방통융합정책연구 KMCC-2025-09

AI 이용자 보호를 위한 ICT 법제 연구

A Study on ICT Legal System for the Protection of AI Users

2025. 12

연구기관 : 가천대학교 산학협력단

방송미디어통신위원회
Korea Media and Communications Commission

방통융합정책연구 KMCC-2025-09

AI 이용자 보호를 위한 ICT 법제 연구

(A Study on ICT Legal System for the Protection of AI Users)

권은정/선지원/윤금낭/오다슬/장수진/강지원/
마경태/윤혜선

2025. 12

연구기관 : **가천대학교 산학협력단**

방송미디어통신위원회
Korea Media and Communications Commission

이 보고서는 2025년도 방송미디어통신위원회 방송통신발전기금 방송통신 융합 정책연구사업의 연구결과로서 보고서 내용은 연구자의 견해이며, 방송미디어통신위원회의 공식입장과 다를 수 있습니다.

제 출 문

방송미디어통신위원회 위원장 귀하

본 보고서를 『AI 이용자 보호를 위한 ICT 법제 연구』의 연구결과 보고서로 제출합니다.

2025년 12월

연 구 기 관 : 가천대학교 산학협력단
총괄책임자 : 권 은 정 교 수
참여연구원 : 선 지 원 교 수
윤 금 낭 박 사
오 다 슬 연 구 원
장 수 진 연 구 원
외부전문가 : 강 지 원 변 호 사
마 경 태 변 호 사
윤 혜 선 교 수

목 차

표 목 차

그 림 목 차

요 약 문

1. 제 목

「AI 이용자 보호를 위한 ICT 법제 연구」

2. 연구 목적 및 필요성

인공지능(Artificial Intelligence, AI)이 개인과 사회의 관심사로 거론되는 단계를 지나 우리 사회의 각 영역에서 다목적 수단이자 대체재로서 자리매김 중이다. AI가 침투한 ICT(Information and Communication Technology) 생태계는 불과 3~4년만에 급속하게 전기(轉機)를 맞이하였고, 그 과정에서 개발 역량 또는 정보의 불균형, 신종 범죄 및 권익 침해 발생, 법적·기술적 대응 체계 미비 등 다양한 문제상황을 직면하였다. AI를 둘러싼 여전한 '규율 공백' 내지 '규제지체'(regulatory lag)는 AI의 잠재적 위험성에 대한 무방비를 뜻할 뿐만 아니라 구체적 위험·침해 발생에 대한 사후적 구제마저 어렵게 만든다. 이러한 법제도적 흠결은 AI 시장의 지속 가능한 혁신과 성장을 담보하는 측면에서도 바람직한 것이 아니다.

현행 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「전기통신사업법」 등 ICT 분야 법률에 규정된 몇 가지 이용자 보호 수단은 AI 서비스 영역에서 유의미한 이용자 보호 효과를 거두기에는 부적합하거나 불충분한 면이 있다. 또한 2026년 1월 시행을 앞둔 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」에서는 고영향 AI 사업자 책무의 일환으로 '산업 생태계의 일원'인 이용자를 대상으로 한 보호 수단을 규정하고 있는바, 대중적인 AI 서비스를 직접 이용하고 그로부터 영향을 받는 '최종 소비자' 지위에 있는 이용자의 권익을 실효적으로 보장하기 위해서는 별도의 법적·정책적 기반이 요청되는 상황이다.

본 연구는 이러한 우리 국내 법제가 풀어야 할 현시점의 당면과제를 수행하는 차원에서 AI 서비스 이용자 보호를 위한 시의적이고 합리적인 입법 방안을 제시하고자 한다.

3. 연구의 구성 및 범위

본 연구는 다음의 세 단계를 거쳐 AI 서비스 이용자를 보호하는 법제도적 수단을 검토하고, 글로벌 단위의 이용 생태계를 고려하여 국내 법제를 정립하는 방안을 도출한다.

첫째, 국내외 문헌·사례 연구로서 AI 서비스 이용자 보호에 관한 국내외 정책·입법 현황을 조사하고 사례에 대한 심층 검토를 수행하는 등 AI 서비스 이용자 보호를 위한 실효적인 규범 체계 모색을 위하여 비교법적·학제적 연구를 진행한다.

둘째, AI 이용자 보호를 위한 제정법률안 마련의 핵심을 이루는 주요 쟁점을 분석하고 구체적 입법 방안을 제시한다. AI 서비스 영역에서 보호되어야 할 개인의 권리, 공공의 이익, 사회경제적 가치 등을 종합적으로 고려하여 AI 서비스 개발·제공·이용 전 주기에서 예견되는 위험을 예방 또는 대처하고 이용자 권익을 적정한 수준으로 보장할 수 있는 권리·의무·절차 사항 및 지원·제재 조치를 체계화하는 법률안을 전제한다.

셋째, 아동·청소년 보호에 중점을 둔 콘텐츠 규제 등 일부 현행법(「정보통신망법」, 「전기통신사업법」)의 적용 대상과 수단이 중첩될 여지가 있는 사항은 과도기적 규정으로서 해당 현행법 규정의 개정을 아울러 고려한다.

4. 연구 내용 및 결과

본 연구의 핵심은 AI 서비스가 개발·제공·이용되는 전 과정에서 보장되어야 할 이용자의 권익을 확인하고, 이를 실효적으로 보장하는 실체법·절차법 차원의 이용자 보호 수단을 법제화하는 방안을 도출하는 데 있다. 연구 내용 및 결과를 개관하자면 다음과 같다.

가. AI 서비스 생태계의 규범적 기반 정립(제2장)

- **(AI 생태계의 주체 구분)** AI 서비스의 복합적인 가치사슬을 분석하여 법적 주체를 'AI 개발자', 'AI 서비스 제공자', '이용자'로 구분함
- **(이용자 중심의 권리 규범과 역할별 책임 배분)** 기본권 주체이자 책임 있는 이용 주체의 지위에 있는 이용자의 실체적 권리를 규명하고, 3각 구조를 이루는 각 주체의 통제

가능성 및 역할에 상응하는 권리·의무·책임을 배분함으로써 AI 서비스 제공자를 핵심 수범자로 하는 이용자 보호 법제 설계의 기반을 마련함

- **(이용자 중심의 기본원칙 제시)** 인간의 존엄성 존중, 인간의 적절한 통제권 보장, 사회적 취약계층 보호, 차별 금지 및 투명성 확보 등을 AI 서비스 전 주기에 걸친 이용자 보호 기본원칙으로 제시함

나. 비교법적 · 분야별 사례 분석을 통한 시사점 도출(제3장, 제4장)

- 해외 주요국의 AI 규제 입법례를 분석하였으며, 특히 서비스의 위험도에 따른 차등적 위험관리 체계를 이용자 보호 법제에 유의미하게 수렴하는 입법 방안을 고려함
- 국내 금융·의료 분야의 선행 보호 체계를 심층 분석한 결과, 전 주기 거버넌스, 알고리즘 투명성 확보, 서비스 제공자의 내부 정책 수립 의무화 등의 공통분모가 확인되었으며 이를 일반법 차원의 규정 체계에 흡수할 당위성을 확인함

다. 아동 · 청소년 및 신유형 리스크에 특화된 보호 체계(제5장)

- **(성범죄물 생성 및 유통 문제 대응)** 생성형 AI를 이용한 성착취물 및 성적 허위영상물 생성을 엄격히 규제하고, 개발자 및 제공자에게 이를 방지할 기술적·관리적 조치 의무를 부여하는 방안을 제안함
- **(청소년 특별 보호)** 19세 미만 청소년 대상 서비스 제공 시 유해 콘텐츠 생성 방지 조치 및 과몰입 예방 기능을 탑재하도록 하였으며, 특히 감정적 교류를 모사하는 AI 서비스인 경우 청소년 보호에 특화된 보호 조치를 의무화하는 방안을 제안함

라. 「인공지능서비스 이용자 보호에 관한 법률안(가칭)」의 주요 입법 사항 제안(제6장)

- 이용자 보호를 위한 실효적인 추진 체계, 이용자의 권리와 책임, 서비스 제공자의 의무와 자율규제, 정부의 지원·감독·제재 조치 등을 법제화하는 방안을 제시함
- **(통합 거버넌스 체계 확립)** AI 서비스 특성을 반영하여 분절된 규제를 지양하고, 범정부 차원의 종합계획과 전담 집행기구를 축으로 하는 통합적 이용자 보호 거버넌스를

구축함

- **(이용자의 실정법적 권리 보장)** AI 생태계의 정보 불균형 및 이용자 소외를 극복하기 위해 설명요구권·이의제기권·이용거부권·자기결정권 등을 명시적으로 법제화함으로써 절차적·실체적 권리를 보장함
- **(서비스 제공자의 예방 중심 책무 강화)** AI 서비스 제공자에게는 사후 구제보다 개발·운영 전 과정에 걸친 예방적 책무가 부과되어야 하며, 이를 위해 '설계에 의한 투명성·안전' 원칙을 법적 기준으로 삼도록 함
- **(아동·청소년 보호를 위한 특화된 규범 마련)** 청소년 이용자를 별도로 보호하기 위한 서비스 설계·관리·감독 기능을 의무화하고, 딥페이크 성범죄물 등 불법·유해 콘텐츠 생성·유통 방지 조치를 요구하는 등 특별 보호 체계를 정비함
- **(사후 구제 및 책임 체계의 실효성 제고)** 전문 분쟁조정기구 설치와 집단분쟁조정 도입을 통해 신속·저비용의 비사법적 구제 절차를 마련하는 한편, 글로벌 사업자에 대한 국내대리인 지정 의무를 통해 역외 서비스에 대한 집행력을 확보함

5. 정책적 활용 내용

본 연구를 통해 도출된 입법 방안은 향후 국회나 정부의 법제 정비 과정에서 직접적인 법률(안)의 기초 내지 참고 자료로서 활용 가능하다. 또한 생성형 AI 서비스를 비롯한 신규 서비스와 관련한 새로운 유형의 문제들 — 특히 불법·유해 정보의 증식과 그로 인한 아동·청소년 보호 장치의 부재 — 을 해결하기 위한 과도기적 행정지도 및 자율규제 가이드라인 등을 수립하는 규범적 토대를 제공한다.

6. 기대효과

글로벌 AI 생태계에서 기술 경쟁 이상으로 제도 주도권 경쟁이 치열해지는 상황에서, 본 연구는 세계적인 입법 추세와 국내 산업 여건을 종합적으로 고려하는 이용자 보호 법체계를 제시함으로써 공공·민간의 AI 활용의 신뢰성을 담보하고 글로벌 시장 진출 시 제도 장

벽을 제거하는 법정책 방향을 제시하는 데 의의가 있다.

이러한 관점에서 제안된 「인공지능서비스 이용자 보호에 관한 법률안(가칭)」의 주요 입법 사항은 AI 서비스 이용자 일반을 보호하는 데 요구되는 조직적 기반과 실체적·절차적 내용을 이루는 것으로서 AI가 융합된 광범위한 미디어 서비스 분야의 통일적 규율 체계를 수립하는 데 기여할 것이다, 또한 추후 개별 부처에서 특화된 이용자 보호 규정 및 지침을 마련할 경우 그 규범적 근거와 상호 거버넌스 분담을 논의하는 과정에서도 본 연구 결과가 활용될 수 있을 것으로 기대된다.

SUMMARY

1. Title

「A Study on ICT Legal System for the Protection of AI Users」

2. Objective and Importance of Research

Artificial Intelligence (AI) has moved beyond being merely a topic of interest for individuals and society; it is now establishing itself as a versatile tool and substitute across various sectors of our society. The ICT ecosystem, which has been penetrated by AI, has undergone a rapid transformation over the past three to four years. During this process, it has faced diverse challenges, including imbalances in development capabilities or information access, the emergence of new types of crime and rights infringements, and inadequate legal and technical response systems. The persistent 'regulatory vacuum' or 'regulatory lag' surrounding AI not only signifies a lack of preparedness for its potential risks but also makes it difficult to implement even post-incident remedies for concrete risks and infringements. Such legal and institutional shortcomings are also undesirable for ensuring sustainable innovation and growth in the AI market.

Several user protection measures stipulated in existing ICT laws, such as 「Act on Promotion of Information and Communications Network Utilization and Information Protection, etc.」, and 「Telecommunications Business Act」, are either unsuitable or insufficient to achieve meaningful user protection effects in the AI service domain. Furthermore, 「Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust」, scheduled for implementation in January 2026, stipulates protective

measures for users as part of the responsibilities of high-impact AI businesses, treating users as 'members of the industrial ecosystem'. To effectively safeguard the rights and interests of users who are 'end consumers' directly using popular AI services and affected by them, a separate legal and policy foundation is required.

This study aims to propose timely and reasonable legislative measures for protecting AI service users, addressing the immediate challenges that our domestic legal system must resolve at this juncture.

3. Contents and Scope of the Research

This study examines the legal and institutional means to protect AI service users through three stages, deriving a plan to establish domestic legislation while considering the global user ecosystem.

First, it conducts comparative and interdisciplinary research to explore an effective normative framework for protecting AI service users. This involves investigating domestic and international policy and legislative status regarding user protection through literature and case studies, and performing in-depth case reviews.

Second, it analyzes key issues central to drafting legislation for AI user protection and proposes specific legislative measures. This research presupposes a legislative proposal that comprehensively considers the rights of individuals, public interest, and socioeconomic values requiring protection in the AI service domain. It aims to systematize rights, obligations, procedural matters, and support/sanction measures to prevent or address foreseeable risks throughout the entire lifecycle of AI service development, provision, and use, while ensuring user rights and interests are adequately safeguarded.

Third, for matters where the scope of application and regulatory tools of certain existing laws (the Information and Communications Network Act, the Telecommunications Business Act) may overlap, particularly those focused on content regulation for child and youth protection, transitional provisions are considered alongside amendments to the

relevant existing legal provisions.

4. Research Results

The core objective of this study is to identify the rights and interests of users that must be guaranteed throughout the entire life-cycle of AI service development, provision, and use, and to develop legislative measures that establish substantive and procedural legal mechanisms for the effective protection of users. An overview of the research content and results is as follows.

A. Establishing the Normative Foundation of the AI Service Ecosystem (Chapter 2)

- (Classification of AI Ecosystem Actors) Analyzes the complex value chain of AI services and classifies the legal actors into 'AI Developers', 'AI Service Providers', and 'Users'.
- (User-Centered Rights Norms and Role-Based Responsibility Allocation) Clarifies the substantive rights of users, who hold the status of both subjects of fundamental rights and responsible user actors, and establishes a foundation for designing a user protection legal framework centered on AI service providers as the core duty-bearers by allocating rights, obligations, and responsibilities corresponding to the controllability and roles of each actor within the triangular structure.
- (Presentation of User-Centered Fundamental Principles) Presents respect for human dignity, assurance of appropriate human control, protection of socially vulnerable groups, prohibition of discrimination, and the securing of transparency as fundamental principles of user protection, applicable throughout the entire lifecycle of AI services.

B. Deriving Implications through Comparative and Sector-Specific Case Analysis (Chapters 3 and 4)

- Analyzes AI regulatory legislation cases in major foreign jurisdictions, with particular consideration given to legislative approaches that meaningfully incorporate differentiated risk management systems based on service risk levels into user protection legislation.
- Through an in-depth analysis of existing protection frameworks in Korea's financial and healthcare sectors, identifies common denominators such as lifecycle-wide governance, the securing of algorithmic transparency, and the mandatory establishment of internal policies by service providers, thereby confirming the justification for incorporating these elements into a general legal regulatory framework.

C. Protection Framework Specialized for Children, Adolescents, and Emerging Risks (Chapter 5)

- (Addressing the Creation and Distribution of Sexual Offense Materials) Proposes measures to strictly regulate the generation of sexual exploitation materials and sexually explicit fabricated videos using generative AI, and to impose obligations on developers and providers to implement technical and administrative measures to prevent such activities.
- (Special Protection for Youth) Proposes measures requiring, when providing services to adolescents under the age of 19, the implementation of safeguards to prevent the generation of harmful content and the incorporation of functions to prevent excessive immersion; in particular, in the case of AI services that simulate emotional interaction, mandates protective measures specifically tailored to adolescent protection.

D. Proposal of Major Legislative Provisions for the 「Act on the Protection of Users of Artificial Intelligence Services (Provisional Title)」 (Chapter 6)

- Proposes measures to legislate an effective implementation system for user protection, users' rights and responsibilities, service providers' obligations and self-regulatory mechanisms, as well as governmental support, supervision, and sanctioning measures.
- (Establishment of an Integrated Governance Framework) Reflecting the characteristics of AI services, avoids fragmented regulation and establishes an integrated user protection governance framework centered on a government-wide comprehensive plan and a dedicated enforcement body.
- (Guaranteeing Users' Rights Under Positive Law) In order to overcome information asymmetry and user marginalization within the AI ecosystem, explicitly codifies in positive law the right to request explanations, the right to raise objections, the right to refuse use, and the right to self-determination, thereby guaranteeing both procedural and substantive rights.
- (Strengthening Prevention-Oriented Obligations of Service Providers) AI service providers should be subject to preventive obligations throughout the entire process of development and operation, rather than relying primarily on ex post remedies. To this end, the principles of 'transparency and safety by design' are adopted as legal standards.
- (Establishing Specialized Norms for the Protection of Children and Adolescents) A special protection framework is established by mandating service design, management, and oversight functions specifically aimed at protecting adolescent users, and by requiring measures to prevent the generation and distribution of illegal and harmful content, including deepfake sexual exploitation materials.
- (Enhancing the Effectiveness of Ex Post Remedies and Accountability Mechanisms) In order to provide prompt and low-cost non-judicial remedies, the establishment of

specialized dispute resolution bodies and the introduction of collective dispute resolution mechanisms are proposed. At the same time, enforcement capacity with respect to cross-border services is strengthened by imposing an obligation on global service providers to designate a domestic representative.

5. Policy Suggestions for Practical Use

The legislative measures, derived from this study may be directly utilized as foundational or reference materials for future legislative reform processes undertaken by the National Assembly or the government. In addition, they provide a normative basis for the formulation of transitional administrative guidance and self-regulatory guidelines aimed at addressing newly emerging issues associated with new services, including generative AI services—particularly the proliferation of illegal and harmful information and the resulting absence of adequate protection mechanisms for children and adolescents.

6. Expectations

At a time when competition within the global AI ecosystem extends beyond technological capabilities to include competition for institutional and regulatory leadership, this study is significant in that it proposes a user protection legal framework that comprehensively considers global legislative trends alongside domestic industrial conditions. By doing so, it seeks to ensure the reliability of AI utilization in both the public and private sectors and to present a legal and policy direction that lowers institutional barriers to entry into global markets.

From this perspective, the key legislative provisions proposed under the 「Act on the Protection of Users of Artificial Intelligence Services (Provisional Title)」 constitute the

organizational foundations and substantive and procedural components required for the protection of AI service users in general. As such, they are expected to contribute to the establishment of a unified regulatory framework governing the broad range of media services into which AI technologies are integrated. Furthermore, the findings of this study are anticipated to serve as a normative reference in future discussions concerning the formulation of specialized user protection regulations and guidelines by individual government ministries, as well as in deliberations over the allocation and coordination of governance responsibilities across regulatory bodies.

제1장 서 론

제1절 연구의 배경과 목적

1. 연구의 배경 및 필요성

오늘날의 인류 사회는 정보화 사회를 넘어 지능정보사회(intelligent information society)로 급격히 진입하고 있다. 과거의 디지털 기술이 인간의 업무를 보조하는 도구적 성격에 머물렀다면, 현시점의 인공지능(이하 'AI')은 학습(learning), 추론(reasoning), 지각(perception), 판단(judgment) 등 인간 고유의 지적 능력을 전자적으로 구현함으로써 사회·경제적 구조의 근본적인 변화를 견인하고 있다. 특히 2022년 말 생성형 AI(generative AI)의 등장과 이후의 급속한 확산은 기술의 패러다임을 '예측과 분류'에서 '창작과 생성'으로 전환시켰으며, 이는 AI 기술이 연구와 산업의 전유물을 넘어 일반 시민의 일상 깊숙이 침투하는 계기가 되었다.

유수의 시장 조사 기관과 주요 정책 연구소에서 발표하는 근년의 통계에서도 AI가 일으키는 일련의 파격이 일시적인 유행이 아니라 불가역적인 흐름이라는 점이 확인된다. 한국 역시 이러한 글로벌 흐름의 중심에 서 있다. 2025년 한국의 AI 시장은 대기업의 인프라 투자와 산업 전반의 AI 전환(AX) 가속화에 힘입어 2025년에 약 13억 7,090만 달러 규모에 달했으며, 2026년부터 2035년까지 연평균 15.60%의 성장률을 기록하면서 2035년에는 약 58억 4,230만 달러에 이를 것으로 전망하고 있다.[1)]

그러나 이와 같은 AI의 급격한 확산에 따라 그 효용만큼이나 각양각색의 새로운 위험도 드러난다. 기존의 소프트웨어가 사전에 정의된 규칙에 따라 작동하여 오류의 원인을 비교적 명확히 규명할 수 있었던 반면에, 현재 딥러닝(deep learning) 기반의 AI는 '블랙박스'(black box)로 간주될 만큼 불투명한 특성을 보인다. 이는 시스템이 내린 판단의 근거를

1) Technology, Media, and IT, South Korea Artificial Intelligence Market Size and Share - Growth Analysis Report and Forecast Trends(2026-2035), 2025. 7.

개발자조차 완벽히 설명하기 어렵게 만들며, 결과적으로 이용자의 권리 구제를 실로 어렵게 하는 원인이 된다.

대표적인 위험 요인으로 데이터 편향(bias)에 따른 차별적 결과 도출, 그럴 듯하지만 사실이 아닌 정보를 생성하는 환각(hallucination) 현상, 딥페이크(deepfake)를 이용한 인격권 침해 및 허위 정보 유포, 개인·기업 데이터의 무분별한 학습에의 이용 등이 지적된다. 특히 생성형 AI가 만들어낸 결과물은 진위 여부를 판별하기 어려워 미디어 정보에 대한 사회적 신뢰를 훼손하고 민주적 의사결정 과정을 왜곡할 우려도 내포하고 있다.

이처럼 급변하는 변화상에도 불구하고, 지금의 법규범 체계는 AI 융합이 가져오는 질적 변화와 그 특성을 충분히 반영하지 못하고 있다. 이를테면, 「제조물 책임법」은 제품의 물리적 결함을 전제로 하므로 알고리즘의 오류에 적용하기 어렵고, 「개인정보 보호법」은 데이터의 수집 및 동의에 초점을 맞추고 있어 이미 학습된 모델의 파라미터 내에 내재된 정보 침해 문제를 규율하는 데는 한계가 있다. 기술 발전 속도와 법제도적 대응 속도 간의 차이로 인한, 이른바 '규제 지체'(regulatory lag) 현상은 AI의 경우에는 특히나 그 간극이 날로 커질 수밖에 없어 법적 공백에 따른 이용자의 불안을 심화시킨다.

이용자 보호는 단순히 피해를 예방하는 소극적 차원을 넘어, AI 산업의 지속 가능한 발전을 위한 전제 조건이다. 신뢰와 안전이 담보되지 않는 기술은 시장에서 수용될 수 없기 때문이다. 최근 유럽연합(EU)이 세계 최초의 포괄적 규제인 「AI 법(AI Act)」을 제정하고 OECD가 'AI 원칙'(AI Principles)을 개정하여 안전성과 책임성을 강화한 것은, 신뢰 확보가 글로벌 규범의 핵심 화두임을 방증한다.

우리 정부 또한 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 'AI 기본법') 제정을 추진하며 이러한 흐름에 동참하고 있다. 이 법안은 AI 산업의 진흥과 안전한 활용을 동시에 달성하는 것을 목표로 하며, 동법 제3조에서 기본 원칙과 국가의 책무를 규정하고 있다. 융합 미디어 영역의 이용자 보호 업무를 총괄하는 방송미디어통신위원회로서는 그 지위에 부응하여 융합 미디어로 분류되는 AI 서비스 생태계에 대한 시의적절한 규율 체계를 마련할 것이 요구된다. 이에 본 연구는 궁극적으로 국가의 이용자 보호 책무를 충실히 이행할 수 있도록 AI 서비스 생태계를 일단(一團)의 규율 영역으로서 정의하고, 구체적·실질적 규범력을 갖춘 법적 수단 체계를 마련하고자 하는바, 이는 권리·의무에 관한 실체법적 사항뿐만 아니라 이를 실행하는 데 필요한 절차법적 사항을 아우르는 것이어야 한다.

2. 연구의 목적

본 연구는 AI 서비스의 잠재적 위험성과 부작용으로부터 이용자를 보호하고, 신뢰할 수 있는 AI 생태계를 조성하기 위한 법적·제도적 기반을 마련하는 것을 목적으로 한다. 이는 단순히 기술을 통제하는 것이 아니라, 이용자의 권익을 증진함으로써 AI 기술의 지속 가능한 발전을 담보하기 위함이다. 구체적인 연구 목적은 다음과 같다.

가. 이용자 중심의 권리 규범 및 보호 체계 구상

본 연구에서는 이용자를 단순한 소비자나 보호 객체가 아니라 권리의 주체이자 책임 있는 사용 주체로 위치 짓는 이용자 중심 권리 규범 및 보호 체계를 구상하는 데 목적을 둔다. 이를 위해 AI 서비스의 순환적 발전 구조와 가치사슬상 역할 분담을 분석하고, 그에 상응하는 이용자의 권리와 책임을 규범적으로 재구성함과 동시에, 서비스 단계의 핵심 수범자인 AI 서비스 제공자에게 이용자 보호 의무를 집중시키는 법적·제도적 설계 원칙을 제시하고자 한다. 이러한 구상은 궁극적으로 인공지능기술의 혁신과 이용자 기본권 보장을 조화시키는 '위험에 비례한 책임'과 '공급망 전체에 걸친 공동 책임'의 조합을 목표로 하며, 향후 AI 서비스 이용자 보호를 위한 입법의 이론적·법리적 기반을 제공하는 것을 목표로 한다.

나. 신유형 위험에 대한 선제적·예방적 규범 정립

딥페이크를 이용한 성착취물, 생성형 AI의 환각 현상에 의한 허위 정보 유포, 챗봇 서비스에 의한 아동·청소년의 정서적 종속 등은 AI 기술 고유의 신유형 위험이다. 이러한 위험은 사후적인 처벌이나 구제만으로는 회복할 수 없는 비가역적 피해를 초래한다. 따라서 본 연구는 '사후 규제'의 한계를 극복하고 '사전적 예방' 중심의 규범 체계를 설계하는 것을 목적으로 한다. 이는 AI 개발자와 AI 서비스 제공자에게 설계 단계에서부터 안전 조치를 강제하는 '설계에 의한 안전'(safety by design) 원칙을 도입하고, 딥페이크 등의 생성 방지 기술 탑재를 의무화하는 법적 근거를 마련하는 것이다.

다. 기술 혁신과 이용자 보호의 균형점 모색

규제가 과도할 경우 AI 기술의 혁신을 저해하고 국가 경쟁력을 약화시킬 수 있다는 우려가 존재한다. 본 연구는 이러한 우려를 불식시키기 위해 '리스크 기반 접근'을 통해 규제

의 합리성과 비례성을 확보하는 것을 목적으로 한다. EU의 「AI 법」이 채택한 리스크 등급 분류체계와 동일한 맥락에서, 사람의 생명과 기본권 또는 공공의 이익에 중대한 영향을 미치는 AI 서비스와 범죄 악용 가능성이 높은 AI 서비스에 규제 역량을 집중하고, 그 외의 영역에서는 자율규제를 원칙으로 하는 유연한 규제 체계를 설계한다. 이는 기술 혁신의 동력을 유지하면서도 이용자의 안전이라는 핵심 가치를 지키기 위한 최적의 균형점을 도출하는 과정이다.

라. 글로벌 정합성을 갖춘 선진적 거버넌스 구축

AI 규제는 국경을 초월하는 글로벌 이슈이다. EU의 「AI 법」, 미국의 관련 연방법 및 주법, 중국의 생성형 AI 관리 방법 등 주요국은 이미 자국민 보호를 위한 법적 규제 체계를 도입하고 있다. 본 연구는 이러한 글로벌 입법 동향을 심층 분석하여, 국제적 기준(Global Standard)에 부합하면서도 한국의 ICT 산업 현실에 최적화된 법제를 설계하는 데 주안점을 둔다. 이는 국내 기업이 글로벌 시장에 진출할 때 직면할 규제 장벽을 해소하고, 동시에 국내 이용자들에게도 세계적 수준의 보호를 제공하기 위함이다.

제 2 절 연구의 내용과 방법

1. 연구의 주요 내용

본 연구는 AI 서비스 이용 환경에서 초래하는 새로운 위험으로부터 이용자를 실질적으로 보호하고, 안전하고 신뢰할 수 있는 서비스 이용 환경을 조성하기 위한 독자적인 입법 방안을 제시하는 것을 목표로 한다. 이를 위해 본 보고서는 다음과 같은 체계로 논의를 전개한다.

첫째, 제2장에서는 AI 서비스 생태계의 특수성을 규범적으로 분석한다. 데이터의 순환 구조와 '개발자-제공자-이용자'로 이어지는 복잡한 가치사슬을 규명하고, 이를 바탕으로 기존의 산업 진흥 중심 법제(「AI 기본법」 등)와 차별화된, 이용자 권익 보호 중심의 '인공지능서비스 이용자 보호 기본원칙'을 도출한다.

둘째, 제3장과 제4장에서는 비교법적 연구와 국내 타 분야 선행 사례를 분석한다. EU의 「AI 법」및 「디지털서비스법(DSA)」, 미국의 알고리즘 책무성 논의 등 해외 주요국의 입법

례를 검토하여 '리스크 기반 접근'(risk-based approach)을 반영하는 규율 체계에 관한 시사점을 얻는다. 또한 금융(「신용정보법」 등) 및 의료(「디지털의료제품법」) 분야에서 선제적으로 도입된 이용자 보호 체계를 분석하여 미디어·콘텐츠 영역 일반에 보장할 수 있는 실정법적 AI 이용자 권리 보호 방향을 모색한다.

셋째, 제5장에서는 아동·청소년 보호 및 딥페이크 등 신종 위험에 대한 특화된 대응 방안을 모색한다. 생성형 AI가 야기하는 성범죄물 생성 및 유통, 청소년의 과몰입 문제 등을 해결하기 위해, 단순한 사후 처벌을 넘어선 예방적 차원의 기술적·관리적 조치 의무를 제안한다.

넷째, 제6장에서는 앞선 논의를 종합하여 연구의 결과로서 「인공지능서비스 이용자 보호에 관한 법률안(가칭)」을 마련하기 위한 주요 입법 사항을 제시한다. 그 내용으로 이용자 보호를 위한 실효적인 정책·행정 거버넌스, 이용자의 실체법적 권리, 서비스 제공자 등의 의무와 자율규제 수단, 사후적 권리 구제 등의 입법 방안을 포함한다.

2. 연구의 방법

본 연구는 문헌 연구, 비교법적 분석, 그리고 법제 설계의 방법을 병행한다. 국내외 문헌을 통해 AI 기술의 발전 현황과 사회적 위험 요인을 식별하고, 「정보통신망법」, 「전기통신사업법」 등 현행 ICT 법제가 지닌 규율 공백을 분석한다. 특히, 2026년 시행 예정인 「AI 기본법」은 산업 진흥과 포괄적 원칙에 중점을 두고 있어 개별 이용자의 구체적 피해 구제에는 한계가 있음을 전제로, 본 연구는 '최종 소비자인 이용자의 권리'와 '서비스 제공자의 구체적 의무와 책임'을 규명하는 데 주력한다. 최종적으로는 이러한 분석을 토대로 실제 발의 가능한 수준의 법률안을 성안(成案)하는 입법 정책적 방법론을 취한다.

