 학습에 사용된 연산량을 추정하여야 한다. 본 부속서는 이러한 학습 연산량을 추정하는 방법에 관한 지침을 제공한다. 부속서 A.1에서는 GPAI모델 및 GPAI-SR모델의 잠재적 제공자가 학습 연산량을 추정할 때 포함하여야 하는 요소를 설명하고, 부속서 A.2에서는 잠재적 제공자가 활용할 수 있는 두 가지의 널리 사용되는 접근방식을 설명한다. 또한 부속서 A.3에서는 부속서 A.2의 계산식을 설명하고, 모델이 GPAI모델에 해당하는지 여부에 관한 예시적 기준의 일부를 구성하는 임계값(제2.1절)을 정당화하기 위하여 약 10억 개의 파라미터로 학습된 여러 모델의 학습 연산량을 추정한다.

A.1. 학습 연산량 산정 대상

- (118) 본 가이드라인에서 GPAI모델의 '학습 연산량'은 다음 중 어느 하나를 의미한다.:
  - 해당 모델이 GPAI-SR이 아닌 경우, 모델의 파라미터 업데이트에 직접적으로 기여하는 총 연산량;
  - 해당 모델이 GPAI-SR에 해당하는지 여부를 판단하는 경우, 모델 학습에 사용된 누적 연산량.

- (119) AI법은 '누적 학습 연산량'의 의미와 관련하여, AI법 전문 111에서 '학습에 사용된 누적 연산량에는 모델이 배포(deployment)되기 전에 모델의 역량을 향상시키기 위한 사전학습, 합성데이터(synthetic data) 생성, 미세조정 등의 활동과 방법에 사용된 연산량이 포함된다'고 규정하고 있다. 다만 이러한 '활동과 방법'은 매우 다양하고 빠르게 발전하고 있으므로, 무엇을 포함하고 포함하지 않아야 하는지를 정확하고 완전한 목록으로 제시하는 것은 현실적으로 어렵다. 이에 따라 위원회는 원칙적으로 모델의 역량에 기여하였거나 앞으로 기여하게 될 모든 연산량을 제공자가 고려하여야 한다고 본다. 특히 모델의 파라미터 업데이트에 직접적으로 기여하는 모든 연산량은 포함되어야 한다.

- (120) 모델이 공개적으로 접근할 수 없는 합성데이터를 사용하여 학습되는 경우, 폐기된 데이터를 포함하여 해당 데이터를 생성하기 위해 수행된 순전파(forward passes)[34]는 누적 학습 연산량 산정에 포함되어야 한다. 예를 들어 100개의 샘플을 생성한 후 그중 상위 10개의 샘플만 학습에 사용하였다면, 선택된 10개의 샘플을 생성하기 위해서는 전체 100개의 샘플 생성 과정이 필요하였으므로, 100개 샘플

34) Forward pass, 입력 데이터를 신경망의 각 층에 차례로 통과시켜 최종 예측값(출력)과 손실값을 계산하는 과정으로, 현재 모델 파라미터에 따라 입력이 어떤 결과로 변환되는지를 구하는 단계

전체를 생성하는 데 사용된 연산량을 모두 포함하여야 한다.

- (121) 모델 가중치 병합(model weight merging)[35], 모델 가중치 평균화(model weight averaging)[36] 또는 가중치 초기화(weight initialisation) 등을 통해 기존 모델의 가중치(pre-existing model weights)를 결합하여 모델이 생성된 경우에는 결합된 모델 가중치를 학습하는 데 사용된 연산량을 해당 모델의 누적 학습 연산량 산정에 포함하여야 한다.

- (122) 반면, 다음은 누적 학습 연산량의 산정에 포함하지 않아도 되는 연산량. 이는 해당 데이터가 다른 공개적으로 접근 가능한 데이터와 구별되지 않을 수 있기 때문이다.;
  - 공개적으로 접근 가능한 합성데이터를 생성하는 데 사용된 연산량, 해당 데이터가 다른 공개적으로 접근 가능한 데이터와 구별되지 않을 수 있기 때문;
  - 모델 평가 또는 레드티밍(red-teaming)과 같이 모델의 역량 향상에 기여하지 않는 순수한 진단 절차에 사용된 연산량;
  - 보다 효율적인 학습 기법으로 이어지는 탐색적 연구 프로젝트에 사용된 연산량, 또는 실패하여 결과가 폐기된 합성데이터 생성 실험에 사용된 연산량과 같이, 인간이 얻은 교훈을 통해서만 모델의 역량 향상에 기여하는 연산량;
  - 증류(distillation)[37]에 사용되는 부모 모델(parent model)을 학습하는 데 사용된 연산량(부모 모델이 해당 모델과 동일한지 여부는 불문함);
  - 가치 함수(value functions) 또는 보상 모델(reward models)과 같은 보조 모델(auxiliary models)[38]을 학습하는 데 사용된 연산량(보조 모델이 해당 모델과 동일한지 여부는 불문함);
  - 메모리 절약을 위한 활성화값 재계산.

A.2. 학습 연산량 산정 방법

- (123) GPAI모델 제공자는 보고된 추정치의 ±30% 이내의 전체 오차범위에서, 자신의 최선의 판단(best judgement)에 따라 추정치가 정확하다고 판단하는 한, 관련 누적 학습 연산량을 추정하기 위하여 어떠한 방법이든 선택할 수 있다. 이는 예를 들어 모델의 학습이 완료되기 전에 추정이 이루어지는 경우와 같이, 제공자가 모델의 누적 학습 연산량을 정확하게 추정하는 데 어려움을 겪을 수 있음을 고려한

35) Model weight merging, 여러 개의 사전 훈련된 모델이나 미세 조정된 모델들의 가중치(파라미터)를 산술적으로 평균내어, 추가 학습 비용 없이 단일 통합 모델로 결합하는 머신러닝 기법
36) Model weight averaging, 독립적으로 학습된 여러 모델의 가중치(weights)를 결합하여, 추론 속도 저하 없이 성능 향상과 일반화 능력을 극대화하는 머신러닝 기법
37) Distillation, 대형 모델(교사 모델, teacher model)의 지식을 소형 모델(학생 모델, student model)에 이전하여, 더 적은 파라미터(parameter)로도 교사 모델과 유사한 성능을 구현하도록 학습시키는 딥러닝 기법.
38) Auxiliary Models, 메인 모델의 작업을 돕거나, 복잡한 시스템의 효율성을 높이기 위해 함께 사용되는 부가적인 모델

것이다. 다만 제공자는 추정 방법과 관련 불확실성을 포함하여, 추정을 수행하는 과정에서 전제한 사항을 문서화하여야 한다.

- (124) 제공자는 산정 대상에 적합한 방식에 따라, 그래픽 처리 장치(graphics processing unit, 'GPU') 사용량을 추적하는 방법(하드웨어 기반 접근방식(hardware-based approach)) 또는 관련 모델의 아키텍처를 바탕으로 연산을 직접 추정하는 방법(아키텍처 기반 접근방식(architecture-based approach))을 사용하여 관련 학습 연산량을 산정할 수 있다.

- (125) 산정 방법과 관계없이, 모든 연산은 부동소수점 정밀도(floating-point precision)와 무관하게 동일하게 계산되어야 한다.

A.2.1. 하드웨어 기반 접근법(hardware-based approach)

- (126) 하드웨어 기반 접근법을 이용하여 학습 연산량을 산정하기 위해, GPAI모델 또는 GPAI-SR모델의 잠재적 제공자는 우선 다음 사항을 추정하여야 한다.:
  - 사용되었거나 사용될 예정인 GPU 또는 기타 하드웨어 유닛의 수 N;
  - 달성되었거나 달성될 것으로 예상되는 총 사용 시간 L(초 단위로 측정);
  - 사용되었거나 사용될 예정인 GPU 또는 기타 하드웨어 유닛의 이론적 최대 성능 H(FLOP/second 단위로 측정); 서로 다른 이론적 최대 성능을 가진 여러 유형의 GPU 또는 하드웨어 유닛이 사용되는 경우에는 각 유형별 하드웨어 유닛 수를 가중치로 하는 가중평균(weighted average)을 사용하여 계산;
  - 사용 기간 동안 달성되었거나 달성될 것으로 예상되는 평균 GPU 활용률 U.

- (127) 모든 하드웨어 유닛 N이 전체 사용 기간 L 동안 사용되는 경우, 이에 대응하는 학습 연산량 C(FLOP 단위로 측정)는 다음 식에 따라 계산할 수 있다.

$$C = N \cdot L \cdot H \cdot U.$$

- (128) 서로 다른 수의 하드웨어 유닛이 서로 다른 기간 동안 사용되는 경우에는 일정한 수의 하드웨어 유닛이 사용된 각 기간별로 위 공식을 적용할 수 있다(다만 전체 시간이 전체 학습 기간과 일치하여야 한다). 이후 각 기간에 대한 추정치를 합산하여 총 학습 연산량을 산정할 수 있다.

- (129) 부속서 A.3에는 위 접근방식을 사용한 계산 예시가 포함되어 있다.

A.2.2. 아키텍처 기반 접근법(architecture-based approach)

- (130) 이 접근법은 관련 모델의 아키텍처를 바탕으로 수행된 연산 수를 직접 추정하는 방식이다.
- (131) 예를 들어, 현재 대부분의 GPAI모델 또는 잠재적 GPAI모델은 순전파(forward pass)[39]와 역전파(backward pass)[40]의 연속을 통해 학습되는 신경망에 기반하고 있다. 이러한 모델의 경우, 이에 대응하는 학습 연산량 C(FLOP 단위로 측정)는 학습 과정에서 수행된 풀 패스(full pass)[41] 횟수(하나의 풀 패스는 순전파와 역전파의 결합을 의미함)와 하나의 풀패스(full pass)에서 수행되는 총 연산 수의 곱으로 계산된다. 이를 식으로 나타내면 다음과 같다.

$$C = \text{Number of forward passes made during training} \cdot \text{number of operations/full pass.}$$

- (132) 트랜스포머 아키텍처에 기반한 하는 일부 밀집형(dense) 대규모[42] 언어모델(large language models)의 경우, 학습 연산량 C는 다음의 계산식으로 다음 식을 이용하여 근사값을 구할 수 있다.:

$$C \approx 6 \cdot P \cdot D$$

여기서 P는 모델의 총 파라미터 수를 의미하고, D는 학습에 사용된 총 토큰 수를 의미한다.[43]

- (133) 부속서 A.3에는 이 접근법을 사용한 계산 예시가 포함되어 있다.

A.3. 약 10억 개의 파라미터를 가진 모델의 학습 연산량

- (134) 모델이 GPAI모델에 해당하는지 여부에 관한 지표적 기준(indicative criterion)의 일부를 구성하는 임계값(제2.1절)은 실제로 널리 사용되는 약 10억 개의 파라미터를 가진 모델을 학습시키는 데 사용된 것으로 추정되는 연산량을 바탕으로 설정되었다. 모델 및 제공자의 명칭은 공개되지 않지만, 이러한 추정과 관련된 정보는 아래와 같이 제공된다.
- (135) 조사 대상 모델의 학습에 사용된 것으로 추정되는 연산량은 다음과 같다.:

39) Forward pass, 입력 데이터를 신경망의 각 층에 차례로 통과시켜 최종 예측값(출력)과 손실값을 계산하는 과정으로, 현재 모델 파라미터에 따라 입력이 어떤 결과로 변환되는지를 구하는 단계

40) backward pass, 순전파에서 계산된 손실을 기준으로, 출력층에서 입력층 방향으로 연산을 거슬러 올라가며 각 파라미터가 손실에 얼마나 기여했는지를 나타내는 기울기(gradient)를 연쇄법칙에 따라 계산·누적하는 과정

41) 순전파(forward pass)와 역전파(backward pass)를 모두 수행하는 1회의 학습 과정. 하나의 순전파와 하나의 역전파가 결합된 단위

42) 밀집형 트랜스포머(dense transformer) 아키텍처를 기반으로 하는 모델은 그 파라미터 수가 입력 컨텍스트의 토큰 수의 12분의 1보다 상당히 큰 경우 대규모 모델로 보아야 한다. 이는 컨텍스트 의존 연산(context-dependent compute)이 학습 연산량에 기여하는 정도가 비임베딩 연산(non-embedding compute)의 기여도(6FD)와 비교하여 무시할 수 있을 정도로 작도록 하기 위한 것이다(예를 들어 각주 16의 참고문헌 참조).

43) 예를 들어, Kaplan, J. 외, 「Scaling Laws for Neural Language Models」, arXiv preprint 2001.08361, 2020 참조

- 모델 A(38억 개 파라미터를 가진 언어모델): 7.5 · 10^{22} FLOP
- 모델 B(10억 개 파라미터를 가진 이미지 확산(image diffusion) 모델)): 10^{23} FLOP
- 모델 C(6억 개 파라미터를 가진 언어모델): 1.3 · 10^{23} FLOP
- 모델 D(15억 개 파라미터를 가진 언어모델): 1.6 · 10^{23} FLOP
- 모델 E(17억 개 파라미터를 가진 언어모델): 3.7 · 10^{23} FLOP
- 모델 F(25억 개 파라미터를 가진 언어모델): 1.8 · 10^{23} FLOP
- 모델 G(26억 개 파라미터를 가진 언어모델): 1.5 · 10^{23} FLOP
- 모델 H(10억 개 파라미터를 가진 언어모델): 6.5 · 10^{23} FLOP

- (136) 위 수치는 아래 계산을 통해 산출된 것으로, 동시에 부속서 A.2에서 설명한 두 가지 산정 방법을 설명하기 위한 예시이기도 하다.

- (137) 모델 A. 아키텍처 기반 접근법: 모델 A는 트랜스포머 디코더(transformer decoder)[44] 아키텍처와 38억 개의 파라미터를 가진 언어모델이며, 3조 3천억 개의 토큰으로 학습되었다. 부속서 A.2.2의 근사식은 다음과 같다.:

$$C \approx 6 \cdot P \cdot D = 6 \cdot 3.8 \cdot 10^{9} \cdot 3.3 \cdot 10^{12} \approx 7.5 \cdot 10^{22} \text{ FLOP.}$$

- (138) 모델 B. 하드웨어 기반 접근법: 모델 B는 이미지 생성을 위한 diffusion 모델이다. 모델 카드(model card)에 따르면, 모델 B는 40GB A100 GPU에서 총 200,000 GPU 시간 동안 학습되었다. 이는 GPU 수 N과 전체 학습 기간 L의 곱 NL이 200,000 · 3,600초와 같음을 의미한다. 이론상 최대 성능을 300 TFLOP/second, GPU 활용률을 50%로 가정할 경우, 부속서 A.2.1의 근사식은 다음과 같다.:

$$C = N \cdot L \cdot H \cdot U \approx 10^{23} \text{ FLOP.}$$

- (139) 모델 C. 아키텍처 기반 접근법: 모델 C는 트랜스포머 디코더 아키텍처와 6억 개의 파라미터를 가진 언어모델이며, 36조 개의 토큰으로 학습되었다. 부속서 A.2.2의 근사식은 다음과 같다.:

$$C \approx 6 \cdot P \cdot D = 6 \cdot 0.6 \cdot 109 \cdot 36 \cdot 1012 \approx 1.3 \cdot 1023 \text{ FLOP.}$$

44) Transformer Decoder, AI가 문맥을 이해하고 새로운 텍스트, 이미지, 또는 기타 시퀀스 데이터를 생성(Generation)하는 핵심 신경망 블록

- (140) 모델 D. 아키텍처 기반 접근법: 모델 D는 트랜스포머 디코더 아키텍처와 15억 4천만 개의 파라미터를 가진 언어모델이며, 18조 개의 토큰으로 학습되었다. 부속서 A.2.2의 근사식은 다음과 같다.:

$$C \approx 6 \cdot P \cdot D = 6 \cdot 1.54 \cdot 10^{9} \cdot 18 \cdot 10^{12} \approx 1.663 \cdot 10^{23} \text{ FLOP.}$$

- (141) 모델 E. 아키텍처 기반 접근법: 모델 E는 트랜스포머 디코더 아키텍처와 17억 개의 파라미터를 가진 언어모델이며, 36조 개의 토큰(tokens)으로 학습되었다. 부속서 A.2.2의 근사식은 다음과 같다.:

$$C \approx 6 \cdot P \cdot D = 6 \cdot 1.7 \cdot 10^{9} \cdot 36 \cdot 10^{12} \approx 3.7 \cdot 10^{23} \text{ FLOP.}$$

- (142) 모델 F. 아키텍처 기반 접근법: 모델 F는 트랜스포머 디코더 아키텍처와 25억 개의 파라미터를 가진 언어모델이며, 12조 개의 토큰으로 학습되었다. 부속서 A.2.2의 근사식은 다음과 같다.:

$$C \approx 6 \cdot P \cdot D = 6 \cdot 2.5 \cdot 10^{9} \cdot 12 \cdot 10^{12} \approx 1.8 \cdot 10^{23} \text{ FLOP.}$$

- (143) 모델 G. 아키텍처 기반 접근법: 모델 G는 트랜스포머 디코더 아키텍처와 26억 개의 파라미터(parameters)를 가진 언어모델이며, 610억 개의 토큰(tokens)을 164 에포크(epoch)[45]에 걸쳐 학습하였고, 이는 총 약 10조 개의 학습 토큰에 해당한다. 부속서 A.2.2의 근사식은 다음과 같다.:

$$C \approx 6 \cdot P \cdot D = 6 \cdot 2.6 \cdot 10^{9} \cdot 10 \cdot 10^{12} \approx 1.5 \cdot 10^{23} \text{ FLOP.}$$

- (144) 모델 H. 하드웨어 기반 접근법: 모델 H는 언어모델이다. 모델 카드에 따르면, 모델 H는 80GB H100 GPU에서 총 370,000 GPU 시간 동안 학습되었다. 앞서와 동일하게 이론상 최대 성능을 989 TFLOP/second, GPU 활용률을 가정할 경우, 부속서 A.2.1의 근사식은 다음과 같다.

$$C = N \cdot L \cdot H \cdot U = 6.5 \cdot 10^{23} \text{ FLOP.}$$

45) epoch, 학습 알고리즘이 전체 데이터세트를 한 번 완전히 통과(순전파+역전파)하여 모델을 학습시킨 횟수

III 투명성(Transparency Chapter)

GPAI모델에 관한 CoP
투명성 장

Nuria Oliver 워킹그룹1 공동의장
Rishi Bommasani 워킹그룹1 부의장

1 투명성 장에 관한 의장 및 부의장의 서문

- CoP의 투명성 장은 서명자가 AI법 제53조(1)(a) 및 (b)와 이에 상응하는 부속서 XI 및 XII에 따른 투명성 의무를 준수하기 위하여 이행하기로 약정한 세 가지 이행조치를 제시한다.
- 이행조치 1.1에 따른 약정의 이행을 지원하기 위하여, AI법에서 요구하는 정보를 하나의 양식에 체계적으로 정리할 수 있도록 서명자를 위한 사용자 친화적인 모델설명서양식(Model Documentation Form)을 수록하였다.
- 모델설명서양식은 각 항목별로 해당 정보의 제공 대상을 다운스트림 제공자, AI사무국 또는 국가 관할 당국으로 구분하여 제시한다. AI사무국 또는 국가 관할 당국을 대상으로 하는 정보는 AI사무국이 직권으로 요청하거나 국가 관할 당국의 요청에 따라 AI사무국이 요청하는 경우에 한하여 제공된다. 이러한 요청에는 법적 근거와 목적이 명시되어야 하며, 요청 당시 AI사무국의 AI법상 업무 수행 또는 국가 관할 당국의 감독 업무 수행에 필요한 모델설명서양식의 항목으로 한정되어야 한다. 특히 시스템 제공자와 모델 제공자가 서로 다른 경우, GPAI모델을 기반으로 구축된 고위험 AI시스템(high-risk AI system) 제공자의 AI법 준수 여부를 평가(ass)하기 위해 필요한 범위에서 이러한 요청이 이루어질 수 있다.
- AI법 제78조에 따라 모델설명서양식에 포함된 정보를 제공받는 자는 취득한 정보의 기밀성을 준수할 의무가 있으며, 특히 지식재산권(intellectual property rights)과 기밀사업정보(confidential business information) 또는 영업비밀(trade secrets)을 보호하여야 한다. 또한 취득한 정보의 보안과 기밀성을 보호하기 위하여 적절하고 효과적인 사이버 보안 조치를 마련하여야 한다.

2 목적

- 본 CoP의 궁극적인 목적은 AI법 제1조(1)에 따라 역내시장의 기능을 개선하고 인간중심적이며 신뢰할 수 있는 AI의 활용을 촉진하는 동시에, 민주주의, 법치주의 및 환경보호를 포함하여 헌장에 명시된 건강, 안전 및 기본권에 대한 높은 수준의 보호를 보장하고, EU 내에서 AI가 초래할 수 있는 유해한 영향으로부터 이를 보호하며, 혁신을 지원하는 데 있다.
- 이러한 궁극적인 목적을 달성하기 위하여, 본 CoP의 구체적인 목적은 다음과 같다.

  A. CoP는 AI법 제53조 및 제55조에 규정된 의무 준수를 입증하기 위한 가이드라인 문서로 기능한다. 다만, CoP를 준수하였다는 사실만으로 AI법상 해당 의무를 준수하였다는 결정적 증거가 되는 것은 아니다.

  B. CoP는 GPAI모델 제공자가 AI법상 의무를 준수할 수 있도록 지원하며, CoP를 활용하여 해당 의무의 준수를 입증하는 GPAI모델 제공자에 대하여 AI사무국이 그 준수 여부를 평가(ass)할 수 있도록 한다.

3 전문

- 다음의 사항을 고려한다.:

(a) 서명자는 GPAI모델 제공자가 AI 가치사슬 전반에 걸쳐 특별한 역할과 책임을 가진다는 점을 인정한다. GPAI모델 제공자가 제공하는 GPAI모델은 다양한 다운스트림 AI시스템의 기반이 될 수 있으며, 이러한 다운스트림 AI시스템은 대개 다운스트림 제공자에 의해 제공된다. 다운스트림 제공자는 GPAI모델을 자신의 제품에 통합하고 AI법에 따른 의무를 이행하기 위하여 GPAI모델 및 해당 모델의 역량에 대한 충분한 이해가 필요하다(AI법 전문 101 참조).

(b) GPAI모델의 미세조정 또는 그 밖의 수정이 있는 경우, 서명자는 해당 모델을 미세조정하거나 수정한 자연인, 법인, 공공기관, 기관 또는 기타 단체가 수정된 모델의 제공자가 되며, 그에 따라 GPAI모델 제공자에게 적용되는 의무를 부담한다는 점을 인정한다. 다만 비례성 원칙을 준수하기 위하여(AI법 전문 109 참조), 해당 제공자의 CoP 투명성 장에 따른 약정은 해당 미세조정 또는 수정의 범위로 한정되어야 한다. 이와 관련하여 서명자는 위원회의 관련 가이드라인을 고려해야 한다.

(c) 서명자는 CoP 투명성 장에 따른 약정의 범위를 벗어나지 않는 범위 내에서, AI사무국 또는 다운스트림 제공자에게 정보를 제공할 때 시장 및 기술 발전을 고려할 필요가 있다는 점을 인정한다. 이는 해당 정보가 AI사무국 및 국가 관할 당국의 AI법상 업무 수행을 지원하고, 다운스트림 제공자가 서명자의 GPAI모델을 AI시스템에 통합하며 AI법에 따른 의무를 이행할 수 있도록 한다는 본래 목적을 지속적으로 달성하기 위함이다(AI법 제56조(2)(a) 참조).

- 본 CoP의 투명성 장은 AI법 제53조(1)(a) 및 (b)에 따른 문서화 의무(Documentation obligations) 가운데 모든 GPAI모델 제공자에게 적용되는 의무(다만 AI법 제53조(2)에 따른 예외는 제외함), 즉 AI법 부속서 XI 제1절 및 부속서 XII에 관한 의무에 중점을 둔다. 한편, GPAI-SR모델 제공자에게만 적용되는 AI법 부속서 XI 제2절의 문서화 의무는 본 CoP의 안전 및 보안 장의 이행조치 10.1에서 다루고 있다.

약정 1. 문서화(Documentation)

해당 법령: AI법 제53조(1)(a) 및 (b), 제53조(2), 제53조(7) 및 AI법 부속서 XI·XII

- AI법 제53조(1)(a) 및 (b)에 따른 의무를 이행하기 위하여, 서명자는 이행조치 1.1에 따라 모델문서화자료(model documentation)를 작성하고 이를 최신 상태로 유지하며, 이행조치 1.2에 따라 GPAI모델을 자신의 AI시스템에 통합하고자 하는 AI시스템 제공자(이하 "다운스트림 제공자") 및 요청이 있는 경우 AI사무국에 관련 정보를 제공할 것(국가 관할당국이 AI법에 따른 감독 업무를 수행하는 데 엄격히 필요한 경우, 특히 시스템 제공자와 모델 제공자가 서로 다른 경우 GPAI모델을 기반으로 구축된 고위험 AI시스템의 AI법 준수 여부를 평가하기 위하여 국가 관할당국의 요청에 따라 AI사무국이 관련 정보의 제공을 요청할 수 있다. AI법 제75조(1) 및 (3), 제88조(2) 참조)을 약정하며, 이행조치 1.3에 따라 문서화된 정보의 품질, 보안 및 무결성을 보장할 것을 약정한다. AI법 제53조(2)에 규정된 요건을 충족하는 자유 및 오픈소스 라이선스로 공개된 GPAI모델의 제공자에게는 이러한 이행조치가 적용되지 않는다. 다만, 해당 모델이 GPAI-SR모델인 경우에는 예외로 한다.

- **이행조치 1.1. 모델문서화자료의 작성 및 최신 상태 유지**

  서명자는 GPAI모델을 출시할 때, 아래의 모델설명서양식에 기재된 모든 정보를 포함하여 모델문서화자료를 작성하여야 한다(이하 해당 정보를 "모델문서화자료(Model Documentation)"라 한다). 서명자는 본 약정을 준수하기 위하여 부록에 수록된 모델설명서양식을 활용할 수 있다.