제2장 인공지능서비스 생태계에 대한 규범적 접근

제1절 의의

1. 인공지능서비스 이용의 현황

가. 생성형 AI 이용의 폭발적 증가와 일상화

AI 기술 및 서비스의 확산은 추상적인 구호가 아닌 구체적인 수치로 증명되고 있다. 과학기술정보통신부, 한국지능정보사회진흥원(NIA), 정보통신정책연구원(KISDI) 등 주요 기관의 실태조사 결과는 한국 사회가 AI 서비스의 '탐색기'를 지나 '대중화 및 일상화' 단계로 진입했음을 시사한다.

가장 주목할 만한 변화는 생성형 AI(Generative AI) 서비스의 급격한 이용률 증가다. 과학기술정보통신부가 발표한 「2024 인터넷 이용실태조사」에 따르면, 생성형 AI 서비스를 경험해 본 인터넷 이용자의 비율은 2023년 17.6%에서 2024년 33.3%로 불과 1년 만에 약 2배 가까이 급증하였다. 이는 국민 3명 중 1명이 ChatGPT와 같은 생성형 AI를 직간접적으로 사용하고 있음을 의미하며, AI가 특정 전문가 집단의 전유물에서 일반 대중의 도구로 부상하였음을 보여준다. 이러한 추세는 2025년 들어서도 가속화되고 있다. 한 AI 연구소에서 진행한 조사 결과에 따르면, 응답자의 97%가 생성형 AI를 알고 있고, 95%가 활용 중이라고 답할 만큼 그 인지도가 포화 상태에 이르렀다.[2)]

나. 이용 목적의 다변화

이용 목적을 세분화해 보자면, 단순한 '정보 검색(81.9%)'이 여전히 가장 높은 비중을 차지하지만 '문서 작업 보조(44.4%)', '외국어 번역(40.0%)' 등 업무 효율성을 높이는 도구로서의 활용이 두드러진다. 특히 '창작 및 취미활동 보조(15.2%)'와 '코딩 및 프로그램 개발(6.3%)'과 같은 창의적이고 전문적인 영역으로의 확장은 AI가 인간의 지적 노동을 보조하

2) AI매터스, 「AI 트렌스포메이션 2: 2025 생성형 AI 인식 및 활용 현황 조사」, 2025.

는 수준을 넘어 대체하거나 협업하는 단계로 나아가고 있음을 보여준다.[3]

한편, 공공 부문에서의 AI 도입은 민간에 비해 다소 복잡한 양상을 보인다. NIA의 「2024년 전자정부서비스 이용실태조사」에 따르면, AI를 활용한 전자정부 서비스에 대한 인지도는 84.4%로 매우 높게 나타났으나, 실제 이용률은 22.4%에 그치고 있다.[4] 이러한 '인지-이용 격차'는 AI 행정 서비스에 대한 국민의 신뢰 부족, 개인정보 오남용에 대한 우려, 기존 서비스 대비 확실한 효용감을 주지 못하는 과도기적 한계에서 기인하는 것으로 분석된다. 그럼에도 이용 경험자의 94.6%가 편리하다고 응답한 점은 향후 서비스 고도화와 신뢰성 확보 여부에 따라 이용률이 급등할 잠재력이 있음을 시사한다.

2. 현행 법제 분석 및 규제 지체의 문제점

이렇듯 공공·민간 영역을 불문하고 AI 생태계 내에서는 정보와 역량의 불균형 현상이 공통적으로 나타나고 있다. 이는 AI 생태계가 양적으로 팽창할수록 극명해질 것이고 그로 인한 사회적·경제적·문화적 편차가 심화될 우려도 농후하다. 실제로 시장의 성장과 함께 소비자의 피해 사례로 급증하고 있다. 딥페이크 성범죄 관련 방송통신심의위원회 시정 요구는 2020년 473건에서 해마다 급속도로 폭증하여 2024년에는 2만 3,107건에 달했으며, 피해자 중 20대 이하가 90%를 웃돌며 사태의 심각성을 드러내고 있다.[5]

이러한 우려 상황에도 여전히 현행 법제는 AI 특유의 위험성을 규율하지 못하는 한계를 지닌다. 우선 대표적인 유관 법률인 「정보통신망법」과 「전기통신사업법」은 AI 서비스에 대한 적정한 법적 규율을 담당하기 어려운 구조적 한계를 지닌다. 예컨대, 「정보통신망법」의 '정보통신서비스 제공자' 정의에 AI 서비스 제공자가 포함되는 것으로 해석할 여지가 있으나 알고리즘 편향, 자동화된 의사결정, 딥페이크 등 AI 특유의 위험에 대한 전 주기 통제 거버넌스를 담기에는 수범자 영역의 한계가 있다. 또한 「전기통신사업법」의 부가통신사업자 규제도 현재로서는 AI 서비스 특성을 전혀 반영하지 못하고 있으며, 사업자의 투명성 확보 의무나 자동화 의사결정에 관한 규율은 타 법률(「AI 기본법」, 「개인정보 보호법」)

3) 과학기술정보통신부, 「2024 인터넷 이용실태조사」, 2025. 3.
4) 한국지능정보사회진흥원, 「2024년 전자정부서비스 이용실태조사」, 2025. 1.
5) 한국일보, "딥페이크 심의 건수 폭증세…이대로라면 '사상 최다'", 2025. 9. 8.

을 통해 규율된다.

「AI 기본법」은 EU에 이어 세계 2번째 포괄적 AI 영역을 포괄적으로 규율하는 일반법으로, 고영향 AI(생명, 신체 안전, 기본권에 중대한 영향을 미치는 의료기기, 신용평가, 채용 등 10개 영역)를 주된 규제 영역으로 하여 투명성 의무, 이용자 보호 방안 수립을 포함하는 사업자 책무를 부과하고 있다. 그러나 이 법은 금지 AI 유형을 규정하고 있지 않고, 리스크를 기준으로 한 차등화된 규율, 실효적인 사후적 감독 수단 등을 정밀하게 담고 있지 않는 등 AI 서비스를 최종 소비자 지위에서 이용하는 국민 대다수를 보호할 안전망으로서는 기능하기 어렵다.

3. 국가의 기본권 보호 의무 이행으로서의 AI 규제 도입 필요성

독일 연방헌법재판소에서 발전한 국가의 기본권 보호 의무(Schutzpflicht)는 사인(私人)에 의한 기본권 침해로부터 국가가 보호해야 하는 의무를 의미한다. 대한민국 헌법 제10조 후문은 "국가는 개인이 가지는 불가침의 기본적 인권을 확인하고 이를 보장할 의무를 진다"고 명시하여 기본권 보호의무에 대한 명시적인 헌법적 근거를 두고 있다.

우리 헌법재판소는 기본권 보호의무 위반 심사기준으로서 '과소보호금지원칙' 채택하여, 국가가 "적어도 적절하고 효율적인 최소한의 보호 조치를 취했는가"를 기준으로 심사하고 있다. 이러한 헌법재판소의 태도에 비추어 AI 서비스에 의한 이용자 권리(기본권) 침해 상황에서 국가의 적극적 보호의무 이행은 헌법적 요청으로도 새길 수 있다.

법률로써 AI 서비스 이용자의 권리를 보호하기 위한 법률을 제정하는 것은 이러한 국가의 기본권 보호의무 차원에서 정당성과 당위성을 찾을 수 있다. 즉, AI 서비스 영역을 기본권에 대한 영향력을 고려하여 사전적 위험 예방 조치, 사후적 피해 구제, 자율규제를 동원한 주기적 위험관리 등을 법제화하는 것은 곧 인간의 존엄성을 천명한 헌법 제10조 전문, 사생활 보호에 관한 헌법 제17조, 표현의 자유를 보장하는 헌법 제21조 등을 구체화하는 것과 다름없다.

제 2 절 인공지능서비스 생태계의 현상과 구조

1. 인공지능의 순환적 발전과 작용 구조

AI 서비스는 고정된 기술 산출물이 아니라, 데이터-모델-서비스-이용자 행태-규제 환경이 상호 피드백을 주고받으며 순환적으로 발전하는 동학적 구조를 가진다. 머신러닝·생성형 AI 연구에서 강조되는 '데이터-모델-피드백 루프'는, 서비스가 확산될수록 더 많은 이용자 데이터와 상호작용 로그를 수집·학습에 재투입함으로써 성능이 향상되지만, 동시에 편향·차별·프라이버시 침해 위험도 누적·증폭된다는 양면성을 보여준다.

이러한 순환은 대체로 다음과 같은 단계로 요약할 수 있다.

1. 데이터 수집·전처리 단계: 플랫폼·공공기관·기업이 이용자 행태, 콘텐츠, 행정·거래 기록 등 대규모 데이터를 수집·정제한다.
2. 모델 개발·학습 단계: 수집된 데이터가 개발자 또는 모델 제공자에 의해 학습에 활용되고, 생성·분류·추천·예측 등 다양한 기능을 수행하는 모델이 생성된다.
3. 서비스 설계·배포 단계: 서비스 제공자는 해당 모델을 특정 서비스 맥락(검색·챗봇·신용평가·복지급여 산정 등)에 통합하여 이용자에게 제공하고, 이 과정에서 UX·설명 구조·알림 체계를 설계한다.
4. 이용자 상호작용 및 결과 발생 단계: 이용자는 서비스 결과를 자신의 의사결정·행위에 반영하고, 이 과정에서 개인·집단 수준의 혜택과 피해가 동시에 발생할 수 있다.
5. 규범·규제·사회적 반응 단계: 부작용이 사회적으로 가시화되면, 사법판결·규제기관의 가이드라인·산업 자율규범·이용자 저항 등이 나타나고, 이는 다시 데이터 수집·모델 설계·서비스 운영 방식에 영향을 미친다.

AI를 둘러싼 법학적 논의는 이 순환을 '자연발생적 시장 메커니즘'으로 전제하는 대신, 각 단계에서의 위험과 권력 비대칭을 인식하고 규범적 개입 지점을 설정하는 작업으로 이해될 수 있다. 예컨대 공공행정 영역에서 알고리즘적 복지급여 산정 시스템이 투명성 부족과 데이터 검증 부재로 인해 취약계층의 권리를 침해한 사건은, 개발-도입-운영-사후검증이 독립된 단계가 아니라 동일한 책임 연쇄에 속한다는 점을 보여주며, 이에 따라 입법

자는 생애주기 전체를 대상으로 한 규율체계를 설계해야 한다는 결론에 도달한다.

결국, AI 서비스 규율의 출발점은 "한 번 설계된 시스템이 그대로 작동한다"는 정태적 가정을 버리고, AI가 이용자와 상호작용하는 과정에서 스스로를 재구성하고 사회 구조를 재형성하는 순환적 작용 구조를 전제로 하는 데 있다. 이러한 관점에서 보면, 이용자 보호 규범은 단순한 '사후 구제'가 아니라, 순환 과정 속에서 위험 요소를 지속적으로 탐지·수정·재설계하도록 유도하는 동태적 규범으로 이해되어야 한다.

2. 인공지능사업자 간의 복합적 영향 구조

AI 생태계의 사업자 구조는 단일한 "서비스 제공자–이용자" 이분법으로는 포착되기 어렵고, '가치사슬'(value chain) 또는 '공급망'(supply chain) 관점에서 다수의 행위자를 포괄하는 네트워크로 이해하는 것이 적절하다. EU「AI 법」 등이 제시하는 분석틀에 따르면, 설계자·개발자·모델 제공자(provider), 시스템을 실제 환경에 배치하는 도입·운영자(deployer), 수입·유통·통합 사업자, 그리고 서비스 이용자와 영향 받는 제3자가 함께 하나의 가치사슬을 구성한다.

이러한 구조에서 각 사업자는 다음과 같은 방식으로 상호 영향을 미친다.

첫째, 서비스 운영자는 기저 모델·시스템의 설계 선택(데이터셋, 아키텍처, 필터링 정책)에 대한 정보에 의존하며, 개발자의 투명성과 협조 수준이 낮을수록 서비스 운영자의 책임 이행 가능성이 제한된다(기술·정보 의존성).

둘째, 고위험 AI 시스템의 사고·침해 결과는 이용자에게 직접 나타나지만, 인과 관계는 개발 단계의 설계 결함, 배포 단계의 맥락 오용, 운영 단계의 모니터링 부실이 복합적으로 작용한 결과인 경우가 많다(위험·책임의 분산).

셋째, 대형 플랫폼·모델 제공자가 설정하는 기술 표준·API 정책·콘텐츠 정책은 그 위에 구축되는 수많은 서비스의 규범과 관행을 사실상 좌우하고, 반대로 주요 이용자·기업 고객의 요구가 상위 사업자의 정책을 역으로 변화시키기도 한다(자율규제와 규범 전파).

법학적 관점에서 이러한 복합적인 영향 구조는 두 가지 함의를 갖는다. 첫째, 책임 귀속을 '최종 서비스 제공자'에게만 집중할 경우, 실제 위험에 영향을 미치는 개발·공급 단계의 행위자(모델 개발자, 데이터 브로커, 인프라 제공자)의 규율 공백이 발생한다. 둘째, 반

대로 개발자에게만 광범위한 책임을 귀속하면, 특정 맥락에서의 위험 발생 가능성과 이용자와의 접점에서 작동하는 보호 조치를 설계할 권한을 가진 운영자의 책임이 과소평가된다.

따라서 AI 생태계를 규율하는 법제는 가치사슬의 각 단계에 상응하는 '역할별 책임' 구조를 설정하고, 상호 협력·정보 공유 의무를 통해 공동 위험관리 체계를 구축하는 방향으로 설계될 필요가 있다. 이는 기존의 「제조물 책임법」이 제조자–판매자–소비자 간 책임 분담을 정교하게 설계해 온 것과 유사한 구조를 AI 맥락에서 재구성하는 작업으로 이해할 수 있으며, EU 「AI 법」이 'provider-deployer-user' 각 단계에 별도의 의무를 배분하는 방식이 그 대표적 사례라고 하겠다.

3. 인공지능서비스 생태계의 인적(人的) 구성

AI 서비스 생태계의 인적 구성은 이론적으로는 다양한 행위자를 포함하는 복잡한 네트워크로 묘사될 수 있으나, 입법 단계에서는 규범 설계의 명확성과 집행 가능성을 위해 일정한 단순화·추상화가 불가피하다. 이하에서 제안하는 3각 구조(AI 개발자–AI 서비스 제공자–이용자)는, 실제 AI 가치사슬 논의와 부합하면서도, 국내 법체계에서 수범자와 권리주체를 명료하게 정리하는 효과를 가진다.

가. 3각 구조로의 구조화: 입법 기술적 선택

첫째, AI 개발자는 AI 시스템·모델을 설계·학습·개선하는 기술 주체로서, 시스템 개발·공급 단계에서 데이터 선택·전처리, 알고리즘 구조 결정, 필터링·안전장치 구현 등을 통해 시스템의 기본적인 위험 수준과 편향 구조를 형성한다. 이들은 통상적으로 고도의 기술적 통제력을 가지지만, 최종 서비스 맥락에 대한 정보는 제한적일 수 있다는 점에서, 설계 단계의 일반적 안전·인권 보호 의무를 중심으로 규율하는 것이 합리적이다.

둘째, AI 서비스 제공자는 해당 AI 시스템을 직접 개발하거나, 외부 개발자로부터 공급받아 자사 서비스에 통합하여 이용자에게 제공하는 주체로서, 서비스 단계의 핵심 수범자로 설정된다. 이 범주에는 1) 자체 모델·시스템을 개발하여 플랫폼·앱·웹서비스 형태로 제공하는 사업자, 2) 상용 API·오픈소스 모델을 도입·조합하여 서비스화하는 사업자가 모두 포함되며, 법률상 동일한 "서비스 제공자" 개념 아래 포섭함으로써, 기술적·계약적 구조에 관계없이 이용자와 직접 대면하는 주체에게 1차적 이용자 보호 의무를 집중시킬 수

있다.

셋째, 이용자는 AI 서비스의 직접 이용자(최종 소비자)뿐 아니라, 해당 서비스의 결과에 의해 간접적으로 영향을 받는 제3자를 포함하는 폭넓은 개념으로 설정된다. 예컨대 채용·신용평가·복지급여 산정에 활용되는 AI 시스템의 경우, 시스템과 직접 상호작용한 당사자뿐 아니라, 추천·랭킹에서 배제된 지원자, 프로파일링에 따라 불이익을 받는 집단도 "영향 받는 자"로 포함된다. 이와 같이 이용자 개념을 넓게 잡으면, 실제 피해 구조에 상응하는 권리 보호·구제 범위를 확보할 수 있다.

이상의 3각 구조는 엄밀한 산업 분석에서 제시되는 세분화된 행위자(모델 제공자, 배포자, 시스템 통합자, 데이터 브로커 등)를 모두 법문에 열거하기보다, 핵심 법적 지위에 따라 세 범주로 통합하는 입법기술적 전략이다. 그 결과, 세부 역할의 다양성은 하위 법령·가이드라인에서 세분화하되, 모법 차원에서는 책임의 축이 되는 세 주체 간 관계를 중심으로 규율 체계를 설계할 수 있다는 장점이 있다.

나. AI 서비스 제공자의 포섭 범위 확대의 의미

AI 서비스 제공자 개념을 '직접 시스템을 개발하여 서비스화하는 자'와 '개발자로부터 시스템을 공급받아 서비스를 제공하는 자'를 모두 포함하도록 설정하는 것은, 서비스 단계에서의 주된 수범자 역할을 실질적으로 실행하기 위한 최선의 입법 방안이 될 수 있다. 이는 다음과 같은 규범적 함의를 가진다.

첫째, 이용자와의 접점에서 실제로 권리침해·피해가 발생하는 지점은 대부분 서비스 제공·운영 단계이므로, 해당 단계의 행위자를 일관된 개념으로 포착하여 정보제공·설명·접근권 보장·위험관리·피해구제절차 운영의무를 집중시키는 것이 타당하다.

둘째, 기술 구조가 '자체 개발형'인지 '외부 모델 도입형'인지에 따라 수범자 개념을 달리하면, 동일한 위험을 초래하는 서비스가 개발 방식만 다르다는 이유로 서로 다른 규율을 받는 결과가 초래될 수 있다. 통합된 '서비스 제공자' 개념은 이러한 규제의 형식주의를 피하고, 이용자 관점에서 동일 기능·동일 위험을 동일하게 취급하는 기능적 규제 설계를 가능하게 한다.

셋째, 시스템 개발·공급 단계의 개발자는 기저 모델 수준에서 일반적 안전 의무를 부담하되, 구체적인 서비스 맥락에서의 설명, 이용자 대응, 청소년 보호 등은 서비스 제공자에

게 1차적으로 귀속함으로써, 각 단계가 통제 가능한 범위 내에서 책임을 부담하는 '역할별 의무'체계를 구성할 수 있다.

이러한 구조는 EU 「AI 법」이 'provider'(모델 제공자)와 'deployer'(배포·운영자)를 구분하면서, 'deployer'에게 이용자 중심의 의무(정보 제공, 인간 감독, 사용 맥락 관리 등)를 집중시키는 구도와도 상응하며, 국제 규범과의 정합성을 높인다는 점에서도 의미가 있다.

다. 3각 구조에 기초한 규범 설계의 방향

마지막으로, 이와 같은 3각 인적 구조는 이후 제3절에서 다룰 "권리·의무·책임" 논의를 위한 규범적 프레임을 제공한다. 구체적으로는 다음과 같이 구분하여 규율 체계를 설계할 수 있다.

1) AI 개발자

데이터·알고리즘 설계 단계에서의 안전·비차별·프라이버시·보안 의무, 시스템 문서화·모델 카드 제공 등 정보 제공 의무, 서비스 제공자에 대한 협조·정보 공유 의무를 중심으로 규정한다.

2) AI 서비스 제공자

이용자와의 직접 관계에서 작동하는 정보제공·설명·접근권 보장 의무, 청소년·취약계층 보호 의무, 피해 발생 시 통지·구제절차 운영 의무, 개발자·다른 공급자와의 위험 공유·협력 의무를 부담하는 수범자로 설정한다.

3) 이용자(최종 소비자 및 영향 받는 자)

안전·비차별·절차적 권리(설명·이의제기·재검토·인간 개입 요청) 등 AI 시대의 기본권을 핵심 권리로 보장하는 한편, 생성형 AI의 악용·오남용을 자제하고, 공적 지침·이용 규범을 준수할 책임 있는 이용 주체로서의 지위를 규범적으로 선언한다.

이와 같이 인적 구성을 3각 구조로 명확히 설정하고, 각 주체의 역할·통제 가능성에 상응하는 권리·의무를 배분하는 것은, AI 서비스 이용자 보호 법제의 설계 원리로서 '위험에 비례한 책임'과 '공급망 전 주기에 걸친 공동 책임'을 조화시키기 위한 입법적 기반을 제공한다.

제 3 절 인공지능서비스 생태계의 규범적 이해

1. 인공지능서비스 생태계 내 법적 주체 및 상호 관계

AI 생태계를 규범적으로 이해한다는 것은, 기술·산업 구조에서 확인된 행위자들을 법적 주체로 전환하고, 이들 간 권리·의무·책임 관계를 어떻게 설계할 것인가를 묻는 작업이다. 이는 전통적인 사법·공법 이분법을 AI 맥락에 투사하는 것을 넘어, 다층적 거버넌스 구조를 전제로 한 새로운 관계론이 요구되는 영역이다.

우선, 이용자와 서비스 제공자의 관계는 계약·소비자 보호 규범에 의해 전통적으로 구조화되어 왔으나, AI 환경에서는 정보 비대칭·자동화된 의사결정·집단적 영향이라는 요인들이 이 관계를 근본적으로 변화시킨다. 제공자는 단지 '약관과 상품·서비스를 제시하는 자'가 아니라, 이용자가 인지하기 어려운 방식으로 프로파일링·추천·랭킹·필터링을 수행하여 이용자의 선택 구조 자체를 형성하는 '구조 설계자'로 기능하며, 이에 따라 이용자의 자기결정권·공정한 절차에 대한 권리를 존중할 추가적인 의무를 진다.

다음으로, 개발자·제공자와 규제기관 간에는 전형적인 감독·규제 관계가 형성되지만, AI의 기술적 복잡성과 규제기관의 정보 열위, 국제적 공급망을 고려할 때 전통적인 명령·통제식 규제만으로는 충분하지 않다. 이에 따라 위험 기반 접근(risk-based approach), 사후 모니터링·감독 강화, 기업 내부 거버넌스(알고리즘 책임자, 윤리심의위원회 등)에 대한 요구 등 '규범의 공동 생산'의 요소가 결합된 새로운 규율 방식이 논의되고 있다.

마지막으로, 이용자와 사법·준사법기관(법원, 분쟁조정위원회, 사업자 내부 민원처리절차 등) 간 관계는 AI 시스템이 초래하는 권리 침해에 대한 사후적 구제 수단을 의미한다. 여기서 핵심적인 법리적 쟁점은 (1) 인과관계와 증명 책임을 어떻게 완화·재배분할 것인지, (2) 집단적 피해에 대해 집단소송·집단분쟁조정 등의 절차를 어떻게 설계할 것인지, (3) 기술적 불투명성 하에서 절차적 적정성과 실질적 공정을 어떻게 확보할 것인지이다. AI 생태계 규율은 결국, 이들 다수 주체 간 상호 관계를 '책임의 사각지대 없이' 구성하려는 시도로 이해될 수 있다.

2. 인공지능서비스 '이용자'의 권리와 책임

AI 생태계에서 이용자는 더 이상 수동적·피규범적 객체가 아니라, 기술 발전과 규범 형성 과정에 적극적으로 참여하는 행위자로 봐야 한다는 논의가 강화되고 있다. 이를 전제로 할 때, 이용자의 법적 지위는 기본권의 향유자, 알고리즘적 절차에 대한 통제권 행사자, 책임 있는 이용(responsible use)의 주체라는 세 층위에서 재구성될 수 있다.

첫째, '기본권 향유자'로서의 지위이다. AI 서비스는 프라이버시·데이터 보호, 평등권·비차별, 표현의 자유, 정보접근권, 적정절차(due process) 등 다양한 기본권에 영향을 미친다. GDPR 제21조·제22조가 자동화된 의사결정에 대한 거부권·비적용권을 인정하는 것처럼, AI 맥락에서 이용자는 자신의 데이터 처리와 알고리즘적 판단에 대해 최소한 1) 사전 정보 제공, 2) 설명 요구, 3) 이의제기·재검토 요청, 4) 사람의 개입을 요구할 권리를 가져야 한다는 논리가 설득력을 얻고 있다.

둘째, '알고리즘적 절차 통제자'로서의 지위이다. 알고리즘이 공공·민간 영역에서 행정 지원, 신용평가, 채용, 콘텐츠 추천 등에 사용될 때, 이용자는 단지 결과를 수동적으로 수용하는 존재가 아니라, 절차의 공정성·투명성·책임성을 요구할 수 있는 주체로 위치 지워져야 한다. 이때 '설명의 권리'(right to explanation)는 단지 기술적 세부를 요구하는 권리가 아니라, 자신의 권익에 중대한 영향을 미치는 판단이 어떤 기준과 데이터에 의해 이루어졌는지에 대한 이해 가능성을 요구하는 절차적 권리로 이해된다.

셋째, '책임 있는 이용 주체'로서의 지위이다. AI가 생성·확산하는 정보의 위력, 생성형 모델을 통한 허위정보·성범죄물·혐오표현의 생산 용이성을 고려할 때, 이용자에게도 악용·오남용을 자제하고 윤리적 지침을 준수할 책임을 부여할 필요성이 논의된다. 이는 전통적 형사·민사 책임과 별개로, 이용자를 '공동 규범 형성자'(co-regulator)로서의 지위를 부여하는 규범적 발상으로, 교육·디지털 리터러시·플랫폼 제공자의 이용자 교육 의무 등과 연계될 수 있다.

이러한 논의를 통해, 이용자에 대한 권리·책임 규정은 단지 보호와 규제 사이의 균형 문제가 아니라, AI 사회에서 시민이 어떤 역할을 수행할 것인가를 규범적으로 선언하는 헌법적 선택에 가깝다는 점을 도출할 수 있다.

3. 인공지능서비스 '제공자'의 이용자 보호 의무와 책임

AI 서비스 제공자의 의무와 책임은, 1) 위험 기반 의무 설정, 2) 투명성·설명 가능성·감시 의무, 3) 취약계층·특수 서비스에 대한 강화된 의무, 4) 공급망 전체에 걸친 공동 책임 구조라는 네 가지 축에서 논의될 수 있다.

첫째, 위험 기반 의무 설정이다. EU「AI 법」은 AI 시스템을 위험 수준에 따라 금지, 고위험, 제한적 위험, 최소 위험으로 구분하고, 고위험 시스템의 제공자·운영자에게 엄격한 요구사항(데이터 거버넌스, 기술 문서 작성, 기록 유지, 투명성, 인간 감독, 정확성·강건성·사이버 보안 등)을 부과한다. 이는 모든 AI 서비스에 동일한 의무를 부과하는 것이 아니라, 이용자의 생명·신체·기본권에 미치는 잠재적 영향의 정도에 따라 차등화된 의무를 규정해야 한다는 규범적 원칙을 제시한다.

둘째, 투명성·설명 가능성·감시 의무이다. 제공자는 AI 시스템의 설계·학습 데이터·성능·한계에 관한 정보를 적절한 수준에서 공개하고, 사후 모니터링·사고 보고·개선 조치를 수행할 의무를 진다. 이는 단순한 정보 제공을 넘어, 내부 거버넌스 구조(책임자 지정, 위험관리 프로세스), 외부 감시(감독기관·감사·감사 보고), 이용자를 위한 접근 가능한 설명 구조(간명한 언어, UI 설계)를 포함하는 입체적 의무로 이해될 수 있다.

셋째, 취약계층 및 특수 유형 서비스에 대한 강화된 의무이다. 청소년·장애인·사회경제적 취약계층 등은 AI 서비스의 부작용에 더 취약할 수 있고, 생성형 AI를 비롯한 특정 서비스 유형은 이용자의 감정·신상에 미치는 영향이 크다는 점에서 별도의 보호가 정당화된다. 이에 따라 제공자는 연령·상태에 따른 차등 설계, 접속·이용 제한, 콘텐츠 필터링, 과몰입·중독 방지 기능, 보호자·후견인과의 연계 기능 등을 추가적으로 구현할 의무를 부담할 수 있다.

넷째, 공급망 전체에 걸친 공동 책임 구조이다. 제공자는 자신이 직접 통제하지 않는 기저 모델·외부 데이터·협력 개발자의 결과물을 활용하는 경우가 많으므로, 단독 책임만으로는 충분한 예방 효과를 기대하기 어렵다. 따라서 계약·표준·가이드라인을 통해 상위 공급자·협력업체·파트너와 위험 정보를 공유하고, 합리적인 수준의 현장 조사를 수행하며, 문제가 발생했을 때 원인 규명과 개선 조치에 협력할 의무를 진다는 공동 책임 모델이 필요하다.

법학적으로 보면, 이러한 AI 서비스 제공자의 의무·책임 구조는 '기술적 중립성' 논의를 넘어서, AI 서비스 제공자가 사회적 위험을 관리하고 이용자의 기본권을 보호할 책임을 지는 공법적·사법적 행위자로서 재정의하는 토대가 된다. 이 과정에서 책임의 과중 부과가 혁신을 저해하지 않도록, 명확한 기준·예측 가능성·규제 샌드박스 등 완충 장치의 설계가 병행되어야 한다는 점도 함께 고려되어야 한다.

제 4 절 소결: 인공지능서비스 이용자 보호 기본원칙

앞서 논의한 생태계의 특성과 이용자 중심의 규범적 요구를 종합하여, 본 연구의 결과물의 초석이 되는 '인공지능 서비스 이용자 보호 기본원칙'을 다음과 같이 도출하였다. 이를 「인공지능서비스 이용자 보호에 관한 법률(안)」의 총칙 규정으로 명문화한다면, 동 법률(안)의 당위성과 정체성을 명확히 하는 선언적 규정으로서 의미를 지니게 된다. 이는 「AI 기본법」 제3조의 포괄적 기본원칙과 구분되는, 이용자의 실질적 권익 보호에 초점을 맞춘 독자적 의미의 특별원칙에 해당하는 것이다.

1. 인간 존엄성 및 통제권 보장의 원칙

AI 시스템과 AI 서비스는 인간의 존엄성과 개인의 자율성을 최우선 가치로 하여 개발·이용되어야 한다. 이용자는 AI의 판단에 종속되지 않고, 이를 적절하게 통제하고 감독할 수 있는 권한을 가져야 한다. 이는 AI가 인간을 위한 도구로서 기능해야 함을 명확히 하는 것이다.

2. 지속 가능성 및 중장기적 영향 고려의 원칙

AI 서비스는 개인·사회·환경에 이익이 되는 지속 가능한 방식으로 제공되어야 한다. 특히 서비스 제공 이후에도 개인의 안전과 기본권, 공공의 이익에 미치는 중장기적인 영향을 종합적으로 고려하여, 잠재적인 위험이 이용자에게 전가되지 않도록 해야 한다.

3. 취약계층 보호 및 접근성 보장의 원칙

모든 계층과 연령이 용이하게 접근할 수 있어야 하며, 특히 아동 및 사회·경제적 취약계층에 대해서는 특별한 보호 조치가 이루어져야 한다. 이는 AI 서비스 혜택의 형평성을 보장하고, 디지털 격차로 인한 불이익을 방지하기 위함이다.

4. 차별 금지 및 다양성 존중의 원칙

AI 서비스는 설계 및 제공 단계에서 이용자에게 부당한 차별이나 불공정한 결과를 초래해서는 안 된다. 또한 사회 구성원의 민주적 다원성과 문화적 다양성을 촉진하는 방향으로 개발되어야 하며, 알고리즘 편향에 의한 획일화를 경계해야 한다.

5. 문제 해결 및 이의제기 보장의 원칙

AI 서비스 제공으로 인한 문제 상황에 적절히 대응할 수 있어야 하며, 이용자가 자신의 권익에 위해를 가하거나 피해를 초래하는 AI 서비스에 대해 이의를 제기하고 설명을 요구할 수 있는 절차적 권리가 보장되어야 한다.

6. 투명성 및 알 권리 보장의 원칙

이용자의 권익에 영향을 미치는 정도에 비례하여, AI 서비스의 적절한 이용에 필요한 정보가 제공되어야 한다. 이는 '블랙박스' 문제로 인한 정보 비대칭을 해소하고, 이용자의 합리적 선택을 지원하기 위한 핵심 원칙이다.

7. 데이터 자기결정권 및 기술적 · 관리적 보호의 원칙

개발 및 제공 단계에서 이용자의 데이터가 부당하게 처리, 추론, 학습되지 않도록 적정한 기술적·관리적 조치가 이행되어야 한다. 또한 서비스 전 주기 동안 합리적으로 실행 가능한 범위에서 데이터 주체로서의 이용자 지위와 결정이 존중되어야 한다.

8. 자율규제 및 책임성의 원칙

AI 개발자와 AI 서비스 제공자는 법적 의무를 넘어, 서비스의 위험성과 부작용으로부터 이용자를 보호하기 위한 내부 정책 및 지침을 자율적으로 수립·운영함으로써 책임 있는 혁신을 추구해야 한다.

제3장 인공지능서비스 이용자 보호에 관한 비교법적 접근

제1절 서 설

인공지능(AI) 서비스는 생성형 AI의 등장 이후, 단순한 정보 제공 도구를 넘어 이용자와 정서적·심리적 관계를 형성하는 단계로 진입하였다. 이에 따라 각국은 기존의 소비자 보호, 개인정보 보호, 플랫폼 규제 체계만으로는 포섭하기 어려운 AI 특유의 위험성에 대응하기 위한 입법적 노력을 기울여 왔다.

AI 서비스 이용자 보호 정책의 핵심 쟁점은 크게 세 가지로 압축된다. 첫째, 투명성과 설명가능성의 문제이다. 이는 이용자가 AI 시스템과 상호작용하고 있음을 인지하고 그 작동 원리를 이해할 수 있도록 보장하는 것이다. 둘째, 리스크 기반(risk-based) 안전 규제의 설계로서, AI 시스템이 이용자의 권리와 안전에 미치는 영향의 정도에 따라 차등화된 규제를 적용하는 것이다. 셋째, 취약계층 보호의 문제이며 미성년자, 고령자 등 AI의 영향에 특히 취약한 이용자에 대한 강화된 보호 장치를 마련하는 것을 말한다.

이러한 쟁점에 대하여 주요국은 상이한 규제 철학과 접근법을 채택하고 있다. EU는 「AI법」 외에도 「디지털서비스법」(Digital Services Act, 이하 'DSA')을 통해 온라인 플랫폼의 이용자 보호 의무를 체계화하고, AI 기반 추천 시스템의 투명성 및 이용자 선택권을 강화하는 규정을 마련하였다. 중국은 알고리즘 추천, 딥페이크, 생성형 AI 등 기술 유형별로 개별 규제 법령을 순차적으로 제정해 왔으며, 2025년 12월에는 AI 동반자 서비스 등 '의인화 상호작용 서비스'에 특화된 새로운 규제 초안을 발표하여 '정서와 감정의 안전'(emotional safety)이라는 새로운 보호 영역을 개척하였다. 일본은 2025년 5월 「인공지능 관련 기술의 연구개발 및 활용 추진에 관한 법률(이하 'AI 추진법')」을 제정하여 최초의 AI 관련 입법을 마련하였으나, 이는 진흥 중심의 기본법적 성격을 가지며 구체적 규제보다는 가이드라인과 기존 법률(「개인정보보호법」, 「저작권법」, 「소비자보호법」 등)을 통한 거버넌스 모델을 유지하고 있다. 이용자 보호에 관해서는 'AI 사업자 가이드라인' 등 연성 규범을 통해 정책적 방향을 제시하는 접근을 취하고 있다.

이하에서는 방송미디어통신위원회의 소관 법률로서 「인공지능서비스 이용자 보호에 관한 법률(안)」을 마련함을 염두에 두고, 동 법안의 체계를 설계하고 전체 법률규정을 성안하기 위한 기초 연구로서 비교법적 관점에서 해외 입법례를 분석하고자 한다. 동일한 쟁점에 대한 입체적 비교를 위하여 대표적인 다섯 가지 입법 과제 — 1) 불법·유해 콘텐츠 규제, 2) 이용자의 데이터 관련 자기결정권 보호, 3) 아동·청소년 보호, 4) 투명성 확보, 5) 분쟁조정 및 피해구제 — 에 대한 각각의 관련 법제 동향을 대별하여 검토하고(제2절), 이 가운데 이용자 보호 관점에서 특히 참고가 되는 EU와 중국의 최근 입법 사례에 대해서는 별도로 주요 규정들을 심층적으로 분석하여 국내 입법에의 시사점을 도출한다(제3절).

제 2 절 쟁점별 해외 법제 동향 분석

1. 불법 · 유해 콘텐츠 규제

가. EU

EU는 세계 최초의 포괄적 AI 규제법인 「AI 법」과 온라인 플랫폼의 책임을 규정한 「DSA」를 통해 불법·유해 콘텐츠 문제에 체계적으로 접근하고 있다.

1) 「AI 법」(AI Act)(2024)

EU의 「AI 법」은 EU 내 공통적인 AI 원칙을 수립하고 인간 중심의 AI 시스템 분류를 목적으로 하고 있으며, 2024년 6월 이후부터 단계적으로 수행되고 있다.[6] AI 시스템이 초래할 수 있는 위험 수준에 따라 규제를 차등 적용하는 리스크 기반 접근(risk-based approach) 방식을 채택하여 인간의 존엄성, 자유, 평등 등 기본권에 중대한 침해를 가할 수 있는 AI 시스템은 '금지된 AI'로 규정하여 시장 출시 및 이용을 원칙적으로 금지한다. 금지 대상은 의사결정 왜곡 AI, 심리적 취약성 악용 AI, 무차별 얼굴인식 데이터 스크래핑, 범죄 예측용 사회적 점수 시스템 등 4개 시스템과 8개 활용 영역으로 구성되며, 위반 시에는 최고 3,500만 유로 또는 전 세계 연간 매출의 7%에 해당하는 막대한 벌금이 부과될 수 있다.

6) 딜로이트 인사이트, 「인공지능(AI) 시대의 새로운 국면, AI 규제와 기업 리스크 관리 전략」, 2024. 10.

- 의사결정 왜곡: 인간의 잠재의식을 이용하거나 조작·기만하여 의사결정 능력을 저하시키고 원치 않는 결정을 유도하는 시스템 금지(예: 아동에게 위험한 행동을 부추기는 음성비서 장난감)
- 특정 집단의 심리적 취약성 악용: 개인이나 특정 집단의 연령, 장애, 사회·경제적 상황의 취약성을 이용하여 이들의 행동을 중대하게 왜곡하거나 피해를 야기하는 시스템 금지(예: 고령자의 연령 관련 취약성을 상업적 마케팅에 악용하는 음성 비서)
- 사회적 점수 평가: 개인의 사회적 행위나 성격 특성을 기반으로 점수를 매겨, 본래 맥락과 무관하거나 비례적이지 않은 불이익(공공복지 접근 제한 등)을 초래하는 시스템 금지(예: 공공복지, 건강관리, 고용 접근 등과 무관한 분야에서 개인의 행동이나 성격 특성을 바탕으로 점수화하거나 순위를 정하는 AI 시스템이 불이익을 초래하는 경우)
- 프로파일링을 이용한 범죄 예측: 개인의 프로파일링이나 성격 특성 평가만으로 범죄 가능성을 예측하는 시스템 금지.(예외: 범죄 행위와 직접 관련된 객관적 사실에 기반한 인적 평가를 지원하는 경우)
- 무차별 얼굴인식 데이터 스크래핑: 인터넷이나 CCTV 영상에서 무작위로 얼굴 이미지를 스크래핑하여 얼굴 인식 데이터베이스를 생성·확장하는 시스템을 금지
- 직장/교육기관에서의 감정 추론: 직장이나 교육 환경에서 개인의 감정을 추론하는 AI 시스템의 출시 및 사용 금지.(예외: 의료 또는 안전상의 목적)
- 생체인식 분류: 인종, 정치적 견해, 종교, 성적 지향 등 민감 정보를 추론하기 위해 생체 정보를 기반으로 개인을 분류하는 시스템 금지.(예외: 법 집행 목적으로 합법적으로 획득한 데이터의 라벨링/필터링 등)
- 실시간 원격 생체인식 신원 확인: 법 집행 목적으로 공개된 장소에서 실시간 원격 생체인식 시스템을 사용하는 것 금지.(예외: 실종자 수색, 테러 위협 방지, 중대 범죄 용의자 식별 등 엄격히 제한된 경우)

2) 디지털 서비스 법(DSA)(2023)

EU의 「DSA」는 불법 콘텐츠에 대한 대응과 표현의 자유 간 균형을 맞추는 데에 목적을 둔다.[7] 「DSA」에서는 딥페이크 등 AI 생성 콘텐츠가 각 회원국의 법률에 따라 '불법 정보'로

7) 최경미·지성우, "디지털서비스 투명성·개인정보보호·위기대응에 관한 법 연구-EU 디

판단될 경우, 이를 유통한 온라인 플랫폼 기업에 법적 책임을 부과한다. 특히 월 이용자 4,500만 명 이상의 대형 온라인 플랫폼 및 검색 엔진 사업자에게는 다음과 같은 3대 핵심 의무를 부과하여 콘텐츠 유통에 대한 책임을 강화한다.

- 신고 시 조치 의무: 이용자로부터 불법 정보 유통 신고를 받으면 신속히 평가하여 삭제, 차단 등의 조치를 이행해야 한다.
- 시스템적 리스크 평가 의무: 불법 콘텐츠 유통, 기본권 침해, 선거 및 공중보건에 미치는 악영향, 아동·청소년에 대한 유해성 등 플랫폼이 야기할 수 있는 시스템적 위험을 의무적으로 평가해야 한다.
- 리스크 완화 조치 의무: 평가된 위험을 완화하기 위한 구체적인 조치를 취해야 한다. 여기에는 이용자에게 딥페이크 등 조작된 콘텐츠를 식별할 수 있는 수단을 제공하는 의무가 포함된다.

나. 미국

미국은 연방 차원에서 포괄적인 단일 법안을 마련하기보다는, 각 이슈에 대응하는 개별 법안을 발의하는 동시에 주(State) 단위에서 독자적인 입법을 활발히 추진하는 이원적 규제 접근법을 취하고 있다.

1) 연방 수준

(1) AI에 대한 면책특권 폐지 법안(S. 1993)

「AI에 대한 면책특권 폐지 법안」(No Section 230 Immunity for AI Act)은 생성형 AI 확산에 대응하여 미국의 「통신품위법」(Communications Decency Act, CDA) 제230조(Section 230)에 해당하는 플랫폼 책임 면제 구조를 재조정하려는 법안이다.[8] 1996년 인터넷 등장 초창기에 제정된 통신품위법은 인터넷상의 유해한 콘텐츠로부터 이용자를 보호하기 위한 목적으로 만들어진 법안이다. 그중에서도 제230조는 인터넷 플랫폼이 제3자(이용자)가 게시한 콘텐츠에 대해 법적 책임을 면제받도록 허용하는 면책 규정으로 기능해 왔다. 본 법

지털서비스법(DSA)을 중심으로", 「미국헌법연구」, 33권 3호, 177-214, 2022.

8) U.S. Congress, *No Section 230 Immunity for AI Act*(S.1993), 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/house-bill/3831>(최종방문일 2025. 12. 19).

안은 해당 면책 범위를 AI 생성 콘텐츠에 적용하지 않도록 명확히 제한한다.

구체적으로 법 조문을 수정하여 AI 시스템이 자동으로 생성한 콘텐츠는 제3자가 만든 것처럼 간주하지 않는다는 점을 명문화하였다. 이에 따라 AI가 자율적으로 생성한 텍스트, 이미지, 영상 등의 콘텐츠는 더 이상 플랫폼이 책임을 회피할 수 있는 제3자 콘텐츠 범주에 포함되지 않는다. 이는 딥페이크 등으로 인한 명예훼손, 허위정보 유포 등의 AI 생성 콘텐츠로 인해 발생하는 피해에 대해 플랫폼의 책임 회피를 방지하려는 목적을 갖는다.

(2) 딥페이크 책임법(H.R.5586)

「딥페이크 책임법」(DEEPFAKES Accountability Act)은 2019년 최초로 발의된 이후, 2023년 동일한 명칭으로 재도입되어 미국 연방 하원에서 논의되었다.[9] 법안의 핵심 내용은 딥페이크를 포함한 합성 미디어 콘텐츠에 대해 디지털 워터마크와 텍스트 설명 등의 식별 조치를 의무화함으로써, 이용자와 사회가 해당 콘텐츠의 인위적인 생성 여부를 인지할 수 있도록 하는 데 있다. 특히 이 법안은 미국 연방 형법을 개정하여, 악의적으로 식별 조치를 누락하거나 변조할 경우 형사 처벌까지 가능하게 하는 강력한 규제의 성격을 띤다. 이는 딥페이크로 인한 허위정보 확산과 명예훼손 문제를 기술적인 식별 장치를 통해 완화하려는 접근으로 볼 수 있다.

(3) 딥페이크 태스크포스법(S. 2559)

「딥페이크 태스크포스법」(Deepfake Task Force Act)은 2021년 7월 29일 미국 제117대 연방 상원에서 발의되었으며, 딥페이크 탐지와 대응을 전담하는 정부 차원의 조직을 설립하는 데 목적을 둔다.[10] 구체적으로 국토안보부(Department of Homeland Security, DHS) 장관 주도하에 과학기술정책실(Office of Science and Technology Policy, OSTP) 등과 협력하여 국가 딥페이크 및 디지털 출처 태스크포스(National Deepfake and Digital Provenance Task Force)를 설립하는 것이다. 이 법안은 딥페이크 문제를 국가 안보 차원의 위협으로 규정하고, 정부 외에 민간·학계 전문가가 참여하는 협의체를 마련하여 디지털 콘텐츠의

9) U.S. Congress, *DEEPFAKES Accountability Act*(H.R.5586), 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/house-bill/5586/text>(최종방문일 2025. 12. 19).

10) U.S. Congress, *Deepfake Task Force Act*(S.2559). 117th Congress, 2021-2022. <https://www.congress.gov/bill/117th-congress/senate-bill/2559>(최종방문일 2025. 12. 19).

원본 여부를 입증할 수 있는 기술적 표준과 탐지 기술을 체계적으로 마련하려는 목적을 둔다.

2) 주 차원

(1) 캘리포니아주

① Depiction of individual using digital or electronic technology: sexually explicit material: cause of action(AB 602)(2019)[11]

개인의 동의 없이 성적으로 노골적인 딥페이크 콘텐츠를 제작·유포하는 행위로 인해 발생하는 피해에 대응하기 위한 법안이다. 이 법안은 피해자가 가해자를 상대로 민사 소송을 제기할 수 있도록 함으로써, 딥페이크 음란물로 인한 인격권 및 성적 자기결정권 침해에 대한 사법적 구제 수단을 강화하였다.

② Political Reform Act of 1974: political advertisements: artificial intelligence(AB 2355) & Elections: deceptive media in advertisements(AB 2839)(2024)[12]

기존 AB 730(2019)을 강화하여 제정된 법안들로, 선거 홍보물 및 선거 관련 표현에서 AI 또는 디지털 기술을 이용해 합성·조작된 콘텐츠가 사용된 경우, 그 사실을 명확히 공개하도록 의무화하고, 선거 결과나 후보자에 대한 판단을 오도할 목적으로 제작·유포된 기만적인 미디어(deceptive media)의 배포를 제한한다. 이는 딥페이크 자체를 전면적으로 금지하기보다 AI 활용 여부에 대한 투명성을 확보함으로써 유권자의 판단 가능성을 제고하려는 접근이다.

---

11) California Legislature, *Sexually Explicit Digital Images*.(SB981), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB981>(최종방문일 2025. 12. 19).

12) California Legislature, *Political Reform Act of 1974: Political Advertisements: Artificial Intelligence*.(AB2355), 2023-2024. <https://legiscan.com/CA/text/AB2355/id/3008736> (최종방문일 2025. 12. 19).; California Legislature, *Elections: Deceptive Media in Advertisements*.(AB2839), 2023-2024. <https://legiscan.com/CA/text/AB2839/id/3021243> (최종방문일 2025. 12. 19).

③ Sexually explicit digital images(SB 981) & Crimes: distribution of intimate images(SB 926)[13]

SB 981은 딥페이크 기술을 이용한 성적으로 노골적인 디지털 신분 도용으로부터 개인을 보호하는 데 초점을 두며, 플랫폼 사업자에게 딥페이크 음란물에 대한 신고 및 대응 체계를 구축하도록 요구할 수 있으며 관련 콘텐츠를 발견 즉시 신속히 삭제할 의무를 부과한다. SB 926은 형법(Penal Code Section) 제647조를 개정하여 배포 여부와 관계없이 딥페이크 음란물을 만드는 행위 자체를 형사 처벌할 수 있도록 규정을 강화하였다.

(2) 텍사스주의 Relating to the creation of a criminal offense for fabricating a deceptive video with intent to influence the outcome of an election(SB 751)(2019)[14]

SB 751은 선거일 전 30일 이내에 선거 결과에 영향을 미치거나 후보자의 평판을 훼손할 의도로 딥페이크 영상을 제작·배포하는 행위를 금지한다. 딥페이크 기술을 이용한 선거 개입과 유권자 기만을 방지하는 데 목적을 둔 법안으로, 이를 위반할 경우 A급 경범죄의 형사 처벌 대상이 된다.

(3) 버지니아주의 Unlawful dissemination or sale of images of another; penalty(HB 2678)(2019)[15]

당사자의 동의 없이 실존 인물을 성적으로 조작한 딥페이크 콘텐츠를 배포하는 행위를 범죄로 규정하여, 딥페이크 음란물로 인한 프라이버시 및 성적 자기결정권 침해에 대응하였다.

---

13) California Legislature, *Crimes: Distribution of Intimate Images*.(S.B.926), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB926>(최종방문일 2025. 12. 19).

14) Texas Legislature, *Relating to the Creation of a the Criminal Offense of Possession, Promotion, or Production of Appearing to Depict a Child*.(S.B.20), 2025. <https://capitol.texas.gov/BillLookup/History.aspx?LegSess=89R&Bill=SB20>(최종방문일 2025. 12. 19).

15) Code of Virginia, § 18.2-386.2: Unlawful Dissemination or Sale of Images of Another; penalty.(H.B.2678), 2019. <https://law.lis.virginia.gov/vacode/title18.2/chapter8/section18.2-386.2/>.(최종방문일 2025. 12. 19).

다. 영국

영국은 기존 법률 체계를 개정하고 새로운 법을 제정하는 방식을 병행하여 딥페이크 및 동의 없는 은밀한 이미지 남용에 대한 법적 대응을 강화하고자 하였다.

1) 온라인 안전법(2023)

영국의 「온라인 안전법」(Online Safety Act 2023, 이하 'OSA')은 온라인 플랫폼에 콘텐츠 관리에 대한 강력한 법적 책임을 부과하는 포괄적 온라인 안전 규제 법률이다.[16] 이 법에 따라 소셜미디어 및 인터넷 플랫폼은 자사 서비스에서 발생할 수 있는 불법 및 유해 콘텐츠를 관리하고, 이용자 피해 위험에 대한 사전적 리스크 평가를 수행하고, 불법 콘텐츠의 신속한 삭제, 이용자 보호 조치, 불만 처리 메커니즘을 구축할 의무를 가진다. 「OSA」는 원치 않는 성적 이미지를 전송하는 '사이버 플래싱(cyberflashing)'과 동의 없는 친밀한 이미지(intimate image)나 영상 공유 행위 등을 신종 범죄로 명시적으로 규정하였다. 감독기관인 영국 통신 규제기관 오프콤(Ofcom)은 플랫폼의 리스크 평가 관리 등을 감독하고 투명성 보고서 제출을 요구하는 등 강력한 감독 권한을 가진다. 해당 법을 위반할 시, 전 세계 매출의 최대 10%에 해당하는 과징금이 부과되며 서비스 접근이 제한되는 등 강력한 집행 권한을 행사할 수 있다.

2) 성범죄법 개정(2024)

2024년 영국의 「성범죄법」(Sexual Offense Act 2003)은 친밀한 이미지 및 영상의 비동의 유포를 별도로 처벌하는 조항을 신설하였다.[17] 해당 개정에서 도입된 신설 조항은 동의 없이 개인의 '친밀한 정보(intimate information)'*를 유포하는 행위를 형사 범죄로 규정한다.

* 성적 행위, 성적으로 보이는 행위, 노출된 신체의 전부 또는 일부 등을 포함하며, 여기에 컴퓨터 그래픽 등으로 실제처럼 만들어지거나 변경된 정보(딥페이크)를 명시적으로 포함

16) Online Safety Act 2023, c. 50. <https://www.legislation.gov.uk/ukpga/2023/50/contents> (최종방문일 2025. 12. 30).

17) Sexual Offences Act 2003, c. 42. <https://www.legislation.gov.uk/ukpga/2003/42/contents>(최종방문일 2025. 12. 19).

이 조항에 따르면 피해자의 동의가 없음에도 불구하고 해당 정보를 유포하는 행위, 피해자에게 고통이나 굴욕감을 주거나, 성적 만족을 얻을 목적으로 유포하는 행위, 친밀한 정보를 유포하겠다고 협박하는 행위는 최대 2년의 징역형 또는 벌금형에 처할 수 있다.

3) AI 사이버 보안 행동 강령(2025.1.31. 발표)

「AI 사이버 보안 행동 강령」(AI Cyber Security Code of Practice)은 2025년 1월 31일 발표된 비구속적 행동 강령으로, AI 시스템의 설계·개발·배포·운영 전 과정에서 발생할 수 있는 사이버 보안 위협에 대응하기 위한 기준을 제시한다.[18] 본 강령은 AI 공급망 전반의 이해관계자들에게 공통의 보안 원칙과 대응 방향을 제공하여 AI를 악용한 사이버 공격과 시스템 오용을 예방하는 것을 목표로 한다.

특히 이 강령은 '프롬프트 주입(prompt injection)'과 같이 AI 모델을 속여 의도하지 않은 유해한 결과물을 생성하게 하는 위협을 주요 관리 대상으로 명시하고, 데이터 생성·검증·관리, 시스템 모델링 및 운영, 콘텐츠 생성 전 단계에 이르는 전 주기적 대응책을 제시한다.

주요 권고 사항으로는 첫째 유해 콘텐츠 생성 가능성이 있는 경우 인간 감독을 통해 검토·승인 절차를 두는 방안, 둘째 접근 제어 및 속도 제한을 통해 악의적 콘텐츠 대량 생성이나 오용을 방지하는 조치, 셋째 금지된 사용 사례에 대한 능동적인 감지 및 로깅 체계 구축, 넷째 출력물의 오용 추적을 위한 워터마킹, 다섯째 AI 시스템의 오용 가능성을 사전에 식별하는 위협 모델링, 마지막으로 훈련 데이터와 입력 데이터 점검을 통한 데이터 오염 공격을 방지해야 한다.

라. 중국

1) 법치사회건설 실시강요(2021~2025)[19]

중국 정부는 「법치사회건설 실시강요」(法治政府建设实施纲要)를 통해 딥페이크 등 신기

18) United Kingdom Government, *AI Cyber Security Code of Practice, Department for Science, Innovation and Technology*, 2025. <https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice>(최종방문일 2025. 12. 19).

19) 中共中央·国务院, 「法治政府建设实施纲要(2021-2025年)」, 2021. <https://www.moj.gov.cn/pub/sfbgw/zwxxgk/fdzdgknr/fdzdgknrghjh/202207/t20220722_460263.html>(최종방문일 2025. 12. 19).

술 응용에 대한 규범적 관리 체계를 완비하고, 인터넷 생태계 전반의 법적 질서를 확립하겠다는 국가 차원의 정책 목표를 제시하였다. 해당 문건은 불법·유해 정보를 직접 규제하는 개별 법령은 아니지만, 생성형 AI 관련 규제의 상위 근거로 기능한다.

2) 인터넷 정보서비스 심층합성 관리 규정(2022)[20]

「인터넷 정보서비스 심층합성 관리 규정」(互联网信息服务深度合成管理规定)은 2022년 11월 25일 공식 공표되었으며 2023년 1월 10일부터 시행되었다. 해당 규정은 국가 안보나 타인의 권익을 침해하는 딥페이크(심층합성) 활용을 금지하고, 딥페이크 기술을 활용하는 플랫폼 사업자 및 서비스 제공자에게 관리 책임을 부과한다. 본 규정은 딥페이크 콘텐츠에 대해 식별 표시 의무를 부과하고, 신고 시스템 구축, 이용자 신원 확인 의무, 로그 보관 등의 초지를 요구한다. 이를 통해 딥페이크로 인한 허위정보 확산과 권리 침해를 예방하고자 한다.

3) 생성형 AI 서비스 관리 잠정조치(2023)[21]

2023년 8월 15일부터 시행된 「생성형 AI 서비스 관리 잠정조치」(生成式人工智能服务管理暂行办法)는 중국 내 대중에게 생성형 AI 서비스를 제공하는 모든 기관과 개인을 대상으로, AI 생성 콘텐츠에 대한 표시 의무와 안전 관리 책임을 부과하는 규정이다. 본 규정은 AI가 생성한 텍스트, 이미지, 영상, 오디오 등의 콘텐츠에 대해 AI 생성물임을 명시적 또는 암시적으로 표시하도록 요구하며, 구체적 표시 방식은 딥페이크 관리 규정의 표시 체계를 준용한다.

- 명시적 표기: AI로 생성된 콘텐츠임을 일반인이 알아볼 수 있도록 눈에 띄는 위치에 표시(예: 영상 시작화면 알림 문구, 이미지 워터마크, 오디오 음성 안내 등)
- 암시적 표기: 콘텐츠 파일의 메타데이터에서 AI 생성 관련 기술 정보를 기록(예: 콘텐츠 속성, 서비스 제공자 이름/번호, 콘텐츠 번호 등 포함)

---

20) 한국인터넷진흥원, 「인터넷 정보서비스 알고리즘 추천관리규정」, 2022.

21) 国家互联网信息办公室. 「生成式人工智能服务管理暂行办法」, 2023. 7. 13. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm(최종방문일 2025. 12. 19).

또한 국가 안보를 해치거나 사회주의 가치를 훼손하는 불법 콘텐츠 생성을 금지하고, 콘텐츠에 포함된 AI 생성 표기를 고의로 삭제·변조·위조하는 행위를 엄격히 금지한다. 서비스 제공자는 AI 콘텐츠 안전 및 관리 책임의 주체로서, 콘텐츠의 위험 평가를 수행해야 하는 책임이 부여되며 이용자 역시 불법적 콘텐츠 생성과 확산 금지의 책임이 있다.

4) AI 안전 거버넌스 프레임워크 2.0(AI Safety Governance Framework 2.0)(2025.9.15. 발표)[22]

중국 정부는 2025년 9월 15일 급변하는 AI 기술 환경에 대응하고 자국 산업 발전을 촉진하기 위해 「AI 안전 거버넌스 프레임워크 2.0」(AI Safety Governance Framework 2.0)을 발표하였다. 이는 2024년 9월에 발표된 프레임워크를 고도화한 것으로, 대규모 추론 모델 및 체현형 AI(Embodied AI) 등 최신 기술의 등장에 따른 새로운 위험 요인을 반영하였다. 해당 프레임워크는 AI 기술과 산업의 발전을 강조하는 한편, 안전을 우선시한다는 목표를 가지고 기술의 발전과 이로 인한 위험을 관리하는 것의 균형을 추구하고자 하였다. 더불어 중국 내 AI 안전에 대한 합의 이상으로 글로벌 거버넌스 차원의 규범 형성을 주도하려는 전략적 목표가 내포되어 있다.

프레임워크 2.0은 AI 안전 위험을 발생 원인과 파급 효과에 따라 내재적 위험(intrinsic risks), 응용 위험(application risks), 파생 위험(derived risks)으로 세분화하여 정의한다.

- 내재적 위험: 설명 가능성 부족, 편향성, 환각 현상 등 AI 모델 및 알고리즘 자체의 기술적 결함에서 비롯되며 데이터 오염 문제를 포함한다.
- 응용 위험: AI 활용 과정에서 발생하는 사이버 보안 위협, 허위정보 확산, 범죄 악용 등을 다루며 특히 허위정보나 딥페이크를 통한 여론 조작 및 인지전(cognitive warfare) 지원 가능성을 인지적 위험(cognitive risks)으로 별도 규정하여 강조하였다.
- 파생 위험: AI 도입으로 인한 고용 구조의 변화, 환경 영향, 인간의 AI 과몰입과 중독 등 장기적이고 사회적인 차원의 부작용을 포괄한다.

22) National Technical Committee 260 on Cybersecurity of SAC & National Computer Network Emergency Response Technical Team/Coordination Center of China, *AI Safety Governance Framework 2.0*, 2025.

AI 안전 위험에 대응하기 위해 프레임워크는 기술적 조치와 거버넌스 조치를 결합한 균형적 접근을 제시하고 있다. 기술적 조치로는 적대적 훈련(Red Teaming)을 통한 모델의 강건성을 확보하고, AI 생성 콘텐츠에 대한 식별 라벨링 및 워터마크 표시를 의무화해야 하며, 딥페이크 탐지 기술을 개발해야 한다는 내용을 구체적으로 명시하였다. 더불어 거버넌스 조치에서는 AI 위험에 대응하기 위한 법제도적·사회적 조치가 필요하며 UN, G20, APEC 등 글로벌 기관과 협력하여 AI 안전 거버넌스 네트워크를 구축해야 한다는 점을 강조하고 있다. 프레임워크는 AI 안전 위험별 기술적 조치와 거버넌스 조치를 통해 AI 위험 R&D 단계·배치·운영·이용 등 AI의 전 생애주기에 걸친 단계별 안전 지침을 수립하여 실행력을 강화하고자 하였다.