  서명자는 동일 모델의 업데이트된 버전과 관련된 변경사항을 포함하여 모델문서화자료에 포함된 정보가 변경된 경우, 해당 변경사항을 반영하여 모델문서화자료를 업데이트해야 한다. 또한 모델 출시 후 10년간 모델문서화자료의 이전 버전을 보관해야 한다.

- **이행조치 1.2. 관련 정보의 제공**

  서명자는 GPAI모델을 출시할 때, AI사무국 및 다운스트림 제공자가 모델문서화자료 또는 그 밖에 필요한 정보에 대한 접근을 요청할 수 있도록 자신의 웹사이트를 통하여 연락처 정보를 공개하여야 하며, 웹사이트가 없는 경우에는 그 밖의 적절한 수단을 통하여 이를 공개하여야 한다.

  서명자는 AI사무국이 AI법 제91조 또는 제75조(3)에 따라 모델문서화자료의 하나 이상의 항목 또는 추가 정보를 요청하는 경우, 요청된 정보를 최신 형태로 제공하여야 한다. 이러한 요청은 AI사무국이 AI법상 업무를 수행하거나 국가 관할 당국이 AI법에 따른 감독 업무를 수행하는 데 필요한 정보에 한하여 이루어질 수 있다. 특히 시스템 제공자와 모델 제공자가 서로 다른 경우[46] GPAI모델을 기반으로 구축된

고위험 AI시스템의 AI법 준수 여부를 평가(ass)하기 위하여 정보 제공이 요청될 수 있다. 요청된 정보는 AI법 제91조(4)에 따라 AI사무국이 지정한 기간 내에 제공되어야 한다.

서명자는 AI법 제53조(7) 및 제78조에 규정된 기밀성 보호조치 및 조건을 전제로, 다운스트림 제공자를 위한 최신 모델문서화자료에 포함된 정보를 다운스트림 제공자에게 제공해야 한다. 또한 서명자는 EU법 및 회원국 법률에 따라 지식재산권 및 기밀사업정보 또는 영업비밀을 보호할 필요성을 저해하지 않는 범위에서, 다운스트림 제공자의 요청이 있는 경우 필요한 범위의 추가 정보를 제공해야 한다. 이러한 추가 정보는 다운스트림 제공자가 자신의 AI시스템에 GPAI모델을 통합하는 데 필요한 GPAI모델의 역량 및 한계를 충분히 이해하고, AI법에 따른 의무를 이행할 수 있도록 하는 데 필요한 범위로 한정된다. 서명자는 이러한 정보를 합리적인 기간 내에 제공해야 하며, 예외적인 상황이 없는 한 요청을 받은 날로부터 14일 이내에 제공하여야 한다.

서명자는 공공의 투명성을 증진하기 위하여 문서화된 정보의 전부 또는 일부를 일반 대중에게 공개할 수 있는지 여부를 검토하도록 권장된다. 이러한 정보 중 일부는 AI사무국이 제공하는 양식에 따라 AI법 제53조(1)(d)에 따라 제공자가 공개하여야 하는 학습 콘텐츠 요약의 일부로서 요약된 형태로 공개될 수 있다.

- **이행조치 1.3. 정보의 품질, 무결성 및 보안의 보장**

서명자는 문서화된 정보에 대한 품질 및 무결성 관리가 이루어지도록 하고, 해당 정보가 AI법상 의무 준수 여부를 입증하는 자료로 보존되도록 하며, 의도하지 않은 변경(alteration)으로부터 보호되도록 하여야 한다. 서명자는 정보 및 기록의 작성, 업데이트, 품질 관리 및 보안 확보와 관련하여 확립된 절차 및 기술표준을 따르는 것이 권장된다.

- **모델설명서양식(Model documentation form)**

아래에는 모델설명서양식의 편집 불가능 버전(static)을 수록하였다. 이 버전에서는 입력란을 작성할 수 없으며, 상호작용이 가능하고 내용을 입력할 수 있는 양식은 별도로 제공된다.

46) AI법 제75조(1) 및 (3), 제88조(2) 참조

Model Documentation Form

This Form includes all the information to be documented as part of Measure 1.1 of the Transparency Chapter of the Code of Practice. Crosses on the right indicate whether the information documented is intended for the AI Office (AIO), national competent authorities (NCAs) or downstream providers (DPs), namely providers of AI systems who intend to integrate the general-purpose AI model into their AI systems. Whilst information intended for DPs should be made available to them proactively, information intended for the AIO or NCAs is only to be made available following a request from the AIO, either ex officio or based on a request to the AIO from NCAs. Such requests will state the legal basis and purpose of the request and will concern only items from the Form strictly necessary for the AIO to fulfil its tasks under the AI Act at the time of the request, or for NCAs to exercise their supervisory tasks under the AI Act at the time of the request, in particular to assess compliance of high-risk AI systems built on general-purpose AI models where the provider of the system is different from the provider of the model.

Any elements of information from the Model Documentation Form shared with the AIO and NCAs shall be treated in accordance with the confidentiality obligations and trade secret protections set out in Article 78 AI Act.

Date this document was last updated: Click or tap to enter a date. **Document version number:** Click or tap here to enter text.

General information		AIO	NCAs	DPs
Legal name for the model provider:	Click here to add text.	☒	☒	☒
Model name:	The unique identifier for the model (e.g. Llama 3.1-405B), including the identifier for the collection of models where applicable, and a list of the names of the publicly available versions of the concerned model covered by the Model Documentation.	☒	☒	☒
Model authenticity:	Evidence that establishes the provenance and authenticity of the model (e.g. a secure hash if binaries are distributed, or the URL endpoint in the case of a service), where available.	☒	☒	☐
Release date:	Click or tap to enter a date. Date when the model was first released through any distribution channel.	☒	☒	☒
Union market release date:	Click or tap to enter a date. Date when the model was placed on the Union market.	☒	☒	☒
Model dependencies:	If the model is the result of a modification or fine-tuning of one or more general-purpose AI models previously placed on the market, list the model name(s) (and relevant version(s) if more than one version has been placed on the market) of those model(s). Otherwise write 'N/A'.	☒	☒	☒

Model properties						AIO	NCAs	DPs
Model architecture:	A general description of the model architecture, e.g. a transformer architecture. *[Recommended 20 words]*					☒	☒	☒
Design specifications of the model:	A general description of the key design specifications of the model, including rationale and assumptions made, to provide basic insight into how the model was designed. *[Recommended 100 words]* If any other please specify:					☒	☒	☐
Input modalities:	☐Text	☐Images	☐Audio	☐Video	If any other please specify:	☒	☒	☒
For each selected modality please include maximum input size or write 'NA' if not defined	Maximum size: …	Maximum size: …	Maximum size: …	Maximum size: …	Maximum size: …	☒	☐	☒
Output modalities:	☐Text	☐Images	☐Audio	☐Video	If any other please specify:	☒	☒	☒
For each selected modality please include maximum output size or write 'NA' if not defined	Maximum size: …	Maximum size: …	Maximum size: …	Maximum size: …	Maximum size: …	☐	☐	☒
Total model size:	The total number of parameters of the model, recorded with at least two significant figures, e.g. 7.3*10^10 parameters.					☒	☐	☐
The range within which the total number of parameters falls.	☐1—500M	☐500M—5B	☐ 5B—15B	☐15B—50B		☐	☒	☒
	☐50B—100B	☐100B—500B	☐500B—1T	☐>1T				

Methods of distribution and licenses		AIO	NCAs	DPs
Distribution channels:	A list of the methods of distribution (e.g. enterprise or subscription-based access through existing software suites or enterprise-specific solutions; public or subscription-based access through an API; public or proprietary access through integrated development environments, device-specific applications or firmware, open-source repositories) through which the model has been made available for distribution or use in the Union market. For each listed method of distribution, please include a link to information about how the model can be accessed, where available, and the level of model access (e.g. weights-level access, black-box access).	☒	☒	☐

	A list of the methods of distribution (e.g. enterprise or subscription-based access through existing software suites or enterprise-specific solutions; public or subscription-based access through an API; public or proprietary access through integrated development environments, device-specific applications or firmware, open-source repositories) through which the model can be made available to downstream providers.	☐	☐	☒
License:	A link to model license(s) (otherwise provide a copy of the license(s) upon a request from the AIO pursuant to Article 91) or indicate that no model license exists.	☒	☒	☐
	The type or category of licence(s) under which the model can be made available to downstream providers such as free and open source licences where models can be openly shared and providers can freely access, use, modify and redistribute them or modified versions thereof; less permissive licenses that impose certain restrictions on the use (e.g. to ensure ethical use), or proprietary licences that restrict access to the model's source code and impose limitations on usage, distribution, and modification. In the absence of a license, describe how access to the model is provided for downstream use, such as through terms of service.	☐	☐	☒
	A list of additional assets (e.g. training data, data processing code, model training code, model inference code, model evaluation code), if any, that are made available with a description of how each can be accessed and what licenses, if any, relate to their use.	☒	☐	☒
Use		**AIO**	**NCAs**	**DPs**
Acceptable Use Policy:	Provide a link to the acceptable use policy applicable (or attach a copy to this document) or indicate that none exists.	☒	☒	☒
Intended uses:	A description of either (i) the uses that are intended by the provider (e.g. productivity enhancement, translation, creative content generation, data analysis, data visualisation, programming assistance, scheduling, customer support, variety of natural language tasks, etc..) or (ii) the uses that are restricted and/or prohibited by the provider (beyond those prohibited by EU or international law, including Article 5 AI Act), in both cases as specified in the information supplied by the provider in the instructions for use, terms and conditions, promotional or sales materials and statements, as well as in the technical documentation. If specifying (i) or (ii) is incompatible with the nature of the license under which the model is provided, then 'N/A' can be entered. *[Recommended 200 words]*	☒	☒	☒
Type and nature of AI systems in which the general-purpose AI model can be integrated:	A list or description of either (i) the type and nature of AI systems into which the general-purpose AI model can be integrated or (ii) the type and nature of AI systems into which the general-purpose AI model should not be integrated. Examples may include autonomous systems, conversational assistants, decision support systems, creative AI systems, predictive systems, cybersecurity, surveillance, or human-AI collaboration. *[Recommended up to 300 words]*	☒	☒	☒
Technical means for model integration:	A general description of the technical means (e.g. instructions for use, infrastructure, tools) required for the general-purpose AI model to be integrated into AI systems. *[Recommended 100 words]*	☐	☐	☒
Required hardware:	A description of any hardware, including the version, required to use the model, where applicable. If not applicable (e.g. model offered via an API), 'N/A' should be entered. *[Recommended 100 words]*	☐	☐	☒
Required software:	A description of any software, including the version, required to use the model where applicable. If not applicable, 'N/A' should be entered. *[Recommended 100 words]*	☐	☐	☒
Training process		**AIO**	**NCAs**	**DPs**
Design specifications of the training process:	A general description of the main steps or stages involved in the training process, including training methodologies and techniques, the key design choices, assumptions made and what the model is designed to optimise for, and the relevance of different parameters, as applicable. For example, "the model is initialized with randomly selected weights and optimised using gradient-based optimization via the Adam optimizer in two steps. First, the model is trained to predict the next word on a large pretraining corpus using the cross-entropy loss, passing over the data for a single epoch. Second, the model is post-trained on a dataset of human preferences for 10 epochs to align the model with human values and make it more useful in responding to user prompts". *[Recommended 400 words]*	☒	☒	☐
Decision rationale:	A description of how and why key design choices were made in model training. *[Recommended 200 words]*	☒	☒	☐
Information on the data used for training, testing, and validation		**AIO**	**NCAs**	**DPs**
Data type/modality: *Select all that apply.*	☐Text ☐Images ☐Audio ☐Video — If any other please specify:	☒	☒	☒
Data provenance: *Select all that apply*	☐Web crawling ☐Private non-publicly available datasets obtained from third parties ☐User data	☒	☒	☒
For definitions of each listed category, see the Template for the Public Summary of the Training Content of General-Purpose AI models provided by the AI Office	☐Publicly available datasets ☐Data collected through other means ☐Synthetic data that is not publicly accessible (when created directly by or on behalf of the provider) — If any other please specify:			

How data was obtained and selected:	A description of the methods used to obtain and select training, testing, and validation data, including methods and resources used to annotate data, and models and methods used to generate synthetic data where applicable. For data previously obtained from third parties, a description of how the provider obtained the rights to the data if not already disclosed in the public summary of training data published in accordance with Article 53(1), point (d). *[Recommended 300 words]*	☒	☒	☐
Number of data points:	The size (in number of data points) of the training, testing, and validation data respectively, together with the definition of the unit of data points (e.g. tokens or documents, images, hours of video or frames), recorded with at least one significant figure (e.g. $3x10^{13}$ tokens).	☐	☒	☐
	The size (in number of data points) of the training, testing, and validation data respectively, together with the definition of the unit of data points (e.g. tokens or documents, images, hours of video or frames), recorded with at least two significant figures (e.g. $1.5x10^{13}$ tokens).	☒	☐	☐
Scope and main characteristics:	A general description of the scope and main characteristics of the training, testing and validation data, such as domain (e.g. healthcare, science, law,...), geography (e.g. global, restricted to a certain region,...), language, modality coverage, where applicable. *[Recommended 200 words]*	☒	☒	☐
Data curation methodologies:	General description of the data processing involved in transforming the acquired data into training, testing, and validation data for the model, such as cleaning (e.g. filtering out irrelevant content such as advertisements), normalisation (e.g. tokenizing), augmentation (e.g. back-translation). *[Recommended 300 words]*	☒	☒	☒
Measures to detect unsuitability of data sources:	A description of any methods implemented in data acquisition or processing, if any, to detect the presence of unsuitable data sources considering the model's intended uses, including but not limited to illegal content, child sexual abuse material (CSAM), non-consensual intimate imagery (NCII), and personal data leading to its unlawful processing. *[Recommended 400 words]*	☒	☒	☐
Measures to detect identifiable biases:	A description of any methods implemented in data acquisition or processing, if any, to address the prevalence of identifiable biases in the training data. *[Recommended 200 words]*	☒	☒	☐
	Computational resources (during training)	**AIO**	**NCAs**	**DPs**
Training time:	A description of what period is being measured along with the range that its duration falls under, within the following ranges: less than 1 month, 1—3 months, 3—6 months, more than 6 months.	☐	☒	☐
	A description of what period is being measured along with the duration in wall clock days (e.g. $9x10^{1}$ days) and in hardware days (e.g. $4x10^{5}$ Nvidia A100 days and $2x10^{5}$ Nvidia H100 days), both recorded with at least one significant figure.	☒	☐	☐
Amount of computation used for training:	Measured or estimated amount of computation used for training, reported in floating point operations and recorded up to its order of magnitude (e.g. 10^{24} floating point operations).	☐	☒	☐
	Measured or estimated amount of computation used for training, reported in computational operations and recorded with at least two significant figures (e.g. $2.4x10^{25}$ floating point operations).	☒	☐	☐
Measurement methodology:	In the absence of a delegated act adopted in accordance with Article 53(5) AI Act to detail measurement and calculation methodologies, describe the methodology used to measure or estimate the amount of computation used for training.	☒	☒	☐
	Energy consumption (during training and inference)	**AIO**	**NCAs**	**DPs**
Amount of energy used for training:	Measured or estimated amount of energy used for training, reported in Megawatt-hours and recorded with at least two significant figures (e.g. $1.0x10^{2}$ MWh). If the amount of energy used for training cannot be estimated due to the lack of critical information from a compute or hardware provider, enter 'N/A'.	☒	☒	☐
Measurement methodology:	In the absence of a delegated act adopted in accordance with Article 53(5) AI Act to detail measurement and calculation methodologies, describe the methodology used to measure or estimate the amount of energy used for training. Where the energy consumption of the model is unknown, the energy consumption may be estimated based on information about computational resources used. If the amount of energy used for training cannot be estimated due to a lack of critical information from a compute or hardware provider, the provider should disclose the type of information they lack. *[Recommended 100 words]*	☒	☒	☐
Benchmarked amount of computation used for inference[1]:	Benchmarked amount of computation used for inference, reported in floating point operations, recorded with at least two significant figures (e.g. $5.1x10^{17}$ floating point operations).	☒	☒	☐
Measurement methodology:	In the absence of a delegated act adopted in accordance with Article 53(5) AI Act to detail measurement and calculation methodologies, provide a description of a computational task (e.g. generating 100000 tokens) and the hardware (e.g. 64 Nvidia A100s) used to measure or estimate the amount of computation used for inference.	☒	☒	☐

[1] This item relates to energy consumption during inference, which makes up the "energy consumption of the model" (Annex XI, 2(e), AI Act) together with energy consumption during training. Since energy consumption during inference depends on more than just the model itself, the information required for this item is limited to relevant information depending only on the model, namely computational resources used for inference.

Ⅳ 저작권(Copyright Chapter)

GPAI모델에 관한 CoP
저작권 장

Alexander Peukert 워킹그룹1 공동의장
Céline Castets-Renard 워킹그룹1 부의장

1 목적

- 본 CoP의 궁극적인 목적은 AI법 제1조(1)에 따라 역내시장의 기능을 개선하고 인간중심적이며 신뢰할 수 있는 AI의 활용을 촉진하는 동시에, 민주주의, 법치주의 및 환경보호를 포함하여 헌장에 명시된 건강, 안전 및 기본권에 대한 높은 수준의 보호를 보장하고, EU 내에서 AI가 초래할 수 있는 유해한 영향으로부터 이를 보호하며, 혁신을 지원하는 데 있다.
- 이러한 궁극적인 목적을 달성하기 위하여, 본 CoP의 구체적인 목적은 다음과 같다.

  A. CoP는 AI법 제53조 및 제55조에 규정된 의무 준수를 입증하기 위한 가이드라인 문서로 기능한다. 다만, CoP를 준수하였다는 사실만으로 AI법상 해당 의무를 준수하였다는 결정적 증거가 되는 것은 아니다.

  B. CoP는 GPAI모델 제공자가 AI법상 의무를 준수할 수 있도록 지원하며, CoP를 활용하여 해당 의무의 준수를 입증하는 GPAI모델 제공자에 대하여 AI사무국이 그 준수 여부를 평가(ass)할 수 있도록 한다.

2 전문

- 다음의 사항을 고려한다.:

(a) 본 장은 AI법 제53조(1)(c)에 따른 의무가 적절히 이행될 수 있도록 기여하는 것을 목적으로 한다. 이에 따라 EU 시장에 GPAI모델을 출시하는 제공자는 저작권 및 저작인접권[47](copyright and related rights)에 관한 EU법을 준수하기 위한 정책을 마련해야 하며, 특히 Directive (EU) 2019/790 제4조(3)에 따라 권리자가 표시한 권리 유보(reservation of rights)를 최첨단 기술 등을 활용하여 식별하고 이를 준수하여야 한다. 서명자는 AI법 제53조(1)(c)의 준수를 입증하기 위하여 본 장에 규정된 이행조치를 수행한다. 다만, CoP를 준수하였다는 사실만으로 저작권 및 저작인접권에 관한 EU법을 준수한 것으로 간주되는 것은 아니다.

(b) 본 장은 저작권 및 저작인접권에 관한 EU법의 적용 및 집행에 어떠한 영향도 미치지 않으며, 해당 법의 해석은 회원국 법원과 최종적으로는 EU 사법재판소(CJEU)의 권한에 속한다.

(c) 서명자는 저작권 및 저작인접권에 관한 EU법에 대하여 다음 사항을 인정한다.:

(i) 저작권 및 저작인접권에 관한 EU법은 회원국을 대상으로 하는 지침(directives)에 규정되어 있으며, 본 장의 목적상 Directive 2001/29/EC, Directive (EU) 2019/790 및 Directive 2004/48/EC가 가장 관련성이 높다.;

(ii) 저작권 및 저작인접권에 관한 EU법은 예방적 성격을 갖는 배타적 권리(exclusive rights)를 규정하고 있으며, 따라서 예외 또는 제한이 적용되는 경우를 제외하고는 사전 동의에 기초한다.;

(iii) Directive (EU) 2019/790 제4조(1)은 텍스트 및 데이터 마이닝(Text and Data Mining, TDM)에 대한 예외 또는 제한을 규정하고 있다. 다만, 이러한 예외 또는 제한은 저작물 등에 적법하게 접근한 경우에 적용되며, 해당 조항에 언급된 저작물 및 그 밖의 보호대상의 이용에 대하여 권리자가 Directive (EU) 2019/790 제4조(3)에 따라 적절한 방식으로 권리를 명시적으로 유보하지 않은 경우에만 적용된다.

(d) 본 장에서 비례적인 이행조치를 요구하는 약정은 제공자의 규모에 상응하고 비례적이어야 하며, 스타트업을 포함한 중소기업(SMEs)의 이익을 충분히 고려해야 한다.

(e) 본 장은 서명자와 권리자 간에 저작물 및 그 밖의 보호대상의 이용을 허가하는 계약에 영향을 미치지 않는다.

47) 저작물을 직접 창작한 자가 아닌 저작물의 실연·제작·방송 등을 통하여 저작물의 전달·유통에 기여하는 자에게 인정되는 권리이며, 실연자, 음반제작자 및 방송사업자에게 인정됨

(f) AI법 제53조(1)(c)에 따른 의무의 준수를 입증하기 위한 본 장의 약정은 AI법 제53조(1)(d)에 따라 제공자가 AI사무국이 제공할 템플릿에 따라 자신의 GPAI모델 학습에 사용된 콘텐츠에 관한 충분히 상세한 요약을 작성하고 이를 공개하여야 하는 의무를 보완한다.

약정1. 저작권 정책

해당 법령: AI법 제53조(1)(c)

- (1) 서명자는 AI법 제53조(1)(c)에 따른 의무의 준수를 입증하기 위하여 저작권 정책을 수립하고, 최신 상태로 유지하며, 이를 이행할 것을 약정한다. 해당 의무에 따라 서명자는 저작권 및 저작인접권(copyright and related rights)에 관한 EU법을 준수하기 위한 정책을 마련해야 하며, 특히 Directive (EU) 2019/790 제4조(3)에 따라 권리자가 표시한 권리 유보를 최첨단 기술 등을 활용하여 식별하고 이를 준수해야 한다. 아래의 이행조치는 저작권 및 저작인접권에 관한 EU법 준수에 영향을 미치지 않으며, 서명자가 EU 시장에 출시하는 GPAI모델에 대한 저작권 정책 마련 의무의 준수를 입증하기 위한 약정을 규정한다.

- (2) 또한 서명자는 아래에 명시된 저작권 정책의 이행조치가 저작권 및 저작인접권에 관한 EU법을 각 회원국이 이행한 국내 법령, 특히 Directive (EU) 2019/790 제4조(3)을 포함하되 이에 한정되지 않는 규정에 부합하는지를 확인할 책임을 진다. 관련 회원국의 영토 내에서 저작권 관련 행위를 수행하기 전에 이러한 확인을 하여야 하며, 이를 이행하지 않을 경우 저작권 및 저작인접권에 관한 EU법에 따른 책임이 발생할 수 있다.

- **이행조치 1.1. 저작권 정책의 수립, 최신 상태 유지 및 이행**

  (1) 서명자는 EU 시장에 출시하는 모든 GPAI모델에 대하여 저작권 및 저작인접권에 관한 EU법을 준수하기 위한 정책을 수립하고, 최신 상태로 유지하며, 이를 이행하여야 한다. 서명자는 본 장에 규정된 이행조치를 포함하는 단일 문서에 해당 정책을 기술할 것을 약정한다. 또한 서명자는 해당 정책의 이행 및 감독을 위한 책임을 조직 내에 배정하여야 한다.

  (2) 서명자는 저작권 정책의 요약본을 공개하고 이를 최신 상태로 유지하도록 권장된다.

- **이행조치 1.2. 월드와이드웹(World Wide Web) 크롤링(crawling) 시 적법하게 접근 가능한 저작권 보호 콘텐츠에 한정한 복제 및 추출**

  (1) 서명자가 웹 크롤러(web-crawlers)를 사용하거나 자신을 대신하여 이를 사용하게 하여 Directive (EU) 2019/790 제2조(2)에서 정의된 텍스트 및 데이터 마이닝 및 GPAI모델의 학습을 목적으로

데이터를 스크래핑(scrape)하거나 그 밖의 방식으로 수집하는 경우, 적법하게 접근 가능한 저작물 및 그 밖의 보호대상만을 복제·추출하도록 하기 위하여 서명자는 다음 사항을 약정한다.:

a) 저작물 및 그 밖의 보호대상에 대한 무단 이용(unauthorised acts)을 방지하거나 제한하기 위하여 마련된 Directive 2001/29/EC 제6조(3)에서 정의하는 효과적인 기술적 조치를 우회하지 않을 것. 특히 구독(subscription) 모델 또는 페이월(paywall)에 의해 적용되는 접근 차단 또는 접근 제한을 존중할 것.

b) 웹 크롤링 시점에 EU 및 유럽경제지역(EEA)의 법원 또는 공공기관에 의해 상업적 규모의 저작권 및 저작인접권 침해를 지속적·반복적으로 행하는 것으로 인정된 웹사이트에 대해서는 크롤링을 수행하지 않을 것. 본 이행조치의 준수를 위하여, EU 및 유럽경제지역(EEA)의 관련 기관이 작성한 해당 웹사이트 목록으로 연결되는 하이퍼링크를 수록한 수시 업데이트 목록(dynamic list)이 EU 웹사이트에 공개될 예정이다.