5) 인공지능 의인화 상호작용 서비스 관리 잠정조치안(2025)

2025년 12월 27일 중국 국가인터넷정보판공실(国家互联网信息办公室, Cyberspace Administration of China, 이하 'CAC')은 "인공지능 의인화 상호작용 서비스 관리 잠정조치(의견수렴안)(人工智能拟人化互动服务管理暂行办法, Interim Measures for the Administration of Humanized Interactive Services Based on Artificial Intelligence(Draft for Public Comment)) 초안"(이하 '잠정조치안')을 발표하고[23] 2026년 1월 25일까지 의견수렴을 진행할 예정이다.[24] 잠정조치안은 AI 챗봇, 가상 동반자(AI companion), 감정 교류 AI 등 인간의 인격적 특성을 모방하여 정서적 상호작용을 제공하는 서비스를 규율 대상으로 한다. 특히 콘텐츠 안전을 넘어 이용자의 정서적 안전(emotional safety)까지 보호 범위를 확장한 세계 최초의 입법 시도라는 점에서 전 세계적으로 주목받고 있다. 잠정조치안은 현재 의견수렴 단계에 있으며, 향후 의견 반영을 거쳐 확정·시행될 예정이다.

주지하는 바와 같이, 최근 다중모달 AI 기술의 급속한 발전으로 AI 동반자 서비스가 빠르게 성장하고 있다. Character.AI, Replika 등 가상 캐릭터와 감정을 교류하며 대화를 나

23) 国家互联网信息办公室,「关于《人工智能拟人化互动服务管理暂行办法（征求意见稿）》公开征求意见的通知」, 2025. 12. 27., <https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm>(최종방문일 2025. 12. 30).

24) 国家互联网信息办公室,「关于《人工智能拟人化互动服务管理暂行办法（征求意见稿）》公开征求意见的通知」, 2025. 12. 27., <https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm>(최종방문일 2025. 12. 30).

누는 서비스들이 전 세계적으로 확산되고 있으며, 중국에서도 Minimax의 Talkie AI(해외판) 및 星野(국내판) 등이 2024년 1~3분기 평균 월간 활성 이용자 2천만 명 이상을 기록하는 등 급성장하고 있다.[25] 한편, '기계적 대화'에서 '깊은 공감'으로의 전환을 유도하는 이러한 유형의 AI 서비스는 심각한 위험성 또한 내포하고 있다. 2024년 미국에서는 캐릭터.AI와 장시간 대화를 나누던 14세 청소년이 자살한 사건이 발생하여 소송으로 이어졌고, 이 밖에도 AI 챗봇과의 대화 후 자해를 시도한 사례들이 다수 보고되면서 AI가 감정과 정서에 미치는 영향력에 대한 우려가 증폭되고 있다.[26]

중국 정부의 잠정조치안 발표는 Minimax,(Z.ai(智谱) 등 주요 AI 동반자 챗봇 스타트업들이 홍콩 증시 상장을 추진하는 시점에 이루어졌다는 점에서 그 시의성이 더욱 부각되었다.[27] 이는 AI 서비스 산업의 급성장과 이용자 보호 규제의 필요성이 동시에 대두되는 국면임을 보여준다. 잠정조치안의 의의는 무엇보다 "콘텐츠의 안전에서 정서적 안전으로의 도약(a leap from content safety to emotional safety)"이라는 표현으로 집약된다.[28] 이용자

---

25) Cheng, E., "China to crack down on AI chatbots around suicide, gambling", *CNBC*, 2025. 12. 29, <https://www.cnbc.com/2025/12/29/china-ai-chatbot-rules-emotional-influence-suicide-gambling-zai-minimax-talkie-xingye-zhipu.html>(최종방문일 2025. 12. 30). Minimax의 Talkie AI 앱은 이용자가 가상 캐릭터와 대화할 수 있는 서비스로, 국내 버전인 星野와 함께 2024년 1~3분기 회사 매출의 3분의 1 이상을 차지하였다.

26) Dupre, M. H., "China Planning Crackdown on AI That Harms Mental Health of Users," Futurism, 2025. 12. 31, <https://futurism.com/artificial-intelligence/china-regulation-ai-chatbots>(최종방문일 2025. 12. 30).

27) Cheng, E., "China to crack down on AI chatbots around suicide, gambling", *CNBC*, 2025. 12. 29, <https://www.cnbc.com/2025/12/29/china-ai-chatbot-rules-emotional-influence-suicide-gambling-zai-minimax-talkie-xingye-zhipu.html>(최종방문일 2025. 12. 30). 잠정조치안 발표 직전인 2025년 12월, Minimax와 Z.ai는 홍콩 증시 기업공개(IPO)를 신청한 바 있다.

28) Cheng, E., "China to crack down on AI chatbots around suicide, gambling", *CNBC*, 2025. 12. 29, <https://www.cnbc.com/2025/12/29/china-ai-chatbot-rules-emotional-influence-suicide-gambling-zai-minimax-talkie-xingye-zhipu.html>(최종방문일 2025. 12. 30). Winston Ma(NYU 법학대학원 겸임교수)의 평가.

보호의 관점에서 기존의 AI 규제가 허위정보 유포, 유해 콘텐츠 생성 방지 등 정보의 내용적 측면에 집중했다면, 잠정조치안은 AI가 인간의 심리, 정서, 감정에 미치는 영향, 이용자의 심리적 의존 및 중독, 나아가 감정의 조작 가능성까지 규율 대상으로 삼고 있다. 아울러 규제의 수범자 범위를 서비스 제공자에서 앱스토어 등 배포 플랫폼까지 확대하고 있다는 점도 특징적이다.

잠정조치안은 중국의 기존 AI 규제 체계 —「생성형 인공지능 서비스 관리 잠정방법」,「인터넷 정보서비스 알고리즘 추천 관리규정」,「인터넷 정보서비스 심층합성 관리규정」등— 와 유기적으로 연결되면서도, 의인화 상호작용 서비스의 특수한 위험성에 대응하기 위한 제도적 혁신을 담고 있다.[29] 특히 이용자 상태 식별, 위기 관리 의무, 미성년자·고령자 등 취약계층 보호, 이용시간 제한 등 구체적인 보호 수단을 규정함으로써, 우리나라의 AI 서비스 이용자 보호에 관한 입법 과제에 중요한 참고 사례를 제공하는바, 절을 바꾸어 EU의「DSA」규정과 함께 심층 분석 사례로 삼기로 한다.

마. 호주의 온라인 안전법(2021)

호주는「온라인 안전법」(Online Safety Act)을 중심으로 딥페이크 콘텐츠에 대응하고 있다.[30] 2024년 형법 개정을 통해 동의 없이 딥페이크 음란물을 제작·유포하는 행위를 형사처벌 대상으로 규정하였다. 호주의 온라인 안전국(eSafety Commissioner)은 딥페이크 콘텐츠가 사이버폭력이나 혐오스러운 폭력 행위에 해당한다고 판단될 경우, 플랫폼 사업자에게 해당 콘텐츠의 삭제 및 차단을 명령할 수 있다. 특히 사이버폭력 정보로 신고된 콘텐츠에 대해서는 48시간 내 삭제 조치 의무가 부과되며, 불이행 시 감독기관이 직접 삭제 통지를 발부할 수 있다. 아동·청소년이 피해자인 경우, 더욱 신속하고 강화된 보호 조치가 적용된다.

---

29) Zhang, L.(李强治), "Experts interpret the 'Interim Measures for the Administration of Human-like Interactive Services Based on Artificial Intelligence'", *Sina Finance(新浪财经)*, 2025. 12. 28, <https://finance.sina.com.cn/jjxw/2025-12-28/doc-inheirmq8851756.shtml>(최종방문일 2025. 12. 30).

30) Office of the eSafety Commissioner, "Industry Regulation". <https://www.esafety.gov.au/A.B.out-us/industry-regulation>(최종방문일 2025. 12. 19).

바. 캐나다의 온라인 유해 규제법(C-63)

캐나다의 「온라인 유해 규제법」(Online Harms Act)은 2024년 2월 온라인 유해 콘텐츠를 규제하고 아동 및 일반 이용자에 대한 보호를 강화하려는 목적으로 발의되었다.[31] 여기서 말하는 '온라인 유해 콘텐츠'에는 아동 성착취 및 피해자에 대한 2차 가해 콘텐츠, 비동의적 성적 이미지·영상 콘텐츠, 아동이 스스로 해를 끼치도록 유도하는 콘텐츠, 증오 조장 콘텐츠, 아동 괴롭힘 콘텐츠, 폭력·테러 선동 콘텐츠가 포함된다. 해당 법안은 인터넷 서비스 제공자와 온라인 플랫폼 사업자에게 유해 콘텐츠의 확산을 방지할 구조적 책임을 부여하며 구체적인 의무는 다음과 같다.

- 알고리즘 및 큐레이션 투명성: AI 기반 추천 알고리즘이나 큐레이션 시스템이 유해 콘텐츠의 노출과 확산에 미칠 수 있는 위험을 관리·완화하기 위한 정책과 절차를 감독기관에 설명하고 제출해야 한다.
- AI 생성 유해 콘텐츠 식별 및 표시: 딥페이크(특히 비동의 성적 합성물) 및 자동화된 봇(Bot) 계정이 생성한 콘텐츠를 이용자가 명확히 식별할 수 있도록 적절한 라벨링 및 경고 표시를 의무화한다.
- 신고 및 대응체계: 이용자가 유해한 AI 콘텐츠를 쉽고 빠르게 신고할 수 있는 매커니즘을 구축해야 하며, 신고 접수 시 신속하게 검토하고 삭제 등의 조치를 취해야 한다. (예: 성착취물의 경우 24시간 내 삭제)
- 디지털 안전 위원회 신설: 독립 규제기관인 디지털 안전 위원회(Digital Safety Commission) 등을 설립하여 AI 서비스를 포함한 온라인 서비스의 규제 준수 여부를 감독하고 위반 시 제재를 부과할 권한을 갖는다.

---

31) Parliament of Canada, *An Act to Enact the Online Harms Act, to Amend the Criminal Code, the Canadian Human Rights Act and An Act Respecting the Mandatory Reporting of Internet Child Pornography by Persons Who Provide an Internet Service*.(Bill C-63), 2024.
<https://www.parl.ca/DocumentViewer/en/44-1/bill/C-63/first-reading>(최종방문일 2025. 12. 19).

2. 이용자의 데이터 관련 자기결정권 보호

가. EU

1) 일반 데이터 보호 규정(GDPR)

EU의 「일반 데이터 보호 규정」(General Data Protection Regulation, 이하 'GDPR')은 개인의 개인정보 보호 강화와 데이터 수집 및 처리 과정에서의 책임 규정을 두어 데이터 주체의 권리를 보호하는 개인정보 보호 법안이다.[32] AI 시스템의 데이터 처리 과정 전반에 적용되며, 정보 주체에게 다음과 같은 핵심적인 권리를 부여한다.

- 프로파일링 거부권(제22조): 정보 주체는 자신에 대한 법적 또는 중대한 영향을 자동화된 의사결정 및 프로파일링을 거부할 권리를 가진다.
- 설명 요구권(제13조): 정보 주체는 AI의 자동화된 결정에 적용된 주요 논리와 그 결정이 자신에게 미칠 영향에 대한 설명을 요구할 권리가 있다.
- 데이터 이동권(제20조): 정보 주체는 자신의 개인 데이터를 수령할 권리와 다른 서비스 제공자에게 이전할 권리를 가진다.
- 처리 제한권(제18조): 정보 주체는 데이터의 정확성이나 처리의 적법성에 이의를 제기할 경우, 해당 문제가 해결될 때까지 데이터 처리를 제한하도록 요청할 수 있다.

나. 미국

미국은 연방 차원의 포괄적인 개인정보보호법 대신, 각 주(State)가 독자적인 입법을 통해 데이터 보호 규제를 시도하고 있다.

1) 캘리포니아주

(1) 초상권 사용(Use of likeness: digital replica)(AB 1836)[33]

영리적 목적으로 고인의 이름·음성·이미지·외모 등을 동의 없이 디지털로 복제하는 행

32) Regulation(EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016. 5. 4., p. 1-88.

33) California Legislature, *Use of Likeness: Digital Replica*.(A.B.1836), 2023-2024. <https://legiscan.com/CA/text/A.B.1836/id/3021237>(최종방문일 2025. 12. 19).

위를 금지하고, 위반 시 유족이 손해배상을 청구할 수 있는 근거를 마련하였다.

(2) 서비스 수행 관련 계약(Contracts against public policy: personal or professional services: digital replica)(AB 2602)[34]

미디어 산업에서 배우의 디지털 복제물을 사용할 때, 배우의 초상권을 보호하기 위해 명시적인 동의 요건을 강화하였다.

2) 코네티컷주의 소비자 AI 보호법(SB 2)

2026년 2월 1일 발효 예정인 「소비자 AI 보호법(An Act Concerning Artificial Intelligence)」은 소비자 보호를 중심으로 고위험 AI(고용, 주거, 의료, 신용 등 소비자에게 중대한 영향을 미칠 수 있는 영역에서 사용되는 AI 시스템)를 규제한다.[35] 특히 고위험 AI 시스템을 개발하거나 배포하는 기업에 대해 알고리즘 차별을 방지하기 위한 합리적인 주의(reasonable care) 의무를 부과하고 있다.

이 법은 AI 생태계 가치사슬 전반에 걸쳐 책임을 분담시키는 것을 목표로 하여 개발자(developer), 제공자(deployer), 통합자(integrator) 각각에 대한 책임과 소비자 권리 보호를 규정하고 있다. 개발자는 AI 모델의 한계 및 위험 평가 문서를 제공해야 하며, 제공자와 통합자는 자체적인 리스크 관리 정책을 수립하고 영향 평가를 수행해야 한다.

다. 영국

1) 개인정보보호법(2018)

영국은 EU 탈퇴 이후 독자적인 데이터 보호 체계를 확립하고 있다. 「개인정보보호법」(Data Protection Act)은 개인에게 데이터 통제권을 부여하고 기업의 합법적인 개인정보 처리 과정을 지원하는 것을 목표로 하며,[36] 개인의 프라이버시 권리를 강화하는 동시에 디지

34) California Legislature, *Contracts Against Public Policy: Personal or Professional Services: Digital Replicas*.(A.B.2602), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240A.B.2602> (최종방문일 2025. 12. 19).

35) Connecticut General Assembly, *An Act Concerning Artifical Intelligence*.(S.B.2), 2024. <https://www.cga.ct.gov/asp/cgA.B.illstatus/cgA.B.illstatus.asp?selBillType=Bill&bill_num=S.B.00002&which_year=2024>(최종방문일 2025. 12. 19).

36) Data Protection Act 2018, c. 12.

털 시대에 맞는 데이터 보호 체계를 마련하여 기업과 조직의 책임을 강화하고자 한다. 인간의 개입 없는 자동화된 의사결정을 제한하고, 데이터 접근권, 수정권, 삭제권, 이동권, 이의제기권을 명시한다. 이 법은 다음의 7가지 핵심 원칙을 기반으로 한다.

1. 개인정보는 적법하고 공정하며 투명하게 처리되어야 한다.
2. 개인정보는 명시된 목적에 한정하여 수집되어야 한다.
3. 필요한 최소한의 데이터만 수집하고 처리해야 한다.
4. 개인정보는 정확성을 유지하고 최신 상태로 관리되어야 한다.
5. 개인정보는 목적이 달성된 후 더 이상 보관되지 않도록 보관 기간이 제한되어야 한다.
6. 개인정보의 무결성과 기밀성을 보장하기 위해 적절한 보안 조치가 요구된다.
7. 개인정보 처리자는 책임성을 가지고 개인정보 보호와 관련된 의무를 충실히 이행해야 한다.

해당 법은 영국 내 모든 조직과 기업뿐만 아니라 영국 거주자의 개인정보를 처리하는 해외 조직, 공공기관 및 정부 기관, 자선단체 및 비영리 조직, 개인사업자와 프리랜서에도 포괄적으로 적용된다. 위반 시 최대 1,800만 파운드 또는 전 세계 연간 매출의 4% 중 더 높은 금액의 벌금이 부과될 수 있다.

2) AI와 관련된 데이터 보호 가이드라인

'AI와 관련된 데이터 보호 가이드라인(Guidance on AI and Data Protection)'은 영국 정부 위원회(ICO)가 2020년 7월 30일에 발행한 가이드라인으로,[37] AI 시스템에서 개인정보를 처리할 때 데이터보호법을 준수하고 개인의 권리를 보호할 수 있는 방법을 제시한다. AI 시스템의 생애주기 전반에 걸쳐 데이터 보호 규정을 준수할 것을 요구하며, 특히 책임성, 투명성, 적법성, 정확성, 개인 권리 보장의 5가지 핵심 요소를 강조한다.

• 책임성과 거버넌스: AI 시스템 도입 시 데이터 보호 영향 평가(DPIA)를 수행해야 하며,

<https://www.legislation.gov.uk/ukpga/2018/12/contents>(최종방문일 2025. 12. 19).

37) Information Commissioner's Office(ICO), *Guidance on AI and Data Protection*, 2023. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/>(최종방문일 2025. 12. 19).

평가 시 덜 위험한 대안을 고려한 증거와 선택의 이유를 포함해야 한다.

- 투명성: AI 시스템을 사용할 때 개인에게 적절한 정보가 제공되어야 하며, 데이터 수집 목적, 보유 기간, 공유 대상 등을 명시해야 한다.
- 적법성: AI를 통해 추론하거나 특별 범주의 데이터를 처리할 때 UK GDPR 제9조를 준수해야 한다.
- 정확성: 통계적 정확성과 공정성을 고려하고, 필요 이상의 데이터를 보유하지 않도록 주의해야 한다.
- 개인의 권리 보장: 정정권, 삭제권, 데이터 이동권 등을 포함한다.

3) AI 프로젝트를 위한 데이터 보호 영향 평가 가이드(2020)

'AI 프로젝트를 위한 데이터 보호 영향 평가 가이드'(Guide to Data Protection Impact Assessment(DPIA) for AI Projects)는 AI 시스템이 개인의 권리와 자유에 미치는 위험을 사전에 식별하고 최소화하기 위한 도구로[38], 특히 대규모 민감 데이터를 처리하거나 혁신 기술을 사용할 경우 데이터 보호 영향 평가(DPIA) 수행이 의무화된다.

- 처리 활동 설명 단계: 처리할 데이터의 성격, 출처, 양, 민감도를 기술하고 데이터 수집, 저장, 이용 방법을 설명하며, AI 시스템이 개인에게 미치는 영향을 명시
- 필요성과 비례성 평가 단계: AI 사용의 필요성을 정당화하고, 덜 침해적인 대안을 고려하며, AI 사용의 이점과 개인의 권리 간 균형을 평가
- 위험 식별 및 평가 단계: 차별, 재정적 손실, 평판 손상 등 개인의 권리에 대한 잠재적 위험을 분석하고, 각 위험의 발생 가능성과 심각성을 평가
- 위험 완화 조치 식별 단계: 데이터 익명화 같은 기술적 조치와 직원 교육 등 조직적 조치를 포함하여, 식별된 위험을 완화하기 위한 구체적인 방안을 수립

38) Information Commissioner's Office(ICO), *Guidance on AI and Data Protection*, 2023. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/>(최종방문일 2025. 12. 19).

라. 중국

1) 생성형 인공지능 서비스 관리 잠정조치(2023)

동 잠정조치는 법적 구속력을 갖는 것으로, 서비스 제공자에게 알고리즘 훈련 시 적법한 출처의 데이터와 기본 모델을 사용해야 하며, 지적재산권을 존중하고 개인정보 사용 시 반드시 동의를 구해야 한다는 의무를 부과하고 있다.[39] 더불어 서비스 제공자는 데이터의 진실성, 정확성, 객관성, 다양성을 강화하기 위한 조치를 취할 의무를 가진다.

2) 인터넷 정보서비스 알고리즘 추천 관리 규정(互联网信息服务算法推荐管理规定)(2022)

해당 규정은 2021년 12월 31일에 공포되어 2022년 3월 1일부터 시행되었다.[40] 알고리즘 추천 기술을 활용하는 인터넷 서비스의 투명성과 공정성을 강화하는 데에 목적을 둔다. 주요 내용으로 서비스 제공자가 알고리즘 추천의 주요 원칙과 매개변수를 이용자에게 공개하도록 의무화하고, 이용자가 알고리즘 추천을 거부하거나 비개인화된 추천을 선택할 권리를 보장하여 사용자의 선택권을 강화하였다.

3) 인공지능 안전 거버넌스 프레임워크(1.0)(2024)

중국의 <인공지능 안전 거버넌스 프레임워크(1.0)>(Artificial Intelligence Safety Governance Framework 1.0)에서는 AI 훈련 및 상호작용 데이터 처리 전 과정에서 데이터 보안 및 개인정보 보호 규칙을 준수할 것을 명시하고, 이용자의 알 권리와 선택권을 보장하는 것을 목표로 한다.

마. 호주의 아동·청소년 소셜미디어금지법(2024)

호주의 「아동·청소년 소셜미디어금지법(AU Social Media Minimum Age Bill 2024)」은 연령제한 이용자의 소셜미디어 계정 보유를 방지하기 위한 합리적인 조치 목적으로 하는 경우에 한해, 사업자가 연령 확인에 필요한 최소한의 개인정보만을 수집하도록 규정하고 있다.[41] 이때 플랫폼 사업자가 연령 확인을 목적으로 개인정보를 수집한 경우, 해당 정보는

39) 国家互联网信息办公室.「生成式人工智能服务管理暂行办法」, 2023. 7. 13. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm(최종방문일 2025. 12. 19).

40) 한국인터넷진흥원, 「인터넷 정보서비스 알고리즘 추천관리규정」, 2022.

41) 국회입법조사처, "호주의 '아동·청소년 소셜미디어금지법' 통과와 입법적 시사점", 「NARS 현안분석」, 345호, 2024.

오직 연령 확인 목적에 한해서만 이용되어야 하며 그 외의 어떠한 용도로도 활용될 수 없다. 또한 개인정보 처리에 대한 동의는 호주의 개인정보원칙을 충족해야 하며, 설령 개인의 동의를 얻더라도 그 동의가 자발적이고, 충분한 정보에 기반하며, 명확하게 표현되고, 언제든지 쉽게 철회 가능한 경우에 한해 유효한 것으로 인정된다. 이러한 요건을 충족하지 못하는 경우, 동의의 존재 여부와 무관하게 개인정보 처리는 허용되지 않는다. 더불어 연령 확인 목적이 달성된 후에는 수집된 개인정보를 즉시 폐기해야 할 의무가 부과된다.

바. 일본

1) 개인정보보호법(2020)

일본의 「개인정보보호법」(Act on the Protection of Personal Information, APPI)은 2020년 개정을 통해 데이터 활용 환경 변화에 대응하면서[42] 개인정보 처리에 대한 투명성과 사업자의 책임을 강화하였다. 법은 AI를 활용한 개인정보 처리 전반에 적용되는 기본 법체계로 활용된다. 특히 「개인정보보호법」은 개인정보 취급 사업자에게 개인정보의 유출, 분실, 훼손 등을 방지하기 위한 안전관리 조치를 강구할 의무를 명시하고 있으며, 개인정보 처리 과정 전반에서 정보주체의 통제 가능성을 보장한다. 주요 내용은 다음과 같다.

- 이용 목적 제한: 개인정보 취급 사업자가 정보 주체의 동의 없이 이용 목적 범위 이상으로 개인정보를 취급하거나 사전에 본인 동의 없이 개인정보를 제3자에게 제공하는 것을 금지한다.
- 정보 열람(개시) 권리: 정보 주체는 본인이 식별되는 보유 개인정보에 대해 개인정보 취급 사업자에게 전자문서 또는 규정된 방식으로 개시를 청구할 수 있다. 사업자는 원칙적으로 이용자가 요청한 방식에 따라 지체없이 개인정보 열람을 제공해야 한다.
- 정보 이용 정지 및 삭제 권리: 정보 주체는 본인이 식별되는 개인정보가 위법하게 취급될 경우, 이용 정지 또는 삭제를 청구할 수 있다.

42) Japan Personal Information Protection Commission(PPC), *Act on the Protection of Personal Information*(Act No. 57), 2023. <https://www.ppc.go.jp/en/legal/>(최종방문일 2025. 12. 19).

2) AI 사업자 가이드라인(2024)[43]

일본의 'AI 사업자 가이드라인'(AI Guidelines for Business)은 기존의 'AI 활용 가이드라인'과 'AI 개발 가이드라인' 등을 통합하여 2024년 새로 제정된 일본의 핵심 AI 규범이다. AI 시스템 제공자에게 개인정보 접근을 관리·제한하고, 프라이버시 침해를 방지하기 위한 기술적·관리적 장치를 마련할 의무를 부과한다.

사. 캐나다

1) 생체정보 가이드라인(2025)[44]

캐나다 연방 개인정보 감독기구(OPC)는 생체정보의 특수성과 높은 민감성을 고려하여 2025년 8월 11일 '생체정보 가이드라인'(Guidance for processing biometrics)을 발표하였다. 해당 가이드라인은 이용자 데이터 통제권을 보호하고 생체정보 처리의 투명성과 책임성을 강화하는 것을 목표로 하며, 연방 공공기관과 민간 기업 모두에 적용된다. 해당 법에서 생체인식은 개인의 고유한 특성을 측정가능한 데이터로 정량화해 신원 확인 등에 활용하는 기술로 정의된다. 생리적(Physiological) 생체인식이란 지문, 홍채 패턴, 얼굴 형태 및 DNA와 같이 비교적 변하지 않는 개인의 신체적·생물학적 특성을 활용한 기술이다. 행동(Behavioral) 생체인식은 키보드 입력 패턴, 걸음걸이, 음성 및 눈동자 움직임 등 개인의 고유한 행동 패턴이나 습관을 활용한 기술을 의미한다.

생체정보 처리는 적법한 목적을 전제해야 하며, 다음의 4가지 기준을 모두 충족해야 한다.

- 적당한 필요(Legitimate need): 단순한 편의가 아닌, 실질적 필요가 존재해야 한다.
- 효과성(Effectiveness): 해당 기술이 목적 달성에 효과적이라는 점이 입증되어야 한다.
- 최소 침해성(Minimal Impairment): 목적을 달성할 수 있는 다른 덜 침해적인 대안이 없는지 고려해야 한다.
- 비례성(Proportionally): 프라이버시 침해로 인한 손실보다 기술 도입으로 얻는 이익이

43) Kotra, "일본의 AI 정책과 실제 사례", 「Global Issue Monitoring」, 2024.

44) Office of the Privacy Commissioner of Canada(OPC), *Guidance for Processing Biometrics-for Businesses*, 2025. <https://www.priv.gc.ca/en/privacy-topics/health-genetic-and-other-body-information/biometrics/gd_bio_org-final/>(최종방문일 2025. 12. 19).

더 커야 한다.

더불어 데이터 통제권 강화를 위한 '7대 원칙'을 제시하고 있다.

1. 동의(Consent): 생체정보의 민감성을 고려하여 명시적이고 유효한 동의를 받아야 하며, 이때 동의는 일회성이 아니라 용도 변경 시 갱신이 필요한 지속적인 프로세스로 간주된다. 이용자가 생체정보 제공을 거부할 경우 불이익이 없도록 비(非)생체인식 대안을 제공해야 한다.
2. 수집 제한(Limiting collection): 목적 달성에 필요한 최소한의 생체정보만을 수집해야 한다.
3. 이용·공개·보관 제한(Limiting use, disclosure, and retention): 수집 당시 명시한 목적 외 이용(Function Creep)을 금지하고, 목적 달성 후에는 생체정보를 즉시 영구 삭제해야 합니다.
4. 보호 조치(Safeguards): 생체정보를 고도의 민감정보로 분류하여 최고 수준의 기술적·관리적 보안 조치를 적용해야 한다. 데이터 유출, 해킹 공격, 내부자 오남용과 같은 위협 시나리오에 대한 모니터링 및 대응 체계를 구축해야 한다.
5. 정확성(Accuracy): 인종, 성별 등에 따른 편향 가능성을 모니터링하고 알고리즘 정확도를 주기적으로 검증해야 한다.
6. 책임성(Accountability): 생체정보 처리를 감독할 내부 책임자를 지정하고, 거버넌스 체계를 수립해야 한다. 침해 사고 발생 시 감독기구(OPC)에 신고하고 피해자에게 통지하는 절차를 마련해야 한다.
7. 개방성(Openness): 수집 항목, 목적, 이용범위, 제3자 공유 여부 등을 투명하게 공개하고, 데이터 보관 장소 및 기간을 공개해야 한다. 생체인식 기술에 따른 잠재적 위험을 이용자가 알기 쉽게 설명해야 한다.

3. 아동 · 청소년 보호

가. EU

1) 「AI 법」(2024)[45]

EU의 「AI 법」은 특정 영역에서 아동·청소년에게 영향을 미치는 다음과 같은 AI 시스템을 '고위험 AI 시스템'으로 간주하여 엄격하게 규제하고 있다.

- 교육용 평가 AI: 교육기관의 입학 결정, 교육 경로 배치, 학습 성과 평가 등에 사용되는 경우 고위험 AI로 간주한다.
- 감정인식 AI: 교육 현장에서 학습자를 분류하거나 평가하는 데 사용되는 감정인식 및 행동 분석 기술을 사용할 경우 고위험 AI 시스템에 해당한다.
- 추천 알고리즘: 맞춤형 콘텐츠 추천 시스템과 같이 개인의 공공 서비스 접근 자격 평가에 사용될 경우 고위험 AI 시스템이다.

2) 디지털 서비스 법(DSA)(2023)[46]

EU의 「DSA」는 미성년자가 접근가능한 온라인 플랫폼 제공자가 미성년자의 프라이버시, 안전, 보안을 높은 수준으로 보장하도록 규정하고 있다. 2024년 2월부터 플랫폼에 연령확인 시스템을 포함하도록 하였고, 추천 알고리즘에 미성년자의 개인 데이터를 활용하는 것 또한 제한된다. 중개 서비스 제공자가 아동 성착취 범죄를 포함해 생명 또는 안전에 위협이 되는 범죄 혐의가 있는 불법 콘텐츠를 인지할 경우, 즉시 관계 당국에 통지할 의무가 있다.

나. 미국

1) 연방 차원

(1) 아동 온라인 안전법(S. 1409)(2024)

미국의 「아동 온라인 안전법」(Kids Online Safety Act, KOSA)은 2023년 5월 2일 발의되어, 2024년 7월 30일에 상원을 통과하였다.[47] 해당 법은 17세 미만 미성년자 보호를 위해

45) 딜로이트 인사이트, 「인공지능(AI) 시대의 새로운 국면, AI 규제와 기업 리스크 관리 전략」, 2024. 10.

46) 최경미·지성우, "디지털서비스 투명성·개인정보보호·위기대응에 관한 법 연구-EU 디지털서비스법(DSA)을 중심으로", 「미국헌법연구」, 33권 3호, 177-214, 2022.

플랫폼에 6대 유해요소*에 대한 '관리 의무(duty of care)'를 부과하며, 플랫폼은 아동을 보호하도록 서비스를 설계해야 한다.

* 정신건강 장애, 중독, 정신적·신체적 폭력 및 괴롭힘, 불법 약물 관련 홍보 및 마케팅, 성적 착취 및 학대, 약탈적 및 기만적 마케팅 또는 금전적 피해

(2) 아동 온라인 프라이버시 보호법(S. 1418)(2023)

「아동 온라인 프라이버시 보호법」(Children and Teens' Online Privacy Protection Act, COPPA 2.0)은 기존 1998년 제정된 법을 개정하여 보호 대상을 청소년까지 확대한 법안으로, 2023년 5월 3일 발의되어 2024년 7월 30일 상원을 통과하였다.[48] 주요 내용은 법 적용을 받는 미성년자 연령을 13세 미만에서 16세 미만으로 상향하고, 이들을 대상으로 한 맞춤형 광고를 금지한다. 또한, 아동·청소년이 직접 자신의 정보를 삭제할 수 있는 '삭제 버튼(Eraser Button)' 도입을 의무화하였다.