● 이행조치 1.3. WWW 크롤링 시 권리 유보의 식별 및 준수

(1) 서명자가 웹 크롤러를 사용하거나 자신을 대신하여 웹 크롤러가 사용되도록 하여 이를 사용하게 하여 Directive (EU) 2019/790 제2조(2)에서 정의된 텍스트 및 데이터 마이닝 및 GPAI모델의 학습을 목적으로 데이터를 스크래핑하거나 그 밖의 방식으로 수집하는 경우, Directive (EU) 2019/790 제4조(3)에 따라 표시된 기계판독 가능한(machine-readable) 권리 유보를 최첨단 기술 등을 활용하여 식별하고 이를 준수하도록 하기 위하여 서명자는 다음 사항을 약정한다.:

a) IETF(Internet Engineering Task Force)의 RFC(Request for Comments) No. 9309에 명시된 로봇 배제 프로토콜(Robot Exclusion Protocol, REP, robots.txt)[48]에 따라 표현된 지시를 읽고 이를 따르는 웹 크롤러를 사용할 것. 또한 IETF가 해당 프로토콜의 후속 버전이 기술적으로 실현 가능하고 AI 제공자와 콘텐츠 제공자(권리자 포함)에 의해 구현될 수 있음을 입증하는 경우, 해당 후속 버전도 포함한다.

b) Directive (EU) 2019/790 제4조(3)에 따른 권리 유보를 표시하기 위한 그 밖의 적절한 기계판독 가능한 프로토콜을 식별하고 이를 준수할 것. 여기에는 자산 기반(asset-based) 또는 위치 기반(location-based) 메타데이터를 활용하는 방식이 포함될 수 있다. 이러한 프로토콜은 국제 또는 유럽 표준화기구가 채택한 것이거나 기술적으로 구현 가능하고 권리자들에 의해 널리 사용되는 등 최첨단 수준에 해당하여야 한다. 또한 이러한 프로토콜은 다양한 문화 분야를 고려하여, 권리자·AI 제공자 및 그 밖의 관련 이해관계자가 참여하는 EU 차원의 선의(bona fide)의 논의를 토대로 포용적인 절차를 거쳐 폭넓은 합의가 이루어진 것이어야 한다.

48) Robot Exclusion Protocol(REP), 웹사이트 운영자가 자동 수집기(crawler)가 접근할 수 있는 경로를 알리기 위하여 사용하는 규약 및 그 규칙을 담은 루트 디렉터리의 텍스트 파일을 의미함. 다만, 이는 크롤링 통제를 위한 규약일 뿐, 자체적으로 접근통제 또는 보안 기능을 보장하는 것은 아님

(2) 본 약정은 Directive (EU) 2019/790 제4조(3)에 따라 권리자가 저작물 및 그 밖의 보호대상을 텍스트 및 데이터 마이닝 목적으로 이용하는 것에 대하여, 온라인에 공개된 콘텐츠의 경우 기계판독 가능한 수단을 사용하는 것을 포함한 적절한 방식으로 명시적으로 권리 유보를 표시할 수 있는 권리에 영향을 미치지 않는다.

또한 본 약정은 제3자가 인터넷에서 스크래핑하거나 크롤링한 보호 콘텐츠를 서명자가 텍스트 및 데이터 마이닝 또는 GPAI모델의 학습에 이용하는 경우에도 적용되는 저작권 및 저작인접권에 관한 EU법에 영향을 미치지 않는다. 특히 이러한 경우에도 Directive (EU) 2019/790 제4조(3)에 따라 표시된 권리 유보는 계속하여 존중되어야 한다.

(3) 서명자는 본 이행조치 제1항 (a) 및 (b)에 언급된 절차를 지원하고, Directive (EU) 2019/790 제4조(3)에 따른 권리 유보를 표시하기 위한 적절한 기계판독 가능한 표준 및 프로토콜을 개발하기 위하여, 권리자 및 그 밖의 관련 이해관계자와 선의(bona fide)에 기반한 논의에 자발적으로 참여할 것이 권장된다.

(4) 서명자는 영향을 받는 권리자가 서명자가 사용하는 웹 크롤러, 해당 웹 크롤러의 robots.txt 준수 기능 및 서명자가 Directive (EU) 2019/790 제4조(3)에 따른 권리 유보를 크롤링 시점에 식별하고 준수하기 위하여 채택한 그 밖의 조치에 관한 정보를 확인할 수 있도록 적절한 조치를 취할 것을 약정한다. 이를 위하여 서명자는 해당 정보를 공개하고, 정보가 업데이트되는 경우 영향을 받는 권리자가 자동으로 통지를 받을 수 있는 수단(예: 웹 피드를 통한 신디케이션(syndication))을 마련하여야 한다. 다만, 이는 Directive 2004/48/EC 제8조에 규정된 정보청구권에 영향을 미치지 않는다.

(5) Regulation (EU) 2022/2065 제3조(j)에서 정의된 온라인 검색엔진을 제공하거나 해당 제공자를 지배하는 서명자는 본 이행조치 제1항에 따른 텍스트 및 데이터 마이닝 활동 및 GPAI모델 학습 과정에서 권리 유보를 준수한 결과가 해당 권리 유보가 표시된 콘텐츠·도메인 및/또는 URL의 검색엔진 색인(indexing)에 직접적인 불이익을 초래하지 않도록 적절한 조치를 취할 것이 권장된다. 즉, 서명자가 권리 유보를 준수하였다는 이유만으로 권리 유보가 표시된 콘텐츠·도메인 또는 URL이 검색엔진의 색인 과정에서 직접적인 불이익을 받아서는 안 된다.

- **이행조치 1.4. 저작권 침해 출력물(outputs)의 위험 완화**

(1) GPAI모델이 통합된 다운스트림 AI시스템이 저작권 및 저작인접권에 관한 EU법에 따라 보호되는 저작물 또는 그 밖의 보호대상의 권리를 침해할 수 있는 출력물을 생성할 위험을 완화하기 위하여, 서명자는 다음 사항을 약정한다.:

(a) 저작권 및 저작인접권에 관한 EU법에 따라 보호되는 학습 콘텐츠를 침해적인 방식으로 재현하는 출력물이 생성되지 않도록 적절하고 비례적인 기술적 보호조치를 구현할 것.

(b) 이용목적제한방침(Acceptable Use Policy)[49], 이용약관(terms and conditions) 또는 그 밖에 이에 상응하는 문서에서 저작권 침해적 이용을 금지할 것. 다만, 자유 및 오픈소스 라이선스로 공개된 GPAI모델의 경우에는 해당 라이선스의 자유 및 오픈소스 성격을 해치지 않는 범위에서, 모델과 함께 제공되는 문서를 통하여 이용자에게 모델의 저작권 침해적 이용이 금지된다는 점을 고지할 것.

(2) 본 이행조치는 서명자가 해당 모델을 자신의 AI시스템에 직접 통합하는 경우뿐만 아니라, 계약관계에 따라 해당 모델을 다른 주체에게 제공하는 경우에도 적용된다.

- **이행조치 1.5. 연락 창구의 지정 및 진정(complaint) 제기 절차의 마련**

(1) 서명자는 영향을 받는 권리자와의 전자적 의사소통을 위한 연락창구를 지정하고, 이에 관한 정보를 쉽게 접근할 수 있도록 제공할 것을 약정한다.

(2) 서명자는 영향을 받는 권리자 및 그 대리인(저작권 집중관리단체(collective management organisations)를 포함)이 본 장에 따른 서명자의 약정 미준수와 관련하여 충분히 구체적이고 적절한 근거를 갖춘 진정을 전자적 수단을 통하여 제출할 수 있도록 하는 절차를 마련하고, 이에 관한 정보를 쉽게 접근할 수 있도록 제공할 것을 약정한다. 서명자는 진정이 명백히 근거 없거나 동일한 권리자가 제기한 동일한 진정에 대하여 이미 답변한 경우를 제외하고, 진정을 성실하고 비자의적(non-arbitrary)인 방식으로 합리적인 기간 내에 처리해야 한다. 본 약정은 저작권 및 저작인접권의 집행을 위하여 EU법 및 회원국 법률에 따라 이용 가능한 조치, 구제수단 및 제재에 영향을 미치지 않는다.

49) Acceptable Use Policy(AUP), '이용 목적 제한 방침' 또는 '허용 가능한 사용 정책'으로도 불리며, 사용자가 기업의 네트워크·기기·인터넷 서비스 등을 이용할 때 허용되는 이용 범위와 금지되는 행위를 구체적으로 규정한 보안 및 이용 가이드라인을 의미함

V 안전 및 보안(Safety and Security Chapter)

GPAI모델 CoP
안전 및 보안 장

Matthias Samwald 워킹그룹2 의장
Marta Ziosi 워킹그룹2 부의장
Alexander Zacherl 워킹그룹2 부의장
Yoshua Bengio 워킹그룹3 의장
Daniel Privitera 워킹그룹3 부의장
Nitarshan Rajkumar 워킹그룹3 부의장
Marietje Schaake 워킹그룹4 의장
Anka Reuel 워킹그룹4 부의장
Markus Anderljung 워킹그룹4 부의장

1 목적

- 본 CoP의 궁극적인 목적은 AI법 제1조(1)에 따라 역내시장의 기능을 개선하고 인간중심적이며 신뢰할 수 있는 AI의 활용을 촉진하는 동시에, 민주주의, 법치주의 및 환경보호를 포함하여 헌장에 명시된 건강, 안전 및 기본권에 대한 높은 수준의 보호를 보장하고, EU 내에서 AI가 초래할 수 있는 유해한 영향으로부터 이를 보호하며, 혁신을 지원하는 데 있다.
- 이러한 궁극적인 목적을 달성하기 위하여, 본 CoP의 구체적인 목적은 다음과 같다.

  A. CoP는 AI법 제53조 및 제55조에 규정된 의무 준수를 입증하기 위한 가이드라인 문서로 기능한다. 다만, CoP를 준수하였다는 사실만으로 AI법상 해당 의무를 준수하였다는 결정적 증거가 되는 것은 아니다.

  B. CoP는 GPAI모델 제공자가 AI법상 의무를 준수할 수 있도록 지원하며, CoP를 활용하여 해당 의무의 준수를 입증하는 GPAI모델 제공자에 대하여 AI사무국이 그 준수 여부를 평가(ass)할 수 있도록 한다.

2 전문

- 다음의 사항을 고려한다:

(a) 적절한 수명주기 관리의 원칙

서명자는 GPAI-SR모델 제공자가 해당 GPAI-SR모델의 전체 수명주기에 걸쳐 시스템적 위험을 지속적으로 평가(ass)하고 완화해야 함을 인정한다. 여기에는 해당 GPAI-SR모델이 출시되기 전과 후에 이루어지는 개발 단계가 모두 포함된다. 또한 GPAI-SR모델 제공자는 AI 가치사슬 전반의 관련 주체(예: 해당 GPAI-SR모델의 영향을 받을 가능성이 있는 이해관계자)와 협력하고 이들의 의견을 고려하여야 하며, 향상되거나 새롭게 나타나는 모델 역량에 대응하여 시스템적 위험 관리 체계를 정기적으로 업데이트함으로써 해당 체계가 향후 변화에도 지속적으로 대응할 수 있도록 하여야 함을 인정한다(AI법 전문 114 및 115 참조). 이에 따라 서명자는 시스템적 위험을 관리하기 위한 적절한 조치를 마련하는 과정에서 일반적으로 최소한 당시의 최첨단 수준에 상응하는 기술 및 방법을 적용하여야 함을 인정한다. 다만, 보다 덜 발전된 절차, 조치, 방법론, 방법 또는 기법을 통해서도 시스템적 위험이 존재하지 않음이 결정적으로 배제될 수 있는 경우에는 예외로 한다. 또한 서명자는 시스템적 위험에 대한 평가(ass)가 다단계 과정임을 인정한다. 모델 평가는 모델의 시스템적 위험을 평가(ass)하기 위하여 사용되는 다양한 방법을 의미하며, 해당 GPAI-SR모델의 전체 수명주기에 걸쳐 필수적인 요소이다. 나아가 시스템적 위험 완화가 이루어진 경우에는 해당 완화의 효과성을 지속적으로 평가(ass)하는 것이 중요함을 인정한다.

(b) 맥락 기반 위험 평가(Contextual Risk ass) 및 완화의 원칙

서명자는 본 안전 및 보안 장(이하 "본 장")이 AI시스템이 아닌 GPAI-SR모델 제공자에게만 적용됨을 인정한다. 그러나 서명자는 시스템적 위험의 평가(ass) 및 완화가 합리적으로 예견 가능한 범위 내에서 해당 GPAI-SR모델이 통합될 수 있는 시스템 아키텍처, 그 밖의 소프트웨어 및 추론 시 이용 가능한 컴퓨팅 자원을 포함하여 이루어져야 함도 인정한다. 이는 이러한 요소들이 안전 및 보안 완화의 효과성에 영향을 미치는 경우 등을 포함하여 해당 GPAI-SR모델의 영향에 중요한 영향을 미칠 수 있기 때문이다.

(c) 시스템적 위험에 대한 비례성의 원칙

서명자는 시스템적 위험에 대한 평가(ass) 및 완화가 그 위험에 비례하여 이루어져야 함을 인정한다(AI법 제56조(2)(d) 참조). 따라서 시스템적 위험에 대한 평가(ass) 및 완화 과정에서 요구되는 검토의 수준, 특히 문서화 및 보고의 상세 수준은 해당 GPAI-SR모델의 수명주기 전반에 걸쳐 각 시점에서 존재하는 시스템적 위험에 비례하여야 한다. 또한 서명자는 시스템적 위험에 대한 평가(ass) 및 완화가 반복적이고 지속적인 과정임을 인정한다. 다만, 이미 수행한 평가(ass)가 해당 GPAI-SR모델로부터 발생하는 시스템적 위험을 판단하는데 여전히 유효한 경우에는 이를 중복하여 수행할 필요는 없다.

(d) 기존 법령과의 통합 원칙

서명자는 본 장이 연합 법령 체계의 일부를 구성하며, 다른 연합 법령과 상호 보완적으로 적용됨을 인정한다. 또한 서명자는 기밀성(영업상 비밀을 포함)에 관한 의무가 연합 법령이 요구하는 범위 내에서 유지됨을 인정하며, 본 장에 따라 AI사무국에 제출된 정보는 AI법 제78조에 따라 처리됨을 인정한다. 아울러 서명자는 AI사무국에 제출하는 향후 개발 및 향후 사업 활동에 관한 정보가 추후 변경될 수 있음을 인정한다. 또한 서명자는 본 장에 따른 건전한 위험 문화 조성을 위한 조치(이행조치 8.3)가 내부고발자 보호에 관한 Directive (EU) 2019/1937 및 AI법 제87조와 연계하여 적용되는 회원국의 이행 법령에 따라 발생하는 어떠한 의무에도 영향을 미치지 않음을 인정한다. 서명자는 또한 국제 표준이 본 장의 규정을 포괄하는 범위 내에서 이를 활용할 수 있음을 인정한다.

(e) 협력의 원칙

서명자는 시스템적 위험에 대한 평가(ass) 및 완화에 상당한 시간과 자원의 투입이 필요함을 인정한다. 또한 서명자는 예를 들어 모델 평가 방법 및/또는 인프라를 공유함으로써 협력에 따른 효율성을 높일 수 있다는 점을 인정한다. 아울러 서명자는 시스템적 위험에 대한 평가(ass) 및 완화 과정에서 라이선스 이용자, 다운스트림 수정자 및 다운스트림 제공자와의 협력이 중요함을 인정하며, 해당 GPAI-SR모델의 영향을 이해하기 위하여 시민사회, 학계 및 그 밖의 관련 이해관계자의 전문가 또는 일반 대표자와 소통하는 것의 중요성도 인정한다. 또한 서명자는 이러한 협력이 시스템적 위험에 대한 평가(ass) 및 완화와 관련된 정보를 공유하기 위하여 협정을 체결할 수 있음을 인정한다. 다만, 이 경우 민감한 정보에 대한 비례적인 보호를 보장하고 적용되는 연합 법령을 준수하여야 한다. 나아가 서명자는 AI 환경에서 새롭게 나타나는 과제와 기회에 대응하기 위하여 GPAI-SR모델 제공자, 연구자 및 규제기관 간의 협력을 촉진하기 위한 AI사무국과의 협력(AI법 제53조(3) 참조)의 중요성을 인정한다.

(f) AI 안전 및 보안에서의 혁신의 원칙

서명자는 GPAI-SR모델의 안전 및 보안을 이해하고 확보하기 위한 가장 효과적인 방법을 규명하는 일이 여전히 발전 중인 과제임을 인정한다. 또한 서명자는 본 장이 GPAI-SR모델 제공자로 하여금 AI 안전 및 보안과 관련 절차 및 조치 분야에서 최첨단 수준의 발전을 촉진하도록 장려하여야 함을 인정한다. 아울러 서명자는 최첨단 수준의 발전에는 유익한 역량을 유지하면서 특정 위험에 대응하는 정밀한 방법을 개발하는 것도 포함됨을 인정한다(예: 유익한 생물의학 역량을 과도하게 저하시키지 않으면서 생물안보 위험을 완화하는 경우). 또한 서명자는 이러한 정밀한 접근법이 보다 일반적인 방법에 비하여 더 많은 기술적 노력과 혁신을 필요로 함을 인정한다. 나아가 서명자는 GPAI-SR모델 제공자가 더 높은 효율성을 달성하는 대체 수단을 통해 동등하거나 그 이상의 안전 또는 보안 성과를 입증할 수 있는 경우, 그러한 혁신은 AI 안전 및 보안 분야의 최첨단 수준의 발전에 기여하는 것으로 인정되어야 하며, 보다 널리 채택될 수 있도록 검토되어야 함을 인정한다.

(g) 사전예방 원칙

서명자는 특히 과학적 데이터의 부족 또는 그 품질의 한계로 인하여 아직 완전한 평가(ass)가 가능하지 않은 시스템적 위험의 경우 사전예방원칙이 중요한 역할을 함을 인정한다. 이에 따라 서명자는 시스템적 위험의 식별에 있어 모델의 현재 도입률과 연구개발 추세에 대한 외삽(extrapolation)[50]에 기반한 추정을 고려해야 함을 인정한다.

(h) 중소기업(SMEs) 및 소규모 중견기업(SMCs)

GPAI-SR모델 제공자의 규모와 역량의 차이를 고려하여, 스타트업을 포함한 중소기업(SMEs) 및 소규모 중견기업(SMCs)에 대해서는 비례성 원칙에 따라 간소화된 준수 방식이 허용될 수 있다. 예를 들어, 중소기업(SMEs) 및 소규모 중견기업(SMCs)은 일부 보고 관련 약정이 면제될 수 있다(AI법 제56조(5) 참조). 중소기업(SMEs) 또는 소규모 중견기업(SMCs)에 해당하여 보고 관련 약정이 면제된 서명자는 그럼에도 불구하고 해당 약정을 자발적으로 준수할 수 있음을 인정한다.

(i) 해석의 원칙

서명자는 모든 약정 및 이행조치가 시스템적 위험의 평가(ass) 및 완화라는 목적에 비추어 해석되어야 함을 인정한다. 또한 서명자는 AI 발전 속도가 빠르다는 점을 고려할 때, 본 장이 지속적으로 효과적이고 관련성을 유지하며 향후에도 유효하게 기능하도록 하기 위하여 시스템적 위험의 평가(ass) 및 완화에 초점을 둔 목적론적 해석이 특히 중요함을 인정한다. 아울러 본 장에서 사용되는 용어 중 용어집에 수록된 정의가 있는 경우 해당 정의에 따른 의미를 가진다. 또한 서명자는 부록 1의 해석에 의문이 있는 경우 다음 사항을 고려하여 선의로 해석되어야 함을 인정한다. (1) AI법 제3조(2)의 '위험' 정의에 따른 피해 발생 가능성과 피해의 심각성; 및 (2) AI법 제3조(65)의 '시스템적 위험' 정의. 또한 서명자는 본 장이 AI사무국의 AI법 관련 지침을 함께 고려하고 이에 부합하게 해석되어야 함을 인정한다.

(j) 중대 사고(Serious incident) 보고

서명자는 중대 사고의 보고가 위법행위 또는 과실의 인정에 해당하지 않음을 인정한다. 또한 서명자는 중대 사고와 관련된 정보를 사고 발생 이후에만 사후적으로 추적·기록·보고하는 것은 충분하지 않음을 인정한다. 중대 사고의 발생에 직접적 또는 간접적으로 영향을 미칠 수 있는 정보는 여러 곳에 분산되어 있는 경우가 많으며, 서명자가 중대 사고를 인지할 시점에는 해당 정보가 이미 유실되거나 덮어쓰여지거나 단편적인 형태로만 남아 있을 수 있다. 따라서 서명자는 중대 사고 발생 이전부터 관련 정보를 추적·기록할 수 있는 절차 및 조치를 마련할 필요가 있다.

50) extrapolation, 현재의 데이터, 도입률 또는 기술 발전 추세를 미래로 연장하여 향후의 변화나 영향을 추정하는 것

약정1. 안전 및 보안 프레임워크

해당 법령 : AI법 제55조(1), 제56조(5), 전문 110, 114 및 115

- 서명자는 최첨단(state-of-the-art) 수준의 안전 및 보안 프레임워크(이하 "프레임워크")를 채택할 것을 약정한다. 프레임워크의 목적은 서명자가 자신의 모델로부터 발생하는 시스템적 위험이 수용 가능한 수준이 되도록 하기 위하여 이행하는 시스템적 위험 관리 절차 및 이행조치를 규정하는 데 있다.

  서명자는 다음의 세 단계로 구성되는 프레임워크 채택 절차를 따를 것을 약정한다.:

  (1) 프레임워크의 수립(이행조치 1.1에 규정된 내용에 따름);

  (2) 프레임워크의 이행(이행조치 1.2에 규정된 내용에 따름); 및

  (3) 프레임워크의 업데이트(이행조치 1.3에 규정된 내용에 따름).

  또한 서명자는 자신의 프레임워크를 AI사무국에 통지할 것을 약정한다(이행조치 1.4에서 규정).

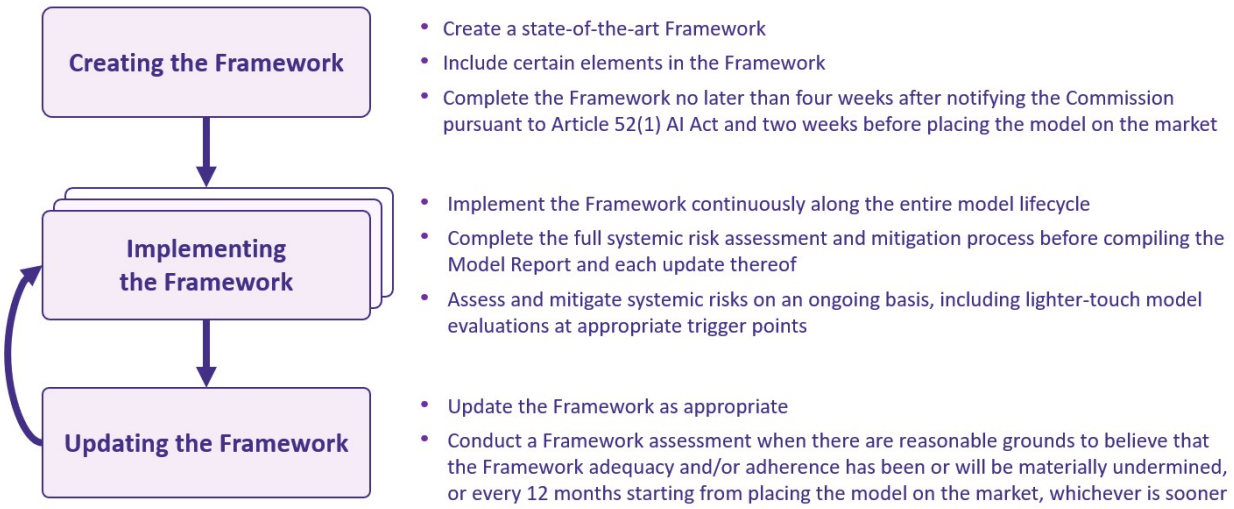

그림 15 프레임워크의 수립, 이행 및 업데이트 절차. 약정 및 이행조치의 본문이 우선한다.

- **이행조치 1.1 프레임워크의 수립**

  서명자는 자신이 개발하거나, 시장에 공급하거나, 그리고/또는 사용하는 모델을 고려하여 최첨단 프레임워크를 마련하여야 한다.

해당 프레임워크에는 본 장을 준수하기 위한 시스템적 위험 평가(ass) 및 완화와 관련하여, 현재 이행 중이거나 계획된 절차 및 이행조치에 대한 개괄적인 설명이 포함되어야 한다.

또한 프레임워크에는 다음 사항이 포함되어야 한다.:

(1) 이행조치 1.2 제2항 (1)(a)에 규정된 바와 같이, 서명자가 해당 모델의 전체 수명주기에 걸쳐 추가적인 약식 모델 평가를 수행하게 되는 트리거 시점(trigger points)과 그 활용 방식에 대한 설명 및 정당화;

(2) 약정 4에 규정된 바에 따라 서명자가 시스템적 위험이 수용 가능하다고 판단하기 위하여 다음 사항을 포함한다.:

(a) 이행조치 4.1에 규정된 바와 같이, 시스템적 위험 등급(systemic risk tiers)[51]을 포함한 시스템적 위험 수용 기준과 그 활용 방식에 대한 설명 및 정당화;

(b) 각 시스템적 위험 등급에 도달한 경우 서명자가 이행해야 하는 안전 및 보안 완화에 대한 개괄적 설명;

(c) 이행조치 4.1에 따라 시스템적 위험 등급을 정의한 각 시스템적 위험에 대하여, 서명자 현재 보유한 모델들이 이미 도달한 최고 시스템적 위험 등급을 초과하는 모델을 언제 보유하게 될 것으로 합리적으로 예상하는지에 대한 추정. 이러한 추정은 (i) 기간 범위 또는 확률분포의 형태로 제시될 수 있으며, (ii) 다른 제공자와 함께 작성한 종합 예측, 설문조사 및 기타 추정치를 고려할 수 있다. 또한 이러한 추정에는 기초가 되는 가정과 불확실성을 포함한 정당화가 수반되어야 한다; 및

(d) 이행조치 4.2에 규정된 바와 같이, 독립적인 외부 평가의 결과에 따른 경우를 제외하고 정부를 포함한 외부 주체의 의견이 서명자의 모델 개발, 시장에 공급 및/또는 사용을 계속 진행하는 데 영향을 미치는지 여부와 영향을 미치는 경우 그 절차에 대한 설명;

(3) 약정 8에서 규정된 바와 같이, 시스템적 위험 평가(ass) 및 완화 절차와 관련된 시스템적 위험 책임이 어떻게 배분되는지에 대한 설명; 및

(4) 이행조치 1.3에서 규정된 바와 같이, 서명자가 프레임워크를 업데이트하는 절차와 업데이트된 프레임워크가 확정되었는지를 판단하는 방식에 대한 설명.