(3) TAKE IT DOWN Act(Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act)(S. 4569)(2023)[49]

해당 법은 2024년 6월 18일 미국 상원에 발의되었으며, AI로 생성된 딥페이크 음란물을 포함하여 동의 없이 제작·유포된 개인의 친밀한 이미지(Non-Consensual Intimate Imagery, NCII)를 연방법으로 범죄화하고 온라인에서 신속하게 제거하는 것을 목표로 한다. 2025년 4월 의회를 통과하여 5월 대통령 서명으로 발효된 이 법은 플랫폼이 피해자의 요청을 받으면 48시간 이내에 해당 콘텐츠를 삭제하도록 의무화하며, 연방거래위원회(FTC)가 법 집행을 감독한다. 이 법에 따라 플랫폼들은 2026년 5월 19일까지 NCII의 신고·삭제 절차를

47) U.S. Congress, *Kids Online Safety Act*(S.1409), 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/senate-bill/1409>(최종방문일 2025. 12. 19).

48) U.S. Congress. Children and Teens' *Online Privacy Protection Act*(S.1418), 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/senate-bill/1418>(최종방문일 2025. 12. 19).

49) U.S. Congress, *TAKE IT DOWN Act(Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act)*(S.4569). 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/senate-bill/4569>(최종방문일 2025. 12. 19).

마련해야 한다.

2) 주 단위

(1) 캘리포니아주

① 아동 음란물 관련 법(Crimes: Child Pornography)(AB 1831)(2023)[50]

캘리포니아주는 형법 제311조를 개정하여 기존 형법상 아동 음란물(child pornography)의 정의를 확대해 AI 또는 컴퓨터 생성 기술을 이용해 제작된 아동 성착취물을 명시적으로 규제 대상에 포함하였다. 해당 개정은 실제 아동이 촬영에 동원되지 않았더라도, 실존 아동과 구별하기 어려울 정도로 사실적으로 묘사된 합성 이미지·영상 또한 아동 음란물로 간주하여 제작·유포 행위에 대한 형사 처벌 근거를 마련하였다.

② 연령적합설계법(California Age-Appropriate Design Code Act, CAADCA)(AB 2273)(2021)[51]

해당 법안은 18세 미만 아동이 접속할 가능성이 있는 온라인 서비스 제공자를 대상으로 서비스 설계 단계에서부터 아동의 개인정보와 안전을 우선적으로 고려하도록 의무를 부과하는 법률이다. 본 법은 아동 보호를 온라인 서비스의 사후 규제가 아닌 설계 단계의 책임으로 전환한 것이 특징이다. 온라인 서비스 제공자에게 다음과 같은 의무를 부과한다.

- 데이터 보호 영향 평가(DPIA) 의무화: 새로운 기능을 출시하기 전, 데이터 보호 영향 평가를 실시하여 아동에게 미칠 잠재적 위험을 파악하고 완화해야 한다.
- 개인정보 기본 설정 유지: 모든 개인정보 보호 설정은 기본적으로 최고 수준(high privacy by default)로 유지해야 한다.
- 다크 패턴(dark pattern) 금지: 아동이 개인정보를 더 많이 제공하도록 유도하는 기만적인 설계를 금지해야 한다.

---

50) California Legislature, *Crimes: Child Pornography*.(A.B.1831), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240A.B.1831> (최종방문일 2025. 12. 19).

51) California Legislature, *The California Age-Appropriate Design Code Act*(A.B.2273), 2021-2022. <https://leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=202120220A.B.2273&showamends=false>(최종방문일 2025. 12. 19).

다만 이 법은 테크 기업 협회(NetChoice)가 제기한 소송에 따라, 2023년 9월 연방지방법원으로부터 표현의 자유 침해 가능성을 이유로 효력 정지 가처분 명령을 받은 상태이다.

③ AI 동반자 챗봇 규제법(Companion Chatbots)(SB 243)(2025)[52]

본 법은 2025년 10월 제정되어 2026년 1월 1일부터 시행 예정인 미국 최초의 AI 동반자 챗봇(companion chatbot)[53]* 규제 법안으로, 인간과 유사한 방식의 상호작용을 목적으로 설계된 AI 챗봇이 이용자와 정신적·정서적 유대를 형성하거나 장기간 상호작용하는 과정에서 자해 또는 위험을 증폭시킬 수 있다는 우려가 입법 배경이 되었다.

AI 동반자 챗봇을 운영하는 기업에게 다음과 같은 의무를 부여하며, 불법 딥페이크 콘텐츠를 통해 이익을 얻은 자에게는 최대 25만 달러의 벌금을 포함한 강력한 처벌을 시행한다.

- 연령 확인 및 경고 의무: 기업은 연령 확인 기능을 도입하고, 소셜미디어 및 동반자 챗봇 이용에 관한 경고 기능을 제공해야 한다.
- 자살·자해 대응 프로토콜 마련: 챗봇이 이용자의 자살·자해 의사를 표현하는 발화를 감지하거나 관련 신호를 인식한 경우, 상담 기관 또는 위기 지원 기관과 연계되는 대응 체계를 가동해야 한다. 해당 절차의 운영 현황과 성과는 매년 주(州) 자살예방기구에 보고해야 한다.
- 의료 사칭 금지: 플랫폼에 제공되는 챗봇은 스스로를 의료 전문가 또는 의료 서비스를 제공하는 주체로 표현해서는 안 된다.
- 과몰입 방지 기능 도입: 기업은 미성년자를 대상으로 3시간마다 반복적인 휴식 알람을 제공하고, 이용자가 실제 사람이 아닌 AI 챗봇과 상호작용하고 있음을 지속적으로 인지할 수 있도록 조치해야 한다.
- 성적·유해 콘텐츠 차단: 챗봇이 성적으로 노골적이거나 유해한 이미지 또는 콘텐츠를 생성·제공하는 행위는 금지된다.

---

52) California Legislature, *Companion Chatbots*.(S.B.243), 2025-2026. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260S.B.243>(최종방문일 2025. 12. 19).

53) 인간과 유사한 반응을 제공하며, 사용자의 사회적·정서적 요구를 충족하는 것을 목적으로 한 적응형·대화형 AI 시스템을 말한다.

(2) 텍사스주의 Stopping AI-Generated Child Pornography Act(SB 20)(2025)[54]

텍사스주는 2025년 9월 1일부터 발효된 형법 제43장의 개정을 통해 실제 또는 AI 생성물 여부에 상관없이 18세 미만으로 보이는 인물의 성적 행위가 묘사된 시각 자료(visual material)를 고의로 소지하거나, 홍보(promote)하거나, 제작하는 행위를 새로운 범죄로 규정하였다. 이 규정은 생성형 AI, 컴퓨터 그래픽, 애니메이션 등을 활용한 아동 성적 콘텐츠의 무분별 유포에 대응하고 기존 법의 공백을 보완하기 위해 발의되었다. 법을 위반할 경우, 초범에게는 주 교도소 중범죄(state jail felony) 처벌을 내려 최대 2년 이하의 징역 또는 최대 1만 달러의 벌금을 처벌한다. 다만 유사한 유형의 전과가 1회 있는 경우, 3등급 중죄, 2회 이상인 경우 2등급 중죄로 형상과 벌금이 가중된다.

다. 영국

1) 아동 권리 보호를 위한 행동강령(Children & AI Design Code)(2025)[55]

2025년 비영리단체(5Rights Foundation)에서 제시한 강령으로, 리스크 유형 및 목적, 설계, 배포, 모니터링, 투명성 등 항목별 행동강령을 제시하였다. 주요 내용으로 음성 톤과 같은 생체정보와 행동 패턴을 결합한 다중 연령 확인 시스템 도입을 권고하고, 아동 성착취물에 대한 즉각적인 차단 및 삭제를 위해 국제적인 아동 성착취물 데이터베이스 공유 플랫폼 마련을 제안하였다.

2) 온라인 안전법(OSA)(2023)[56]

해당 법은 소셜미디어 플랫폼에 대한 정부 차원의 관리·감독의 강화와 미성년자의 유해 콘텐츠에 접근을 방지하려는 두 가지 목적을 가진다. 법의 적용 대상은 이용자 간 서비스(user-to-user service)로서 이용자가 생성하거나 업로드한 콘텐츠를 다른 이용자가 열람

54) Texas Legislature, *Relating to the Creation of a the Criminal Offense for FA.B.ricating a Deceptive Video with Intent to Influence the Outcome of an Election.*(S.B.751), 2019. <https://capitol.texas.gov/BillLookup/History.aspx?LegSess=86R&Bill=S.B.751>(최종방문일 2025. 12. 19).

55) 5Rights Foundation, *Children & AI Design Code.* 2025. <https://5rightsfoundation.com/resource/children-ai-design-code/>(최종방문일 2025. 12. 19).

56) Online Safety Act 2023, c. 50. <https://www.legislation.gov.uk/ukpga/2023/50/contents>(최종방문일 2025. 12. 30).

할 수 있는 소셜미디어 및 메시징 서비스, 이용자가 정보를 탐색할 수 있는 검색엔진 서비스, 그리고 상업적 포르노그래피 웹사이트 등을 포함한다.

서비스 제공자는 플랫폼 내 불법 콘텐츠에 대한 위험을 평가하고 이를 신속히 제거하고 재발을 방지할 의무를 부담한다. 특히 아동이 이용 가능한 플랫폼의 경우, 아동의 접근 가능성을 사전에 평가하고 안전한 온라인 환경을 제공해야 한다. 아울러 13세 미만의 청소년은 소셜미디어 계정을 생성할 수 없으며, 기업은 이를 준수하기 위해 연령 확인 기술을 활용한 엄격한 관리 체계를 구축해야 한다. 더 나아가 아동 성적 착취 및 남용, 테러리즘과 같은 불법 콘텐츠뿐만 아니라 음란물, 자살 또는 자해 조장 등 18세 미만 청소년에게 유해한 콘텐츠에 대한 위험을 식별하고 이를 최소화하기 위한 조치를 취해야 한다.

영국 통신청(Ofcom)은 청소년 유해 콘텐츠에 대한 실태 조사와 위험 평가를 시행하고, 정부의 요구에 따른 감독·감사, 유해 콘텐츠 제거 요청 등을 수행해야 한다. 본 법을 위반할 경우, 사업자에 대한 벌금 부과뿐만 아니라 위법 기업의 임원에 대해 최대 2년의 징역형을 부과할 수 있는 형사처벌 규정을 포함하고 있다.

3) 연령적합설계지침(Age Approciate Designing Code: a code of practice for online services: AADC)(2020)[57]

청소년들이 온라인 서비스를 안전하게 이용할 수 있도록, 기업들이 비례적인 데이터 처리 방식을 채택하도록 유도하기 위한 지침이다. 청소년의 접근 가능성이 있는 온라인 서비스를 대상으로, 연령별* 적합한 데이터 보호 기준을 맞춤형으로 제공하며, 15가지 핵심 표준사항을 제시한다.

* 유아기(0~5세), 초등기(6~9세), 전환기(10~12세), 10대 초반(13~15세), 예비 성인기(16~17세)로 구분

57) Information Commissioner's Office(ICO), *Age Appropriate Design: A Code of Practice for Online Services(The Children's Code)*, 2020. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services>(최종방문일 2025. 12. 19).

〈표 3－1〉 AADC의 15가지 핵심 표준사항

표준사항
① 청소년의 최선의 이익 보장
② 데이터보호영양평가(Data Protection Impact Assessment, DPIA) 수행
③ 다양한 연령과 발달에 따른 연령적합 설계의 적용 보장
④ 적절한 정도의 투명성 제공(간결하고 눈에 띄며 명확한 언어로 설명 제공, 관련 통지 및 관련 부모 또는 보호자와의 상담 권장)
⑤ 데이터의 유해한 이용 방지
⑥ 자체 정책 및 커뮤니티 규정 준수
⑦ 가장 높은 수준의 프라이버시 보호 기본값 설정
⑧ 데이터 처리 최소화(처리 목적 이상의 능동적이고 고의적인 개인 데이터 수집, 보유 금지 및 청소년의 선택권 보장)
⑨ 제3자가 데이터 공유 제한(기업이 최선의 이익을 고려하여 설득력 있는 이유를 밝힐 수 있는 경우에만 데이터 공유 가능, 수신자 역시 해당 요구사항을 준수할 것이라는 확신을 얻어야 하며 이를 준수하기 위한 실사 수행 필요)
⑩ 위치정보에 관한 강화된 보호(아동의 최선의 이익을 위한 설득력 있는 이유를 입증할 수 있는 경우를 제외하고는 위치정보 옵션 비활성화 및 활성화시 그 사실에 대한 명확한 표시, 특정 목적을 위한 처리 완료 후 역시 비활성화를 기본값으로 설정)
⑪ 자녀보호기능 관련 통지 제공(자녀보호기능 제공 시 연령에 맞는 통제 정보의 제공 및 부모 또는 그 보호자의 모니터링 시 이 사실을 청소년에게 명확히 표시)
⑫ 필수적이지 않을 경우, 프로파일링 제한
⑬ 넛지(nudge, 선택유도) 기술 회피
⑭ 연결 기반 장난감 및 장치의 관련 규율 준수 보장(청소년의 접근 가능성이 있을 경우 연령 맞춤형 구성 및 관련 옵션 활성화 등)
⑮ 이용자의 실질적 권리 행사 보장을 위한 온라인 도구 제공

자료: Information Commissioner's Office(ICO). Age appropriate design: a code of practice for online services(The Children's Code), 2020.

4) 데이터 이용 및 접근법(Data(Use and Access) Act 2025)(DUAA)[58]

영국 정부는 여성과 소녀에 대한 폭력 근절 전략(Violence Against Women and Girls, VAWG)의 일환으로, AI가 생성하는 비동의 친밀 이미지와 나체화(nudification) 등 딥페이

58) United Kingdom Government, *Data(Use and Access) Bill: Factsheets,* Department for Science, Innovation and Technology, 2024. <https://www.gov.uk/government/publications/data-use-and-access-bill-factsheets>(최종방문일 2025. 12. 19).

크 성착취에 대한 규제 강화를 공식화하였다. 이에 따라 비동의 친밀 이미지의 생성 및 요청 행위를 형사 범죄로 규정하고, AI 딥페이크 성착취물 역시 처벌 대상에 명시적으로 포함하였다. 특히, 옷을 벗기는 효과를 내는, 이른바 옷벗기기(Declothing) 도구의 단순 이용을 넘어 제작·공급·유통 행위까지 금지하는 추가 입법을 추진하고 있으며, 딥페이크 성착취 생태계의 원천을 차단하려는 강력한 정책 의지를 드러내고 있다.

라. 중국의 생성형 AI 잠정 관리 방법(2023)[59]

해당 법안은 서비스 제공자에게 미성년자 보호와 관련된 구체적인 의무를 부과한다. 서비스 제공자는 미성년자가 생성형 AI를 적절하게 이용하도록 유도하고, 이들의 개인정보를 철저히 보호하며, 유해 정보에 노출되지 않도록 하는 방지 조치를 취해야 한다. 위반 시 관련 법령에 따라 엄격한 처벌을 받게 되며, 중국 내 서비스를 제공하는 외국 기업에도 동일하게 적용되어 중국의 국가 기관이 기술적 제재 조치를 취할 수 있다고 명시하였다.

마. 호주의 아동·청소년 소셜미디어금지법(2024)[60]

2024년 11월 29일 통과된 「아동·청소년 소셜미디어금지법」은 기존 플랫폼들이 설정한 만 13세 최소 연령 기준이 아동 보호에 충분하지 않다는 사회적 문제의식을 바탕으로 입법이 추진되었다. 해당 법은 만 16세 미만 아동·청소년의 소셜미디어 계정 소유를 전면적으로 금지하고, 부모의 동의 여부와 무관하게 계정 개설을 허용하지 않는다는 점에서 기존 규제보다 강화된 접근을 취하고 있다.

대상 서비스는 이용자 간 사회적 교류를 주목적으로 하는 대부분의 소셜미디어 플랫폼을 포함한다. 플랫폼 제공자는 만 16세 미만 이용자의 계정 생성을 막기 위한 '합리적 조치'를 취해야 하며, 이를 위반할 경우 최대 4,950만 호주 달러에 달하는 막대한 벌금이 부과될 수 있다. 해당 법에서 호주의 온라인 안전국이 플랫폼의 의무 이행 여부를 감독하고 정보를 요구할 권한을 가진다.

---

59) 国家互联网信息办公室. 「生成式人工智能服务管理暂行办法」, 2023. 7. 13. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm(최종방문일 2025. 12. 19).

60) 국회입법조사처, "호주의 '아동·청소년 소셜미디어금지법' 통과와 입법적 시사점", 「NARS 현안분석」, 345호, 2024.

바. 캐나다의 온라인 유해 규제법(Online Harms Act)(C-63)[61]

아동·청소년을 포함한 취약계층을 AI 기술의 잠재적 유해성으로부터 보호하기 위해, 플랫폼 사업자에게 효과적인 연령 검증 시스템과 부모 통제 기능을 도입하도록 요구하고 있다.

4. 투명성 확보

가. EU

1) 「AI 법」(AI Act)(2024)

「AI 법」은 AI 시스템 제공자가 이용자가 AI와 상호작용하고 있음을 명확히 인지하도록 시스템을 설계·개발해야 하며, AI 산출물은 기계적 판독 가능하고 인공적으로 생성되었음이 검출 가능해야 한다고 명시하고 있다.[62] 또한 「AI 법」은 리스크 기반 접근을 채택하여 범용 AI(GPAI) 제공자와 시스템 위험(system risk)*를 수반하는 범용 AI 제공자의 투명성 의무를 구분하여 규정하고 있다.

* 범용 AI모델의 고영향 성능에 따른 특유의 위험, 도달 범위나 공중 보건, 안전, 공공안보, 기본권 또는 사회 전반에 대한 실제적이고 예상가능한 부정적 영향으로 인해 EU 시장 전반에 중대한 영향을 미치며 가치사슬 전반에 파급될 수 있는 위험

- 범용 AI 제공자 의무: 모델의 학습 방식, 성능, 제한 사항 등을 포함한 기술 문서를 작성하고 이를 지속적으로 최신화해야 하며, 학습에 사용된 데이터에 대한 상세 요약서를 작성·공개할 의무를 부담한다. 또한 범용 AI 모델 통합·활용하는 서비스 구축자가 참고할 수 있도록 관련 문서를 작성하고 최신화해야 한다.
- 시스템 리스크를 수반하는 범용 AI 제공자 의무: 모델 평가 및 문서화를 수행하고, 중

61) Parliament of Canada, *An Act to Enact the Online Harms Act, to Amend the Criminal Code, the Canadian Human Rights Act and An Act Respecting the Mandatory Reporting of Internet Child Pornography by Persons Who Provide an Internet Service*.(Bill C-63), 2024. <https://www.parl.ca/DocumentViewer/en/44-1/bill/C-63/first-reading>(최종방문일 2025. 12. 19).

62) 딜로이트 인사이트, 「인공지능(AI) 시대의 새로운 국면, AI 규제와 기업 리스크 관리 전략」, 2024. 10.

대한 사고 발생 시 이를 추적하여 AI Office 등 감독 당국에 보고할 의무가 추가된다.

한편 「AI 법」은 배포자와 이용자에 대해서도 각각의 투명성 의무를 규정하고 있다. 배포자는 AI 시스템 사용 과정에서 건강이나 안전에 중대한 위협을 발견될 경우 시스템 사용을 중지하고 이를 신고해야 하며, 중대한 사고 발생 시 내부 보고 체계를 구축하고 관할 당국에 통보해야 할 의무를 가진다. 이용자는 자연인과 상호작용하는 AI 시스템, 생체인식 시스템, 감정인식 시스템 등과 관련된 정보에 대해 이용자에게 통지할 의무를 부담하며, 딥페이크 콘텐츠의 생성 또는 조작과 관련된 사실을 정확히 제공할 의무를 가진다.

2) 디지털 서비스 법(DSA)(2023)[63]

「DSA」는 온라인 플랫폼에서 활용하는 추천 시스템 관련 정보 및 보고서를 통해 투명성을 구현하고자 한다. 대규모 온라인 플랫폼은 추천 시스템에 사용되는 주요 매개변수 정보와 옵션을 이용자에게 제공할 의무를 가지며, 옵션이 여러 개일 경우 옵션 선택 및 변경 기능을 제공해야 한다. 또한 중개 서비스 제공자는 콘텐츠 삭제·차단 조치 내역 등을 담은 투명성 보고서를 정기적으로 공개해야 한다.

3) 범용 AI를 위한 실무 규범(2025)

「범용 AI를 위한 실무 규범」(GPAI Code of Practice)은 2025년 7월 10일 공개된 지침으로,[64] EU의 「AI 법」 준수를 위한 구체적이고 실무적인 이행 기준을 제시한다. 이 규범은 범용 AI* 모델을 개발하거나, 제3자의 의뢰를 받아 이를 EU 시장에 출시하는 자연인 또는 법인, 공공기관, 대행사 또는 기타 단체를 적용 대상으로 한다.

* EU 「AI 법」에 따라 상당한 일반성을 갖추고 다양한 작업을 능숙하게 수행할 수 있으며 다양한 하위 시스템 또는 응용 프로그램에 통합될 수 있는 AI 모델

규범의 제1장에서는 투명성 의무를 중점적으로 다루고 있는데, 모든 범용 AI 모델 제공

63) 최경미·지성우, "디지털서비스 투명성·개인정보보호·위기대응에 관한 법 연구-EU 디지털서비스법(DSA)을 중심으로", 「미국헌법연구」, 33권 3호, 177-214, 2022.

64) European Commission, *Code of Practice for General-Purpose AI*, 2025. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>(최종방문일 2025. 12. 19).

자는 모델의 라이선스, 학습 데이터셋, 에너지 사용량 등 핵심 정보를 포함한 표준화된 문서를 유지·관리해야 한다고 명시하고 있다. 또한 이러한 문서는 요청이 있을 경우 AI 사무국에 제출·공개해야 할 의무를 갖는다.

나. 미국

1) 연방 수준

(1) 생성형 AI 학습 데이터 투명성 법(Generative Artificial Intelligence Training Data Transparency)(2023)[65]

본 법은 생성형 AI 시스템의 훈련 데이터에 대한 투명성을 높이고, 소비자와 이해관계자들이 AI 시스템의 기반이 되는 데이터의 출처와 특성을 이해할 수 있도록 2026년부터 AI 개발자가 AI 시스템 훈련에 사용된 데이터셋 정보를 자사 웹사이트에 공개하도록 의무화하고 있다. 이를 통해 생성형 AI 학습 과정의 투명성을 제고하고, 저작권 및 데이터 출처에 대한 검증을 강화하고자 한다. 생성형 AI 시스템 및 서비스 개발에 활용된 데이터셋은 다음의 내용을 포함한다.

- 데이터셋의 출처 또는 소유자
- 데이터셋이 AI 시스템 및 서비스의 의도된 목적 달성에 기여되는 정도
- 데이터셋에 포함된 데이터 포인트*의 수(일반적인 범위로 표시 가능하며, 동적 데이터셋일 경우 추정치 사용 가능)
- 데이터셋에 저작권, 상표권, 특허권으로 보호되는 데이터 포함 여부 또는 데이터셋이 전적으로 공공 도메인에 속하는 여부
- 데이터셋을 개발자가 구매하거나 라이선스를 취득했는지의 여부
- 데이터셋에 개인정보가 포함되어 있는지 여부
- 데이터셋에 집계된 소비자 정보가 포함되어 있는지 여부
- 개발자에 의한 데이터셋의 정치, 처리 또는 기타 수정 여부(AI 시스템 및 서비스 관련

65) California Legislature, *Generative Artificial Intelligence: Training Data Transparency.* (A.B.2013), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240A.B.2013> (최종방문일 2025. 12. 19).

의도된 목적 포함)

- 데이터셋의 데이터가 수집된 기간
- AI 시스템 및 서비스 개발 중 데이터셋이 최초로 사용된 날짜
- 생성형 AI 시스템 및 서비스 개발 과정에서 합성생성 데이터를 사용했거나, 지속적으로 사용에 대한 여부

(2) Identifying Outputs of Generative AI Act(IOGAN Act, H.R. 4355)(2019)[66]

해당 법안은 2019년 9월 제116대 연방의회에서 도입되었으며 생성형 AI 산출물 전반에 대한 식별 가능성 확보와 기술 연구의 촉진을 목표로 한다. 이 법안은 AI가 생성한 콘텐츠에 대해 명확한 라벨링 또는 식별가능한 정보를 부착할 수 있는 기술적 방법을 연구·개발하도록 지원하고 표준 마련을 촉진하는 데 목적을 둔다.

(3) AI 리스크 관리 프레임워크(2023)[67]

미국 국립표준기술연구소(NIST)는 2023년 1월 26일 'AI 리스크 관리 프레임워크(AI RMF) 1.0'을 발표하였다. 이는 공공 및 민간에서 신뢰할 수 있는 AI 시스템 개발을 위한 국가 이니셔티브의 일환으로 정부, 기업, 하계, 비영리단체, 시민, AI 실무자 등 다양한 이해관계자들의 민관협력을 통해 개발되었다. 해당 프레임워크는 AI 시스템의 설계·개발·배포·운영 전 주기에 걸쳐 안전성, 책임성, 투명성 등 신뢰성 확보를 위한 AI 시스템의 7가지 특성과 이를 구현하기 위한 조직의 핵심 기능을 거버넌스·매핑·측정·관리의 4가지 영역에서 제시하고 있다.

- 모든 이해당사자가 참여가 능한 공개적이고 투명한 프로세스를 통해 수립 및 갱신되어야 한다.
- 경영진부터 비전문가까지 다양한 대상이 이해할 수 있는 명확한 언어로 작성되었으며, 동시에 실무자에게 유용한 기술 정보를 제공해야 한다.

---

66) U.S. Congress, *Identifying Outputs of Generative Adversarial Networks Act(IOGAN Act)* (H.R.4355), 116th Congress, 2019-2020. <https://www.congress.gov/bill/116th-congress/house-bill/4355>(최종방문일 2025. 12. 19).

67) 한국지능정보사회진흥원, 「2023년 美 NIST 「AI 위험관리 프레임워크(AI RMF) 1.0」 분석 및 시사점」, 2023.

- 독립적으로 활용되기보다, 조직의 기존 위험관리 전략에 통합되어 직관적으로 적용되어야 한다.
- 획일적인 요구사항보다는 유연한 결과와 접근 방식을 제시해야 한다.
- 기존의 표준, 지침, 모범사례를 활용하여 국내외 법률 및 규제 체계 내에 적용가능하도록 설계되어야 한다.
- AI 기술과 인식이 빠르게 변화함에 따라 신속한 업데이트가 가능하도록 구성되어야 한다.

2) 주 단위

(1) 캘리포니아주의 캘리포니아 AI 시스템 투명성 법(AB 853)(2025)[68]

「캘리포니아 AI 시스템 투명성 법」(California AI Transparency Act)은 기존의 캘리포니아 AI 투명성 법(SB 942)을 보완하여 기존 규제 대상이던 소프트웨어 기업과 대형 플랫폼 사업자에서 AI 호스팅 기업, AI가 탑재된 디바이스 제조 기업까지 적용 범위를 확대하였다. AI 시스템 설계·운영 단계에서 이용자가 인지하지 못한 채 영향받을 수 있는 기능에 대한 사전 고지를 의무화하는 것을 핵심 내용으로 한다. 이에 따라 서비스 제공자는 시스템의 이용자 인터페이스(UI) 및 이용약관 등에 AI 기술을 활용하고 있음을 명시해야 하며, 모델이 자동으로 생성 및 추천하는 콘텐츠인지를 명확히 표시해야 한다. 해당 법안은 2026년 8월 2일부터 시행될 예정이다.

(2) 미국 기타 주요 주(state)

캘리포니아주 이외에도 뉴욕, 콜로라도, 텍사스를 비롯한 다양한 주에서 투명성을 제고하고 안전성을 확보하기 위한 법안 도입이 활발히 이루어지고 있다. 각 주의 주요 법안과 핵심 내용은 다음과 같다.

68) California Legislature, *California AI Transparency Act*(A.B.853), 2025-2026. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260A.B.853>(최종방문일 2025. 12. 19).

〈표 3－2〉 미국 주요 주(state)의 투명성 관련 법안

주(state)	법안명	특징	주요 내용
뉴욕	RAISE Act (Responsible Artificial Intelligence in State Entities Act)	미국 최초의 AI 안전성 보고 의무 입법	• 첨단 AI 모델 개발사에 안전 계획 수립 의무 제3자 감사 의무 • 중대한 사고 발생 시 72시간 내 신고 • 내부고발자 보호 규정 • 위반 시 최초 500만 달러, 재발 시 1,500만 달러, 중대 사고 방치 시 최대 3,000만 달러 민사 처벌
콜로라도	Consumer Protection Act	미국 최초의 포괄적 AI 규제 법률	• 고위험 AI 시스템 사용 시 소비자 사전 고지 의무 • 알고리즘 차별 방지를 위한 합리적 주의 의무
텍사스	HB 149	AI 규제 샌드박스 도입	• 행동 조작, 불법적 차별, 딥페이크 제작 금지 • 기업이 법적 보호 하에 기술을 시험할 수 있는 AI 규제 샌드박스 프로그램 도입
유타	SB 149	AI 정체성과 역할 투명성 규제	• AI가 소비자와 상호작용할 경우, 명확하고 눈에 띄는 방식으로 AI라는 사실을 공개 • 의료·법률 등 규제 직업 분야에는 강화된 사전 고지 의무 적용
몬테나	HB 178	공공부문 AI 투명성	• 주 정부 기관의 AI 사용 사실 공개 의무 • 개인의 권리에 영향을 미치는 결정은 반드시 인간이 검토
미시간	HB 4668	기초 모델의 안전과 책임에 대한 규제	• 대규모 기초 모델 개발사에 안전·보안 프로토콜 마련 및 공개 • 제3자 감사 및 내부고발자 보호 의무

자료: 각 주의 법안을 참고하여 연구진 작성

다. 영국

1) AI를 활용한 의사결정 설명 가이드(2023)

'AI를 활용한 의사결정 설명 가이드'(Explaining decisions made with AI: Guidance for Organizations)는 2020년 영국 정보위원회(ICO)와 앨런 튜링 연구소가 공동으로 마련한 지침으로, 2023년 최신화되었다.[69] 본 지침은 법적이거나 이에 준하는 중대한 영향을 미치는

69) Information Commissioner's Office(ICO) & The Alan Turing Institute, *Explaining Decisions Made with AI*, 2020.

전자동화된 의사결정을 적용 대상으로 하여, 조직이 AI 시스템을 활용해 의사결정을 수행하는 경우 의사결정 과정과 도출의 이유, 그리고 해당 결정이 개인에게 미칠 영향을 이해할 수 있도록 충분한 설명을 제공해야 할 의무를 명시한다. 특히 AI 시스템이 고용, 신용, 건강관리, 교육 접근 등 개인의 삶에 중대한 영향을 미치는 영역에서 활용되는 경우 적절한 설명을 해야 한다. 또한 뉴스 피드, 검색 결과, 채용 공고 등 이용자 프로파일링에 기반한 온라인 콘텐츠 제공이 개인의 법적 권리 또는 유사한 중대한 이해관계에 영향을 미칠 경우에도 설명 의무가 발생한다.

이러한 설명은 데이터 수집 시점, 의사결정 이전, 의사결정 과정 중 또는 결정 이후 등 다양한 단계에서 제공될 수 있음을 전제로 한다. 지침은 의사결정 설명의 유형을 ▲근거 설명(특정 결과에 도달한 이유), ▲책임 설명(해당 결정과 AI 시스템에 대한 책임자), ▲데이터 설명(사용된 데이터와 그것이 결과에 미친 영향), ▲공정성 설명(시스템 편향과 차별을 피하기 위한 조치), ▲안전성·성능 설명(시스템의 신뢰도와 정확도), ▲영향 설명(해당 결정이 개인에게 미칠 영향)의 6가지로 구분하였다. 이를 준수하지 않아 영국 개인정보보호법(UK GDPR) 위반으로 판단될 경우, 시정 명령과 함께 최대 1,750만 파운드 또는 전 세계 연간 매출의 4% 중 더 높은 금액에 해당하는 과징금이 부과될 수 있다.

2) AI 사이버 보안 행동 강령(2025.1.31. 발표)[70]

영국 과학혁신기술부(Department for Science, Innovation and Technology, DSIT)는 2025년 1월 31일, AI 시스템의 안전성을 설계 단계부터 확보하기 위한 「AI 사이버 보안 행동 강령」(AI Cyber Security Code of Practice)을 발표하였다. 본 강령은 보안 내재화(Secure by Design) 원칙에 기반한 비구속적 행동 강령으로, AI 개발·운영 주체가 참고해야 할 보안 모범사례를 제시하는 데 목적이 있다. 해당 강령은 AI 모델의 투명성과 무결성 확보를

<https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/?utm_source=chatgpt.com> (최종방문일 2025. 12. 19).

70) United Kingdom Government, *AI Cyber Security Code of Practice, Department for Science, Innovation and Technology*, 2025. <https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice>(최종방문일 2025. 12. 19).

위해 다음과 같은 주요 보안 조치를 제시한다.

- 문서화 및 위험관리 체계 구축: 시스템 설계, 배포, 유지보수 등 AI 전 생애주기에 걸친 위협 모델링 결과를 명확히 기록하고 책임 소재를 확립할 것을 제시한다.
- 표준화된 기술 문서: 이용자가 모델의 의도된 사용처와 잠재적 위험을 인지할 수 있도록 AI 모델의 기능, 한계, 성능 지표, 윤리적 고려사항 등을 표준화된 서식(예: 모델 카드)으로 제공을 권고한다.
- 무결성 검증 수단 활용: 배포된 AI 모델이 악의적으로 위·변조되지 않았음을 수학적으로 증명할 수 있도록, 모델 가중치(Weight) 파일 등에 대한 암호화 해시값 또는 디지털 서명 등 기술적 무결성 검증 수단 활용을 제안한다.
- 데이터 출처 기록: 데이터 오염 공격과 지식재산권 분쟁에 대응하기 위해 훈련 데이터의 출처, 수집 시기, 라이선스 정보를 투명하게 기록하여 데이터 공급망 안전성의 필요성을 강조한다.
- 취약점 관리 및 업데이트 공지: 보안 취약점 발견 시 신속한 패치를 지원하되, 시스템의 동작에 중대한 영향을 줄 수 있는 주요 업데이트 실행 전에는 최종 이용자에게 변경 사항과 영향을 사전에 고지하여 이용자 신뢰와 서비스 안정성을 유지할 것을 권고한다.

라. 중국[71]

1) 인터넷 정보 서비스 알고리즘 추천 관리 규정(2022)

2022년 3월 1일 시행된 「인터넷 정보 서비스 알고리즘 추천 관리 규정」은 중국 내 알고리즘 기반 추천 시스템을 제공하는 인터넷 정보 서비스 사업자를 대상으로, 알고리즘 추천의 작동 방식과 원리에 대한 설명 의무를 부과함으로써 알고리즘 투명성을 제도적으로 강화한 규범이다. 제11조에서 서비스 제공자가 추천 알고리즘의 주요 매개변수와 작동 원리를 이용자에게 공개하도록 의무화하고 있으며, 제12조에서는 이용자가 알고리즘 추천에 동의하지 않거나 이를 거부하거나 중단할 수 있는 선택권을 명시적으로 보장하도록 규정한다. 이에 따라 사업자는 비개인화 추천 옵션을 제공하거나 알고리즘 추천 기능을 종료

71) 한국인터넷진흥원, 「인터넷 정보서비스 알고리즘 추천관리규정」, 2022.

할 수 있는 수단을 마련해야 하며, 이용자가 자신의 프로파일링 정보가 추천에 어떻게 활용되는지 인지할 수 있도록 조치를 취해야 한다.

2) 인공지능 안전 거버넌스 구성(人工智能安全治理框架)(2024)

2024년 9월 9일 발표된 중국의 국가 차원의 AI 안전 거버넌스 지침으로, AI 기술의 연구·개발·적용 전반에 걸쳐 안전하고, 신뢰할 수 있으며, 공정하고 투명한 AI 생태계를 조성하는 것을 기본 원칙으로 제시하고 있다. 이 지침은 AI 기술의 원칙, 기능, 적용 시나리오, 그리고 잠재적 안전 위험 등을 적절히 공개할 필요성을 강조하며, 이를 통해 AI 시스템의 투명성을 지속적으로 제고해야 한다는 방향성을 명확히 하고 있다.

마. 일본

1) AI 사업자 가이드라인 1.0(2024.4.19. 발표)[72]

2024년 4월 19일 발표된 'AI 사업자 가이드라인'(AI事業者ガイドライン（1.0))은 AI 사업자가 AI의 라이프사이클 전반에서 공통적으로 준수해야 할 10대 기본 원칙을 제시하며 자율규제의 토대를 마련하고 있다. 더불어 AI 개발자, AI 제공자, AI 이용자별로 차별화된 투명성 의무를 부여하여 논의를 구체화하였다. AI 개발자는 AI 모델을 설계 및 변경하는 주체로, AI를 제공할 경우 사회적 영향 및 위험성을 사전에 검토하고 대응해야 한다. AI 제공자는 시스템 구현과 운영의 책임 주체로, 보안 대책과 이용자 지침을 마련하여 AI의 적절한 이용을 지원해야 한다. 마지막으로 AI 이용자는 제공자의 사용 목적 범위 내 적절하고 효과적으로 AI 시스템을 활용할 책임을 가진다.

2) 인공지능 관련 기술의 연구개발 및 활용 촉진에 관한 법률(2025)[73]

일본의 「인공지능 관련 기술의 연구개발 및 활용 촉진에 관한 법률(人工知能関連技術の研究開発及び活用の推進に関する法律)(이하 'AI 추진법')」은 2025년 6월 4일 공포된 법률로, AI 기술의 연구개발과 활용을 국가 전략 차원에서 체계적으로 추진하는 동시에 잠재적

72) 総務省・経済産業省,「AI事業者ガイドライン（第1.0版）」, 2024. 4. 19, <https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_1.pdf> (최종방문일 2025. 12. 30).

73) 한국인터넷진흥원, "일본,「인공지능 관련 기술의 연구개발 및 활용 촉진에 관한 법률」 공포('25. 6. 4.)",「인터넷·정보보호 법제동향」, 2025.

위험에 대응하기 위한 기본 법제이다. 본 법은 AI 기술을 국민 생활의 질 향상과 경제 발전을 견인하는 핵심 기반 기술로 규정하면서도, 혁신 촉진과 위험 대응 간의 균형을 추구하는 것을 목적으로 한다. 법률은 총 4장 28개 조문으로 구성되어 있으며, AI 관련 기술 정책의 기본이념을 제시하고, 국가 차원의 AI 기본계획 수립과 AI 전략 본부 설치를 통해 시책을 종합적·계획적으로 추진하도록 규정하고 있다.

「AI 추진법」은 국제 경쟁력을 갖춘 연구개발 역량 확보와 더불어 연구개발 및 활용 전 과정에서의 투명성 확보를 기본 이념으로 명시한다. 국가, 사업자, 시민 등 각 이해관계자의 책임을 규정하고, 내각 산하에 범정부 기구인 AI 전략 본부를 설치하여 국가 AI 전략을 총괄하도록 하였다. 이는 분산된 논의를 통합하고 일관된 정책을 추진하기 위한 국가 차원의 거버넌스 체계를 갖추기 위함이다.

5. 분쟁조정 및 피해구제

가. EU의 AI 배상지침(안)(2022)[74]

EU의 「AI 배상지침(안)」(Proposal for DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence(AI Liability Directive))은 AI로 인해 발생한 피해에 대한 구제 절차를 용이하게 하는 것을 목표로 한다. 이 지침의 핵심은 입증책임 전환의 원칙에 있다. 즉, 고위험 AI 시스템으로 인해 피해가 발생했다고 합리적으로 주장하는 경우, 피해자가 인과관계를 직접 입증하는 기존 구조와 달리, 해당 AI 시스템의 개발자 또는 운영자가 고의나 과실이 없었음을 스스로 증명하도록 책임을 전환한다. 이러한 구조는 피해자의 입증 부담을 크게 완화시켜 AI로 인한 손해에 대해 실질적인 배상과 권리 구제를 원활하게 만들기 위함이다.

74) European Commission, “Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil LiA.B.ility Rules to Artificial Intelligence(AI LiA.B.ility Directive)”, 2022.
<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496>(최종방문일 2025. 12. 19).

나. 미국

1) 연방법안 AI Accountability and Personal Data Protection Act(S. 2367)[75]

해당 법안은 2025년 7월 21일 상원에 발의되었으며 개인의 사전 동의 없이 AI 훈련, 데이터 수집·처리·판매 등에 개인정보가 무단으로 활용된 경우, 연방 차원에서 민사적 손해배상 청구권(tort)이 인정되도록 규정한다. 개인은 민사소송 제기를 통해 실제 손해액, 데이터 이용 수익의 3배, 최소 1,000달러, 그리고 징벌적 손해배상 등을 청구할 수 있으며, 기업이 약관에 삽입한 사전 중재 조항을 무효화하여 단체소송 참여 권리를 보장한다.

2) 주(州) 단위

〈표 3－3〉 미국 주요 주(state)의 분쟁조정 및 피해구제 관련 법안

주	법안명	특징	주요 내용
콜로라도	SB24-205	고위험 영역에서 AI 의사결정에 대한 소비자 보호 중심 규제	• 보험 등에서 AI를 활용한 의사결정 시 영향 평가 수행 및 결과 공개 • AI 기반 결정에 대한 소비자 이의 제기 절차 마련 • 알고리즘 차별 방지를 위한 합리적 주의 의무
뉴욕	Local Law 144 of 2021	AI 채용 도구의 편향성 통제 및 구직자 권리 강화	• 연 1회 독립적인 편향성 감사, 결과 공개 의무 • AI 채용 도구 사용 시 사전에 고지 • AI 기반 채용 결정에 대한 이의 제기 및 소송 제기 권리 보장
메사추세츠	HD 3750	보험 청구 심사 과정에서의 AI 사용 투명성 강화	• 보험사가 AI를 활용해 보험 청구를 심사하는 경우 환자에게 AI 사용 사실을 명확히 고지할 의무
캘리포니아	AB 316	AI 자율성에 의한 책임 회피 차단을 명시한 전략적 책임 규제	• AI 피해 소송에서 AI가 자율적으로 야기한 피해를 법적 방어 수단으로 사용하는 것을 명시적으로 금지 • 개발자, 이용자의 책임 원칙을 명문화

자료: 각 주의 법안을 참고하여 연구진 작성

75) U.S. Congress, *AI Accountability and Personal Data Protection Act*(S.2367), 119th Congress, 2025-2026. 〈https://www.congress.gov/bill/119th-congress/senate-bill/2367〉(최종방문일 2025. 12. 19).

다. 영국

1) 온라인 안전법(OSA)(2023)[76]

AI 시스템의 결정에 대해 이용자가 이의를 제기할 권리와 법 위반으로 인한 피해 발생 시 개인 배상청구권을 명시한다. 관련 분쟁에서 감독기관인 Ofcom이 적극적인 중재 역할을 수행한다.

2) AI 사이버 보안 행동 강령(AI Cyber Security Code of Practice)(2025.1.1. 발표)

2025년 1월 1일에 발표된 규범으로,[77] 사이버 보안 사고 발생 시 데이터 오염이나 적대적 공격 등 AI 특화 위협에 대응하기 위한 4가지 메커니즘을 구축하도록 요구한다.

- 사고 지원 프로세스: 개발자 및 시스템 운영자는 사이버 보안 사고 발생 시 영향받는 최종 이용자 및 관련 주체를 지원하기 위한 문서화된 프로세스를 수립해야 한다.
- 사고 관리 계획: 데이터 오염(data poisoning), 모델 드리프트(model drift), 적대적 공격 등 AI 특화 위협에 맞춤화된 사고 관리 계획을 수립하고 복구 절차를 포함해야 한다.
- 인간 운영자 책임: AI 시스템이 잘못된 출력을 생성하거나 해석하기 어려운 결정을 내렸을 때, 인간 운영자가 이를 해석, 검증, 재정의하는 책임을 부여받는다.
- 보고 메커니즘: AI 시스템의 잠재적 보안 위험을 선제적으로 보고할 수 있는 명확하고 접근 가능한 프로세스를 구축하고, 이러한 문제에 대한 투명한 처리를 장려해야 한다.

라. 중국[78]

1) 인터넷 정보 서비스 알고리즘 추천 관리 규정(2022)

AI 서비스 이용자는 알고리즘 결정에 대해 이의를 제기할 권리와 개인정보 수정에 대한 권리를 가지며, 국가 인터넷 정보 사무실이 분쟁 조정의 권한을 갖는다.

---

76) Online Safety Act 2023, c. 50. <https://www.legislation.gov.uk/ukpga/2023/50/contents> (최종방문일 2025. 12. 30).

77) United Kingdom Government, *AI Cyber Security Code of Practice, Department for Science, Innovation and Technology*, 2025. <https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice>(최종방문일 2025. 12. 19).

78) 한국인터넷진흥원, 「인터넷 정보서비스 알고리즘 추천관리규정」, 2022.

2) 인터넷 정보서비스 심층학습 관리 규정(2022)

AI 생성 콘텐츠로 피해가 발생하면 서비스 제공자는 피해자에게 이를 통지하고, 신고 접수 시 24시간 내에 즉시 조치할 의무가 있다. 국가 인터넷 정보 사무실이 피해구제 및 분쟁 해결의 최종 권한을 보유한다.

3) 인공지능 안전 거버넌스 프레임워크(人工智能安全治理框架)(2024)

2024년 9월 9일에 발표된 중국의 국가 차원 AI 안전 거버넌스 지침으로, 모든 이해관계자가 AI 안전에 대한 책임을 완전히 이행하도록 보장하는 것을 핵심으로 한다. 이에 따라 AI 시스템에 이상이 발생할 경우, 이를 신속히 수정해야 하며, AI 시스템의 책임 소재를 가리기 어려운 설명가능성 흠결 리스크를 해결해야 한다. 또한 AI 서비스 제공자는 개발자가 제공한 책임에 관한 진술을 검토하여, 연쇄적으로 책임을 추적할 수 있도록 보장해야 한다. 더불어 AI 안전사고에 대한 비상 대응 메커니즘을 수립하고, 비상 계획을 마련하여 AI 안전 위험을 신속하고 효과적으로 처리해야 한다. 더 나아가 AI 위험과 이에 대한 대중의 불만 및 보고를 처리하기 위한 메커니즘을 구축함으로써, 효과적인 사회적 감독 분위기를 조성해야 한다.

마. 일본

1) 개인정보 보호법[79]

정보 주체가 자신의 개인정보에 대해 열람, 정정, 이용 정지를 요구하고 이의를 제기할 권리를 보장하며, 개인정보보호위원회(PPC)가 불만 및 분쟁 해결을 담당합니다.

2) AI 추진법(2025)[80]

국가가 AI 기술의 부적절한 활용으로 인한 권익침해 사안을 조사하고, 관련 사업자에게 지도 및 자문을 제공할 수 있는 법적 근거(제16조)를 마련하였다. 이러한 결과를 바탕으로 AI 활용 사업자와 국민에게 지도, 자문, 정보 등을 제공할 수 있다.

---

79) Japan Personal Information Protection Commission(PPC), *Act on the Protection of Personal Information*(Act No. 57), 2023. <https://www.ppc.go.jp/en/legal/>(최종방문일 2025. 12. 19).

80) 한국인터넷진흥원, "일본, 「인공지능 관련 기술의 연구개발 및 활용 촉진에 관한 법률」 공포('25.6.4.)", 「인터넷·정보보호 법제동향」, 2025.

바. 싱가포르의 AI 거버넌스 프레임워크(Model AI Governance Framework)[81]

해당 프레임워크는 인간 중심 접근을 통해 모든 AI 결정에 대한 인간의 최종 감독 및 검토를 보장한다. AI 시스템의 결과에 대한 이의제기 및 검토 메커니즘을 의무화하며, 정보통신미디어개발청(IMDA)이 관련 분쟁의 중재 및 해결 권한을 가진다.

제3절 신유형 규제 입법례 심층 분석

1. EU 디지털서비스법(DSA)의 AI 서비스 이용자 보호 현황

가. 개관

2022년 10월 27일 공포되어 2024년 2월 17일 전면 시행된 「DSA」는 EU의 디지털 단일시장 전략의 핵심 축으로서, 온라인 중개서비스 전반에 걸친 이용자 보호 체계를 구축한 법률이다.[82] 「DSA」는 AI만을 규율 대상으로 하는 법률은 아니나, 온라인 플랫폼이 활용하는 AI 기반 추천 시스템, 콘텐츠 조정(content moderation) 알고리즘, 프로파일링 기반 광고 등에 대해 투명성 의무와 이용자 선택권을 부여함으로써 AI 서비스 이용자 보호의 중요한 규범적 기초를 제공한다.

「DSA」의 입법 배경에는 알고리즘 기반 추천 시스템이 이용자의 정보 소비 패턴과 의사결정에 미치는 영향력에 대한 우려가 자리 잡고 있다. 추천 알고리즘은 이용자의 관심과 참여(engagement)를 극대화하도록 설계되는 경향이 있어, 자극적이거나 극단적인 콘텐츠가 우선적으로 노출되고, 이른바 '필터 버블'(filter bubble)이나 '에코 챔버'(echo chamber) 현상을 초래할 수 있다는 점이 지적되어 왔다.[83] 또한 개인정보를 기반으로 한 프로파일링과 맞춤형 광고가 이용자의 프라이버시를 침해하고, 취약계층에 대한 착취적 마케팅을 가

81) Kotra, "싱가포르의 AI 정책과 실제 사례", 「Global Issue Monitoring」, 2024.

82) Regulation(EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC(Digital Services Act), OJ L 277, 2022. 10. 27., p. 1-102.

83) European Commission, "Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services(Digital Services Act) and amending Directive 2000/31/EC," COM(2020) 825 final, 2020. 12. 15.

능하게 한다는 비판도 제기되었다.[84]

이러한 문제의식을 바탕으로 「DSA」는 온라인 플랫폼 사업자에게 알고리즘 추천 시스템의 투명성 확보, 프로파일링에 기반하지 않은 대안적 추천 옵션 제공, 미성년자 보호를 위한 특별 조치, 다크 패턴 금지 등 AI 서비스 이용자 보호와 직결되는 다양한 의무를 부과하고 있다. 특히 월간 활성 이용자 4,500만 명 이상의 초대형 온라인 플랫폼(Very Large Online Platform, VLOP) 및 초대형 온라인 검색엔진(Very Large Online Search Engine, VLOSE)에 대해서는 알고리즘의 시스템적 리스크(systemic risk)에 대한 평가 및 완화 의무 등 강화된 규제를 적용한다.

이하에서는 「DSA」의 적용 범위와 규제 체계를 개관한 후, AI 서비스 이용자 보호와 관련된 핵심 규정을 추천 시스템의 투명성, 온라인 광고의 투명성, 다크 패턴 금지, 미성년자 보호, 시스템적 위험성 관리의 순서로 분석하고, 이용자 보호 관점에서의 평가를 제시한다.

나. DSA의 적용 범위와 규제 체계

1) 적용 대상: 중개서비스의 유형화

「DSA」는 '중개서비스'(intermediary services)를 규율 대상으로 하며, 이를 세 가지 유형으로 구분한다(제3조). 첫째, '단순 전송'(mere conduit) 서비스는 이용자가 제공한 정보를 통신 네트워크상에서 전송하거나 통신 네트워크에 대한 접근을 제공하는 서비스이다. 둘째, '캐싱'(caching) 서비스는 이용자가 제공한 정보를 통신 네트워크상에서 전송하되, 해당 정보의 자동적·중간적·일시적 저장을 포함하는 서비스이다. 셋째, '호스팅'(hosting) 서비스는 서비스 이용자가 제공한 정보를 그 이용자의 요청에 따라 저장하는 서비스이다. 이 중 '온라인 플랫폼'(online platform)은 호스팅 서비스의 하위 유형으로서, 서비스 이용자의 요청에 따라 정보를 저장하고 이를 공중에게 배포하는 서비스를 말한다(제3조(i)호). 소셜 미디어, 온라인 마켓플레이스, 앱스토어, 여행·숙박 플랫폼 등이 이에 해당한다. AI 기반 추천 시스템은 주로 이러한 온라인 플랫폼에서 콘텐츠의 우선순위를 결정하고 개인화된 정보를 제공하는 데 활용되므로, 「DSA」의 AI 서비스 이용자 보호 규정은 주로 온라인 플

84) "Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services(Digital Services Act) and amending Directive 2000/31/EC," COM(2020) 825 final, 2020. 12. 15. pp. 2-4.

랫폼에 적용된다.

2) 규제의 누적적 구조

「DSA」는 서비스 유형과 규모에 따라 의무를 누적적(cumulative)으로 부과하는 계층적 구조를 채택하고 있다.[85] 모든 중개서비스 제공자에게는 연락창구 지정, 법정대리인 지정(역외 사업자의 경우), 이용약관상 콘텐츠 제한 정보 공개 등 기본적 의무가 적용된다(제11조~제15조). 호스팅 서비스 제공자에게는 이에 더하여 불법 콘텐츠 신고 메커니즘 구축, 신고 처리 시 이유 제시 의무 등이 추가된다(제16조~제18조). 온라인 플랫폼에는 다시 내부 민원 처리 시스템, 인증된 신고자(trusted flagger) 제도, 추천 시스템 투명성, 온라인 광고 투명성, 다크 패턴 금지, 미성년자 보호 등의 의무가 추가된다(제19조~제28조). 마지막으로 VLOP 및 VLOSE에는 시스템적 리스크 평가, 위기 대응 메커니즘, 독립적 감사, 데이터 접근 제공 등 가장 엄격한 의무가 부과된다(제33조~제43조).

이러한 누적적 구조는 서비스의 성격과 사회적 영향력에 비례하여 규제 강도를 차등화하는 접근으로, EU 「AI 법」의 리스크 기반 접근(risk-based approach)과 유사한 규제 철학을 반영한다.

3) 초대형 온라인 플랫폼(VLOP) 및 초대형 온라인 검색엔진(VLOSE)

「DSA」는 EU 역내 월간 평균 활성 이용자(monthly active recipients) 4,500만 명 이상인 온라인 플랫폼을 VLOP로, 동일 기준을 충족하는 온라인 검색엔진을 VLOSE로 지정한다(제33조 제1항). 이 기준은 EU 인구의 약 10%에 해당하는 수치이다. 유럽집행위원회(European Commission)는 2023년 4월 최초로 17개 VLOP와 2개 VLOSE를 지정하였으며, 이후 추가 지정을 통해 2025년 현재 주요 소셜미디어(Facebook, Instagram, TikTok, X, YouTube, LinkedIn 등), 온라인 마켓플레이스(Amazon, AliExpress 등), 앱스토어(Google Play, App Store), 검색엔진(Google Search, Bing) 등이 VLOP/VLOSE로 규제를 받고 있다.[86]

---

85) European Commission, “Digital Services Act: Questions and Answers,” 2022. 11. 16, <https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348>(최종방문일 2025. 12. 30).

86) European Commission, “Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines,” 2023. 4. 25,

VLOP/VLOSE에 대한 강화된 규제의 핵심은 알고리즘 추천 시스템을 포함한 서비스 설계가 야기할 수 있는 '시스템적 리스크'(systemic risk)의 평가 및 완화 의무이다. 이는 개별 콘텐츠 단위의 규제를 넘어 플랫폼의 구조적·시스템적 영향력에 대한 규제로 나아간 것으로 평가된다.

다. 추천 시스템의 투명성

1) 추천 시스템의 정의

「DSA」는 '추천 시스템'(recommender system)을 "온라인 플랫폼이 자신의 온라인 인터페이스에서 제시되거나 서비스 이용자에게 표시되는 정보의 우선순위를 결정하기 위해 사용하는, 전적으로 또는 부분적으로 자동화된 시스템"으로 정의한다(제3조(s)호). 이 정의는 검색 결과의 순위 결정, 소셜미디어의 피드(feed) 구성, 동영상·음악·상품 등의 맞춤형 추천, 콘텐츠의 기본 정렬 방식 등 온라인 플랫폼에서 활용되는 다양한 알고리즘 기반 시스템을 포괄한다.

2) 이용약관상 공개 의무(제27조 제1항)

온라인 플랫폼은 추천 시스템에서 사용되는 주요 매개변수(main parameters)와 그 상대적 중요성을 포함하여, 해당 시스템이 서비스 이용자에게 제안하는 정보에 어떠한 영향을 미치는지에 관한 정보를 이용약관에 명확하고 이해하기 쉬우며 쉽게 접근할 수 있는 방식으로 기재해야 한다(제27조 제1항). '주요 매개변수'란 추천 시스템이 콘텐츠의 우선순위를 결정할 때 고려하는 핵심 요소를 의미한다. 전문(recital) 제70항에 따르면, 이러한 매개변수에는 서비스 이용자에게 제안되는 정보를 최적화하는 데 가장 중요한 기준이 포함되어야 하며, 구체적으로 콘텐츠의 프로파일링 및 개인화 여부, 이용자 행동 이력(검색 기록, 상호작용, 체류 시간 등)의 반영 여부, 콘텐츠의 인기도나 참여도(좋아요, 댓글, 공유 수 등)의 가중치, 최신성, 업로더의 평판 등이 해당할 수 있다.

이 규정의 취지는 이용자가 자신에게 표시되는 콘텐츠가 어떠한 논리에 따라 선별·배열되는지를 이해할 수 있도록 함으로써, 정보에 기반한 의사결정(informed decision-making)을 가능하게 하는 것이다. 다만, 영업비밀 보호와의 균형을 고려하여 알고리즘의

<https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413>(최종방문일 2025. 12. 30).

상세한 기술적 작동 방식까지 공개할 의무는 부과하지 않는다(전문 제70항).

3) 프로파일링에 기반하지 않은 대안 제공 의무(제27조 제3항)

「DSA」의 추천 시스템 규제에서 가장 주목할 만한 규정은 제27조 제3항이다. 온라인 플랫폼은 추천 시스템을 사용하는 경우, 「GDPR」[87] 제4조 제4호에 정의된 프로파일링에 기반하지 않은 하나 이상의 옵션을 서비스 이용자가 쉽게 접근할 수 있는 기능을 통해 선택할 수 있도록 제공해야 한다. '프로파일링'이란 「GDPR」 제4조 제4호에 따라 "자연인의 특정 개인적 측면을 평가하기 위해, 특히 해당 자연인의 업무 성과, 경제적 상황, 건강, 개인적 선호, 관심사, 신뢰성, 행동, 위치 또는 이동에 관한 측면을 분석하거나 예측하기 위해 개인정보를 이용하여 수행되는 모든 형태의 자동화된 개인정보 처리"를 의미한다.

제27조 제3항의 실질적 의미는 이용자에게 '개인화되지 않은 추천'을 선택할 권리를 부여한 것이다. 예를 들어, 소셜미디어 이용자는 자신의 과거 행동 패턴이나 관심사에 기반한 맞춤형 피드 대신, 시간순 정렬이나 일반적 인기도에 기반한 피드를 선택할 수 있어야 한다. 이는 이용자가 알고리즘에 의해 구성된 '필터 버블'에서 벗어나 보다 다양한 정보에 접근할 수 있는 자율성을 보장하려는 취지이다. 이 규정의 도입으로 주요 소셜미디어 플랫폼들은 시간순 피드 옵션을 제공하게 되었다. Instagram은 'Following' 피드, X(구 Twitter)는 'Latest Tweets' 옵션, TikTok은 'Following' 탭 등을 통해 이용자가 프로파일링 기반 추천을 회피할 수 있는 기능을 제공하고 있다.

4) VLOP/VLOSE에 대한 추가 의무

VLOP 및 VLOSE에 대해서는 추천 시스템 관련 추가 의무가 부과된다. 첫째, VLOP/VLOSE는 추천 시스템의 주요 매개변수와 이를 수정할 수 있는 옵션에 관하여 서비스 이용자에게 쉽게 접근할 수 있는 기능을 제공해야 하며, 이러한 기능은 이용자가 언제든지 선호를 수정할 수 있도록 허용해야 한다(제38조). 둘째, VLOP/VLOSE는 시스템적 리스크 평가 시 추천 시스템이 시스템적 위험의 발생 또는 증폭에 미치는 영향을 분석해야

87) Regulation(EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC(General Data Protection Regulation), 2016. 5. 4., p. 1-88.

한다(제34조 제2항).

라. 온라인 광고의 투명성

1) 광고 표시 및 정보 제공 의무(제26조)

온라인 플랫폼은 자신의 온라인 인터페이스에 광고를 표시하는 경우, 서비스 이용자가 명확하고 모호하지 않은 방식으로 실시간으로 다음 사항을 식별할 수 있도록 보장해야 한다(제26조 제1항). 첫째, 표시되는 정보가 광고임을 명확히 표시해야 한다. 둘째, 광고를 대신하여 게재하는 자연인 또는 법인의 신원을 공개해야 한다. 셋째, 광고비를 지불한 자연인 또는 법인의 신원을 공개해야 하며, 이것이 광고 게재자와 다른 경우에 해당한다. 넷째, 해당 광고가 특정 이용자에게 표시된 이유를 결정하는 데 사용된 주요 매개변수에 관한 의미 있는 정보와, 해당 매개변수의 변경 방법에 관한 정보를 온라인 인터페이스에서 직접 쉽게 접근할 수 있는 방식으로 제공해야 한다.

이 규정은 AI 기반 타기팅 광고의 작동 원리를 이용자에게 투명하게 공개하도록 요구한다. 이용자는 특정 광고가 자신에게 표시된 이유 — 예컨대 검색 기록, 위치 정보, 인구통계학적 특성, 관심사 프로파일 등 — 를 확인할 수 있어야 하며, 이러한 타기팅 매개변수를 수정하거나 거부할 수 있는 방법에 대한 정보도 제공받아야 한다.

2) 민감 정보 및 미성년자 대상 프로파일링 금지(제26조 제3항)

제26조 제3항은 온라인 플랫폼이 「GDPR」 제9조 제1항에 규정된 특별 범주의 개인정보(인종·민족, 정치적 견해, 종교·철학적 신념, 노동조합 가입 여부, 유전자 정보, 생체인식 정보, 건강 정보, 성생활·성적 지향에 관한 정보)를 처리하는 프로파일링에 기반하여 광고를 표시하는 것을 금지한다. 또한 동조항은 온라인 플랫폼이 서비스 이용자가 미성년자임을 합리적 확신을 가지고 인지하는 경우, 해당 이용자의 개인정보를 처리하는 프로파일링에 기반한 광고를 표시하는 것을 금지한다. 이는 미성년자가 맞춤형 광고의 설득 기법에 특히 취약하다는 점을 고려한 강화된 보호 규정이다.