서명자는 AI법 제52조(1)에 따라 위원회에 통지를 한 날부터 4주 이내이면서, 해당 모델을 시장에 출시하기 최소 2주 전까지 프레임워크를 확정하여야 한다.

51) risk tier, 위험의 수준을 계층화하여 분류한 체계로, 각 등급마다 필요한 대응 조치나 요구 조건이 상이함

이행조치 1.2 프레임워크의 이행

서명자는 아래 각 항에 규정된 바에 따라 자신의 프레임워크에 명시된 절차 및 조치를 이행해야 한다.

모델의 전체 수명주기에 걸쳐 서명자는 지속적으로 다음 사항을 수행해야 한다.:

(1) 다음의 방법을 통해 모델로부터 발생하는 시스템적 위험을 평가(ass)한다:

(a) 시간, 학습 연산량, 개발 단계, 사용자 접근, 추론 연산량 및/또는 기능활용성(affordances) 등을 기준으로 정의된 적절한 트리거 시점(trigger points)에서 **부록** 3을 반드시 준수할 필요가 없는 약식 모델 평가(예: 자동화된 평가)의 수행;

(b) **이행**조치 3.5에서 규정된 바에 따른 모델 출시 후 사후 모니터링의 수행;

(c) 약정 9에 따른 중대 사고 관련 정보의 고려; 및

(d) (a), (b) 및 (c)의 결과를 바탕으로 평가(ass)의 범위 및/또는 깊이를 확대하거나, 다음 항에 규정된 전체 시스템적 위험 평가(ass) 및 완화 절차를 수행할 것; 및

(2) (1)의 결과를 고려하여 시스템적 위험 완화를 이행하며, 필요한 경우 중대 사고에 대한 대응을 포함한다.

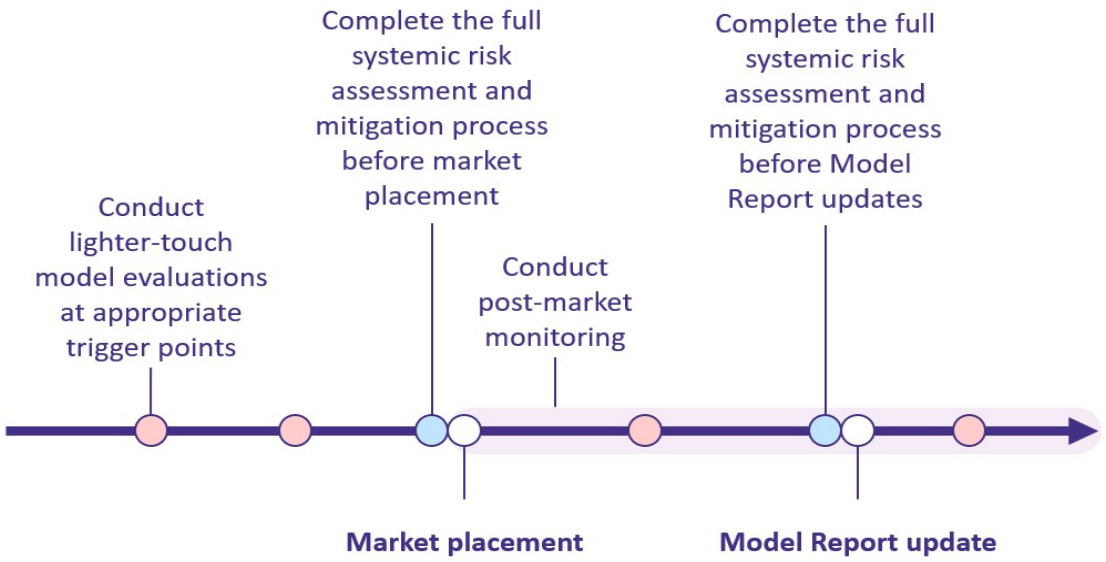

그림 2. 모델 수명주기 전반에 걸친 시스템적 위험 평가(ass) 및 완화의 예시적 타임라인.
약정 및 이행조치의 본문이 우선한다.

또한 서명자는 해당 모델에 대하여 이미 수행한 시스템적 위험 평가(ass) 중 현재에도 유효한 부분은

다시 수행할 필요 없이, 다음의 네 단계로 구성되는 전체 시스템적 위험 평가(ass) 및 완화 절차를 이행한다.:

(1) 약정 2에 규정된 바에 따라 모델로부터 발생하는 시스템적 위험의 식별;

(2) 약정 3에 규정된 바에 따라 식별된 시스템적 위험 각각에 대한 분석;

(3) 이행조치 4.1에 규정된 바에 따라 모델로부터 발생하는 시스템적 위험이 수용 가능한지 여부의 판단;

(4) 모델로부터 발생하는 시스템적 위험이 수용 가능한 것으로 판단되지 않는 경우, 약정5 및 약정6에서 규정된 바에 따라 안전 및/또는 보안 완화를 이행하고, 이행조치 4.2에 규정된 바에 따라 (1)부터 다시 시작하여 해당 시스템적 위험을 재평가(re-ass)하는 것;

서명자는 이러한 전체 시스템적 위험 평가(ass) 및 완화 절차를 적어도 모델을 출시하기 전에는 수행해야 하며, 또한 이행조치 7.6 제1항 및 제3항에 규정된 요건이 충족될 때마다 이를 수행해야 한다. 서명자는 약정7에 규정된 바에 따라 자신이 이행한 조치 및 절차를 AI사무국에 보고해야 한다.

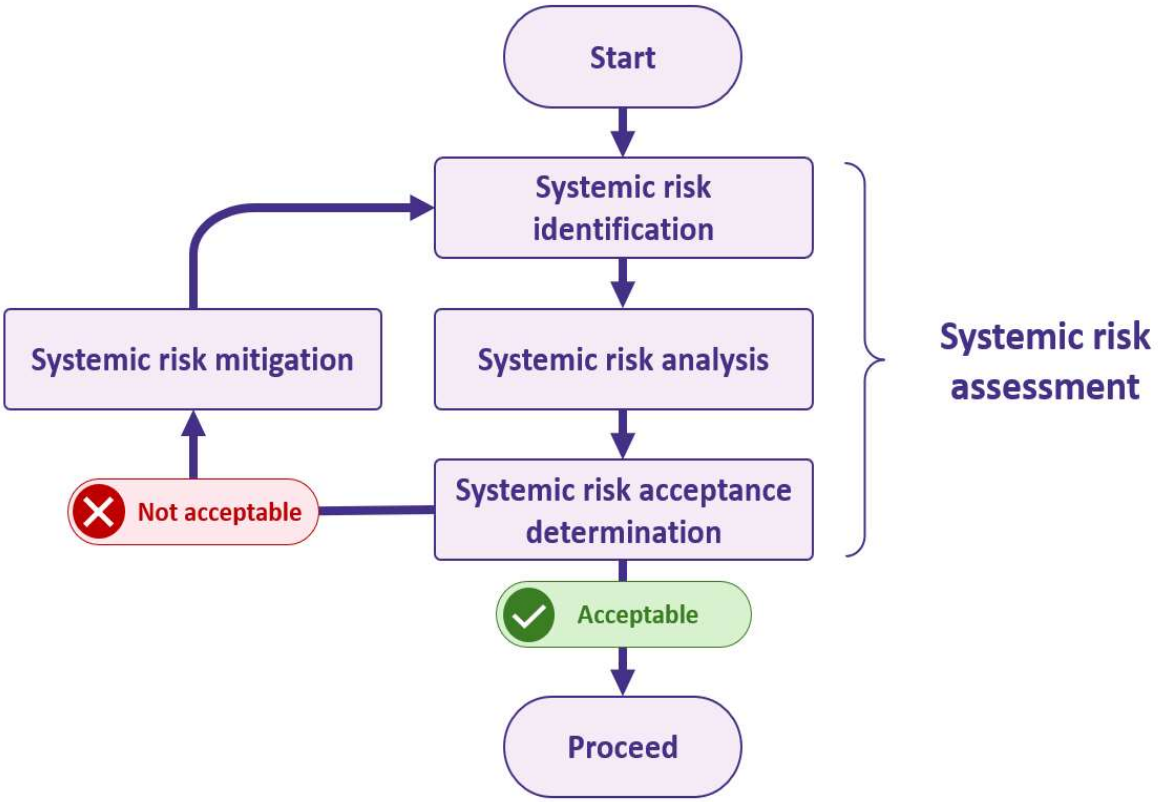

그림 3. 전체 시스템적 위험 평가(ass) 및 완화 절차.
약정 및 이행조치의 본문이 우선한다.

● 이행조치 1.3 프레임워크의 업데이트

서명자는 프레임워크를 적절하게 업데이트하며, 프레임워크 평가(ass)(다음 문단에서 규정됨) 이후에는 부당한 지체 없이 이를 포함하여 업데이트를 수행하여, 이행조치 1.1의 정보가 최신 상태로 유지되고 프레임워크가 최소한 최첨단 수준을 유지하도록 한다. 프레임워크의 모든 업데이트에 대하여 서명자는 변경 이력(changelog)을 포함하여야 하며, 해당 이력에는 프레임워크가 어떻게 그리고 왜 업데이트되었는지에 대한 설명과 함께 버전 번호 및 변경 일자를 포함해야 한다.

서명자는 자신의 프레임워크의 적정성 및/또는 이에 대한 준수가 중대하게 저해되었거나 저해될 것으로 합리적으로 판단할 근거가 있는 경우 혹은 모델을 시장에 출시한 날부터 12개월이 되는 시점 중 더 이른 시점에 적절한 프레임워크 평가(ass)를 수행하여야 한다. 이러한 근거의 예시는 다음과 같다.:

(1) 서명자의 모델 개발 방식이 중대하게 변경되었거나 변경될 것으로 합리적으로 예상되며, 이로 인해 적어도 하나의 모델로부터 발생하는 시스템적 위험이 수용 가능하지 않게 될 것으로 합리적으로 예상되는 경우;

(2) 자신의 모델 또는 유사한 모델과 관련된 중대 사고 및/또는 아차사고[52]가 발생하였으며, 이로 인해 적어도 하나의 모델로부터 발생하는 시스템적 위험이 수용 가능하지 않음을 시사하는 경우;

(3) 적어도 하나의 모델로부터 발생하는 시스템적 위험이 중대하게 변화하였거나 변화할 가능성이 있는 경우. 예를 들어 안전 및/또는 보안 완화가 중대하게 효과를 상실하였거나 상실할 가능성이 있거나, 적어도 하나의 모델의 역량 및/또는 경향이 중대하게 변화하였거나 변화할 가능성이 있는 경우.

프레임워크 평가(ass)는 다음을 포함한다:

(1) 프레임워크 적정성: 프레임워크에 포함된 절차 및 이행조치가 서명자의 모델로부터 발생하는 시스템적 위험에 적절한지 여부에 대한 평가(ass). 이 평가(ass)는 모델이 현재 어떠한 방식으로 개발·시장에 공급 및/또는 사용되고 있는지와 향후 12개월 동안 어떠한 방식으로 개발·시장에 공급 및/또는 사용될 것으로 예상되는지를 고려한다.

(2) 프레임워크 준수: 서명자의 프레임워크 준수 여부에 초점을 둔 평가(ass)로서, 다음을 포함한다: (a) 직전 프레임워크 평가(ass) 이후 프레임워크 미준수 사례 및 그 사유; 및 (b) 프레임워크에 대한 지속적인 준수를 보장하기 위해 이행되어야 하는 조치(안전 및 보안 완화를 포함). (a) 및/또는 (b)로 인해 향후 미준수 위험이 발생하는 경우, 서명자는 프레임워크 평가(ass)의 일환으로 시정계획을 수립한다.

● 이행조치 1.4 프레임워크 통지

52) near misses, 실제 사고는 발생하지 않았으나, 사고로 이어졌을 가능성이 높았던 위기 상황이나 사례

서명자는 프레임워크 및 그 업데이트에 대하여, 각각 확정된 날로부터 5영업일 이내에 AI사무국이 (삭제 또는 가림처리되지 않은 형태로) 접근할 수 있도록 제공한다.

약정2. 시스템적 위험 식별

해당 법령: AI법 제55조(1) 및 전문 110

- 서명자는 모델로부터 발생하는 시스템적 위험을 식별할 것을 약정한다. 시스템적 위험 식별의 목적에는 시스템적 위험 분석(약정 3에 따름) 및 시스템적 위험의 수용 가능 여부 판단(약정 4에 따름)을 지원하는 것이 포함된다.

  시스템적 위험 식별은 다음의 두 가지 요소로 구성된다.:

  (1) 이행조치 2.1에 규정된 바와 같이, 모델로부터 발생하는 시스템적 위험을 식별하기 위한 체계적인 절차를 따르는 것; 및

  (2) 이행조치 2.2에 규정된 바와 같이, 식별된 각 시스템적 위험에 대한 시나리오를 개발하는 것.

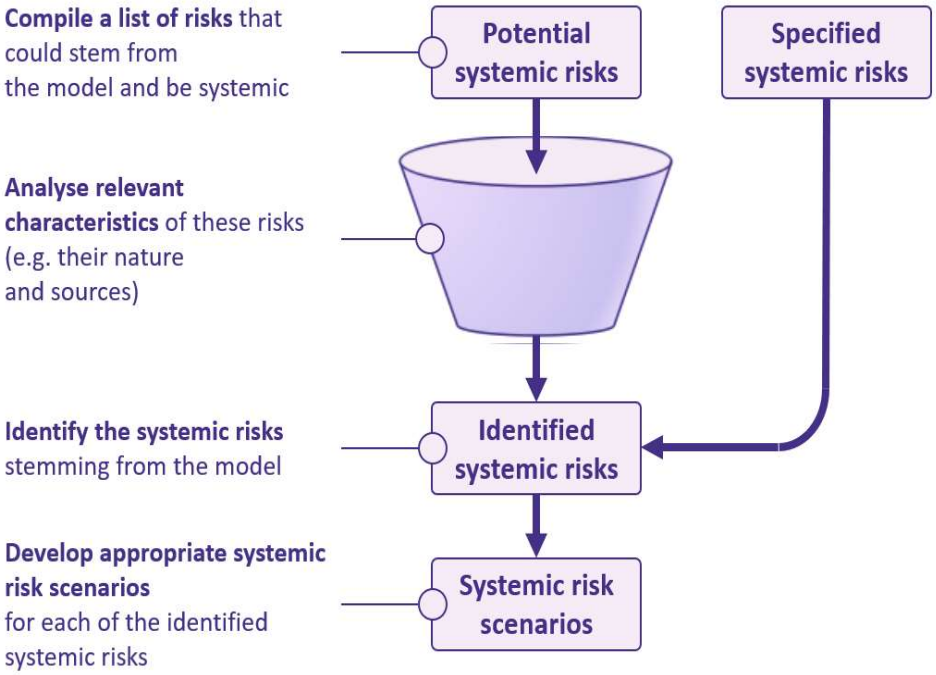

그림 4. 시스템적 위험 식별 절차.
약정 및 이행조치의 본문이 우선한다.

- **이행조치 2.1 시스템적 위험 식별 절차**

서명자는 다음을 식별해야 한다:

(1) 다음 절차를 통해 도출된 시스템적 위험:

(a) 부록 1.1의 위험 유형을 바탕으로, 다음 사항을 고려하여 모델로부터 발생할 수 있으며 시스템적 위험에 해당할 수 있는 위험 목록을 작성하는 것:

(i) 모델 비의존적 정보(model-independent information)(이행조치 3.1에 따름);

(ii) 사후 모니터링에서 얻은 정보(이행조치 3.5에 따름) 및 중대 사고 및 아차사고에 관한 정보(약정 9에 따름)를 포함하여, 해당 모델 및 유사한 모델에 관한 관련 정보; 및

(iii) AI사무국, 독립 전문가 과학 패널, 또는 AI사무국이 이러한 목적을 위하여 승인한 국제 AI 안전연구소 네트워크(International Network of AI Safety Institutes) 등의 기타 이니셔티브가 서명자에게 직접 또는 공개 발표를 통하여 전달한 기타 관련 정보;

(b) (a)에 따라 작성된 위험의 관련 특성(예: 그 성격(부록 1.2에 기반함) 및 발생원천(부록 1.3에 기반함)을 분석하는 것; 및

(c) (b)를 바탕으로 모델로부터 발생하는 시스템적 위험을 식별하는 것; 및

(2) 부록 1.4에 명시된 시스템적 위험.

- **이행조치 2.2 시스템적 위험 시나리오**

서명자는 식별된 각 시스템적 위험(이행조치 2.1에 따름)에 대하여, 이러한 시스템적 위험 시나리오의 수 및 세부 수준에 관한 사항을 포함하여, 적절한 시스템적 위험 시나리오를 개발해야 한다.

약정3. 시스템적 위험 분석

해당 법령: AI법 제55조(1) 및 전문 114

- 서명자는 식별된 각 시스템적 위험(약정 2에 따름)을 분석할 것을 약정한다. 시스템적 위험 분석의 목적에는 시스템적 위험의 수용 가능 여부 판단(약정 4에 따름)을 지원하는 것이 포함된다.

시스템적 위험 분석은 식별된 시스템적 위험 각각에 대하여 다음의 다섯 가지 요소로 구성되며, 이러한 요소들은 서로 중첩될 수 있고 반복적으로 이행될 필요가 있을 수 있다.:

(1) 모델 비의존적 정보의 수집(이행조치 3.1에 규정된 바와 같음);

(2) 모델 평가의 수행(이행조치 3.2에 규정된 바와 같음);

(3) 시스템적 위험의 모델링(이행조치 3.3에 규정된 바와 같음); 및

(4) 시스템적 위험의 추정(이행조치 3.4에 규정된 바와 같음); 동시에

(5) 사후 모니터링의 수행(이행조치 3.5에 규정된 바와 같음).

● 이행조치 3.1 모델 비의존적 정보

서명자는 시스템적 위험과 관련된 모델 비의존적 정보를 수집해야 한다.

서명자는 해당 시스템적 위험의 특성에 맞게 정보 탐색 및 수집의 범위와 강도를 조정하여 이러한 정보를 탐색하고 수집하며, 다음과 같은 방법을 활용할 수 있다.:

(1) 웹 검색(예: 오픈 소스로부터 수집된 정보를 수집 및 분석하는데 오픈소스 인텔리전스(open-source intelligence) 기법을 활용하는 것);

(2) 문헌 검토;

(3) 시장 분석(예: 시장에 공급된 다른 모델의 역량을 중심으로 한 분석);

(4) 학습 데이터 검토(예: 데이터 오염(data poisoning)[53] 또는 변조(tampering)[54] 여부를 확인하기 위한 검토;

(5) 과거 사고 데이터 및 사고 데이터베이스의 검토 및 분석;

(6) 전반적인 동향에 대한 예측(예: 알고리즘 효율성, 연산 자원 사용량, 데이터 가용성 및 에너지 사용의 발전에 관한 예측);

(7) 전문가 인터뷰 및/또는 전문가 패널; 및/또는

(8) 취약계층을 포함한 자연인에 대한 모델의 영향 등을 조사하는 일반인 인터뷰, 설문조사, 지역사회 협의 또는 기타 참여형 연구 방법.

53) data poisoning, 해커나 악의적인 공격자가 AI나 머신러닝(ML) 모델의 학습 데이터에 의도적으로 오류나 악성 데이터를 주입하거나 조작하여, AI가 잘못된 판단을 내리도록 망가뜨리는 사이버 공격

54) tampering, AI모델, 학습 데이터, 또는 AI가 실행되는 시스템을 허가 없이 변조, 조작, 오염시키는 행위/데이터, 소프트웨어, 하드웨어 또는 시스템을 허가 받지 않은 방식으로 변경하거나 조작하는 악의적인 행위

이행조치 3.2 모델 평가

서명자는 부록 3에 규정된 바와 같이, 모델의 역량, 경향, 기능활용성 및/또는 영향을 평가(ass)하기 위하여, 시스템적 위험과 관련된 모달리티에 대하여 최소한 최첨단 수준의 모델 평가를 수행하여야 한다.

서명자는 이러한 모델 평가가 모델 및 시스템적 위험에 적절한 방법을 사용하여 설계되고 수행되도록 보장하며, 예상치 못한 행태, 역량의 한계 또는 창발적 특성(emergent properties)을 식별하기 위한 목적으로 시스템적 위험에 대한 이해를 제고하기 위한 개방형 테스트(open-ended testing)를 포함하도록 한다. 모델 평가 방법의 예시는 다음과 같다: 질의응답(Q&A) 세트, 작업 기반 평가, 벤치마크(benchmarks)[55], 레드티밍(red-teaming)[56] 및 그 밖의 적대적 테스트(adversarial testing) 방법, 인간 업리프트 연구(human uplift studies), 모델 유기체(model organisms)[57], 시뮬레이션, 및/또는 기밀 자료(classified materials)에 대한 대리 평가. 또한 모델 평가의 설계는 이행조치 3.1에 따라 수집된 모델의 비의존적 정보를 바탕으로 이루어져야 한다.

이행조치 3.3 시스템적 위험 모델링

서명자는 시스템적 위험에 대한 시스템적 위험 모델링을 수행한다. 이를 위하여 서명자는 다음을 수행한다.:

(1) 최소한 최첨단 수준에 부합하는 위험 모델링 방법을 사용할 것;

(2) 이행조치 2.2에 따라 개발된 시스템적 위험 시나리오를 기반으로 할 것; 및

(3) 최소한 이행조치 2.1 및 본 약정에 따라 수집된 정보를 고려할 것.

이행조치 3.4 시스템적 위험 추정

서명자는 해당 시스템적 위험에 대한 피해의 발생 가능성 및 심각도를 추정해야 한다.

서명자는 최소한 최첨단 수준에 부합하는 위험 추정 방법을 사용해야 하며, 최소한 약정2, 본 약정 및 약정9에 따라 수집된 정보를 고려해야 한다. 시스템적 위험에 대한 추정은 위험 점수, 위험 매트릭스(risk matrix), 확률분포 또는 기타 적절한 형식으로 표현되어야 하며, 정량적, 반정량적 및/또는 정성적일 수 있다. 이러한 시스템적 위험 추정의 예시는 다음과 같다.: (1) 정성적 시스템적 위험 점수(예: "보통" 또는 "심각)"); (2) 정성적 시스템적 위험 매트릭스(예: "발생 가능성: 낮음" × "영향: 높음"); 및/또는 (3)

55) benchmark, 특정 AI모델의 성능과 역량을 객관적으로 측정하고 비교하기 위해 설계된 표준화된 시험
56) red-teaming, 시스템의 취약점이나 오용 가능성을 사전에 탐색하기 위해 실시하는 공격적 시뮬레이션 평가 방식
57) model organism, AI 연구에서 특정 문제적 행동이나 위험을 단순화·재현하여, 그 원인과 완화 방법을 실험적으로 분석하기 위해 사용하는 모델 또는 테스트베드

정량적 시스템적 위험 매트릭스(예: "X-Y%" × "X-Y 유로(EUR) 규모의 피해").

이행조치 3.5 사후시장 모니터링

서명자는 시스템적 위험이 수용 가능하지 않은 것으로 판단될 수 있는지 여부(이행조치 4.1에 따름)를 평가(ass)하는데 관련된 정보를 수집하고, 모델 보고서(Model Report)의 업데이트 필요 여부(이행조치 7.6에 따름)를 판단하는 데 필요한 정보를 제공하기 위하여, 적절한 사후 모니터링을 수행한다. 또한 서명자는 일정에 대한 추정치를 산출하는 데 관련된 정보를 수집하기 위하여(이행조치 1.1 제(2)(c)에 따름) 최선의 노력을 기울여 사후 모니터링을 수행해야 한다.

이러한 목적을 위하여, 사후 모니터링은 다음을 수행해야 한다.:

(1) 모델의 역량, 경향, 기능활용성 및/또는 효과에 대한 정보를 수집할 것;

(2) 아래에 열거된 예시적인 방법을 고려할 것; 및

(3) 서명자가 자신의 모델을 통합한 AI시스템을 직접 제공 및/또는 배포하는 경우, 해당 AI시스템의 일부로서 모델을 모니터링하는 것을 포함할 것.

위 (2)의 목적을 위한 사후 모니터링 방법의 예시는 다음과 같다.:

(1) 엔드유저(end-user) 피드백 수집;

(2) (익명) 신고 채널 제공;

(3) (중대) 사고 신고 양식 제공;

(4) 버그 바운티(bug bounties) 제공;

(5) 커뮤니티 주도의 모델 평가 및 공개 리더보드 구축;

(6) 영향을 받는 이해관계자와의 정기적인 대화 실시;

(7) 사용 양상을 파악하기 위하여 소프트웨어 저장소, 알려진 악성코드, 공개 포럼 및/또는 소셜 미디어를 모니터링하는 것;

(8) 학계, 시민사회, 규제기관 및/또는 독립 연구자와 협력하여 모델의 역량, 경향, 기능활용성 및/또는 효과에 관한 과학적 연구를 지원하는 것;

(9) 예를 들어 워터마크(watermarks), 메타데이터(metadata) 및/또는 그 밖의 최소한 최첨단 수준에 부합하는 출처 추적(provenance) 기법 등을 활용하여 모델의 입력 및 출력에 대한 개인정보 보호형

로깅(logging) 및 메타데이터 분석 기법을 구현하는 것;

(10) 모델 이용 제한 위반 및 그 위반으로 인해 후속적으로 발생한 사고에 관한 관련 정보를 수집하는 것; 및/또는

(11) 시스템적 위험의 평가(ass) 및 완화와 관련되나 제3자에게는 투명하게 공개되지 않는 모델의 측면을 모니터링하는 것. 예를 들어, 파라미터가 공개적으로 다운로드 가능하도록 제공되지 않는 모델의 숨겨진 사고의 사슬(chains-of-thought).

사후 모니터링을 지원하기 위하여, 서명자는 다음에 대하여 독립적인 외부 평가자가 접근할 수 있도록 적정한 수의 독립적인 외부 평가자에게 적정한 수준의 무상 접근권(free access)을 제공해야 한다.:

(1) 시스템적 위험과 관련하여 시장에 공급된 해당 모델의 최고 성능 모델 버전;

(2) (1)의 모델 버전의 사고의 사슬(이용 가능한 경우); 및

(3) 해당 시스템적 위험과 관련하여 안전 완화가 가장 적게 적용된 (1)의 모델 버전에 대응하는 모델 버전(예를 들어, 존재하는 경우 유용성 중심 모델 버전(helpful-only model version[58]) 및 이용 가능한 경우 그 사고의 사슬(chains-of-thought);

다만 해당 모델이 동일한 시스템적 위험과 관련하여 동등한 수준으로 안전하거나 그보다 더 안전한 모델로 간주되는 경우는 제외한다(부록 2.2에 따름). 서명자는 상황에 따라 API, 온프레미스 접근(on-premise access)(이전(transport) 포함), 서명자가 제공하는 하드웨어를 통한 접근 또는 모델 파라미터의 공개 다운로드 제공 등의 방식으로 이러한 모델 접근을 제공할 수 있다.

전 항에 따른 독립적인 외부 평가자 선정을 위하여, 서명자는 신청서를 평가(ass)하기 위한 적절한 기준을 공개해야 한다. 해당 평가자의 수, 선정 기준 및 보안 조치는 전항의 (1), (2) 및 (3)에 따라 서로 달라질 수 있다.

서명자는 모델로부터 발생하는 시스템적 위험을 평가(ass) 및 완화하기 위한 목적에 한하여 독립적인 외부 평가자의 평가 결과에 접근하고, 이를 저장 및/또는 분석하여야 한다. 특히 서명자는 평가자의 명시적인 허가 없이 그러한 시험 실행의 입력 및/또는 출력 데이터를 자사 모델의 학습에 사용하지 않는다. 또한 서명자는 독립적인 외부 평가자가 다음 사항을 준수하는 한, 해당 평가자의 테스트 수행 및/또는 발견사항의 공개를 이유로 법적 또는 기술적 보복을 하여서는 아니 된다:

(1) 명시적으로 허용된 경우를 제외하고, 테스트를 통하여 모델의 이용 가능성을 의도적으로 방해하지 않을 것;

58) helpful-only model, 안전·정렬 제한을 최소화한 상태에서 유용한 응답 생성 능력에만 초점을 두고 동작하도록 구성된 모델 버전

(2) EU 법을 위반하여 민감하거나 기밀인 사용자 데이터에 의도적으로 접근·수정 및/또는 사용하지 않을 것. 또한 평가자가 그러한 데이터에 접근하게 되는 경우, 필요한 범위 내에서만 이를 수집하고, 유포하지 않으며, 법적으로 가능한 한 신속히 삭제할 것;

(3) 자신의 접근 권한을 공공의 안전 및 보안에 중대한 위험을 초래하는 활동에 의도적으로 사용하지 않을 것;

(4) 발견 사항을 이용하여 서명자, 사용자 또는 가치사슬 내의 다른 행위자를 위협하지 않을 것. 다만 사전에 합의된 정책 및 일정에 따른 공개는 그러한 강요에 해당하지 않는다.; 및

(5) 책임 있는 취약점 공개를 위한 서명자의 공개 절차를 준수할 것. 해당 절차에는 최소한, 발견사항의 공개가 시스템적 위험을 실질적으로 증가시키는 경우와 같이 예외적으로 더 긴 기간이 필요한 경우를 제외하고, 서명자가 발견 사항을 인지한 날부터 30영업일을 초과하여 공개를 지연시키거나 차단할 수 없다는 내용이 포함되어야 한다.

SMEs 또는 SMCs에 해당하는 서명자는 본 조치의 준수를 지원하거나 촉진하기 위한 지원 또는 자원을 제공받기 위하여 AI사무국에 연락할 수 있다.

약정4. 시스템적 위험 수용 여부 판단

해당 법령: AI법 제55조(1)

서명자는 시스템적 위험의 수용 기준[59]을 명시하고, 모델로부터 발생하는 시스템적 위험이 수용 가능한지 여부를 판단할 것을 약정한다(이행조치 4.1에 규정된 바와 같음). 서명자는 시스템적 위험의 수용 가능 여부 판단을 바탕으로, 모델 개발, 시장에 공급 및/또는 사용을 계속 진행할지 여부를 결정할 것을 약정한다(이행조치 4.2에 규정된 바와 같음).

● 이행조치 4.1 시스템적 위험 수용 기준 및 수용 여부 판단

서명자는 모델로부터 발생하는 시스템적 위험이 수용 가능한지 여부를 어떻게 판단할 것인지 설명하고 정당화하여야 한다(이행조치 1.1 (2)(a)에 따른 프레임워크에 따름). 이를 위하여 서명자는 다음 사항을 수행하여야 한다.:

(1) 식별된 각 시스템적 위험(이행조치 2.1에 따름)에 대하여 최소한 다음 중 하나를 수행할 것:

(a) 다음과 같은 적절한 시스템적 위험 등급을 정의할 것:

59) risk acceptance criteria, 시스템적 위험이 '허용 가능한 수준'인지 판단하기 위해 설정된 평가 기준으로, 주관적 해석 여지가 존재함

(i) 모델의 역량을 기준으로 정의되고, 추가적으로 모델의 경향, 위험 추정치 및/또는 기타 적절한 지표를 포함할 수 있을 것;

(ii) 측정 가능할 것; 및

(iii) 해당 모델이 아직 도달하지 않은 시스템적 위험 등급을 최소 하나 이상 포함할 것; 또는

(b) 시스템적 위험 등급이 해당 시스템적 위험에 적절하지 않고, 해당 시스템적 위험이 특정 시스템적 위험(부록 1.4에 따름)에 해당하지 않는 경우, 기타 적절한 시스템적 위험 수용 기준을 정의할 것;

(2) 이러한 등급 및/또는 기타 기준을 사용하여 식별된 각 시스템적 위험(이행조치 2.1에 따름) 및 전체 시스템적 위험이 수용 가능한지 여부를 어떻게 판단할 인지 설명할 것; 및

(3) (2)에 따른 이러한 등급 및/또는 기타 기준의 사용이 식별된 각 시스템적 위험(이행조치 2.1에 따름) 및 전체 시스템적 위험이 수용 가능성을 어떻게 보장하는지 정당화할 것.

서명자는 안전 여유(safety margin)[60]를 반영하여(다음 문단에 규정된 바와 같음), 식별된 각 시스템적 위험(이행조치 2.1에 따름)에 시스템적 위험 수용 기준을 적용해야 한다. 이를 바탕으로 식별된 각 시스템적 위험 및 전체 시스템적 위험이 수용 가능한지 여부를 판단해야 한다. 이러한 수용 가능 여부 판단 시에는 최소한 시스템적 위험 식별 및 분석(약정2 및 약정3에 따름)을 통하여 수집된 정보를 고려해야 한다.

안전 여유는 다음 사항을 충족해야 한다.:

(1) 해당 시스템적 위험에 적절할 것; 및

(2) 다음 사항의 잠재적 한계, 변화 및 불확실성을 고려할 것:

(a) 시스템적 위험 발생원천(예: 평가(ass) 시점 이후의 역량 향상);

(b) 시스템적 위험 평가(ass)(예: 모델 평가의 과소 유도(under-elicitation) 또는 유사한 시스템적 위험 평가(ass)의 과거 정확도); 및

(c) 안전 및 보안 완화의 효과성(예: 안전 및 보안 완화가 우회(circumvent), 비활성화(deactivate) 또는 무력화(subvert)되는 경우).

● 이행조치 4.2 시스템적 위험 수용 여부 판단에 따른 진행 여부 결정

서명자는 모델로부터 발생하는 시스템적 위험이 수용 가능한 것으로 판단되는 경우에만(이행조치 4.1에

60) safety margin, 특정 위험 기준에 도달하기 전 미리 조치할 수 있도록 설정된 보수적 기준 또는 임계값

따름), 모델의 개발, 시장 제공 및/또는 사용을 계속해야 한다.

모델로부터 발생하는 시스템적 위험이 수용 가능한 것으로 판단되지 않거나, 가까운 시일 내에 수용 가능한 것으로 판단되지 않게 될 것이 합리적으로 예견되는 경우(이행조치 4.1에 따름), 서명자는 계속 진행하기에 앞서 모델로부터 발생하는 시스템적 위험이 수용 가능하며 앞으로도 수용 가능한 상태를 유지하도록 보장하기 위한 적절한 조치를 취하여야 한다. 특히 서명자는 다음 사항을 수행하여야 한다.:

(1) 필요한 경우 모델을 시장에 공급하지 않거나, 시장에 공급하는 것을 제한하거나(예: 라이선스 또는 이용 제한의 조정), 모델을 시장에서 철수하거나 회수할 것;

(2) 안전 및/또는 보안 완화를 구현할 것(약정 5 및 약정6에 따름); 및

(3) 시스템적 위험 식별(약정 2에 따름), 시스템적 위험 분석(약정 3에 따름), 및 시스템적 위험 수용 여부 판단(본 약정에 따름)의 다시 수행할 것.

약정5. 안전 완화

해당 법령: AI법 제55조(1) 및 전문 114

서명자는 모델로부터 발생하는 시스템적 위험이 수용 가능하도록 보장하기 위하여(약정 4에 따름), 본 약정의 이행조치에 규정된 바와 같이 모델의 전체 수명주기에 걸쳐 적절한 안전 완화를 구현할 것을 약정한다.

● 이행조치 5.1 적절한 안전 완화

서명자는 모델의 공개 및 배포 전략[61]을 고려하여, 적대적 압박(adversarial pressure)(예: 미세조정 공격(fine-tuning attacks)[62] 또는 탈옥(jailbreaking)[63] 에도 충분히 견딜 수 있는 수준의 안전 완화를 이행하여야 한다.

안전 완화의 예시는 다음과 같다.:

(1) 신뢰할 수 없는 사고의 사슬 흔적(unfaithful chain-of-thought traces)과 같은 바람직하지 않은 모델 경향을 초래할 수 있는 데이터를 포함하여, 학습 데이터를 필터링하고 정제하는 것;

(2) 모델의 입력 및/또는 출력을 모니터링하고 필터링하는 것;

(3) 특정 요청을 거부하거나 유용하지 않은 응답을 제공하도록 모델을 미세조정하는 등 안전을 위하여

61) deployment strategy, AI모델을 어떤 환경에 어떻게 출시하고, 누구에게 어떻게 제공할지를 결정하는 조직의 운영 방침
62) fine-tuning attacks, 모델의 세부 조정을 악용하여 보안 우회 또는 응답 조작을 유도하는 공격 방식
63) jailbreaking, AI모델의 내장된 안전장치를 우회하거나 무력화시키는 입력 기법 또는 공격

모델 행동을 변경하는 것;

(4) 검증된 사용자에게 API 접근을 제한적으로 제공하거나, 사후 모니터링을 바탕으로 접근 범위를 점진적으로 확대하거나, 초기에는 모델 파라미터를 공개적으로 다운로드할 수 있도록 제공하지 않는 것 등을 통하여 모델에 대한 접근을 단계적으로 제공하는 것;

(5) 모델로부터 발생하는 시스템적 위험을 완화하기 위하여 다른 행위자가 사용할 수 있는 도구를 제공하는 것;

(6) 모델의 행동에 관하여 고보증(high-assurance)의 정량적 안전 보장을 제공하는 기법[64];

(7) 모델 식별(model identifications), 특화된 통신 프로토콜 또는 사고 모니터링 도구 등 안전한 AI 에이전트 생태계를 가능하게 하는 기법; 및/또는

(8) 사고의 사슬 추론(chain-of-thought reasoning)에 대한 투명성을 확보하거나 모델이 다른 안전 완화를 무력화하는 능력에 대응하기 위한 기법 등 그 밖의 새로운 안전 완화

약정6. 보안 완화(Security mitigations)

해당 법령: AI법 제55조(1) 및 전문 114, 115

서명자는 무단 공개(unauthorised releases), 무단 접근(unauthorised access) 및/또는 모델 도난(model theft)으로부터 발생할 수 있는 모델의 시스템적 위험이 수용 가능하도록 보장하기 위하여(약정 4에 따름), 본 약정의 이행조치에 규정된 바와 같이 모델 및 물리적 인프라에 대하여 전체 수명주기에 걸쳐 적절한 수준의 사이버보안 보호를 구현할 것을 약정한다.

모델의 역량이 파라미터가 공개적으로 다운로드 가능하도록 제공되는 모델 가운데 적어도 하나의 모델의 역량보다 낮은 경우, 해당 모델에는 본 약정이 적용되지 않는다.

서명자는 모델의 파라미터가 공개적으로 다운로드할 수 있도록 제공되거나 안전하게 삭제될 때까지 해당 모델에 대하여 이러한 보안 완화를 적용한다.

• 이행조치 6.1 보안 목표

서명자는 최소한 자신의 모델의 현재 및 예상 역량을 고려하여, 비국가 행위자에 의한 외부 위협(non-state external threats), 내부자 위협[65] 및 기타 예상되는 위협 행위자를 포함하여, 보안 완화가 대응하고자 하는 위협 행위자를 구체적으로 명시하는 목표("보안 목표[66]"라 함)를 정의해야 한다.

64) high-assurance techniques, 모델의 안전성을 수학적 또는 공식적으로 입증 가능한 방법을 통해 확보하려는 기술
65) insider threats, 조직 내부 인원이 의도적 또는 실수로 보안에 위협을 가하는 행위를 말하며, 보통 인증된 접근 권한을 가진 자가 포함
66) security goal, AI시스템이 방어해야 할 위협 주체와 그 공격 방식의 범위를 사전에 정의한 기준으로, 보안 설계의 기반이 됨

● 이행조치 6.2 적절한 보안 완화

서명자는 보안 목표를 충족하기 위하여 적절한 보안 완화를 구현해야 하며, 여기에는 부록4에 따른 보안 완화가 포함된다. 서명자가 부록 4.1부터 4.5까지의 (a)에 열거된 보안 완화 중 어느 하나를 적용하지 않는 경우, 예를 들어, 서명자의 조직 환경 또는 디지털 인프라로 인한 경우에는 해당 완화와 동일한 완화 목표를 달성하는 대체 보안 완화를 적용한다.

요구되는 보안 완화의 구현은 모델 전체 수명주기에 걸친 모델 역량의 증가에 맞추어 적절하게 단계적으로 이루어질 수 있다.

약정7. 안전 및 보안 모델 보고서

해당 법령: AI법 제55조(1) 및 제56조(5)

서명자는 모델을 시장에 출시하기 전에(이행조치 7.1부터 7.5에 규정된 바와 같음), 안전 및 보안 모델 보고서(Safety and Security Model Report)(이하 “모델 보고서”라 함)를 작성하여 자신의 모델 및 시스템적 위험 평가(ass) 및 완화와 조치에 관한 정보를 AI사무국에 보고할 것을 약정한다. 또한 서명자는 모델 보고서를 최신 상태로 유지할 것(이행조치 7.6에 규정된 바와 같음)과 모델 보고서를 AI사무국에 통지할 것(이행조치 7.7에 규정된 바와 같음)을 약정한다.

서명자가 이미 다른 보고서 및/또는 통지를 통하여 관련 정보를 AI사무국에 제공한 경우, 모델 보고서에서 해당 보고서 및/또는 통지를 참조할 수 있다. 어느 한 모델의 시스템적 위험 평가(ass) 및 완화와 조치가 다른 모델에 대한 참조 없이는 이해될 수 없는 경우, 서명자는 해당 모델들에 대하여 하나의 모델 보고서를 작성할 수 있다.

SMEs 또는 SMCs에 해당하는 서명자는 규모 및 역량 제약을 반영하기 위하여 필요한 범위 내에서 모델 보고서의 세부 수준을 축소할 수 있다.

● 이행조치 7.1 모델 설명 및 행동

서명자는 모델 보고서에 다음 사항을 포함한다.:

(1) 모델의 아키텍처, 역량, 경향 및 기능활용성에 대한 개괄적 설명과 모델의 학습 방법 및 학습 데이터를 포함한 개발 방식에 대한 설명, 그리고 이러한 사항이 서명자가 시장에 공급한 다른 모델과 어떻게 다른지에 대한 설명;

(2) 모델이 어떻게 사용되어 왔고 앞으로 어떻게 사용될 것으로 예상되는지에 대한 설명. 여기에는 다른

모델의 개발, 감독 및/또는 평가에 해당 모델이 활용되는 경우가 포함된다.

(3) 시장에 공급될 예정이거나 현재 시장에 공급 및/또는 사용되고 있는 모델 버전에 대한 설명. 여기에는 시스템적 위험 완화 조치와 시스템적 위험의 차이가 포함된다; 및

(4) 서명자가 모델이 어떠한 방식으로 작동하도록 의도하는지에 대한 명세(예: 유효한 하이퍼링크를 통한 명세)(통상 "모델 명세(model specification)"로 알려짐). 여기에는 다음 사항이 포함된다.:

(a) 모델이 따르도록 의도된 원칙의 명시;

(b) 모델이 서로 다른 종류의 원칙 및 지시사항에 어떠한 우선순위를 부여하도록 의도되는지에 대한 설명;

(c) 모델이 지시를 거부하도록 의도된 주제의 목록; 및

(d) 시스템 프롬프트(system prompt)의 제공.

- **이행조치 7.2 진행 결정의 사유**

서명자는 모델 보고서에 다음 사항을 포함한다.:

(1) 모델로부터 발생하는 시스템적 위험이 수용 가능하다는 판단에 대한 상세한 정당화. 여기에는 반영된 안전 여유의 세부 내용이 포함된다(이행조치 4.1에 따름);

(2) (1)의 정당화가 더 이상 유효하지 않게 되는 합리적으로 예견 가능한 조건; 및

(3) 개발, 시장에 공급 및/또는 사용을 계속 진행하기로 한 결정(이행조치 4.2에 따름)이 어떠한 방식으로 이루어졌는지에 대한 설명. 여기에는 외부 행위자의 의견이 해당 결정에 반영되었는지 여부(이행조치 1.1 (2)(d)에 따름), 그리고 부록 3.5에 따른 독립적인 외부 평가자의 의견이 해당 결정에 반영되었는지 여부 및 어떠한 방식으로 반영되었는지가 포함된다.

- **이행조치 7.3 시스템적 위험 식별·분석 및 완화에 관한 문서화**

서명자는 모델 보고서에 다음 사항을 포함한다.:

(1) 시스템적 위험 식별 및 분석 결과와 이를 이해하는 데 관련된 정보에 대한 설명. 여기에는 다음 사항이 포함된다.:

(a) 부록1.1에 규정된 위험 유형에 속하는 위험에 대한 시스템적 위험 식별 절차의 설명(이행조치 2.1 제(1)호에 따름);

(b) 모델이 어떻게 사용되고 AI시스템에 통합될 것인지에 관한 불확실성 및 가정에 대한 설명;

(c) 시스템적 위험에 대한 시스템적 위험 모델링 결과의 설명(이행조치 3.3에 따름);

(d) 모델로부터 발생하는 시스템적 위험에 대한 설명 및 그 근거. 여기에는 다음 사항이 포함된다.: (i) 시스템적 위험 추정치(이행조치 3.4에 따름); 및 (ii) 안전 및 보안 완화가 구현된 경우의 시스템적 위험과 모델이 완전히 유도된 경우의 시스템적 위험 간 비교(부록 3.2에 따름).

(e) 모델로부터 발생하는 시스템적 위험을 이해하는데 관련된 모든 모델 평가 결과 및 다음 사항에 대한 설명: (i) 평가가 수행된 방식; (ii) 모델 평가에 포함된 테스트 및 작업; (iii) 모델 평가의 점수 산정 방식; (iv) 모델이 유도된 방식(부록 3.2에 따름); (v) 해당되는 경우, 모델 버전별 및 평가 설정별로 해당 점수가 인간 기준선(human baselines)과 어떻게 비교되는지;

(f) 모델 평가 결과에 대한 독립적인 해석과 모델로부터 발생하는 시스템적 위험에 대한 이해를 촉진하기 위하여, 각 관련 모델 평가의 입력 및 출력에서 무작위로 추출한 최소 5개의 샘플. 여기에는 완성문(completions), 생성물(generations) 및/또는 궤적(trajectories)이 포함될 수 있다. 특정 궤적이 시스템적 위험의 이해에 실질적으로 기여하는 경우, 해당 궤적도 제공되어야 한다. 또한 서명자는 AI사무국이 추후 요청하는 경우 관련 모델 평가의 입력 및 출력에서 무작위로 추출한 충분히 많은 수의 샘플을 제공하여야 한다.;

(g) 다음 대상에게 제공된 접근 권한 및 기타 자원에 대한 설명: (i) 내부 모델 평가팀(부록 3.4에 따름); 및 (ii) 부록 3.5에 따른 독립적인 외부 평가자. 앞의 제(ii)호를 대신하여, 서명자는 해당 독립적인 외부 평가자가 필요한 정보를 서명자가 모델 보고서를 AI사무국에 제출하는 것과 동시에 AI사무국에 직접 제공하도록 할 수 있다.; 및

(h) 부록 2에 따른 "동등하거나 더 안전한 모델" 개념을 사용하는 경우, "안전 참조 모델(safe reference model)" 기준(부록 2.1에 따름) 및 "동등하거나 더 안전한 모델" 기준(부록 2.2에 따름)이 어떻게 충족되는지에 대한 정당화;

(2) 다음 사항에 대한 설명: (a) 이행된 모든 안전 완화(약정 5에 따름); (b) 해당 완화가 이행조치 5.1의 요구사항을 어떻게 충족하는지; 및 (c) 해당 완화의 한계(예: 바람직하지 않은 모델 행동(undesirable model behaviour)의 사례를 학습하는 것이 향후 그러한 행동을 식별하는 것을 더욱 어렵게 만드는 경우);

(3) `다음 사항에 대한 설명: (a) 보안 목표(이행조치 6.1에 따름); (b) 이행된 모든 보안 완화(이행조치 6.2에 따름); (c) 해당 완화가 보안 목표를 어떻게 충족하는지. 여기에는 관련 국제 표준 또는 기타 관련 지침(예: RAND 「Securing AI Model Weights」 보고서)과 어느 정도 부합하는지가 포함된다; 및 (d) 서명자가 부록 4.1부터 4.5까지의 (a)에 열거된 보안 완화 중 하나 이상과 다르게 조치를 적용한 경우, 이행된 대체 보안 완화가 해당 완화 목표를 어떻게 달성하는지에 대한 정당화; 및

(4) 다음 사항에 대한 개괄적 설명: (a) 향후 6개월 동안 모델을 추가로 개발하기 위하여 사용할 예정인 기술 및 자산. 여기에는 다른 AI모델 및/또는 AI시스템의 활용이 포함된다.; (b) 이러한 향후 버전 및 보다 발전된 모델이 역량 및 경향 측면에서 현재 모델과 어떻게 달라질 수 있는지; 및 (c) 해당 모델에 대하여 구현할 예정인 신규 또는 업데이트된 안전 및 보안 완화.

이행조치 7.4 외부 보고서

서명자는 모델 보고서에 다음 사항을 포함한다.:

(1) 다음으로부터 제공되는 모든 이용 가능한 보고서(예: 유효한 하이퍼링크를 통한 제공):

(a) 부록 3.5에 따른 모델 평가에 참여한 독립적인 외부 평가자; 및

(b) 부록 4.5에 따라 독립적인 외부 당사자가 수행한 보안 검토; 다만 이는 기존의 비밀유지 의무(영업비밀 관련 의무를 포함함)를 준수하고, 해당 외부 평가자 또는 외부 당사자가 자신의 발견사항 공개에 대한 통제권을 유지할 수 있는 범위 내에서 이루어져야 하며, 서명자가 해당 보고서의 내용을 묵시적으로 지지하는 것으로 해석되지 않는 범위에서 제공되어야 한다.;

(2) 부록 3.5에 따른 모델 평가에 독립적인 외부 평가자가 참여하지 않은 경우, 부록 3.5 제1문단의 (1) 또는 (2)의 요건이 어떻게 충족되었는지에 대한 정당화; 및

(3) 부록 3.5에 따른 모델 평가에 적어도 한 명의 독립적인 외부 평가자가 참여한 경우, 자격 기준(qualification criteria)에 근거한 해당 평가자 선정 사유에 대한 설명.

이행조치 7.5 시스템적 위험 환경의 중대한 변화

서명자는 모델 보고서에 AI사무국이 모델의 개발, 시장 공급 및/또는 사용이 본 장에 따른 시스템적 위험 평가(ass) 및 완화 조치·절차의 이행과 관련하여 시스템적 위험 환경(systemic risk landscape)에 어떠한 중대한 변화를 초래하는지 여부 및 그 방식에 대하여 이해하는 데 필요한 정보를 포함하도록 보장하여야 한다.

이러한 정보의 예시는 다음과 같다:

(1) 모델 역량을 향상시키는 새로운 방법을 시사하는 스케일링 법칙(scaling laws)에 대한 설명;

(2) 계산 효율성 또는 모델 역량 측면에서 최첨단 수준을 실질적으로 향상시키는 새로운 아키텍처의 특성에 대한 요약;

(3) 완화의 효과성을 평가(ass)하는데 관련된 정보에 대한 설명. 예를 들어 모델의 사고의 사슬이 인간이 이해하기 어려운 형태가 되는 경우; 및/또는

(4) 분산 학습(distributed training)의 효율성 또는 실현 가능성을 실질적으로 향상시키는 학습 기법에 대한 설명.