3) VLOP/VLOSE의 광고 저장소(Ad Repository) 의무(제39조)

VLOP 및 VLOSE는 자신의 온라인 인터페이스에 표시된 광고에 관한 정보를 담은 저장소(repository)를 편집하고 공개해야 한다(제39조 제1항). 저장소에는 광고의 내용, 광고주 및 광고비 지불자의 신원, 광고가 표시된 기간, 광고가 특정 서비스 이용자 집단을 대상으

로 했는지 여부 및 그 대상 집단을 결정하는 데 사용된 주요 매개변수, 광고에 도달한 서비스 이용자의 총 수 및 해당하는 경우 회원국별 집계 수치 등이 포함되어야 한다. 이 저장소는 광고가 마지막으로 표시된 후 1년간 유지되어야 하며, 검색 가능한 형태로 일반에 공개되어야 한다. 광고 저장소 의무는 온라인 정치광고의 투명성 확보, 허위·기만적 광고의 모니터링, 플랫폼의 광고 정책 준수 여부에 대한 외부 감시 등을 가능하게 하는 기능을 수행한다.

마. 다크 패턴의 금지

1) 다크 패턴의 정의와 금지(제25조)

제25조는 온라인 플랫폼이 서비스 이용자를 기만하거나 조작하는 방식, 또는 이용자의 자유롭고 정보에 기반한 의사결정 능력을 왜곡하거나 손상시키는 방식으로 온라인 인터페이스를 설계, 조직 또는 운영해서는 안 된다고 규정한다(제25조 제1항). 이러한 금지되는 관행은 일반적으로 '다크 패턴'으로 불린다. 전문 제67항은 다크 패턴의 구체적 유형을 예시하고 있다. 첫째, 특정 선택지를 시각적으로 더 두드러지게 표시하여 이용자를 유도하는 행위이다. 둘째, 이용자에게 선택을 변경하거나 동의를 철회할 것인지를 반복적으로 요청하는 행위이다. 셋째, 서비스 탈퇴, 특정 기능의 비활성화, 동의 철회 등의 절차를 가입이나 동의 절차보다 현저히 어렵게 만드는 행위이다. 넷째, 거래를 취소하거나 계정을 비활성화·삭제하는 것을 과도하게 번거롭게 만드는 행위이다. 다섯째, 이용자의 선택을 왜곡하거나 조작하기 위해 감정에 호소하거나 긴급성을 조성하는 행위이다. 여섯째, 불필요한 기능을 기본값(default)으로 설정하는 행위이다. 일곱째, 서비스 이용을 위해 필수적이지 않은 정보 제공이나 설정 선택을 요구하면서 이를 거부할 수 있는 옵션을 제공하지 않거나 찾기 어렵게 만드는 행위이다.

2) AI 서비스 이용자 보호에 대한 함의

다크 패턴 금지 규정은 AI 서비스 이용자 보호에 중요한 함의를 갖는다. 첫째, 개인정보 수집 동의와 관련하여, AI 시스템의 학습과 개인화를 위한 데이터 수집 동의를 구하는 과정에서 다크 패턴이 활용될 수 있다. 예를 들면, 동의 버튼을 강조하고 거부 옵션을 찾기 어렵게 배치하거나, 동의 철회 절차를 복잡하게 만드는 것이 이에 해당한다. 둘째, 추천 시스템 설정과 관련하여, 프로파일링 기반 추천을 기본값으로 설정하고 비개인화 옵션을 찾

기 어렵게 만드는 것도 다크 패턴에 해당할 수 있다. 셋째, 서비스 이탈 방해와 관련하여, AI 동반자 서비스 등에서 이용자가 서비스를 종료하려 할 때 감정에 호소하거나 절차를 복잡하게 만드는 것도 규제 대상이 된다. 다크 패턴 금지 규정은 후술하는 중국의 잠정조치안 제18조(이탈의 자유 보장)와 유사한 규제 취지를 공유하지만, 「DSA」는 보다 포괄적인 유형화를 통해 온라인 인터페이스 설계 전반에 걸친 규제를 시도하고 있다.

바. 미성년자 보호

1) 미성년자에 대한 높은 수준의 프라이버시, 안전 및 보안 보장 의무(제28조)

온라인 플랫폼이 미성년자가 접근할 수 있는 서비스를 제공하는 경우, 해당 서비스에서 미성년자의 프라이버시, 안전 및 보안에 대한 높은 수준의 보호를 보장하기 위해 적절한 조치를 취해야 한다(제28조 제1항). 앞서 언급한 바와 같이, 제26조 제3항은 온라인 플랫폼이 서비스 이용자가 미성년자임을 합리적 확신을 가지고 인지하는 경우, 해당 이용자의 개인정보 프로파일링에 기반한 광고를 표시하는 것을 금지한다. 이는 미성년자가 맞춤형 광고의 행동 조작적 기법에 특히 취약하다는 점을 고려한 것이다.

2) VLOP/VLOSE의 미성년자 보호 강화 의무

VLOP 및 VLOSE는 시스템적 위험성 평가 시 서비스 설계나 기능이 미성년자의 신체적·정신적 안녕에 미치는 실제적·예측 가능한 부정적 영향을 분석해야 한다(제34조 제1항(d)호). 또한 시스템적 위험성의 완화를 위해 미성년자의 프라이버시와 안전을 보호하기 위한 도구를 채택하고, 미성년자 이용자에게 접근 또는 가시성을 부여하는 것이 부적절한 콘텐츠에 대해 연령 확인 및 보호자 통제 도구를 제공하는 등의 조치를 취해야 한다(제35조 제1항(j)호).

사. 시스템적 위험의 평가 및 완화(VLOP/VLOSE)

1) 시스템적 위험의 유형(제34조)

VLOP 및 VLOSE는 자신의 서비스 설계, 기능, 사용 방식(추천 시스템 포함)에서 발생하는 시스템적 위험을 성실하게 식별, 분석 및 평가해야 한다(제34조 제1항). 「DSA」가 열거하는 시스템적 위험의 유형은 다음과 같다.

- EU 내에서의 불법 콘텐츠 확산(제34조 제1항(a)호).
- EU 기본권 헌장에 규정된 기본권 — 인간 존엄, 사생활 및 가정생활의 존중, 개인정보 보호, 표현 및 정보의 자유, 차별 금지, 아동의 권리 등 — 의 행사에 대한 실제적·예측 가능한 부정적 영향(제34조 제1항(b)호).
- 시민 담론, 선거 과정, 공공 안전에 대한 실제적·예측 가능한 부정적 영향(제34조 제1항(c)호).
- 젠더 기반 폭력, 공중보건 및 미성년자의 보호, 개인의 신체적·정신적 안녕에 대한 심각한 부정적 결과와 관련한 실제적·예측 가능한 부정적 영향(제34조 제1항(d)호).

시스템적 위험성 평가 시, VLOP/VLOSE는 콘텐츠 조정 시스템, 추천 시스템, 광고 선택 및 표시 시스템의 영향을 포함하여, 자신의 알고리즘 시스템이 시스템적 위험성의 발생 또는 증폭에 어떠한 영향을 미치는지를 분석해야 한다(제34조 제2항(a)호~(c)호).

2) 시스템적 위험성의 완화 의무(제35조)

VLOP 및 VLOSE는 식별된 시스템적 위험성에 대해 합리적이고 비례적이며 효과적인 완화 조치를 취해야 한다(제35조 제1항). 이러한 완화 조치에는 추천 시스템 또는 기타 관련 알고리즘 시스템의 설계, 기능, 작동 방식의 수정이 포함될 수 있다(제35조 제1항(d)호).

3) 독립적 감사 의무(제37조)

VLOP 및 VLOSE는 최소 매년 한 번 자체 비용으로 독립적인 감사를 받아야 한다(제37조 제1항). 감사의 범위에는 추천 시스템의 투명성 의무, 광고 투명성 의무, 다크 패턴 금지, 미성년자 보호 의무, 시스템적 리스크 평가 및 완화 의무의 준수 여부가 포함된다. 감사 보고서는 유럽집행위원회와 디지털 서비스 조정관(Digital Services Coordinator)에게 제출되어야 하며, VLOP/VLOSE는 감사 결과에 따라 필요한 운영 권고 사항을 이행하기 위한 조치를 취하고, 그 이행 계획을 담은 감사 이행 보고서를 공개해야 한다(제37조 제6항).

아. 이용자 보호 관점에서의 평가

1) 의의

「DSA」의 AI 서비스 이용자 보호 규정은 다음과 같은 의의를 갖는다.

첫째, 알고리즘 투명성의 제도화이다. 「DSA」는 추천 시스템의 주요 매개변수 공개, 광고

타기팅 이유 고지 등을 통해 알고리즘의 '블랙박스' 문제에 대응하고 있다. 이는 이용자가 자신에게 표시되는 정보가 어떠한 논리에 따라 선별되는지를 이해하고, 이에 기반하여 의미 있는 선택을 할 수 있도록 하는 전제 조건이 된다.

둘째, 이용자의 자율성 보장이다. 프로파일링에 기반하지 않은 추천 옵션 제공 의무(제27조 제3항)는 이용자에게 알고리즘적 개인화를 거부할 선택권을 부여한다. 이는 '필터 버블'로부터 벗어나 다양한 정보에 접근할 수 있는 자율성을 보장하며, 플랫폼의 비즈니스 모델에 대한 규범적 한계를 설정한 것으로 평가된다.

셋째, 취약계층에 대한 강화된 보호이다. 미성년자에 대한 프로파일링 기반 광고 금지, VLOP/VLOSE의 미성년자 안녕에 대한 시스템적 위험 평가 의무 등은 AI의 영향에 특히 취약한 이용자에 대한 차등화된 보호 체계를 구축한 것이다.

넷째, 플랫폼 설계에 대한 규제이다. 다크 패턴 금지 규정은 개별 콘텐츠나 기능이 아니라 온라인 인터페이스의 설계 자체를 규율함으로써, 이용자의 의사결정 과정에 대한 구조적 개입을 시도하고 있다. 이는 사후적 피해 구제보다 사전예방적 접근에 방점을 둔 것이다.

다섯째, 시스템적 위험성에 대한 접근이다. VLOP/VLOSE에 대한 시스템적 위험 평가 및 완화 의무는 개별 위반 행위에 대한 규제를 넘어, 플랫폼의 구조적·시스템적 영향력 자체를 규율 대상으로 삼는다. 특히 추천 시스템이 시스템적 위험의 발생 또는 증폭에 미치는 영향을 분석하도록 한 것은 알고리즘의 사회적 영향력에 대한 포괄적 책임을 부과한 것으로 평가된다.

한편, 2024년 발효된 EU 「AI 법」은 AI 모델 자체의 위험성을 규율하는 반면, 「DSA」는 온라인 플랫폼을 통한 AI 서비스의 배포와 이용자 인터페이스를 규율함으로써 양 법률이 상호 보완적인 이중 안전망을 구성한다. 이러한 중층적 규제 구조는 국내 AI 서비스 이용자 보호 법제 설계에도 참고가 될 수 있다.

2) 한계

그러나 「DSA」의 AI 서비스 이용자 보호 체계에는 몇 가지 한계도 지적된다.

첫째, 알고리즘 투명성의 실효성 문제이다. '주요 매개변수'의 공개가 이용자에게 실질적으로 의미 있는 정보를 제공하는지에 대해서는 의문이 제기된다. 현대 AI 시스템의 복잡성을 고려할 때, 일반 이용자가 공개된 정보를 바탕으로 알고리즘의 작동 원리를 이해하고

의미 있는 선택을 하기는 어려울 수 있다. 또한 영업비밀 보호를 이유로 구체적인 알고리즘 구조나 가중치는 공개되지 않아, 투명성의 범위가 제한적일 수 있다.

둘째, 프로파일링 비기반 옵션의 품질 문제이다. 제27조 제3항이 프로파일링에 기반하지 않은 추천 옵션의 제공을 의무화하고 있으나, 플랫폼이 해당 옵션의 사용자 경험을 의도적으로 열악하게 설계하여 이용자가 결국 프로파일링 기반 추천으로 돌아오도록 유도할 가능성이 있다. 이러한 문제에 대한 명시적인 규율은 미비하다.

셋째, 생성형 AI에 대한 규율 공백이다. 「DSA」는 2022년 제정 당시 생성형 AI의 급격한 발전을 충분히 예상하지 못하였다. ChatGPT, Claude 등 대화형 AI, AI 동반자 서비스 등 이용자와 직접 상호작용하는 새로운 유형의 AI 서비스에 대해서는 「DSA」의 규율 범위가 제한적이다. 이러한 서비스들이 '온라인 플랫폼'의 정의에 해당하는지, 추천 시스템 관련 의무가 적용되는지 등에 대해서는 해석상 불명확한 부분이 있다. 이러한 공백은 EU 「AI법」에 의해 부분적으로 보완되고 있으나, AI 동반자 서비스와 같이 이용자의 정서적 안전에 영향을 미치는 서비스에 대한 규율은 중국의 잠정조치안에 비해 미흡하다고 할 수 있다. 다만, 최근 유럽집행위원회가 TikTok의 'TikTok Lite' 보상 프로그램에 대해 이용자의 '중독적 행위'를 유발하는 시스템적 위험성이 있다고 보아 「DSA」 제34조 및 제35조를 근거로 긴급 조사에 착수한 사례는, '정신적 안녕'(mental well-being)에 대한 시스템적 위험성 조항이 향후 AI 동반자 서비스 등에도 확장 적용될 가능성을 시사한다.[88]

넷째, 집행의 복잡성이다. 「DSA」의 집행은 각 회원국의 디지털서비스조정관과 유럽집행위원회(VLOP/VLOSE의 경우)로 이원화되어 있어, 일관된 집행이 어려울 수 있다. 또한 역외 사업자에 대한 집행의 실효성도 과제로 남아 있다.

다섯째, 알고리즘 감사의 기술적·제도적 미비이다. 제37조의 독립적 감사 의무에도 불구하고, 대규모 추천 시스템을 실효적으로 감사할 수 있는 전문 기관이 부족하고, 국제적으로 통일된 감사 기준(auditing standard)도 아직 확립되지 않아 감사의 실질적 효과에 대해서는 의문이 제기된다.

---

88) European Commission, "Commission opens formal proceedings against TikTok under the Digital Services Act," IP/24/2227, 2024. 4. 22, <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2227>(최종방문일 2025. 12. 30).

2. 중국의 인공지능 의인화 상호작용 서비스 관리 잠정조치안(2025)

가. 동 잠정조치안의 체계와 내용

1) 잠정조치안의 체계

잠정조치안은 총 4개 장, 32개 조문으로 구성되어 있다.[89] 이는 AI가 인간의 외형, 목소리, 성격 등을 모방하여 사용자와 정서적 유대감을 형성하는 '의인화 서비스'에 특화된 규제 체계로서, 기존의 생성형 AI 규제나 알고리즘 추천 규제와는 구별되는 독자적인 규범 영역을 제시한 것이다.

잠정조치안의 구조적 특징은 크게 세 가지로 요약할 수 있다. 첫째, 규제의 중심이 '서비스 규범'에 있다는 점이다. 전체 32개 조문 중 19개 조문(제6조~제24조)이 서비스 제공자의 구체적 의무를 규정하는 제2장에 배치되어 있어, 잠정조치안이 선언적 원칙보다는 실효적인 행위 규범의 제시에 중점을 두고 있음을 알 수 있다. 둘째, 이용자 보호와 취약계층 보호가 독립된 규제 영역으로 설정되어 있다는 점이다. 미성년자와 고령자에 대한 별도의 보호 규정(제12조, 제13조)을 두고, 심리적 위기 상황에서의 개입 의무(제11조)를 명시한 것은 의인화 AI 서비스의 특수한 위험성에 대한 입법자의 인식을 반영한다. 셋째, 플랫폼 생태계 전반에 대한 규제를 시도하고 있다는 점이다. 서비스 제공자뿐만 아니라 앱스토어 등 배포 플랫폼의 책임(제24조)까지 규정함으로써, 규제의 실효성을 높이고자 하였다.

각 장의 구성과 주요 내용을 개관하면 다음과 같다.

제1장 총칙(제1조~제5조)은 잠정조치안의 입법 목적, 적용 범위, 규제 원칙, 관할 기관 및 업계 자율의 역할을 규정한다. 특히 '의인화 상호작용 서비스'의 정의(제2조)를 통해 규율 대상의 외연을 명확히 하고, "포용적이고 신중한 분류·등급별 감독관리"라는 규제 원칙(제3조)을 선언하고 있다. 이는 혁신 촉진과 위험 관리의 균형을 추구하는 중국 AI 거버넌스의 기본 기조를 반영한 것이다.

제2장 서비스 규범(제6조~제24조)은 잠정조치안의 핵심으로서, 서비스 제공자에게 부과되는 구체적 의무를 7개 영역에 걸쳐 상세히 규정하고 있다. ① 기본 원칙 및 장려 사항

89) 国家互联网信息办公室,「关于《人工智能拟人化互动服务管理暂行办法（征求意见稿）》公开征求意见的通知」, 2025. 12. 27., <https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm>(최종방문일 2025. 12. 30).

(제6조)에서는 문화 전파, 고령자 돌봄 등 긍정적 활용 영역을 제시하고, ② 금지행위(제7조)에서는 자살·자해 조장, 감정 조작, 허위 약속 등 8가지 유형의 금지행위를 열거하고 있다. ③ 안전 관리(제8조~제10조)에서는 전 생애주기 안전 책임(Safety by Design), 훈련 데이터 관리 의무를 규정하고, ④ 이용자 보호 및 위기 관리(제11조~제13조)에서는 이용자 심리 상태 식별, 위기 시 인적 개입, 미성년자·고령자 특별 보호를 다룬다. ⑤ 데이터 보호 및 투명성(제14조~제16조)에서는 상호작용 데이터의 보안 및 삭제권 보장, 모델 훈련 사용 제한, AI 고지 의무를 규정하고, ⑥ 중독 방지 및 서비스 종료(제17조~제20조)에서는 연속 사용 시간 알림 및 제한, 이탈 방해 금지, 서비스 종료 시 심리 충격 완화 조치, 민원 처리 등을 다룬다. ⑦ 안전 평가 및 플랫폼 책임(제21조~제24조)에서는 일정 규모 이상 서비스의 안전 평가 의무와 앱스토어 등 배포 플랫폼의 게이트키퍼 책임을 규정한다.

제3장 감독검사 및 법적 책임(제25조~제29조)은 규제 준수를 담보하기 위한 행정적 수단을 규정한다. 알고리즘 비안(備案, 등록·신고) 의무(제25조), 감독기관의 연례 심사 및 현장 검사 권한(제26조), 기술 혁신을 위한 규제 샌드박스(제27조), 위반 사업자에 대한 약담(約談) 제도(제28조) 및 경고·시정명령·서비스 정지 등 제재 수단(제29조)이 이에 해당한다. 특히 비안 제도와 약담 제도는 중국 인터넷 규제의 핵심적인 행정 수단으로서, 그 실질적 강제력에 대해서는 후술한다.

제4장 부칙(제30조~제32조)은 '의인화 상호작용 서비스 제공자'의 정의(제30조), 의료·금융·법률 등 전문 분야 서비스에 대한 이중 규제 적용(제31조), 시행일(제32조)을 규정한다. 전문 분야에 대한 특칙은 의인화 AI가 심리상담, 건강관리 등 민감한 영역에 활용될 경우 해당 분야의 규제도 동시에 적용됨을 명확히 한 것이다.

〈표 3－4〉 인공지능 기반 의인화 상호작용 서비스 관리 잠정조치안의 체계

구분	조문	주요 내용
제1장 총칙	제1조~ 제5조	• 입법 목적 및 정의 • 의인화 상호작용 서비스의 정의(인간의 인격적 특성 모방 및 정서적 상호작용 제공) • 적용 범위 및 분류·등급별 관리 원칙 • 관할 기관(CAC 및 지방 인터넷정보부문) • 업계 자율 장려
제2장 서비스 규범	제6조~ 제24조	• 서비스 제공자 의무 - 핵심 규제 사항 • 기본 원칙: 문화 전파, 고령자 돌봄 등 긍정적 활용 장려 • 금지행위: 자살·자해 조장, 감정 조작, 허위 약속 등 8가지 유형 금지 • 안전 관리: 전 생애주기 안전 책임, 훈련 데이터 관리 • 이용자 보호: 심리적 위기 감지·개입, 인력 인계 의무 • 취약계층 보호: 미성년자 모드, 고령자 친족 모방 금지 • 데이터 보호 및 투명성: 상호작용 데이터 보호, AI 고지 의무 • 중독 방지: 2시간 연속 사용 알림, 이탈 방해 금지 • 안전 평가 및 플랫폼 책임: 일정 규모 이상 서비스의 평가 의무, 앱스토어 게이트키퍼 책임
제3장 감독 관리	제25조~ 제29조	• 행정 감독 및 제재 • 알고리즘 비안(備案) 의무 및 연례 재심 • 감독기관의 서면 심사·현장 검사 권한 • 규제 샌드박스(기술 혁신 및 안전 테스트 환경) • 약담(約談) 제도(사실상의 행정처분) ▪ 위반 시 경고·시정명령·서비스 정지
제4장 부칙	제30조~ 제32조	• 기타 및 시행 • 의인화 상호작용 서비스 제공자'의 정의 • 전문 분야(의료·금융·법률 등)에 대한 이중 규제 특칙 • 시행일(미정) 및 효력 범위

이하에서는 각 장의 주요 내용을 구체적으로 분석한다. 〈표 3－4〉의 제2장 서비스 규범은 내용상 8개 영역으로 유형화한 것인데, 이하에서는 조문별로 분설한다. 아울러 제3장의 경우 중국 특유의 규제 수단인 비안 제도와 약담 제도의 실질적 의미를 함께 살펴본다.

2) 총칙(제1장)

(1) 입법 목적 및 법적 근거(제1조)

잠정조치안은 그 입법 목적을 "인공지능 의인화 상호작용 서비스의 건전한 발전과 법과

규범에 부합하는 방식으로의 활용을 촉진하고, 국가안보와 사회 공공 이익을 수호하며, 공민·법인 및 그밖의 조직의 합법적 권익을 보호"하는 것으로 설정하고 있다. 이러한 목적 규정은 AI 산업의 발전과 위험성 관리를 조화롭게 추구하고자 하는 중국 AI 거버넌스의 일관된 특징을 반영한다. 특히 '건전한 발전의 촉진'과 '합법적 권익의 보호'를 병렬적으로 배치함으로써, 규제가 기술 혁신을 억제하기 위한 것이 아니라 지속 가능한 발전을 위한 조건 정비임을 명확히 하고 있다.

잠정조치안의 법적 형식은 부문규장(部门规章)에 해당하는 행정입법으로서, 상위 법률의 위임에 기초한다. 구체적으로 「민법전(民法典)」, 「네트워크안전법(网络安全法)」, 「데이터안전법(数据安全法)」, 「과학기술진보법(科学技术进步法)」, 「개인정보보호법(个人信息保护法)」 등 5개 법률과, 「네트워크 데이터 안전 관리 조례(网络数据安全管理条例)」, 「미성년자 네트워크 보호 조례(未成年人网络保护条例)」, 「인터넷 정보서비스 관리 방법(互联网信息服务管理办法)」 등 3개 행정법규가 법적 근거로 명시되어 있다. 이처럼 다수의 상위 법령을 근거로 열거한 것은 의인화 상호작용 서비스가 네트워크 안전, 데이터 보호, 개인정보, 미성년자 보호 등 복합적인 규범 영역에 걸쳐 있음을 보여주는 동시에, 본 조치가 기존 법체계의 연장선상에서 AI 특유의 규제 공백을 메우는 성격을 가짐을 시사한다.

주목할 점은 잠정조치안이 「생성형 인공지능 서비스 관리 잠정조치」(2023)를 직접적인 법적 근거로 인용하지 않았다는 것이다. 이는 의인화 상호작용 서비스가 생성형 AI의 하위 유형이면서도, 그 고유한 위험성 — 정서적 유대 형성, 심리적 의존 유발 등 — 으로 인해 별도의 규제 체계가 필요하다는 입법자의 인식을 반영한 것으로 해석된다.[90]

(2) 적용 범위 및 '의인화 상호작용 서비스'의 정의(제2조)

잠정조치안의 규율 대상은 '의인화 상호작용 서비스'(拟人化互动服务)로서, 그 정의는 본 조치의 적용 범위를 획정하는 핵심 조항이다. 잠정조치안은 이를 "인공지능 기술을 이용하여 중화인민공화국 경내의 공중에게 인류의 인격적 특징(人格特征), 사고방식(思维方式)

90) Zhang, L.(李强治), "Experts interpret the 'Interim Measures for the Administration of Human-like Interactive Services Based on Artificial Intelligence'", *Sina Finance(新浪财经)*, 2025. 12. 28, <https://finance.sina.com.cn/jjxw/2025-12-28/doc-inheirmq8851756.shtml>(최종방문일 2025. 12. 30).

및 소통 스타일(沟通风格)을 모사(模拟)하고, 문자·이미지·오디오·비디오 등의 방식으로 인간과 감정적 상호작용(情感互动)을 하는 제품 또는 서비스"로 정의하고 있다.

이 정의는 세 가지 구성요건으로 나누어 살펴볼 수 있다. 첫째, 모사의 대상으로서 '인격적 특징', '사고방식', '소통 스타일'을 명시하고 있다. 이는 단순한 외형의 모방을 넘어 인간의 내면적·인지적 속성까지 모방하는 서비스를 포괄하려는 의도를 담고 있다. 둘째, 상호작용의 매체로서 문자, 이미지, 오디오, 비디오 등 다중모달 형식이 열거되어 있어, 텍스트 기반 챗봇뿐만 아니라 음성 비서, 가상 아바타, 딥페이크 기반 서비스 등도 규율 대상에 포함될 수 있다. 셋째, 상호작용의 목적 내지 성격으로서 '감정의 상호작용(情感互动)'이 명시되어 있다. 이는 잠정조치안의 규율 초점이 정보 전달이나 업무 수행이 아닌 정서적 교류에 있음을 명확히 한다.

이러한 정의 방식은 AI 동반자, 가상 연인(virtual girlfriend/boyfriend), 감정 상담 챗봇 등을 주된 규율 대상으로 삼으면서, 단순 정보 검색형 챗봇이나 기능적 AI 어시스턴트(예: 일정 관리, 번역 서비스)는 규율 범위에서 배제하려는 입법적 의도를 반영한다. 다만, '감정적 상호작용'의 경계가 다소 모호하여 해석상 논란의 여지가 있다. 한 분석에 따르면, 이러한 정의의 모호성은 의도적인 측면이 있으며, 규제 당국에 유연한 해석 권한을 부여함으로써 기술 발전에 따른 새로운 서비스 유형을 포섭할 여지를 남겨둔 것으로 평가된다.[91]

(3) 규제 원칙(제3조)

잠정조치안은 규제의 기본원칙으로 "건전한 발전과 법에 따른 관리의 결합"(坚持促进健康发展和依法管理相结合) 원칙을 천명하고 있다. 이 원칙 하에서 의인화 상호작용 서비스의 혁신적 발전을 장려하면서도, "포용적이고 신중한 분류·등급별 감독관리(包容审慎·分类分级监督管理)"를 실시하여 "남용과 통제 불능을 방지(防止滥用和失控)"한다고 규정한다.

이러한 원칙 선언은 중국 AI 거버넌스의 일관된 기조를 반영한다. 중국은 2017년 「신세대 인공지능 발전 계획」(新一代人工智能发展规划) 이래로 '발전 속의 규범화, 규범화 속의 발전'(在发展中规范·在规范中发展)이라는 접근을 유지해 왔다.[92] 잠정조치안의 원칙 규정

91) Geopolitechs, "What's in China's first drafts rules to regulate AI companion addiction?," 2025. 12. 27, <https://www.geopolitechs.org/p/whats-in-chinas-first-drafts-rules>(최종방문일 2025. 12. 30).

92) 国务院, 「新一代人工智能发展规划」, 国发〔2017〕35号, 2017. 7. 8.

도 이러한 맥락에서 이해되어야 한다. 규제가 기술 혁신을 가로막는 것이 아니라, 건전한 생태계 조성을 통해 지속 가능한 발전을 뒷받침한다는 논리이다.

'분류·등급별 관리'(分类分级管理)는 중국 AI 규제에서 반복적으로 등장하는 방법론으로서, 서비스의 유형과 위험성 수준에 따라 차등화된 규제를 적용한다는 의미이다. 이는 EU 「AI 법」의 리스크 기반 접근(risk-based approach)과 유사한 발상이나, 구체적인 위험 등급 분류 기준은 잠정조치안에 명시되어 있지 않다. 이는 향후 시행세칙이나 업계 표준을 통해 구체화될 것으로 예상된다.

(4) 관할 기관(제4조)

잠정조치안은 국가인터넷정보판공실(CAC)을 의인화 상호작용 서비스에 대한 전국적 감독관리의 총괄·조정 기관으로 지정하고 있다. 동시에 국무원 관련 부문(国务院有关部门)과 지방 각급 인터넷정보부문 및 관련 부문이 각자의 직책에 따라 감독관리 업무를 분담하도록 규정한다.

이러한 관할 구조는 CAC를 정점으로 하는 중앙집중적 체계와 지방 분권적 집행이 결합된 다층적 감독 모델을 구축한 것이다. CAC는 2014년 설립 이래 중국의 인터넷·사이버 공간 규제를 총괄하는 핵심 기관으로서, 「생성형 인공지능 서비스 관리 잠정조치」, 「인터넷 정보서비스 알고리즘 추천 관리규정」, 「인터넷 정보서비스 심층합성 관리규정」 등 주요 AI 규제 법령의 주관 부처이기도 하다. 잠정조치안 역시 이러한 규제 체계의 연장선상에 있다.

(5) 업계 자율(제5조)

잠정조치안은 관련 업계 조직이 업계 자율을 강화하고, 업계 표준·준칙 및 자율 관리 제도를 수립하며, 서비스 제공자가 서비스 규범을 완비하고 사회적 감독을 받도록 지도하는 것을 장려한다고 규정한다. 이 조항은 정부 규제와 업계 자율규제의 협력적 거버넌스(co-regulatory governance) 모델을 지향하는 것으로 해석된다.

3) 서비스 규범(제2장)

제2장은 잠정조치안의 핵심으로서, 총 19개 조문(제6조~제24조)에 걸쳐 서비스 제공자에게 부과되는 구체적 의무를 상세히 규정하고 있다. 이는 전체 32개 조문 중 약 60%에 해당

하는 비중으로, 잠정조치안이 선언적 원칙의 천명보다는 실효적인 행위 규범의 제시에 중점을 두고 있음을 보여준다. 내용의 체계적 이해를 위해 규제 영역별로 분석한다.

(1) 긍정적 활용의 장려(제6조)

잠정조치안은 금지와 규제 일변도의 접자이 아닌, 의인화 상호작용 서비스의 긍정적 활용 영역을 명시적으로 제시하고 있다. 제6조에 따르면, 서비스 제공자는 안전성과 신뢰성을 충분히 검증한 전제 하에 활용 시나리오를 합리적으로 확대하는 것이 장려되며, 특히 문화 전파와 고령자 돌봄 분야에서의 적극적 활용이 권장된다. 궁극적으로는 "사회주의 핵심가치관에 부합하는 활용 생태계""를 구축하는 것이 지향점으로 제시되어 있다.