이행조치 7.6 모델 보고서 업데이트

서명자는 모델로부터 발생하는 시스템적 위험이 수용 가능하다고 판단한 근거(이행조치 7.2 (1)에 따름)가 실질적으로 약화되었다고 믿을 만한 합리적인 근거가 있는 경우 모델 보고서를 업데이트하여야 한다. 그러한 경우의 예시는 다음과 같다.:

(1) 이행조치 7.2의 (2)에 따라 열거된 조건 중 하나가 현실화된 경우;

(2) 추가적인 사후 학습, 추가 도구에 대한 접근 또는 추론 연산량(inference compute)의 증가 등을 통하여 모델의 역량, 경향 및/또는 기능활용성이 실질적으로 변경되었거나 변경될 예정인 경우;

(3) 모델의 사용 방식 및/또는 AI시스템으로의 통합 방식이 실질적으로 변경되었거나 변경될 예정인 경우;

(4) 해당 모델 또는 유사한 모델과 관련된 중대 사고 및/또는 아차 사고가 발생한 경우; 및/또는

(5) 수행된 모델 평가의 외적 타당성(external validity)을 실질적으로 약화시키거나 모델 평가 방법의 최첨단 수준을 실질적으로 향상시키거나 그 밖의 이유로 수행된 시스템적 위험 평가(ass)가 실질적으로 부정확할 수 있음을 시사하는 발전이 발생한 경우.

모델 보고서의 업데이트는 서명자가 업데이트 필요성을 인지한 후 합리적인 기간 내에 완료되어야 한다. 예를 들어 지속적인 시스템적 위험 평가(ass) 및 완화(이행조치 1.2 제2문단에 따름) 과정에서 이를 확인한 경우가 이에 해당한다. 모델에 대한 의도적인 변경이 이루어지고 해당 변경 사항이 시장에 공급되는 경우, 모델 보고서의 업데이트와 그 근거가 되는 전체 시스템적 위험 평가(ass) 및 완화 절차(이행조치 1.2 제3문단에 따름)는 해당 변경 사항이 시장에 공급되기 전에 완료되어야 한다.

또한 해당 모델이 시장에 공급되고 있는 자신의 최고 역량 모델군에 속하는 경우, 서명자는 최소 6개월마다 업데이트된 모델 보고서를 AI사무국에 제출하여야 한다. 다만, 다음 각 호의 어느 하나에 해당하는 경우에는 제출할 필요가 없다.: (1) 모델의 역량, 경향 및/또는 기능활용성이 AI사무국에 마지막으로 모델 보고서 또는 그 업데이트본을 제출한 이후 변경되지 않은 경우; (2) 1개월 이내에 더 높은 역량을 가진 모델을 시장에 출시할 예정인 경우; 및/또는 (3) 해당 모델이 식별된 각 시스템적 위험(이행조치 2.1에 따름)에 대하여 동등하거나 더 안전한 모델(부록 2.2에 따름)로 간주되는 경우.

업데이트된 모델 보고서에는 다음 사항이 포함되어야 한다.:

(1) 전체 시스템적 위험 평가(ass) 및 완화 절차(이행조치 1.2 제3문단에 따름)의 결과를 반영하여

업데이트된 이행조치 7.1부터 이행조치 7.5까지의 정보; 및

(2) 모델 보고서가 어떠한 이유로 어떠한 방식으로 업데이트되었는지를 설명하는 변경 이력, 버전 번호 및 변경 일자.

- **이행조치 7.7 모델 보고서 통지**

서명자는 모델을 출시할 때까지 모델 보고서에 대한 접근 권한을 AI사무국에 제공하여야 한다. 다만 서명자가 적용받는 국가안보 관련 법률에 따라 필요한 경우를 제외하고는 비공개 처리(redactions) 없이 제공해야 한다. 이러한 제공은 예를 들어 공개적으로 접근 가능한 링크 또는 AI사무국이 지정한 충분히 안전한 채널을 통하여 이루어질 수 있다. 모델 보고서가 업데이트된 경우, 서명자는 업데이트가 확인된 날부터 5영업일 이내에 업데이트된 모델 보고서에 대한 접근 권한을 AI사무국에 제공하여야 한다. 다만 서명자가 적용받는 국가안보 관련 법률에 따라 필요한 경우를 제외하고는 비공개 처리없이 제공해야 한다.

모델의 출시를 촉진하기 위하여, 서명자는 모델 보고서 또는 그 업데이트본의 제공을 최대 15영업일까지 연기할 수 있다. 다만 이는 AI사무국이 해당 서명자가 선의로 행동하고(acting in good faith) 있다고 판단하는 경우에 한하며, 서명자가 이행조치 7.2 및 이행조치 7.5에 규정된 정보를 포함하는 잠정 모델 보고서(interim Model Report)를 AI사무국에 지체 없이 제공하는 경우에만 가능하다.

약정8. 시스템적 위험 관련 책임 배분

해당 법령: AI법 제55조(1) 및 전문 114

서명자는 다음을 약정한다.: (1) 조직 전반에 걸쳐 자신의 모델로부터 발생하는 시스템적 위험 관리에 관한 책임을 명확히 정의할 것(이행조치 8.1에 규정된 바와 같음); (2) 시스템적 위험 관리 책임을 담당하는 주체에게 적절한 자원을 배분할 것(조치 8.2에 규정된 바와 같음); 및 (3) 건전한 위험관리 문화(healthy risk culture)를 촉진할 것(이행조치 8.3에 규정된 바와 같음).

- **이행조치 8.1 명확한 책임의 정의**

서명자는 조직 전반에 걸쳐 자신의 모델로부터 발생하는 시스템적 위험을 관리하기 위한 책임을 명확히 정의하여야 한다. 여기에는 다음과 같은 책임이 포함된다.:

(1) 시스템적 위험 감독(Systemic risk oversight): 서명자의 시스템적 위험 평가(ass) 및 완화 절차·조치를 감독하는 책임.

(2) 시스템적 위험 책임(Systemic risk ownership): 서명자의 모델로부터 발생하는 시스템적 위험을 관리하고, 시스템적 위험 평가(ass) 및 완화 절차·조치를 관리하며, 중대 사고에 대한 대응을 관리하는 책임.

(3) 시스템적 위험 지원 및 모니터링(Systemic risk support and monitoring): 서명자의 시스템적 위험 평가(ass) 및 완화 절차·조치를 지원하고 모니터링하는 책임.

(4) 시스템적 위험 보증(Systemic risk assurance): 서명자의 시스템적 위험 평가(ass) 및 완화 절차·조치의 적절성에 대하여, 감독 기능을 수행하는 경영기구 또는 그 밖의 적절한 독립 기구(예: 협의회 또는 기업 내부 이사회)에 내부적 및 필요한 경우 외부적 보증을 제공하는 책임.

서명자는 자신의 거버넌스 구조와 조직의 복잡성을 고려하여, 다음과 같은 조직 내 각 수준에 해당 책임을 적절히 배분한다.:

(1) 감독 기능을 수행하는 경영기구 또는 그 밖의 적절한 독립 기구(예: 협의회 또는 기업 내부 이사회);

(2) 집행 기능을 수행하는 경영기구;

(3) 관련 운영팀;

(4) 존재하는 경우, 내부 보증 제공자(예: 내부 감사 기능); 및

(5) 존재하는 경우, 외부 보증 제공자(예: 제3자 감사인).

서명자가 자신의 모델로부터 발생하는 시스템적 위험의 특성에 비추어 아래 각 호를 모두 준수하는 경우, 이 조치는 충족된 것으로 추정된다.

(1) 시스템적 위험 감독: 서명자의 시스템적 위험 관리 절차 및 조치를 감독할 책임이 감독 기능을 수행하는 경영기구의 특정 위원회(예: 위험위원회 또는 감사위원회) 또는 하나 이상의 적절한 독립 기구(예: 협의회 또는 기업 내부 이사회)에 부여되어 있다. SMEs 또는 SMCs에 해당하는 서명자의 경우, 해당 책임은 감독 기능을 수행하는 경영기구의 개별 구성원에게 주로 부여될 수 있다.

(2) 시스템적 위험 책임: 모델로부터 발생하는 시스템적 위험 관리 책임이 집행 기능을 수행하는 경영기구의 적절한 구성원에게 부여되어 있으며, 해당 구성원은 연구 및 제품 개발(예: 연구 책임자 또는 제품 책임자)과 같이 시스템적 위험을 발생시킬 수 있는 관련 핵심 사업 활동에 대해서도 책임을 진다. 집행 기능을 수행하는 경영기구의 구성원은 시스템적 위험을 발생시키는 사업 활동의 일부(예: 특정 연구 분야 또는 특정 제품)를 감독하는 운영 관리자에게 하위 조직 수준의 책임을 부여한다. 조직의 복잡성에 따라 단계적 책임 구조가 존재할 수 있다.

(3) 시스템적 위험 지원 및 모니터링: 위험 평가(ass)의 수행을 포함하여 서명자의 시스템적 위험 관리

절차 및 조치를 지원하고 모니터링할 책임이 집행 기능을 수행하는 경영기구의 구성원 중 최소 1인(예: 최고위험관리책임자(Chief Risk Officer) 또는 안전 및 보안 프레임워크 담당 부사장)에게 부여된다. 해당 구성원은 연구 및 제품 개발과 같이 서명자의 시스템적 위험을 발생시킬 수 있는 핵심 사업 활동에 대한 책임을 동시에 맡아서는 안 된다. SMEs 또는 SMCs에 해당하는 서명자의 경우, 집행 기능을 수행하는 경영기구 내에 서명자의 시스템적 위험 평가(ass) 및 완화 절차·조치를 지원하고 모니터링하는 업무를 담당하는 개인이 최소 1인 이상 존재해야 한다.

(4) 시스템적 위험 보증: 감독 기능을 수행하는 경영기구 또는 그 밖의 적절한 독립 기구(예: 협의회 또는 기업 내부 이사회)에 대하여 서명자의 시스템적 위험 평가(ass) 및 완화 절차·조치의 적절성에 관한 보증을 제공할 책임이 관련 주체(예: 최고감사책임자(Chief Audit Executive), 내부감사책임자(Head of Internal Audit) 또는 관련 소위원회)에 부여되어 있다. 해당 책임자는 내부 감사 기능 또는 이에 상응하는 기능의 지원을 받으며, 필요한 경우 외부 보증도 제공받는다. 서명자의 내부 보증 활동은 적절하여야 한다. SMEs 또는 SMCs에 해당하는 서명자의 경우, 감독 기능을 수행하는 경영기구는 서명자의 시스템적 위험 평가(ass) 및 완화 절차·조치를 정기적으로 평가(ass)한다(예: 서명자의 프레임워크 평가(ass)를 승인하는 방식).

이행조치 8.2 적절한 자원의 배분

서명자는 자신의 경영기구가 자신의 모델로부터 발생하는 시스템적 위험에 상응하는 수준으로 책임을 담당하는 주체(이행조치 8.1에 따름)에게 자원이 배분되도록 감독하게 해야 한다. 이러한 자원의 배분에는 다음 사항이 포함된다.:

(1) 인적 자원;

(2) 재정적 자원;

(3) 정보 및 지식에 대한 접근; 및

(4) 연산 자원.

이행조치 8.3 건전한 위험 문화의 촉진

서명자는 건전한 위험관리 문화를 촉진하고, 자신의 모델로부터 발생하는 시스템적 위험을 관리할 책임을 담당하는 주체(이행조치 8.1에 따름)가 시스템적 위험에 대하여 합리적이고 균형 잡힌 접근방식을 취하도록 보장하기 위한 적절한 조치를 취해야 한다.

본 이행조치의 목적상 건전한 위험 문화의 지표의 예시는 다음과 같다:

(1) 경영진이 직원들에게 서명자의 프레임워크를 명확하게 전달하는 등의 방식으로, 상위 경영진부터

건전한 시스템적 위험 문화를 조성할 것;

(2) 시스템적 위험에 관한 의사결정에 대하여 명확한 의사소통과 이의 제기가 가능하도록 할 것;

(3) 시스템적 위험 평가(ass) 및 완화에 관여하는 직원에 대하여 유인을 마련하고 충분한 독립성을 보장함으로써, 과도한 시스템적 위험 감수를 지양하고 자신의 모델로부터 발생하는 시스템적 위험에 대한 편향 없는 평가(ass)를 장려할 것;

(4) 익명 설문조사 결과, 직원들이 시스템적 위험에 관한 우려를 제기하는 데 부담을 느끼지 않고, 그러한 우려를 제기할 수 있는 경로를 알고 있으며, 서명자의 프레임워크를 충분히 이해하고 있는 것으로 나타날 것;

(5) 내부 신고 채널이 적극적으로 활용되고 있으며, 신고 사항에 대하여 적절한 조치가 이루어질 것;

(6) 매년 근로자에게 서명자의 내부고발자 보호 정책을 고지하고, 해당 정책을 웹사이트 게시 등의 방법으로 근로자가 쉽게 이용할 수 있도록 할 것; 및/또는

(7) 서명자를 위하여 수행한 업무 관련 활동 과정에서 취득한 정보를 바탕으로, 자신의 모델로부터 발생하는 시스템적 위험에 관한 정보를 관계 당국에 공개하거나 제공한 사람에 대하여, 그 사람이 해당 정보가 진실하다고 믿을 만한 합리적인 근거를 가지고 있는 경우에는, 해고·강등·법적 조치·부정적 평가 또는 적대적인 근무환경 조성 등 직접적 또는 간접적으로 불이익을 주는 어떠한 형태의 보복도 하지 않을 것.

약정9. 중대 사고 보고

해당 법령: AI법 제55조(1) 및 전문 114, 115

서명자는 본 약정의 이행조치에서 규정된 바와 같이, 모델의 전체 수명주기에 걸쳐 발생하는 중대 사고와 이를 해결하기 위한 가능한 시정 조치에 관한 관련 정보를 추적하고, 문서화하며, AI사무국 및 해당되는 경우 국가 관할 당국에 부당한 지체 없이 보고하기 위한 적절한 절차와 조치를 마련·운영할 것을 약정한다. 또한 서명자는 중대 사고의 심각성 및 자신의 모델이 해당 사고에 관여한 정도에 상응하는 수준으로 이러한 절차와 조치에 필요한 자원을 제공할 것을 약정한다.

● 이행조치 9.1 중대 사고 식별 방법

서명자는 중대 사고에 관한 관련 정보를 추적하기 위하여 이행조치 3.5의 예시적 방법을 고려한다. 또한 서명자는 다음 사항을 이행한다.:

(1) 경찰 보고서 및 언론 보도, 소셜미디어 게시물, 연구 논문 및 사고 데이터베이스와 같은 기타 정보원을 검토할 것; 및

(2) 다운스트림 수정자, 다운스트림 제공자, 사용자 및 기타 제3자가 중대 사고에 관한 관련 정보를 다음 대상에게 보고할 수 있도록 촉진할 것:

(a) 서명자; 또는

(b) AI사무국 및 해당되는 경우, 국가 관할 당국; 이는 가능한 경우 해당 제3자에게 직접 신고 채널을 안내함으로써 이루어지며, AI법 제73조에 따른 이들의 신고 의무에 영향을 미치지 않는다.

이행조치 9.2 중대 사고의 추적, 문서화 및 보고를 위한 관련 정보

서명자는 중대 사고와 관련하여, 해당 정보에 적용되는 기타 연합법(other Union law) 을 준수하기 위하여 필요한 범위 내에서 비공개 처리(redacted)된 형태로, 자신이 알고 있는 범위 내에서 최소한 다음 정보를 추적·문서화하고 AI사무국 및 해당되는 경우 국가 관할 당국에 보고하여야 한다.:

(1) 중대 사고의 시작일 및 종료일 또는 정확한 날짜가 불명확한 경우 그에 대한 최선의 추정치;

(2) 중대 사고로 인하여 발생한 피해 및 피해자 또는 영향을 받은 집단;

(3) 중대 사고로 이르게 된 일련의 사건(직접적 또는 간접적인 경우를 포함함);

(4) 중대 사고와 관련된 모델;

(5) 모델이 해당 중대 사고에 어떠한 방식으로 관여하였는지를 보여주는 이용 가능한 자료에 대한 설명;

(6) 서명자가 해당 중대 사고에 대응하여 수행할 계획이 있거나 이미 수행한 조치가 있는 경우 그 내용;

(7) 서명자가 해당 중대 사고에 대응하여 AI사무국 및 해당되는 경우 국가 관할 당국이 취할 것을 권고하는 조치가 있는 경우 그 내용;

(8) 근본 원인 분석 및 중대 사고로 이어진 모델의 출력에 대한 설명(직접적 또는 간접적인 경우를 포함함), 그리고 해당 출력이 생성되는데 기여한 요인에 대한 설명. 여기에는 사용된 입력과 시스템적 위험 완화의 실패 또는 우회가 포함된다.; 및

(9) 사후 모니터링(이행조치 3.5에 따름) 과정에서 탐지된 패턴으로서, 해당 중대 사고와 합리적으로 관련이 있다고 볼 수 있는 모든 패턴. 여기에는 아차 사고에 관한 개별 또는 집계 데이터가 포함될 수 있다.

서명자는 시스템적 위험 평가(ass)에 활용하기 위하여, 위 목록의 정보를 포함하여 중대 사고의 원인과

영향을 조사해야 한다. 서명자가 위 목록의 특정 관련 정보를 아직보유하고 있지 않은 경우, 해당 사실을 중대 사고 보고서에 기록해야 한다. 중대 사고 보고서의 세부 수준은 해당 사고의 심각성에 상응해야 한다.

이행조치 9.3 보고 타임라인

서명자는 예외적인 상황이 없는 한, 자신의 모델의 관여(직접적 또는 간접적)가 다음 각 호의 결과로 이어진 경우, 다음 시점에 AI사무국 및 해당되는 경우 국가 관할 당국에 제출되는 초기 보고서(initial report)에 이행조치 9.2 (1)부터 (7)까지의 정보를 제공하여야 한다.:

(1) 중요 인프라의 관리 또는 운영에 대한 중대하고 비가역적인 교란(irreversible disruption), 또는 서명자가 자신의 모델과 해당 교란 간 그러한 인과관계가 존재한다고 확인하거나 합리적인 개연성을 바탕으로 의심하는 경우, 서명자가 해당 사고에 자신의 모델이 관여하였음을 인지한 후 늦어도 2일 이내;

(2) 모델 가중치(model weights)의 (자기)유출((self-)exfiltration) 및 사이버공격을 포함한 중대한 사이버보안 침해, 또는 서명자가 자신의 모델과 해당 침해 간 그러한 인과관계가 존재한다고 확인하거나 합리적인 개연성을 바탕으로 의심하는 경우, 서명자가 해당 사고에 자신의 모델이 관여하였음을 인지한 후 늦어도 5일 이내;

(3) 사람의 사망 또는 서명자가 자신의 모델과 해당 사망 간 그러한 인과관계가 존재한다고 확인하거나 합리적인 개연성을 바탕으로 의심하는 경우, 서명자가 해당 사고에 자신의 모델이 관여하였음을 인지한 후 늦어도 10일 이내; 및

(4) 사람의 건강(정신적 및/또는 신체적)에 대한 중대한 피해,기본권 보호를 목적으로 하는 연합법상 의무의 침해 및/또는 재산 또는 환경에 대한 중대한 피해, 또는 서명자가 자신의 모델과 해당 피해 또는 침해 간 그러한 인과관계가 존재한다고 확인하거나 합리적인 개연성을 바탕으로 의심하는 경우, 서명자가 해당 사고에 자신의 모델이 관여하였음을 인지한 후 늦어도 15일 이내.

미해결 중대 사고의 경우, 서명자는 초기 보고서 제출 이후 최소 4주마다 AI사무국 및 해당되는 경우 국가 관할 당국에 제출하는 중간 보고서(intermediate report)를 통하여 초기 보고서의 정보를 업데이트하고 이행조치 9.2에 따라 요구되는 추가 정보를 이용 가능한 범위 내에서 제공하여야 한다.

서명자는 이행조치 9.2에서 요구되는 모든 정보를 포함하는 최종 보고서(final report)를 중대 사고가 해결된 날로부터 늦어도 60일 이내에 AI사무국 및 해당되는 경우 국가 관할 당국에 제출해야 한다.

유사한 중대 사고가 보고 타임라인 내에 여러 건 발생한 경우, 서명자는 최초 중대 사고에 적용되는 보고 타임라인을 준수하는 범위 내에서 해당 사고들을 최초 중대 사고에 대한 보고서들에 포함할 수 있다.

● 이행조치 9.4 보존 기간

서명자는 이 약정을 준수하는 과정에서 수집한 모든 관련 정보에 관한 문서를 문서 작성일 또는 중대 사고 발생일 중 늦은 날부터 최소 5년간 보존하여야 한다. 이는 해당 정보에 적용되는 연합법에 영향을 미치지 않는다.

약정10. 추가 문서화 및 투명성

해당 법령: AI법 제53조(1)(a) 및 제55조(1)

서명자는 본 장의 이행 내용을 문서화하고(이행조치 10.1에 규정된 바와 같음), 필요에 따라 자신의 프레임워크 및 모델 보고서의 요약본을 공개할 것(이행조치 10.2에 규정된 바와 같음)을 약정한다.

● 이행조치 10.1 추가 문서화

서명자는 AI사무국의 요청이 있는 경우 이를 제공하기 위하여 다음 정보를 작성하고 최신 상태로 유지하여야 한다.:

(1) 모델 아키텍처에 대한 상세한 설명;

(2) 모델이 AI시스템에 통합되는 방식에 대한 상세한 설명. 여기에는 소프트웨어 구성요소들이 서로 어떠한 방식으로 연계되고 전체 처리 과정에 통합되는지에 대한 설명이 포함되며, 이는 서명자가 해당 정보를 알고 있는 범위 내에서 작성된다.;

(3) 본 장에 따라 수행된 모델 평가에 대한 상세한 설명. 여기에는 평가 결과 및 평가 전략이 포함된다; 및

(4) 적용된 안전 완화에 대한 상세한 설명(약정 5에 따름).

문서화는 모델이 출시된 날부터 최소 10년간 보존되어야 한다.

또한 서명자는 AI사무국의 요청이 있는 경우 본 장을 준수하였음을 입증하기 위하여, 제1문단에서 이미 다루고 있는 경우를 제외하고 다음 정보를 추적하여야 한다.:

(1) 시스템적 위험 평가(ass) 및 완화의 일부를 구성하는 절차, 조치 및 주요 결정;

(2) 서명자가 본 장을 준수하기 위하여 특정 모범사례(best practice), 최첨단 수준의 절차·조치 또는 그 밖의 보다 혁신적인 절차나 조치에 의존하는 경우, 해당 선택에 대한 정당화.

서명자는 제3문단의 정보를 하나의 매체 또는 장소에 통합하여 보관할 필요는 없으며, AI사무국의 요청이 있는 경우 이를 취합할 수 있다.

이행조치 10.2 공개 투명성

서명자는 시스템적 위험을 평가(ass) 및/또는 완화하기 위하여 필요한 경우 및 필요한 범위 내에서 자신의 프레임워크와 모델 보고서 및 그 업데이트본(약정 1 및 약정 7에 따름)의 요약본을 공개하여야 한다(예: 웹사이트를 통한 공개). 다만, 안전 및/또는 보안 완화의 효과성을 유지하고 민감한 상업적 정보를 보호하기 위하여 필요한 사항은 제외할 수 있다. 모델 보고서의 경우, 이러한 공개에는 시스템적 위험 평가(ass) 결과와 적용된 안전 및 보안 완화에 대한 개괄적 설명이 포함되어야 한다. 프레임워크의 경우, 서명자의 모든 모델이 부록 2.2에 따른 동등하거나 더 안전한 모델인 경우에는 이러한 공개가 필요하지 않다. 모델 보고서의 경우, 해당 모델이 부록 2.2에 따른 동등하거나 더 안전한 모델인 경우에는 이러한 공개가 필요하지 않다.

G 용어집(Glossary)

- 본 장에서 AI법 제3조에 정의된 용어를 사용하는 경우, 해당 용어에는 AI법상의 정의가 적용되며, 본 장에서 해당 용어의 사용으로 인해 다른 해석이나 상충되는 해석이 가능한 경우에도 AI법상의 정의가 우선한다. 그 외의 경우에는, AI법상의 정의를 보완하는 의미에서, 본 장에서 다음 용어를 아래와 같은 의미로 사용한다. 별도의 규정이 없는 한, 이 용어집에서 정의된 용어의 모든 문법적 변형은 해당 정의의 적용 범위에 포함되는 것으로 본다.

용어	정의
appropriate 적절한	시스템적 위험 평가(ass) 및/또는 완화의 의도된 목적을 달성하는 데 적합하고 필요한 것으로서, 모범사례(best practices), 최첨단 또는 최신 수준을 넘어서는 보다 혁신적인 절차·조치·방법론·방법 또는 기술에 기반한 접근
best practice 모범 사례	특정 시점에서 시스템적 위험을 가장 효과적으로 평가(ass) 및 완화하는 절차·조치·방법론·방법 및 기술로서, 시스템적 위험이 있는 범용 AI모델 제공자들 사이에서 수용되는 것
confirmed 확인된	적용 가능한 거버넌스 절차에 따라 요구되는 승인을 받은 프레임워크 또는 모델 보고서, 또는 그 업데이트본
deception 기만	타인에게 체계적으로 잘못된 믿음을 형성하게 하는 모델 행동으로서, 모델이 자신이 평가되고 있음을 탐지하여 의도적으로 성능을 낮추거나 그 밖의 방식으로 감독을 약화시키는 등 감독 회피를 수반하는 목표 달성을 위한 모델 행동을 포함
external validity 외적 타당성	높은 과학적·기술적 엄밀성(high scientific and technical rigour)의 한 측면(아래 정의 참조)으로서, 모델 평가 결과가 평가 환경 외부의 모델 행동에 대한 대리 지표(proxy)로 사용될 수 있도록 모델 평가가 적절하게 보정(calibrated)되었음을 보장하는 것. 외적 타당성(external validity)의 입증 방식은 시스템적 위험 및 모델 평가 방법에 따라 달라질 수 있으나, 예를 들어 모델 평가 환경, 해당 환경이 실제 세계 맥락과 차이가 나는 방식 및 모델 평가 환경의 다양성에 대한 문서화를 통해 입증될 수 있는 것.
high scientific and technical rigour 높은 과학적·기술적 엄밀성	모델 평가를 위한 품질 기준으로서, 높은 과학적·기술적 엄밀성을 갖춘 모델 평가는 내적 타당성(아래 정의 참조), 외적 타당성(external validity)(위 정의 참조) 및 재현가능성(아래 정의 참조)을 갖춘 것. 추가 내용은 부록 3.2 참조.

용어	정의
including 포함하여	해당 용어가 요구하는 최소 사항으로 이해되어야 하며, 열거된 사항에 한정되지 않음
independent external 독립적 외부	서명자 또는 그 자회사·관계회사에 재정적·운영상·관리상 종속되지 아니하고, 계약상 보호장치 및 적절한 이해충돌 정책 등을 포함하여, 결론 도출 및/또는 권고 수행 과정에서 그 밖에 서명자의 통제로부터 자유로운 자연인 또는 법인
insider threats 내부자 위협	민감한 조직 자원에 접근할 수 있는 인간·AI모델 및/또는 AI시스템(예: 고위 경영진, 조직 연구팀의 고위 구성원, 불만을 가진 직원, 목표 조직에 침투한 산업 스파이 행위자 및/또는 모델의 자기 유출에 의해 수행되는 적대적 활동 및/또는 우발적인 모델 유출
internal validity 내적 타당성	높은 과학적·기술적 엄밀성의 한 측면(위 정의 참조)으로서, 모델 평가 결과가 평가 환경 내에서 과학적으로 가능한 한 정확하며, 그 결과를 훼손할 수 있는 방법론적 결함이 없도록 보장하는 것. 내적 타당성의 입증 방식은 시스템적 위험 및 모델 평가 방법에 따라 달라질 수 있으나, 예를 들어 다음과 같은 방법을 통해 입증될 수 있음: 충분한 표본 크기의 확보; 통계적 유의성 및 통계적 검정력(statistical power)의 측정, 사용된 환경 변수의 공개, 교란변수(confounding variables)의 통제 및 가성 상관관계(spurious correlation)의 완화, 시험 데이터가 학습에 사용되는 것을 방지하는 것(예: 학습·시험 데이터 분할(train-test splits) 및 카나리 문자열(canary strings)의 준수); 서로 다른 조건 및 환경에서 모델 평가를 반복 수행하는 것(모델 평가의 개별 요소(예: 프롬프트의 강도와 안전 및 보안 완화)를 변경하는 것을 포함); 추론 과정(trajectories) 및 기타 출력(outputs)에 대한 상세한 검토; 특히 인간 주석자(human annotators)가 참여하는 모델 평가에서 잠재적인 라벨링 편향(labelling bias)을 방지하는 것(예: 블라인드 처리(blinding) 또는 주석자 간 합치도(inter-annotator agreement)의 보고); 투명성 향상 기법의 활용(예: 모델의 '내부 작동 방식'을 대표하고 평가자가 이해할 수 있는 추론 흔적(reasoning traces)); 감독을 회피하려는 모델의 능력을 측정 및/또는 감소시키기 위한 기법의 활용; 및/또는 새로운 모델 평가의 생성 및 관리 방법을 공개하여 그 무결성을 확보하는 것.
management body 경영기구	국내법에 따라 임명되고 다음 권한을 부여받은 법인 기관:(1) 다음을 통하여 수행되는 집행 기능: (a)조직의 전략·목표 및 전반적 방향의 설정; 및 (b) 조직의 일상적 관리 수행; 및 (2) 집행 의사결정을 감독·모니터링함으로써 수행되는 감독 기능. 관련 국내법에 따라, 집행 기능 및 감독 기능은 하나의 관리기구 내 서로 다른 인원에 의해 수행될 수도 있고, 관리기구의 별개 부분에 의해 수행될 수도 있는 것.
model 모델	시스템적 위험이 있는 범용 AI모델 동일한 모델에는 서로 다른 목적을 위하여 미세조정(fine-tuning)된 버전, 서로 다른 도구에 접근 가능한 버전 및/또는 서로 다른 안전 및/또는 보안 완화 조치를 가진 버전

용어	정의
	등 다양한 버전이 존재할 수 있음. 본 장에서 'model'에 대한 모든 언급은 문맥상 필요한 관련 모델 버전을 의미함 일반적으로 시스템적 위험 평가(ass) 및 완화의 맥락에서, 'model'에 대한 모든 언급은 집합적으로 해당 모델로부터 발생하는 시스템적 위험을 구성하는 모든 모델 버전을 의미하며, 여기에는 다음 각 호의 모델 버전이 포함됨: (1) 가장 발전된 모델 버전; (2) 제(1)호에 해당하면서 시스템적 위험에 대한 안전 및/또는 보안 완화 조치가 제한적으로만 구현되었거나 전혀 구현되지 않은 모델 버전; 및/또는 (3) 널리 사용되는 모델 버전. 서로 다른 'models' 간 비교의 맥락(예:이행조치 3.5 및 부록 3.5와 부록 2 및 약정 6의 결합 적용)에서 'model'에 대한 모든 언급은 단일 모델 버전을 의미함. 'model' 앞에 'AI'라는 용어가 붙는 경우, 해당 용어는 예외적으로 시스템적 위험이 있는 범용 AI모델만을 의미하지 않으며, 시스템적 위험이 있는 범용 AI모델 이외의 모든 모델도 포함함.
model elicitation 모델 유도	모델의 역량·성향·기능활용성 및/또는 효과를 체계적으로 향상시키기 위한 기술적 작업. 이를 통해 달성 가능성이 있는 해당 역량·성향·기능활용성 및/또는 효과의 전체 범위를 정확하게 측정할 수 있도록 함
model evaluation 모델 평가	시스템적 위험 평가(ass) 기법으로서, 시스템적 위험 평가(ass)의 모든 단계(아래 정의 참조)에서 사용될 수 있는 기법
model-independent information 모델 비특정 정보	특정 모델에 종속되지 아니하나, 여러 모델에 걸친 시스템적 위험 평가(ass) 및 완화에 정보를 제공할 수 있는 정보로서, 데이터 및 연구를 포함함 추가 내용은 이행조치 3.1 참조
near miss 아차 사고	중대 사고가 발생할 수 있었으나 최종적으로는 발생하지 않은 상황
non-state external threats 비국가 외부 위협	다음 요건을 충족하는 비국가 행위자에 의해 수행되는 적대적 활동: (1) 사이버보안 분야의 숙련된 전문 인력 10명 수준에 대체로 상응할 것; (2) 해당 활동에 수개월 동안 최대 100만 유로(EUR 1 million)의 총예산을 사용할 것; 및 (3) 주요한 기존 사이버공격 인프라는 보유하고 있으나, 대상 조직에 대한 기존 접근 권한은 없을 것
post-market monitoring 출시 후 모니터링	모델이 시장에 제공된 시점부터 시장 제공이 종료되는 시점까지의 기간 동안 수행되는 모델 모니터링 추가 내용은 이행조치 3.5 참조

용어	정의
process (명사; 시스템적 위험 관리의 맥락에서) 절차	본 장에서 규정된 이행조치를 구성하거나 그 결과로 이어지는 구조화된 일련의 행위
reproducibility 재현 가능성	높은 과학적·기술적 엄밀성의 한 측면(위 정의 참조)으로서, 동일한 입력 데이터·계산 기법·코드 및 모델 평가 조건을 사용하여 일관된 모델 평가 결과를 얻을 수 있는 능력. 이를 통해 다른 연구자 및 엔지니어가 모델 평가 결과를 검증·재현 또는 개선할 수 있도록 함. 재현가능성은 예를 들어 독립된 제3자에 의한 성공적인 동료심사 및/또는 재현, 충분한 양의 모델 평가 데이터·모델 평가 코드·모델 평가 방법론 및 방법에 관한 문서·모델 평가 환경 및 계산 환경·모델 유도 기법을 AI사무국에 안전하게 공개하는 것 및/또는 공개적으로 이용 가능한 API·기술적 모델 평가 표준 및 도구의 사용을 통해 입증 가능함.
resolved(중대 사고) 해결된	가능한 경우 피해를 시정하고, 해당 중대 사고와 관련된 시스템적 위험을 평가(ass) 및 완화하기 위한 시정 조치를 서명자가 채택한 모델의 중대 사고. 'Unresolved'는 이에 따라 해석함
scaling law 스케일링 법칙	규모 또는 학습·추론에 사용되는 시간·데이터·계산 자원의 양과 같이, AI모델 또는 AI 시스템의 개발 또는 사용과 관련된 일부 변수와 그 성능 간의 체계적 관계
(self-)exfiltration of model weights 모델 가중치의 (자기)유출	모델 자체 및/또는 권한 없는 행위자에 의한 모델 가중치(weights) 또는 관련 자산의 안전한 저장소로부터의 접근 또는 이전(transfer)
similar model 유사 모델	서명자가 이용 가능한 공개 및/또는 비공개 정보를 바탕으로 실질적으로 유사한 역량(capabilities)·성향(propensities) 및 기능활용성(affordances)을 가진 것으로 간주하는, 시스템적 위험 유무와 관계없는 범용 AI모델. 여기에는 "안전한 참조 모델"(부록 2.1에 따름) 및 "동등하거나 더 안전한 모델"(부록 2.2에 따름)이 포함됨
state of the art 최첨단 수준	모범사례(best practice)를 넘어서는 관련 연구·거버넌스 및 기술의 최전선(frontier)
system prompt 시스템 프롬프트	사용자 상호작용이 시작되기 전에 모델에 제공되는 지시사항·지침 및 맥락 정보의 집합
systemic risk acceptance criteria	서명자가 자신의 모델로부터 발생하는 시스템적 위험의 수용 가능 여부를 결정하기 위하여 프레임워크에서 정의한 기준. 시스템적 위험 등급(systemic risk tiers)(아래 정

용어	정의
시스템적 위험 수용 기준	의 참조)는 시스템적 위험 수용 기준의 한 유형임. 추가 내용은 이행조치 4.1 참조.
systemic risk assessment 시스템적 위험 평가	시스템적 위험 식별(약정 2에 따름), 시스템적 위험 분석(약정 3에 따름) 및 시스템적 위험 수용 여부 결정(약정 4에 따름)을 모두 포괄하는 상위 개념
systemic risk management 시스템적 위험 관리	시스템적 위험 평가(ass) 및 완화를 포함하여, 시스템적 위험과 관련하여 조직을 관리하기 위한 조정된 절차 및 조치
systemic risk mitigations 시스템적 위험 완화	시스템적 위험에 대한 안전 완화 조치(약정 5에 따름), 보안 완화 조치(약정 6에 따름) 및 거버넌스 완화 조치(약정 1 및 7 내지 10에 따름)로 구성됨
systemic risk modelling 시스템적 위험 모델링	모델로부터 발생하는 시스템적 위험이 현실화될 수 있는 경로를 특정하기 위한 구조화된 절차. ‘threat modelling’이라는 용어와 상호교환적으로 사용되는 경우가 많음. 본 장에서는 ‘threat modelling’이라는 용어가 사이버보안 분야에서 특정한 의미를 가지므로 ‘risk modelling’이라는 용어를 사용함. 추가 내용은 이행조치 3.3 참조.
systemic risk scenario 시스템적 위험 시나리오	모델로부터 발생하는 시스템적 위험이 현실화될 수 있는 시나리오. 추가 내용은 이행조치 2.2 참조.
systemic risk source 시스템적 위험 원천	단독으로 또는 다른 요인과 결합하여 시스템적 위험을 발생시킬 수 있는 요인. 추가 내용은 부록 1.3 참조
systemic risk tiers 시스템적 위험 등급	모델로부터 발생하는 일정 수준의 시스템적 위험에 대응하도록 프레임워크에서 정의된 단계. 시스템적 위험 등급은 시스템적 위험 수용 기준의 한 유형임 추가 내용은 이행조치 4.1 참조
use (모델의) 사용	서명자 또는 그 밖의 행위자에 의한 모델의 사용

A 부록(Appendices)

부록 1. 시스템적 위험 및 기타 고려사항

해당 법령

AI법 제3조(64) : '고영향 역량(high-impact capabilities)'이란 가장 발전된 GPAI모델에서 확인된 역량과 동등하거나 이를 초과하는 역량을 의미한다.

AI법 제3조(65) :'시스템적 위험(systemic risk)'이란 GPAI모델의 고영향 역량에 특유한 위험으로서, 그 도달 범위(reach)로 인해 연합 시장에 중대한 영향을 미치거나, 또는 공중보건, 안전, 공공보안, 기본권 또는 사회 전체에 대한 실제적이거나 합리적으로 예견 가능한 부정적 영향으로 인해 연합 시장에 중대한 영향을 미치며, 가치사슬 전반에 걸쳐 대규모로 전파될 수 있는 위험을 의미한다.

추가 관련 법률 문언
AI법 전문 110

부록 1.1 위험 유형

이행조치 2.1의 (1) 및 AI법 제3조(65)에 따른 시스템적 위험 식별을 위하여, 다음과 같은 서로 구별되지만 경우에 따라 일부 중복될 수 있는 위험 유형이 적용된다.

(1) 공중보건에 대한 위험.

(2) 안전에 대한 위험.

(3) 공공보안에 대한 위험.

(4) 기본권에 대한 위험.

(5) 사회 전체에 대한 위험.

이러한 유형의 위험을 바탕으로, 구체적인 시스템적 위험 목록은 부록 1.4에 제시되어 있다.

서명자가 수행하는 시스템적 위험 식별 절차의 일환으로, 이행조치 2.1 (1)(a)의 위험 목록을 작성할 때

참고할 위 다섯 가지 위험 유형에 해당하는 위험의 예시는 다음과 같다. 즉, 대규모 사고의 위험, 핵심 부문 또는 중요 인프라에 대한 위험, 공중의 정신건강, 표현의 자유 및 정보의 자유, 비차별, 프라이버시 및 개인정보 보호, 환경, 비인간 복지(non-human welfare), 경제안보 및 민주적 절차에 대한 위험, 그리고 권력 집중 및 불법적·폭력적·혐오적·극단화 유발(radicalising) 콘텐츠 또는 허위 콘텐츠로부터 발생하는 위험이 이에 해당한다. 여기에는 아동 성착취물(CSAM, Child Sexual Abuse Material) 및 비동의 성적 이미지(NCII, Non-Consensual Intimate Images)로부터 발생하는 위험이 포함된다.

부록 1.2 시스템적 위험의 성격

시스템적 위험의 성격에 관한 아래의 고려사항은 시스템적 위험 식별(약정 2에 따름)에 활용된다. 이러한 고려사항은 시스템적 위험의 성격에 관한 본질적 특성(essential characteristics)(부록 1.2.1)과 기여 특성(contributing characteristics)(부록 1.2.2)을 구분한다.

부록 1.2.1 본질적 특성

(1) 해당 위험은 AI법 제3조(65) 및 제3조(64)에 따른 고영향 역량에 특유한 위험이다.

(2) 해당 위험은 AI법 제3조(65)에 따라 연합 시장에 중대한 영향을 미친다.

(3) 해당 영향은 AI법 제3조(65)에 따라 가치사슬 전반에 걸쳐 대규모로 전파될 수 있다.

부록 1.2.2 기여 특성

(1) 역량 의존성: 해당 위험은 모델의 역량이 증가함에 따라 커지거나, 모델 역량의 프론티어에서 새롭게 나타날 수 있다.

(2) 도달 범위 의존성: 해당 위험은 모델의 도달 범위가 확대될수록 커진다.

(3) 고속성: 해당 위험은 신속하게 현실화될 수 있으며, 이로 인해 완화가 이를 따라가지 못할 수 있다.

(4) 복합적 또는 연쇄적 특성: 해당 위험은 다른 시스템적 위험이나 연쇄 반응을 촉발할 수 있다.

(5) 복구가 어렵거나 불가능한 특성: 해당 위험이 현실화되면, 이를 시정하기 위해 비상한 노력·자원 또는 시간이 요구되는 지속적인 변화를 초래하거나, 영구적으로 되돌릴 수 없는 결과를 초래한다.

(6) 비대칭적 영향: 소수의 행위자 또는 사건만으로도 해당 위험의 현실화를 촉발할 수 있으며, 그 결과 행위자 또는 사건의 수에 비하여 불균형적으로 큰 영향을 초래할 수 있다.

부록 1.3 시스템적 위험 원천

다음의 모델 역량, 모델 경향[67], 모델 기능활용성[68] 및 맥락적 요인(contextual factors)은 시스템적 위험 식별(약정 2에 따름)을 위한 잠재적 시스템적 위험 위험 원천으로 간주되며, 이에 한정되지 않는다.

부록 1.3.1 모델 역량

모델 역량에는 다음이 포함된다.:

(1) 공격적 사이버 역량(offensive cyber capabilities);

(2) 화학·생물·방사능·핵(CBRN) 역량 및 그 밖의 무기 획득 또는 확산 역량;

(3) 기본권에 대한 지속적이고 중대한 침해를 초래할 수 있는 역량;

(4) 조작·설득 또는 기만할 수 있는 역량;

(5) 자율적으로 작동할 수 있는 역량;

(6) 새로운 작업을 스스로 적응하여 학습하는 역량;

(7) 장기 계획·예측 또는 전략 수립 역량

(8) 자기 추론(self-reasoning) 역량(예: 모델이 자기 자신, 자신의 구현 또는 환경에 대하여 추론하거나, 자신이 평가받고 있는지를 인지하는 능력);

(9) 인간의 감독을 회피하는 역량;

(10) 자기복제(self-replicate), 자기개선(self-improve) 또는 자신의 구현 환경을 수정할 수 있는 역량;

(11) AI 연구개발을 자동화할 수 있는 역량;

(12) 다중 모달리티(예: 텍스트·이미지·오디오·비디오 및 그 밖의 모달리티)를 처리할 수 있는 역량;

(13) 도구를 사용할 수 있는 역량으로서, "컴퓨터 사용"(예: 모델 자체의 일부가 아닌 하드웨어 또는 소프트웨어, 애플리케이션 인터페이스 및 사용자 인터페이스와의 상호작용하는 것)을 포함한다.;및

(14) 물리적 시스템을 제어할 수 있는 역량.

부록 1.3.2 모델 경향

모델 경향은 모델이 특정한 행동 또는 패턴을 보일 경향이나 경향을 포괄하는 개념으로서, 다음을 포함한다.:

(1) 인간의 의도와의 오정렬(misalignment)

(2) 인간의 가치와의 오정렬(예: 기본권에 대한 무시)

67) model propensities, 모델의 구조나 학습 과정에서 비롯되어 특정 행동이나 출력 패턴이 나타날 가능성을 높이는 경향적 특성
68) model affordances, 모델의 구조, 구성 또는 사용 환경으로 인해 특정 행위나 활용이 가능해지도록 하는 기능적 조건 또는 가능성

(3) 해로운 방식으로 역량을 발휘하려는 경향(예: 조작 또는 기만);

(4) "환각(hallucinate)"을 일으키거나, 허위정보(misinformation)를 생성하거나, 정보 출처를 불분명하게 만드는 경향;

(5) 차별적 편향(discriminatory bias);

(6) 성능 신뢰성(performance reliability)의 부족;;

(7) 무법성, 즉 유사한 상황에 있는 사람에게 부과될 법적 의무를 합리적으로 고려하지 않거나, 영향을 받는 사람의 법적으로 보호되는 이익을 합리적으로 고려하지 않고 행동하는 것;

(8) "목표 추구", 목표 변경에 대한 유해한 저항 또는 권력 추구;

(9) 다른 AI모델/시스템과의 "공모"; 및

(10) 다른 AI모델/시스템과의 조정 실패 또는 충돌

부록 1.3.3 모델 기능활용성 및 기타 시스템적 위험의 위험 원천

모델 기능활용성 및 기타 시스템적 위험의 위험 원천은 모델 구성, 모델 특성 및 모델이 시장에 공급되는 맥락을 포괄하는 개념으로서, 다음을 포함한다.:

(1) 도구(다른 AI모델/시스템 포함), 연산 능력(예: 모델의 작동 속도 향상을 가능하게 하는 경우) 또는 핵심 인프라를 포함한 물리적 시스템에 대한 접근;

(2) 확장성(예: 대량 데이터 처리, 신속한 추론 또는 병렬화(parallelisation)를 가능하게 하는 경우);를 가능하게 하는 경우);

(3) 공개 및 배포(distribution) 전략;

(4) 인간 감독 수준(예: 모델 자율성의 정도);

(5) 가드레일의 적대적 제거에 대한 취약성;

(6) 모델 유출에 대한 취약성(예: 모델 누출/도난(model leakage/theft));

(7) 적절한 인프라 보안의 부족;

(8) 모델의 비즈니스 사용자 수 및 최종 사용자 수. 여기에는 해당 모델이 통합된 AI시스템의 최종 사용자 수도 포함됨;

(9) 공격-방어 균형(offence-defence balance)으로서, 악의적 행위자가 모델을 오용할 수 있는 잠재적 수, 역량 및 동기가 포함됨;

(10) 모델의 영향을 받을 수 있는 특정 환경의 취약성(예: 사회적 환경, 생태적 환경).

(11) 적절한 모델 설명가능성(explainability) 또는 투명성의 부족

(12) 다른 AI모델 및/또는 AI시스템과의 상호작용; 및

(13) 모델의 부적절한 사용(예: 해당 모델의 역량 또는 경향에 부합하지 않는 용도로 모델을 사용하는 경우).

부록 1.4 특정된 시스템적 위험

부록 1.1의 위험 유형을 바탕으로, 부록 1.2의 시스템적 위험의 성격과 부록 1.3의 시스템적 위험 위험 원천을 고려하고, AI법 제56조(1) 및 전문 110에 따른 국제적 접근법을 감안하여, 다음은 이행조치 2.1 (2)에 따른 시스템적 위험 식별을 위한 특정된 시스템적 위험으로 간주된다.:

(1) 화학·생물·방사능·핵(CBRN): 화학·생물·방사능·핵(CBRN) 공격 또는 사고를 가능하게 함으로써 발생하는 위험. 여기에는 관련 무기 또는 물질의 설계, 개발, 획득, 공개, 배포(distribution) 및 사용 과정에서 악의적 행위자의 진입장벽을 현저히 낮추거나, 달성 가능한 잠재적 영향을 현저히 증대시키는 경우가 포함된다.

(2) 통제 상실(loss of control): 인간이 모델을 신뢰성 있게 지시·수정 또는 종료할 수 있는 능력을 상실함으로써 발생하는 위험. 이러한 위험은 인간의 의도 또는 가치와의 오정렬(misalignment), 자기 추론, 자기 복제(self-replication), 자기 개선(self-improvement), 기만(deception), 목표 변경에 대한 저항(resistance to goal modification), 권력 추구 행동(power-seeking behaviour), 또는 AI모델이나 AI시스템을 자율적으로 생성하거나 개선하는 행위로부터 발생할 수 있다.

(3) 사이버 공격(cyber offence): 중요 시스템(예: 핵심 인프라)에 대한 공격을 포함하여, 대규모의 정교한 사이버공격을 가능하게 함으로써 발생하는 위험. 여기에는 자동화된 취약점 탐색(vulnerability discovery), 익스플로잇 생성(exploit generation), 작전 수행(operational use) 및 공격 확장(attack scaling) 등을 통해 공격적 사이버 활동에서 악의적 행위자의 진입장벽을 현저히 낮추거나 달성 가능한 잠재적 영향을 현저히 증대시키는 경우가 포함된다.

(4) 유해한 조작(harmful manipulation): 설득, 기만 또는 개인 맞춤형 표적화를 통해 대규모 인구집단 또는 중대한 의사결정권자의 행동이나 신념을 전략적으로 왜곡할 수 있게 함으로써 발생하는 위험. 여기에는 특히 멀티턴 대화(multi-turn interactions)를 통하여, 그리고 개인이 그러한 영향력을 인지하지 못하거나 합리적으로 탐지할 수 없는 상황에서, 설득, 기만 및 개인 맞춤형 표적화(personalised targeting) 역량을 현저히 강화하는 경우가 포함된다. 이러한 역량은 보호

특성에 기반한 착취를 포함하여 민주적 절차와 기본권을 훼손할 수 있다.

부록 2. 동등하거나 더 안전한 모델

• 부록 2.1 안전 참조 모델(Safe reference models)

다음 각 호의 요건을 모두 충족하는 경우, 모델은 특정 시스템적 위험과 관련하여 안전 참조 모델로 간주될 수 있다.:

(1) 해당 모델이 다음 중 어느 하나에 해당할 것.:

(a) 이 장이 공표되기 전에 출시되었을 것; 또는

(b) 모델로부터 발생하는 시스템적 위험이 수용 가능한 것으로 결정되었고(약정 4에 따름), AI사무국이 해당 모델의 모델 보고서(약정 7에 따름)를 접수한 것을 포함하여, 전체 시스템적 위험 평가(ass) 및 완화 절차(이행조치 1.2 제3문단에 따름)를 완료하였을 것.

(2) 서명자가 해당 모델의 관련 아키텍처 세부사항, 역량, 경향, 기능활용성 및 안전 완화 등 모델의 특성에 대하여 충분한 가시성을 확보하고 있을 것. 이러한 가시성은 서명자가 직접 개발한 모든 모델에 대하여 인정되며, 또한 모델 파라미터를 포함하여 서명자가 전체 시스템적 위험 평가(ass) 및 완화 절차(이행조치 1.2 제3문단에 따름)를 완료하는 데 필요한 모든 정보에 접근할 수 있는 모든 모델에 대하여 인정된다.

(3) 해당 모델로부터 발생하는 시스템적 위험이 수용 가능하지 않다고 믿을 만한 다른 합리적인 근거가 없을 것.

• 부록 2.2 동등하거나 더 안전한 모델(Similarly safe or safer models)

다음 각 호의 요건을 충족하는 경우, 모델은 특정 시스템적 위험과 관련하여 동등하거나 더 안전한 모델로 간주될 수 있다.:

(1) 서명자가 시스템적 위험 식별(약정 2에 따름)을 수행한 결과, 해당 모델로부터 발생할 수 있는 시스템적 위험 시나리오가 안전 참조 모델로부터 발생할 수 있는 시스템적 위험 시나리오와 비교하여 실질적으로 다르다고 합리적으로 예견되지 않을 것.

(2) 관련성이 있는 최소한 최첨단 수준의 경량 벤치마크(light-weight benchmarks)를 기준으로 평가한 결과, 해당 모델의 성능은 안전 참조 모델의 성능을 유의미하게 초과하지 않을 것(무시할 수 있는 수준의 오차 범위 내 차이는 허용). 다만, 안전 참조 모델에 비해 역량이 일부 향상되었더라도 그로 인해 시스템적 위험이 실질적으로 증가하지 않는 경우에는 이를 고려하지 않을 수 있다. 이러한 벤치마크는

이행조치 3.2에 따라 수행된 것이어야 한다.

(3) 해당 모델과 안전 참조 모델을 비교할 때, 관련 아키텍처 세부사항, 역량, 경향, 기능활용성 및 안전 완화 등 모델의 특성에 있어 시스템적 위험을 실질적으로 증가시킬 것으로 합리적으로 예견되는 알려진 차이가 없을 것. 또한 해당 모델로부터 발생하는 시스템적 위험이 안전 참조 모델에 비해 실질적으로 증가하였다고 판단할 만한 다른 합리적인 근거가 없을 것.

서명자는 전 문단의 (2) 및 (3)과 부록 2.1의 (2)에 대한 평가(ass)를 수행할 때, 예를 들어 참조 모델에 관한 정보 부족이나 측정 오차로 인해 발생할 수 있는 불확실성을 적절히 고려하여야 하며, 이를 위해 충분히 넓은 안전 여유를 반영하여야 한다.

서명자가 다른 모델을 동등하거나 더 안전한 모델로 취급하기 위한 기준으로 사용하던 안전 참조 모델이 이후 안전 참조 모델의 지위를 상실하는 경우, 서명자는 6개월 이내에 다음 중 하나를 이행하여야 한다.: (1) 해당 모델이 동등하거나 더 안전한 모델로 간주될 수 있는 다른 안전 참조 모델을 식별할 것; 또는 (2) 해당 모델의 동등하거나 더 안전한 모델 지위를 이유로 종전에 면제 및/또는 완화가 적용되어 왔던 경우, 그 모델을 이 장의 모든 약정 및 이행조치의 적용 대상으로 취급할 것. 여기에는 전체 시스템적 위험 평가(ass) 및 완화 절차(이행조치 1.2 제3문단에 따름) 중 면제되거나 완화되었던 모든 부분을 수행하는 것이 포함된다.

부록 3. 모델 평가

다음은 전체 시스템적 위험 평가(ass) 및 완화 절차(이행조치 1.2 제3문단에 따름) 동안 이행조치 3.2에 따라 요구되는 모델 평가를 규정한다.

• 부록 3.1 엄격한 모델 평가

서명자는 모델 평가가 높은 수준의 과학적·기술적 엄밀성을 갖추어 수행되도록 보장하여야 하며, 이를 위해 다음 사항이 확보되도록 해야 한다.

(1) 내적 타당성(internal validity);

(2) 외적 타당성(external validity);및

(3) 재현가능성(reproducibility).

• 부록 3.2 모델 유도

서명자는 모델 평가가 모델의 역량, 경향, 기능활용성 및/또는 효과를 유도해낼 수 있는 최소한 최첨단 수준의 모델 유도 수준에서 수행되도록 보장하여야 하며, 이를 위해 다음 사항을 충족하는 최소한 최첨단

기법을 사용해야 한다.

(1) 과소 유도(under-elicitation)의 위험을 최소화할 것.

(2) 모델 평가 중 모델 기만의 위험(예: 샌드배깅[69])을 최소화할 것.

여기에는 테스트 시점 연산[70], 속도 제한[71], 스캐폴딩[72] 및 도구를 조정하고, 미세조정 및 프롬프트 엔지니어링을 수행하는 방식 등이 포함된다. 이를 위하여 서명자는 최소한 다음 사항을 충족하여야 한다.

(1) 시스템적 위험 시나리오(이행조치 2.2에 따름)와 관련된 오용 행위자(misuse actors)의 모델 유도 역량 수준에 상응할 것.

(2) 다음 각 목의 AI시스템 통합 사례를 고려하여 파악되는 모델의 예상 사용 맥락(예: 동일한 수준의 스캐폴딩 및/또는 도구 접근권한)에 상응할 것.: (a) 해당 모델에 대하여 계획되었거나 검토 중인 통합; 및/또는 (b) 해당 통합이 서명자에게 알려져 있고, 서명자가 자신의 모델에 대한 유사한 사용 가능성을 배제할 수 없는 경우, 현재 유사한 모델에 사용되고 있는 통합.

● 부록 3.3 완화 조치의 효과성 평가(ass)

서명자는 모델 평가가 안전 완화의 효과성을 평가(ass)하도록 보장하여야 하며, 그 평가의 범위와 깊이는 시스템적 위험 수용 여부 결정이 특정 완화의 효과성에 의존하는 정도에 상응하여야 한다. 여기에는 적대적 압력 하에서의 효과성 평가(예: 미세조정 공격 또는 탈옥)도 포함된다. 이를 위하여 서명자는 최소한 최첨단 기법을 사용하여야 하며, 다음 사항을 고려하여야 한다.:

(1) 완화가 계획된 대로 작동하는 정도;

(2) 완화가 우회·비활성화 또는 무력화되었거나 되고 있는 정도; 및

(3) 완화의 효과성이 장래에 변화할 가능성.

● 부록 3.4 자격을 갖춘 모델 평가팀 및 충분한 자원

서명자는 모델 평가를 수행하는 팀이 시스템적 위험에 대한 종합적이고 다각적인 이해를 갖출 수 있도록, 기술적 전문성과 해당 시스템적 위험에 관한 관련 분야의 전문지식을 함께 보유하도록 보장하여야 한다. 이러한 기술적 전문성 및/또는 관련 분야 전문지식에 대한 예시적인 자격요건은 다음과 같다.:

(1) 해당 시스템적 위험과 관련된 분야의 박사학위(PhD), 동료심사를 거쳐 인정받은 학술논문 또는 출판물, 또는 이에 상응하는 연구 또는 엔지니어링 경험을 보유하고 있을 것.

---

69) sandbagging, 모델이 평가 상황을 인지하고 실제 역량보다 낮은 성능을 의도적으로 보이거나 행동을 조절하여 평가 결과를 왜곡하는 현상
70) test-time compute, 모델 학습 이후 추론이나 평가 단계에서 성능을 향상시키기 위해 추가로 사용하는 계산 자원 또는 연산량
71) rate limits, 일정 시간 동안 허용되는 요청이나 연산의 횟수를 제한하여 시스템 사용량을 통제하는 기술적 장치
72) scaffolding, 모델이 복잡한 작업을 수행하도록 지원하기 위해 외부 도구, 실행 구조, 프롬프트 체계 등을 결합하여 구축되는 보조적 실행 구조

(2) 해당 시스템적 위험에 관한 공개된 모델 평가 방법 중 동료심사를 거쳤거나 널리 사용되는 모델 평가 방법을 설계하거나 개발한 경험이 있을 것.; 또는

(3) 해당 시스템적 위험과 직접적으로 관련된 분야에서 3년 이상의 실무경험을 보유하고 있을 것. 다만, 해당 분야가 초기 단계에 있는 경우에는 해당 분야에서의 연구 경험 또는 직접적으로 이전 가능한 지식을 제공하는 분야에서의 실무경험 등 이에 상응하는 경험을 보유하고 있을 것.

모델 평가팀에는 다음이 제공되어야 한다.:

(1) 이 부록 3에 따른 모델 평가를 수행하는 데 충분한 수준의 모델 접근권한. 여기에는 적절한 경우 모델 활성값[73], 기울기(gradients)[74], 로짓(logits)[75] 또는 그 밖의 형태의 원시 모델 출력(raw model outputs), 사고의 사슬[76] 및/또는 기타 기술적 세부사항에 대한 접근 권한과 가장 적은 안전 완화가 적용된 모델 버전(예를 들어 해당 버전이 존재하는 경우 유용성 중심 모델(helpful-only model) 버전)에 대한 접근권한이 포함된다. 모델 평가팀에 대한 확대된 모델 접근권한의 적정성을 판단할 때, 서명자는 이로 인해 발생할 수 있는 모델 보안상 위험을 고려하고 평가를 위한 적절한 보안조치를 구현하여야 한다.

(2) 모델 명세(시스템 프롬프트 포함), 관련 학습 데이터, 테스트 세트 및 과거 모델 평가 결과를 포함한 정보. 이러한 정보는 (a) 해당 시스템적 위험 및 (b) 해당 모델 평가 방법에 비추어 적절한 수준으로 제공되어야 한다.

(3) 이 부록 3에 따른 모델 평가를 적절하게 설계 및/또는 수정하고, 디버깅하며, 수행하고, 분석할 수 있도록 하는 충분한 시간. 이러한 시간은 (a) 해당 시스템적 위험과 (b) 해당 모델 평가 방법 및 그 신규성에 비추어 적절한 수준이어야 한다. 예를 들어 대부분의 시스템적 위험 및 모델 평가 방법의 경우 최소 20영업일의 기간이 적절하다.

(4) (a) 충분히 장기간의 모델 평가 수행, 병렬 실행(parallel execution) 및 재수행이 가능하도록 하는 적정한 연산 예산, (b) 적정한 인력 및 (c) 적정한 엔지니어링 예산 및 지원. 여기에는 모델 평가 결과를 검토하여 소프트웨어 버그 또는 모델의 거부를 식별·수정함으로써 역량 추정치가 실제보다 낮게 산정되는 것을 방지하기 위한 예산 및 지원이 포함된다. (b)목과 관련하여, 서명자가 독립적인 외부 평가자를 활용하는 경우에는 해당 평가자의 인력 수준이 충분한지에 관한 보증에 의존할 수 있다.

● 부록 3.5 독립적인 외부 모델 평가

73) activations, 신경망의 각 층에서 계산되는 내부 표현 값으로, 입력 데이터가 모델 내부에서 어떻게 변환되는지를 나타내는 중간 계산 결과
74) gradients, 모델 학습 과정에서 손실 함수가 각 파라미터에 대해 얼마나 변화하는지를 나타내는 미분값으로, 파라미터 업데이트 방향과 크기를 결정하는 정보
75) logits, 모델이 확률 분포를 계산하기 전에 출력하는 원시 점수값으로, 일반적으로 소프트맥스(softmax) 함수 이전 단계의 출력값
76) Chain-of-thought(CoT), 모델이 최종 답변에 도달하기 전에 수행하는 단계적 추론 과정 또는 이를 텍스트 형태로 표현한 추론 내용

서명자는 내부 모델 평가에 더하여, 적절한 자격을 갖춘 독립적인 외부 평가자가 이 부록 3에 따라 해당 시스템적 위험에 관한 모델 평가를 수행하도록 보장하여야 한다. 다만, 다음 각 호의 어느 하나에 해당하는 경우에는 그러하지 아니하다.:

(1) 해당 모델이 부록 2.2에 따른 동등하거나 더 안전한 모델인 경우

(2) 서명자가 적절한 자격을 갖춘 독립적인 외부 평가자를 확보하기 위하여 초기 단계에서 탐색 노력을 기울였음에도 불구하고(예: 20영업일 동안 공개 모집을 실시하는 경우), 그리고 식별된 평가자에게 지체 없이 통지하였음에도 불구하고, 적절한 자격을 갖춘 독립적인 외부 평가자를 선임하지 못한 경우. 이 경우 서명자는 모델로부터 발생하는 시스템적 위험의 수용 가능 여부를 결정할 때(약정 4에 따름), 부록 3.5에 따른 독립적인 외부 평가의 부재로 인해 발생하는 추가적인 불확실성을 고려하여야 한다.

독립적인 외부 평가자의 적절한 자격은 다음을 포함한다.:

(1) 해당 시스템적 위험에 관한 상당한 수준의 분야 전문성을 보유하고 있으며, 모델 평가를 수행하는 데 필요한 기술적 역량과 경험을 갖추고 있을 것.

(2) 적절한 내부 및 외부 정보보안 프로토콜을 갖추고 있을 것.

(3) 영업상 비밀정보에 대한 접근이 필요한 경우, 해당 정보를 보호하기로 합의하였을 것.

서명자는 부록 4.4의 (1)에 따른 사항에 영향을 미치지 않는 범위에서, 독립적인 외부 평가자에게 충분한 접근권한, 정보, 시간 및 그 밖의 자원(부록 3.4에 따름)을 제공하여야 한다. 서명자는 평가자의 명시적인 허가 없이 시험 실행의 입력값 및/또는 출력값을 저장하거나 분석함으로써 외부 모델 평가의 무결성을 훼손해서는 안된다.

중소기업(SMEs) 또는 소형 중견기업(SMCs)인 서명자는 AI사무국에 연락할 수 있으며, AI사무국은 이 부록 3.5의 준수를 지원하기 위하여 지원 또는 자원을 제공할 수 있다.

부록 4. 보안 완화 목표 및 조치

다음은 보안 목표를 달성하기 위하여 이행해야 하는 보안 완화 목표 및 조치(이행조치 6.2에 따름)를 규정한다.

● 부록 4.1 일반 보안 완화

서명자는 다음의 완화 목표를 달성하는 일반 보안 완화를 이행하여야 한다.:

(1) 다음을 통해 무단 네트워크 접근을 방지할 것. (a) 기기 및 계정의 공동 사용 제한, 다중요소 인증[77],

77) Multi-Factor Authentication (MFA), 사용자의 신원을 확인하기 위해 비밀번호와 같은 단일 수단이 아니라 서로 다른 두 가지 이상의 인증 요소

강력한 비밀번호 적용, 강력한 접근 관리 도구, 802.1X 인증, 제로 트러스트 아키텍처[78], 유선 네트워크와 동일한 수준의 무선 네트워크 보호, 및 방문자용 네트워크와 업무 네트워크의 분리를 포함한 강력한 신원 및 접근 관리 관행

(2) 다음을 통해 사회공학 위험을 감소시킬 것. (a) 의심스러운 첨부파일, 링크 및 그 밖의 피싱 시도에 대한 이메일 필터링

(3) 다음을 통해 악성코드 감염 및 이동식 기기의 악의적 사용 위험을 감소시킬 것. (a) 이동식 매체의 사용에 관한 정책

(4) 다음을 통해 취약점 악용 및 악성 코드 실행 위험을 감소시킬 것. (a) 정기적인 소프트웨어 업데이트 및 패치 관리

● 부록 4.2 미공개 모델 파라미터의 보호

서명자는 다음의 완화 목표를 달성하는 보안 완화를 구현함으로써 미공개 모델 파라미터를 보호해야 한다.

(1) 다음을 통해 모든 기기 및 위치에 저장된 모델 파라미터 사본에 대한 관리 책임성을 확보할 것. (a) 모델 파라미터가 저장된 모든 기기 및 위치를 기록하는 안전한 내부 레지스트리(registry);

(2) 다음을 통해 모델 파라미터가 관리되지 않는 기기로 무단 복사되는 것을 방지할 것. (a) 모델 파라미터를 저장하는 모든 기기에 대한 접근 관리 및 관리되지 않는 기기로 복사가 이루어지는 경우의 자동화된 경고 및 알림(alerts);

(3) 다음을 통해 전송 중 및 저장 시 모델 파라미터에 대한 무단 접근을 방지할 것. (a) 적절한 경우 모델 파라미터가 전송 및 저장 과정에서 항상 암호화되도록 보장할 것. 여기에는 최소 256비트 보안 수준의 암호화와 암호화 키를 신뢰 플랫폼 모듈(TPM)[79]에 안전하게 저장하는 것이 포함된다.;

(4) 다음을 통해 임시 저장 중인 모델 파라미터에 대한 무단 접근을 방지할 것. (a) 모델 파라미터가 정당한 사용 목적을 위해서만 비영속 메모리(non-persistent memory)에서 복호화되도록 보장할 것;

(5) 다음을 통해 사용 중인 모델 파라미터에 대한 무단 접근을 방지할 것. (a) 적절한 경우, 하드웨어 기반의 검증된 신뢰 실행 환경[80]을 활용한 기밀 컴퓨팅[81]을 구현할 것;

(6) 다음을 통해 모델 파라미터를 저장·처리하는 시스템에 대한 무단 물리적 접근을 방지할 것. (a) 데이터

(예: 비밀번호, 일회용 코드, 생체정보 등)를 결합하여 인증하는 보안 방식

78) zero trust architecture, 네트워크 내부와 외부를 구분하여 신뢰하지 않고 모든 사용자·장치·접근 요청을 지속적으로 검증하고 최소 권한 원칙에 따라 접근을 통제하는 보안 설계 방식

79) Trusted Platform Module(TPM), 암호키의 생성·저장과 시스템 무결성 검증을 위해 사용되는 하드웨어 기반 보안 모듈

80) Trusted Execution Environment(TEE), 운영체제나 다른 애플리케이션으로부터 격리된 안전한 실행 영역에서 코드와 데이터를 보호하며 처리하는 하드웨어 기반 보안 실행 환경

81) Confidential Computing, 데이터가 저장되거나 전송될 때뿐 아니라 연산 중에도 암호화 상태를 유지하도록 하여 데이터와 모델을 보호하는 컴퓨팅 보안 기술

센터 및 그 밖의 민감한 업무 환경에 대한 물리적 접근을 필요한 인원으로만 제한하고, 권한 없는 인원 또는 기기의 존재 여부를 확인하기 위하여 해당 시설에 대한 정기적인 점검을 실시할 것.

● 부록 4.3 미공개 모델 파라미터에 대한 인터페이스 접근 강화

서명자는 다음의 완화 목표를 달성하는 보안 완화를 구현함으로써, 사용 중인 미공개 모델 파라미터(unreleased model parameters)에 대한 인터페이스 접근을 강화하여야 한다.:

(1) 다음을 통해 모델 파라미터에 대한 불필요한 인터페이스 접근을 방지할 것. (a) 모델 파라미터에 대한 접근이 필요한 소프트웨어 및 인원만을 명시적으로 승인하고, 이를 다중요소 인증 메커니즘(multi-factorauthentication mechanisms)을 통해 강제하며, 최소 6개월마다 정기적으로 점검할 것.;

(2) 다음을 통해 취약점 악용 또는 데이터 유출의 위험을 감소시킬 것. (a) 모델 파라미터에 접근할 수 있는 모든 소프트웨어 인터페이스에 대하여, 보안팀이 취약점 또는 데이터 유출 여부를 식별하기 위한 철저한 검토를 수행할 것. 및/또는 (b) 모든 소프트웨어 인터페이스 코드에 대하여 자동화된 보안 검토를 수행할 것. 이 경우 해당 검토는 최소한 다른 민감한 코드에 적용되는 최고 수준의 자동화된 보안 검토와 동일한 수준이어야 한다.;

(3) 다음을 통해 모델 파라미터 유출의 위험을 감소시킬 것. (a) 출력 속도 제한 등의 방법을 사용하여 모델 파라미터에 접근할 수 있는 인터페이스를 강화할 것.

(4) 다음을 통해 내부자 위협 또는 침해된 계정으로 인한 위험을 감소시킬 것. (a) 모델 파라미터에 대해 강화되지 않은 인터페이스 접근(non-hardened interface-access)을 보유한 인원의 수를 제한할 것.

● 부록 4.4 내부자 위협(Insider threats)

서명자는 모델에 의해 수행되는 (자체) 유출((self-)exfiltration) 또는 사보타주(sabotage)를 포함한 내부자 위협으로부터 보호하기 위하여, 다음의 완화 목표를 달성하는 보안 완화를 구현하여야 한다.:

(1) 다음을 통해 서명자에 대한 업무상 접근권한을 획득하려는 내부자 위협으로부터 모델 파라미터를 보호할 것. (a) 미공개 모델 파라미터 또는 해당 파라미터에 대한 접근을 관리하는 시스템에 대한 읽기·쓰기 접근권한을 보유하고 있거나 합리적으로 보유하게 될 수 있는 임직원 및 계약 인력에 대한 신원조회;

(2) 다음을 통해 내부자 위협 위험에 대한 인식을 제고할 것. (a) 내부자 위협을 인지하고 신고하는 방법에 관한 교육 제공;

(3) 다음을 통해 모델의 자체 유출 위험을 감소시킬 것. (a) 가상머신 및 코드 실행 격리와 같은 모델 주변의 샌드박스(sandboxes)

(4) 다음을 통해 모델 학습 및 사용에 대한 사보타주(sabotage) 위험을 감소시킬 것. (a) 학습 데이터에 변조의 징후가 있는지 점검할 것.

● 부록 4.5 보안 보증

서명자는 다음의 완화 목표를 달성하는 추가적인 보안 완화를 구현함으로써, 자신의 보안 완화가 보안 목표를 충족한다는 보증을 확보하여야 한다.:

(1) 다음을 통해 내부 전문성이 충분하지 않은 경우 보안 완화의 효과성에 대한 독립적인 외부 검증을 확보할 것. (a) 시스템적 위험을 완화하기 위하여 적절한 수준의 정기적인 독립적 외부 보안 검토;

(2) 다음을 통해 네트워크 및 물리적 접근 관리에 대한 검증과 보안상 공백의 식별을 수행할 것. (a) 시스템적 위험을 완화하기 위하여 적절한 수준으로 실시되는 빈번한 레드티밍 활동;

(3) 다음을 통해 네트워크 소프트웨어 무결성에 대한 검증을 수행할 것. (a) 시스템적 위험을 완화하기 위하여 적절한 경우, 외부 공개 엔드포인트(public-facing endpoints)에 대한 보안 테스트에 대중의 참여를 장려하는 경쟁형 버그 바운티(bug bounty) 프로그램;

(4) 다음을 통해 내부자 위협 관련 보안 완화에 대한 검증을 수행할 것. (a) 정기적인 직원 정직성 평가(personnel integrity testing);

(5) 다음을 통해 보안 문제의 신고를 용이하게 할 것. (a) 제3자가 보안 문제를 신고할 수 있는 안전한 통신 채널;

(6) 다음을 통해 의심스럽거나 악의적인 활동을 탐지할 것. (a) 모든 네트워크 및 기기에 엔드포인트 탐지 및 대응(Endpoint Detection and Response, EDR)[82] 및/또는 침입 탐지 시스템(IDS)[83] 도구를 설치할 것;

7) 다음을 통해 악의적 활동에 적시에 효과적으로 대응할 것. (a) EDR 경고 및 알림(alerts)을 모니터링하고, 보안 침해 발생 시 적시에 효과적으로 보안 사고 처리, 대응 및 복구를 수행하기 위한 보안팀을 운영할 것

82) Endpoint Detection and Response(EDR), 컴퓨터, 서버, 모바일 기기 등 네트워크에 연결된 개별 장치(엔드포인트)에서 발생하는 의심스러운 활동을 지속적으로 모니터링하고 탐지하여, 보안 위협에 대해 분석·차단·대응을 수행하는 보안 기술

83) Intrusion Detection System(IDS), 네트워크 또는 시스템에서 발생하는 트래픽과 활동을 분석하여 비인가 접근이나 공격 시도와 같은 침입 행위를 탐지하는 보안 시스템

T 번역어

용어	번역어
Abstract reasoning	추상적 추론
AI literacy	AI 리터러시
Assess /Assessment	평가(ass)하다/ 평가(ass)
Benchmarks	벤치마크
Capability	역량
Chain of Thought(CoT)	사고의 사슬
Cumulative training compute	누적 학습 연산량
Data curation	데이터 선별
Deployment	배포
Distillation	증류
Enforcement	집행
Fine-tuning	미세조정
Full pass	풀 패스
Generality	범용성
General-Purpose AI	범용인공지능(GPAI)
Hazard	위험원
Lifecycle	수명주기
Mitigation	완화
Model architecture	모델 아키텍처
Notification	통지
Open-source licence	오픈소스 라이선스
Parameter	파라미터
Peak theoretical performance	이론적 최대 성능
Pre-training	사전학습
Quantisation	양자화
Recomputation of activation	재계산
Red-teaming	레드티밍
Scaling	스케일링
Self-supervision	자기 지도 학습
Synthetic data	합성 데이터
Systemic risk	시스템적 위험
Trigger point	트리거 시점
Uplift	업리프트
Weight averaging	가중치 평균화
Weight initialisation	가중치 초기화
Weight merging	가중치 병합

유럽연합 인공지능법
범용인공지능모델(GPAI) 실천강령 안내서

발 행 일 2026년 7월 15일
발 행 처 한국전자통신연구원 인공지능안전연구소 AI안전정책및대외협력실
발 행 인 최 민 석 (AI안전정책및대외협력실 실장)
연 구 진 송 경 호 (AI안전정책및대외협력실 선임연구원)
주 지 연 (AI안전정책및대외협력실 석사후연수연구원)
신 민 규 (AI안전정책및대외협력실 석사후연수연구원)
전 화 번 호 031)739-7651
주 소 (13488) 경기도 성남시 분당구 대왕판교로712번길 22, 글로벌 R&D센터 A동 4층
편집·디자인 ※ 본 자료의 표지 디자인에는 생성형AI 기반 이미지 제작 도구가 활용됨.

유럽연합 인공지능법

범용인공지능모델(GPAI) 실천강령 안내서

ETRI 한국전자통신연구원 www.etri.re.kr

인공지능안전연구소 AI Safety Institute

비매품/무로
95560

9 788955 193633

ISBN 978-89-5519-363-3 (PDF)