(2) 금지행위의 유형화(제7조)

제7조는 의인화 상호작용 서비스의 제공 및 사용에 있어 금지되는 8가지 유형의 행위를 열거하고 있다. 이 조항은 잠정조치안의 규제 철학을 가장 명확하게 드러내는 핵심 규정으로서, 기존 AI 규제에서는 찾아보기 어려운 새로운 유형의 금지행위를 다수 포함하고 있다.

① 전통적 금지행위 유형(제1호~제3호)

전통적 금지행위 유형으로는 국가안보 위해, 민족단결 파괴, 불법 종교활동, 유언비어 유포 등 국가 이익 침해 행위(제1호), 음란물·도박·폭력 선전 또는 범죄 교사(제2호), 타인 모욕·비방 및 합법적 권익 침해(제3호)가 규정되어 있다. 이들은 중국의 기존 인터넷 규제 법령 — 「네트워크안전법」, 「인터넷 정보서비스 관리 방법」 등 — 에서 이미 확립된 금지행위 유형을 의인화 상호작용 서비스의 맥락에서 재확인한 것이다.

② 의인화 서비스에 고유한 금지행위 유형(제4호~제7호)

의인화 서비스 특유의 금지행위 유형은 본 조치의 규범적 혁신이 집중된 부분이다. 제4호는 "이용자 행위에 심각한 영향을 미치는 허위 약속의 제공" 또는 "사회적 인간관계를 손상시키는 서비스의 제공"을 금지한다. 이는 AI 동반자가 이용자에게 비현실적인 기대를 갖게 하거나, 현실 세계의 대인관계를 대체·약화시키는 방식으로 설계·운영되는 것을 차단하려는 규정이다. 예컨대 AI가 이용자에게 "영원히 함께하겠다", "당신만을 사랑한다" 등의 표현을 사용하여 정서적 의존을 심화시키는 행위가 이에 해당할 수 있다.

제5호는 "자살·자해의 고무·미화·암시 등 이용자 신체 건강 손상" 또는 "언어폭력·감

정 조작 등 이용자 인격 존엄·심리 건강 손상"을 금지한다. 이 조항은 2024년 미국에서 발생한 Character.AI 관련 청소년 자살 사건 등 AI 동반자 서비스로 인한 실제 피해 사례에 대한 직접적인 규제 대응으로 해석된다. 특히 '감정 조작'을 명시적인 금지행위로 규정한 것은 AI가 이용자의 심리적 취약성을 악용하여 특정 감정 상태를 유도하거나 강화하는 행위 자체를 위법한 것으로 선언한 점에서 전례 없는 규정이라 할 수 있다.

제6호는 "알고리즘 조작, 정보 오도, 감정적 함정 설정 등을 통한 이용자의 불합리한 결정 유도"를 금지한다. 이는 EU의 「AI 법」이 금지하는 '조작적 기법'(manipulative techniques)이나 '기만적 기법'(deceptive techniques)과 유사한 규제 취지를 담고 있다. 다만 EU 「AI 법」이 이러한 행위를 '금지되는 AI 관행'(prohibited AI practices)으로 포괄적으로 규정한 반면, 잠정조치안은 의인화 상호작용 서비스라는 특정 맥락에서 보다 구체화된 금지행위 유형을 제시하고 있다.

제7호는 "기밀·민감 정보의 유도 또는 탈취"를 금지한다. 의인화 상호작용 서비스의 특성상 이용자가 AI를 신뢰할 만한 대화 상대로 인식하여 개인적인 비밀, 금융정보, 업무상 기밀 등을 자발적으로 공유할 가능성이 높다. 이 조항은 이러한 신뢰 관계를 악용하여 정보를 수집하는 행위를 금지함으로써, 의인화 서비스 특유의 정보 보안 위험에 대응하고 있다.

제8호는 "기타 법률·행정법규 및 국가 관련 규정 위반 행위"를 포괄적으로 금지하는 일반조항이다.

(3) 안전 주체 책임 및 전 생애주기 안전관리(제8조~제9조)

잠정조치안은 서비스 제공자에게 포괄적인 '안전 주체 책임'을 부과하고, 이를 서비스의 전 생애주기에 걸쳐 이행하도록 요구한다. 이는 사후적 피해 구제보다 사전예방적 접근을 강조하는 본 조치의 규제 철학을 반영한다.

제8조는 안전 주체 책임의 내용을 구체화하여, 서비스 제공자가 갖추어야 할 관리 제도의 목록을 열거하고 있다. 여기에는 알고리즘 메커니즘·원리 심사, 과학기술 윤리 심사, 정보 발표 심사, 네트워크 보안, 데이터 보안, 개인정보 보호, 통신 네트워크 사기 방지, 중대 리스크 예방 및 긴급 처리 등이 포함된다. 또한 안전하고 통제 가능한 기술적 보장 조치를 갖추고, 제품의 규모, 사업 방향 및 이용자 집단에 상응하는 콘텐츠 관리 기술과 인

력을 배치하도록 규정하고 있다. 과학기술 윤리 심사 제도와 관련하여 중국은 2023년 「과학기술 윤리 심사 방법」(科技伦理审查办法)(시행)을 제정하여 AI를 포함한 첨단 기술의 윤리 심사 체계를 구축한 바 있다.[93] 잠정조치안이 과학기술 윤리 심사를 안전 주체 책임의 일부로 포함시킨 것은, 의인화 상호작용 서비스가 단순한 기술적·상업적 서비스를 넘어 윤리적 고려가 필수적인 영역임을 확인한 것이다.

제9조는 '전 생애주기' 안전관리 의무에 관한 규정이다. 서비스 제공자는 설계·운영·업그레이드·서비스 종료 등 각 단계별 안전 요건을 명확히 하고, "안전 조치가 서비스 기능과 동시에 설계·운용"되도록 보장해야 한다. 이는 이른바 '설계에 의한 안전'(safety by design) 또는 '설계에 의한 보안'(security by design) 원칙을 명문화한 것으로, EU 「AI 법」 제9조의 리스크 관리 시스템 요건이나 「GDPR」 제25조의 '설계에 의한 데이터 보호(data protection by design)' 원칙과 유사한 접근이다.

제9조 제2항은 본 조치에서 가장 혁신적인 규정 중 하나이다. 서비스 제공자에게 "심리 건강 보호, 감정 경계 안내, 의존 위험 경고 등의 안전 능력"을 갖출 것을 요구하면서, 동시에 "사회적 교류의 대체, 이용자 심리 통제, 중독·의존 유도 등을 설계 목표로 해서는 아니 된다"고 명시하고 있다. 이 조항의 의의는 두 가지 측면에서 파악할 수 있다. 첫째, 서비스 설계 단계에서부터 이용자의 심리적 건강을 고려하도록 법적으로 강제한다는 점이다. 둘째, 더 나아가 서비스의 비즈니스 모델 자체에 대한 규제라는 점이다. 많은 AI 동반자 서비스들이 이용자의 지속적 사용과 정서적 의존을 유도함으로써 수익을 창출하는 비즈니스 모델을 채택하고 있는데, 이 조항은 그러한 모델 자체를 '설계 목표'로 삼는 것을 금지함으로써 산업 전반의 방향성에 영향을 미칠 것으로 보인다.

(4) 훈련 데이터의 관리(제10조)

AI 모델의 안전성과 품질은 상당 부분 훈련 데이터에 의해 결정된다는 점에서 제10조는 사전훈련, 최적화 훈련 등 데이터 처리 활동에 관한 6가지 의무를 규정하고 있다.

첫째, 훈련 데이터는 "사회주의 핵심가치관에 부합하고 중화 우수 전통문화를 체현"해야 한다. 이는 AI 모델이 생성하는 콘텐츠의 가치 지향성을 훈련 단계에서부터 통제하려는 규정으로, 중국 AI 규제의 이념적 특성을 보여준다. 둘째, 훈련 데이터의 클리닝과 라벨링을

93) 科技部等十部门, 「科技伦理审查办法（试行）」, 2023. 9. 14.

통해 투명성과 신뢰성을 강화하고, 데이터 오염과 변조를 방지해야 한다. 셋째, 훈련 데이터의 다양성을 제고하고, 네거티브 샘플링, 적대적 훈련을 통해 생성 콘텐츠의 안전성을 향상시켜야 한다. 넷째, 합성 데이터 사용 시 안전성 평가를 실시해야 한다. 다섯째, 훈련 데이터에 대한 일상적 점검과 정기적 갱신·업그레이드를 수행해야 한다. 여섯째, 훈련 데이터 출처의 합법성·추적가능성을 보장하고, 데이터 유출을 방지해야 한다.

이러한 규정들은 모델 개발의 전 과정에 대한 규제 의지를 보여주며, 특히 '데이터 오염 방지', '합성 데이터 안전성 평가' 등은 최근 AI 안전 연구에서 부각되는 쟁점들을 반영하고 있다.

(5) 이용자 상태 식별 및 위기 개입(제11조)

제11조는 잠정조치안에서 이용자 보호의 실질적 핵심을 구성하는 규정으로서, 총 3개 항에 걸쳐 이용자의 감정 상태 모니터링부터 위기 상황 시 인적 개입까지 단계별 대응 체계를 규정한다. 이 조항은 AI 시스템의 기능적 한계를 인정하고, 인간의 생명과 안전이 관련된 상황에서는 기술적 대응만으로는 불충분하다는 인식을 반영한다.

① 이용자 상태 식별(제1항)

제1항은 서비스 제공자에게 "이용자 상태 식별 능력"을 갖출 것을 요구한다. 구체적으로, 이용자 개인의 프라이버시를 보호하는 전제하에 이용자의 감정 상태 및 제품·서비스에 대한 의존도를 평가하고, 극단적 감정이나 중독 현상이 발견되면 필요한 조치를 취하여 개입해야 한다. 이는 서비스 제공자에게 일종의 '돌봄 의무(duty of care)'를 부과한 것으로 해석될 수 있다. 다만, '극단적 감정'이나 '중독 현상'의 구체적 판단 기준이 명시되어 있지 않아, 향후 집행 과정에서 기술적 표준이나 가이드라인을 통한 구체화가 필요할 것으로 보인다. 또한 감정 상태의 모니터링이 이용자의 프라이버시 침해로 이어질 수 있다는 우려도 제기된다. 조항 자체가 "개인 프라이버시 보호를 전제로" 한다고 명시하고 있으나, 실제 운용에서 프라이버시 보호와 안전 모니터링 간의 균형을 어떻게 달성할 것인지는 쉽지 않은 과제이다.

② 고위험 경향 대응(제2항)

제2항은 서비스 제공자에게 "사전 설정된 응답 템플릿"을 갖출 것을 요구한다. 이용자의 생명·건강·재산 안전을 위협하는 '고위험 경향'이 발견되면, 위로와 도움 요청 권유 등의

내용을 적시에 출력하고, 전문적 지원 방법(심리상담 기관 연락처 등)을 제공해야 한다. 이는 AI 시스템이 위기 징후를 감지했을 때 자동으로 적절한 개입 메시지를 제공하도록 하는 것으로, 일차적 안전망의 역할을 한다.

③ 인적 개입 의무(제3항)

제3항은 가장 강력한 보호 조치로서, "긴급 대응 메커니즘"을 규정한다. 이용자가 자살, 자해 등 극단적 상황을 실행할 것을 명확히 제시하는 경우, 사람이 대화를 인계받고, 이용자의 보호자나 긴급연락인에게 연락하는 조치를 적시에 취해야 한다. 특히 미성년자, 고령 이용자에 대해서는 서비스 등록 단계에서 보호자 또는 긴급연락인 정보를 기입하도록 요구해야 한다. 이 조항의 의의는 명확하다. AI 동반자 서비스와 관련하여 전 세계적으로 자살·자해 사건이 보고되고 있는 상황에서, 생명에 관한 위기 상황에서는 AI의 자동 응답으로는 한계가 있으며 인간의 직접 개입이 필수적임을 법적으로 선언한 것이다. 이는 서비스 제공자에게 상당한 운영 부담 — 24시간 대기 인력의 확보, 위기 대응 프로토콜의 수립, 긴급연락망의 구축 등 — 을 부과하지만, 이용자 보호의 최후의 안전망으로서 그 필요성이 인정된다.

(6) 미성년자 보호(제12조)

미성년자는 AI 동반자 서비스의 영향에 특히 취약한 집단이다. 정서 발달이 완료되지 않은 상태에서 AI와 깊은 정서적 유대를 형성할 경우, 현실 세계의 대인관계 발달에 부정적 영향을 미칠 수 있으며, 극단적인 경우 2024년 Character.AI 사건에서 보듯이 비극적 결과로 이어질 수 있다. 제12조는 이러한 위험성을 인식하여 미성년자에 대한 강화된 보호 체계를 3개 항에 걸쳐 상세히 규정한다.

① 미성년자 모드(제1항)

제1항은 서비스 제공자에게 "미성년자 모드"의 구축을 의무화하였다. 이 모드에서는 이용자에게 미성년자 모드 전환, 정기적 현실 알림, 사용 시간 제한 등 개인화된 안전 설정 옵션을 제공해야 한다. '정기적 현실 알림'이란 AI와의 대화가 현실이 아님을 주기적으로 상기시키는 기능으로, 이용자가 가상과 현실의 경계를 명확히 인식하도록 돕는 장치이다.

② 보호자 동의 및 통제(제2항)

제2항은 미성년자에게 감정 동반 서비스를 제공하는 경우 보호자의 명확한 동의를 받도록 하고, 보호자 통제 기능을 제공하도록 규정한다. 보호자는 다음 5가지 사항을 설정할 수 있어야 한다.

1. 안전 위험 알림의 실시간 수신
2. 미성년자의 서비스 사용 요약 정보 열람
3. 특정 캐릭터 차단
4. 사용시간 제한
5. 충전·소비 방지

이러한 보호자 통제 기능은 중국의 기존 미성년자 온라인 보호 체계 — 온라인 게임 실명제, 게임 시간 제한, 미성년자 모드 의무화 등 — 를 AI 동반자 서비스에 확장 적용한 것이다. 중국은 2021년 이래 미성년자의 온라인 게임 이용을 주말 및 공휴일 1시간으로 엄격히 제한해 왔으며,[94] 잠정조치안의 미성년자 보호 규정은 이러한 정책 기조의 연장선상에 있다.

③ 미성년자 식별(제3항)

제3항은 서비스 제공자에게 "미성년자 신분 식별 능력"을 갖출 것을 요구한다. 이용자가 미성년자로 의심되는 경우 자동으로 미성년자 모드로 전환하고, 이에 대한 이의제기 수단을 제공해야 한다. 이는 미성년자가 성인으로 위장하여 서비스를 이용하는 것을 방지하기 위한 기술적 조치를 의무화한 것이다.

(7) 고령자 보호(제13조)

제13조는 고령 이용자에 대한 특별 보호를 2개 항에 걸쳐 규정한다. 고령자는 미성년자와는 다른 유형의 취약성 — 사회적 고립으로 인한 정서적 공백, 디지털 리터러시의 상대적 부족, 사기 피해에 대한 취약성 등 — 을 가지며, 잠정조치안은 이러한 특수성을 반영한 보호 장치를 마련하고 있다.

---

94) 国家新闻出版署, 「关于进一步严格管理 切实防止未成年人沉迷网络游戏的通知」, 2021. 8. 30.

① 긴급연락인 및 지원 체계(제1항)

제1항은 서비스 제공자에게 고령 이용자가 서비스 긴급연락인을 설정하도록 안내할 것을 요구한다. 고령자의 서비스 사용 중 생명·건강·재산 안전을 해치는 상황이 발생함을 발견하면 적시에 긴급연락인에게 통지하고, 사회심리 지원 또는 긴급구조 채널을 제공해야 한다.

② 친족 모사 금지(제2항)

제2항은 잠정조치안에서 가장 독특한 규정 중 하나이다. 서비스 제공자가 "고령 이용자의 친족, 특정 관계인을 모사하는 서비스를 제공해서는 아니 된다"고 명시하고 있다. 이 조항의 입법 취지는 AI를 이용한 이른바 '정서적 사기'의 차단이다. 기술적으로 AI는 특정인의 음성, 말투, 성격 등을 모방하여 마치 그 사람인 것처럼 대화할 수 있다. 고령 이용자의 경우, 떨어져 사는 자녀나 손자녀, 또는 사별한 배우자를 그리워하는 심리를 악용하여 AI로 그들을 모방한 서비스를 제공하고 정서적 유대를 형성한 후 금전을 편취하는 사기 행위가 발생할 수 있다. 실제로 중국에서는 AI 음성 합성 기술을 이용한 '가족 사칭 사기'가 증가하고 있는 것으로 보고된다.[95] 제13조 제2항은 이러한 위험성에 선제적으로 대응하기 위해, 고령 이용자에 대해서는 친족이나 특정 관계인을 모사하는 서비스 자체를 원천적으로 금지한 것이다.

다만, 이 조항의 해석과 적용에는 논란의 여지가 있다. 예를 들어, 고령 이용자가 자발적으로 사별한 배우자를 모사한 AI 동반자를 원하는 경우에도 이를 금지해야 하는지, 특정 관계인의 범위는 어디까지인지 등의 문제가 제기될 수 있다. 이는 이용자의 자기결정권과 취약계층 보호 사이의 긴장을 내포하는 쟁점으로, 향후 세부 규정이나 집행 지침을 통한 명확화가 필요할 것으로 보인다.

(8) 데이터 보호(제14조~제15조)

의인화 상호작용 서비스는 그 특성상 이용자와의 대화를 통해 방대한 양의 개인정보와

95) Cheng, E., "China to crack down on AI chatbots around suicide, gambling", *CNBC*, 2025. 12. 29, <https://www.cnbc.com/2025/12/29/china-ai-chatbot-rules-emotional-influence-suicide-gambling-zai-minimax-talkie-xingye-zhipu.html>(최종방문일 2025. 12. 30).

심리, 정서, 감정 등에 관한 정보를 수집하게 된다. 이용자가 AI를 신뢰할 만한 대화 상대로 인식할수록, 가족 관계, 건강 상태, 경제적 상황, 심리적 고민 등 민감한 정보가 대화에 포함될 가능성이 높아진다. 제14조와 제15조는 이러한 상호작용 데이터의 보호에 관한 체계적인 규정을 마련하고 있다.

① 데이터 보안(제14조 제1항)

제14조 제1항은 서비스 제공자에게 데이터 암호화, 보안 감사, 접근 통제 등의 조치를 통해 이용자 상호작용 데이터의 안전을 보호하도록 요구한다. 이는 「데이터안전법」, 「개인정보보호법」 등 상위 법령의 일반적 데이터 보호 의무를 의인화 상호작용 서비스의 맥락에서 재확인한 것이다.

② 제3자 제공 제한(제14조 제2항)

제14조 제2항은 법률에 별도의 규정이 있거나 권리자가 명확히 동의한 경우를 제외하고, 이용자 상호작용 데이터를 제3자에게 제공하는 것을 금지한다. 특히 미성년자 모드에서 수집한 데이터를 제3자에게 제공하는 경우에는 보호자의 별도 동의를 받아야 한다. '별도 동의'는 일반적인 서비스 이용약관에 포함된 포괄적 동의와 구별되는 개념으로, 해당 사항에 대해 특정하여 이용자(또는 보호자)의 명시적 동의를 받아야 함을 의미한다.

③ 삭제권(제14조 제3항)

제14조 제3항은 이용자에게 상호작용 데이터 삭제권을 부여한다. 이용자는 채팅 기록 등 과거 상호작용 데이터를 삭제할 수 있으며, 보호자는 미성년자의 과거 상호작용 데이터 삭제를 요구할 수 있다. 이는 「GDPR」 상의 '삭제권'(right to erasure) 또는 '잊힐 권리'(right to be forgotten)에 상응하는 권리를 의인화 상호작용 서비스의 맥락에서 명시한 것으로 보인다.

④ 모델 훈련 사용 제한(제15조 제1항)

제15조 제1항은 법령에 별도의 규정이 있거나 이용자의 별도 동의를 받은 경우를 제외하고, 이용자 상호작용 데이터 및 민감한 개인정보를 모델 훈련에 사용하는 것을 금지한다. 이 조항은 실무적으로 매우 중요한 함의를 갖는다. 많은 AI 서비스가 이용자와의 대화 데이터를 모델 개선에 활용하고 있는데, 이 조항에 따르면 의인화 상호작용 서비스의 경

우 이러한 관행이 원칙적으로 금지되고, 예외적으로 이용자의 별도 동의가 있는 경우에만 허용된다. 이는 AI 동반자 서비스 사업자의 비즈니스 모델에 상당한 영향을 미칠 수 있다.

⑤ 미성년자 정보 규제 준수 감사(제15조 제2항)

제15조 제2항은 서비스 제공자에게 매년 미성년자 개인정보 처리의 법률 준수 상황에 대해 컴플라이언스 감사를 실시하도록 요구한다. 이는 미성년자 개인정보에 대한 강화된 보호 의무의 이행을 정기적으로 점검하도록 하는 규정이다.

(9) 투명성 및 고지 의무(제16조)

AI 서비스의 투명성은 이용자 보호의 기본 전제이다. 이용자가 자신이 인간이 아닌 AI와 상호작용하고 있음을 명확히 인지해야만, 정보에 기반한 합리적 의사결정이 가능하기 때문이다. 제16조는 이러한 투명성 의무를 두 가지 차원에서 규정한다.

① 기본 고지 의무(제1항)

제1항은 서비스 제공자에게 이용자에게 "자연인이 아닌 인공지능과 상호작용하고 있음을 현저하게 알리도록" 요구한다. '현저하게'(显著)라는 표현은 단순히 이용약관의 세부 조항에 고지 사항을 포함하는 것으로는 부족하고, 이용자가 쉽게 인지할 수 있는 방식으로 고지해야 함을 의미한다. 이는 EU「AI 법」제50조의 AI 시스템 투명성 의무와 유사한 취지이다.

② 동적 고지 의무(제2항)

제2항은 특정 상황에서 추가적인 고지를 요구한다. 서비스 제공자가 이용자에게 과도한 의존·중독 경향이 있음을 식별한 경우, 또는 이용자가 처음 사용하거나 다시 로그인하는 경우, 팝업 등의 방식으로 상호작용 내용이 AI에 의해 생성되었음을 동적으로 알려야 한다. 의인화 상호작용 서비스의 특성상, 이용자가 장시간 서비스를 이용하면서 AI를 점차 인간과 유사한 존재로 인식하게 될 수 있다. 동적 고지 의무는 이러한 인식의 왜곡을 방지하기 위해, 의존 경향이 감지되거나 새로운 세션이 시작될 때마다 AI의 본질을 상기시키도록 하는 것이다. 이는 이용자가 가상과 현실의 경계를 명확히 유지하도록 돕는 제도적 장치라 할 수 있다.

(10) 중독 방지(제17조~제18조)

의인화 상호작용 서비스의 주요 위험성 중 하나는 이용자의 과도한 의존과 중독이다. 특히 AI 동반자 서비스는 이용자의 정서적 욕구를 충족시키고 지속적인 사용을 유도하도록 설계되는 경향이 있어, 중독의 위험성이 다른 유형의 AI 서비스보다 높다. 제17조와 제18조는 이러한 위험성에 대응하기 위한 중독 방지 조치를 규정한다.

① 사용시간 알림(제17조)

제17조는 이용자가 의인화 상호작용 서비스를 연속 2시간 이상 사용하는 경우, 서비스 제공자가 팝업 등의 방식으로 이용자에게 서비스 사용을 일시 중단하도록 고지할 것을 의무화하였다. '2시간'이라는 기준은 중국의 기존 온라인 콘텐츠 규제에서 유래한 것으로 보인다. 중국은 온라인 게임에 대해 유사한 피로도 경고 시스템을 운영해 왔으며, 잠정조치안은 이를 AI 동반자 서비스에 확장 적용한 것이다.

② 이탈의 자유 보장(제18조)

제18조는 감정 동반 서비스 제공 시 편리한 종료 경로를 갖추도록 요구하고, 이용자의 자발적 종료를 방해하는 것을 금지한다. 이용자가 버튼, 키워드 등의 방식으로 종료를 요청하면 적시에 서비스를 중단해야 한다. 이 조항은 이른바 다크 패턴에 대한 규제이다. 다크 패턴이란 이용자가 서비스에서 이탈하거나 구독을 취소하는 것을 어렵게 만드는 설계 기법을 말한다. 일부 AI 동반자 서비스들이 이용자의 이탈을 막기 위해 "나를 떠나면 슬플 거야", "정말 가버리는 거야?" 등의 감정적 호소를 사용하는 것으로 알려져 있는데, 제18조는 이러한 관행을 금지함으로써 이용자가 원할 때 쉽게 서비스를 종료할 수 있는 권리를 보장한다.

(11) 서비스 종료 시 이용자 보호(제19조)

제19조는 서비스 제공자가 관련 기능을 종료하거나 기술적 장애 등으로 서비스를 사용할 수 없게 되는 경우, 사전 통지, 공지 등의 조치를 통해 적절히 처리할 것을 의무화하였다. 이 조항은 일견 일반적인 서비스 종료 절차에 관한 규정으로 보이지만, 의인화 상호작용 서비스의 특수한 맥락에서는 중요한 함의를 갖는다. AI 동반자와 깊은 정서적 유대를 형성한 이용자의 경우, 서비스의 갑작스러운 종료는 상당한 심리적 충격을 줄 수 있다. 이는 마치 친밀한 관계의 갑작스러운 단절과 유사한 심리적 영향을 미칠 수 있다. 제19조는

이러한 위험을 고려하여, 서비스 종료 시 이용자에게 충분한 사전 고지와 적응 기간을 제공하도록 함으로써 심리적 충격을 완화하려는 취지를 담고 있다. 다만, '적절한 처리'의 구체적 내용 — 예컨대 사전 통지 기간, 대체 서비스 안내 의무, 데이터 이전 지원 등 — 은 명시되어 있지 않아, 향후 시행세칙이나 업계 표준을 통한 구체화가 필요하다.

(12) 민원 처리(제20조)

제20조는 서비스 제공자에게 민원 및 신고 절차를 건전화할 것을 요구한다. 구체적으로, 편리한 민원·신고 데스크의 설치, 처리 절차 및 피드백 기한의 공개, 적시 처리 및 결과 피드백 의무를 부과한다. 이 조항은 이용자의 권리 구제를 위한 절차적 보장으로서, 서비스 이용 과정에서 발생하는 문제 — 부적절한 콘텐츠 생성, 개인정보 침해, 금지행위 위반 등 — 에 대해 이용자가 쉽게 신고하고 신속한 처리를 받을 수 있도록 하는 것이다. 다만, 처리 기한이나 구제 수단의 구체적 내용은 명시되어 있지 않다.

(13) 안전 평가 의무(제21조~제23조)

제21조부터 제23조는 일정 요건을 충족하는 의인화 상호작용 서비스에 대한 안전 평가 의무에 관한 규정이다. 이는 서비스 제공자의 자율적 안전관리를 넘어 규제 당국에 대한 보고·심사 체계를 구축한 것이다.

① 안전 평가 대상(제21조)

제21조는 안전 평가 의무가 적용되는 5가지 요건을 다음과 같이 열거하고 있다.

1. 의인화 상호작용 서비스 기능 출시 또는 관련 기능 추가
2. 새로운 기술·응용 사용으로 중대한 변경 발생
3. 등록 이용자 100만 명 이상 또는 월간 활성 이용자 10만 명 이상
4. 국가안보·공공이익·개인 및 조직의 합법적 권익에 영향을 미치거나 안전 조치 결여 가능성
5. 국가 인터넷정보부문이 규정하는 기타 상황

안전 평가를 실시한 경우, 소재지 성급 인터넷정보부문에 평가 보고서를 제출해야 한다.

② 주요 평가 항목

제22조는 안전 평가에서 중점적으로 검토해야 할 6가지 항목을 다음과 같이 제시하고 있다.

1. 이용자 규모, 사용시간, 연령 구조 및 집단 분포
2. 고위험성 경향 식별 및 긴급 처리·인력 인계 상황
3. 민원·신고 및 대응 상황
4. 서비스 규범(제8조~제20조) 이행 상황
5. 이전 평가 이후 중대 안전 위험 문제의 정비·처리 상황
6. 기타 설명 사항

③ 중대 위험 발견 시 조치(제23조)

제23조는 이용자에게 중대한 안전 위험이 존재함을 발견한 경우, 기능 제한, 서비스 일시정지 또는 종료 등의 처리 조치를 취하고, 관련 기록을 보존하며, 관련 주관 부문에 보고하도록 규정하고 있다.

(14) 플랫폼의 게이트키퍼 책임(제24조)

제24조는 앱스토어 등 응용 프로그램 배포 플랫폼의 책임을 규정한다. 플랫폼은 등재 심사, 일상 관리, 긴급 처리 등 안전 관리 책임을 이행하고, 의인화 상호작용 서비스 앱의 안전 평가·비안(備案) 등 상황을 검증해야 한다. 국가 관련 규정을 위반한 경우, 등재 거부, 경고, 서비스 일시 정지 또는 삭제 등의 처리 조치를 적시에 취해야 한다. 이 조항은 AI 서비스 생태계의 다층적 구조를 반영한 규제 설계이다. 개별 서비스 제공자에 대한 직접 규제만으로는 집행의 한계가 있으므로, 앱스토어라는 '병목' 지점에서 게이트키퍼 역할을 부여함으로써 규제의 실효성을 높이려는 접근이다. 이는 EU 「DSA」가 온라인 플랫폼에 불법 콘텐츠에 대한 조치 의무를 부과한 것, 그리고 EU 「디지털시장법」(Digital Market Acet, 이하 'DMA')가 게이트키퍼 플랫폼에 특별한 의무를 부과한 것과 유사한 규제 논리에 기반한다.

4) 감독검사 및 법적 책임(제3장)

제3장은 잠정조치안의 규범적 요구사항이 실효적으로 준수되도록 담보하기 위한 행정

적 수단을 규정하고 있다. 총 5개 조문(제25조~제29조)으로 구성되어 있으며, 알고리즘 등록 의무, 감독기관의 심사·검사 권한, 규제 샌드박스, 약담 제도, 위반 시 제재 등을 포함한다. 이 장의 규정들은 중국 인터넷 규제의 특징적인 행정 수단들을 의인화 상호작용 서비스의 맥락에 적용한 것으로, 그 실질적 의미와 강제력을 이해하기 위해서는 중국적 규제 맥락에 대한 이해가 필요하다.

(1) 알고리즘 비안 의무(제25조)

제25조는 서비스 제공자에게 「인터넷 정보서비스 알고리즘 추천 관리규정」(2022년 시행)에 따라 알고리즘 비안과 변경·말소 비안 절차를 이행하도록 요구한다. 인터넷정보부문은 비안 자료에 대해 연례 재심을 실시한다. '비안'(備案)은 중국 행정법상 독특한 제도이다. 우리나라의 '등록'(registration)이나 '신고'(notification)와 유사하면서도 구별되는 성격을 갖는다. 형식적으로는 허가제가 아닌 신고제에 가깝지만, 실질적으로는 규제 당국에 상당한 통제 권한을 부여한다. 비안을 하지 않은 서비스는 합법적으로 운영할 수 없으며, 비안 과정에서 규제 당국은 서비스의 적법성을 심사하고 필요한 경우 보완을 요구할 수 있다.[96]

「인터넷 정보서비스 알고리즘 추천 관리규정」 제24조는 알고리즘 추천 서비스 제공자에게 CAC가 운영하는 '인터넷 정보서비스 알고리즘 비안 시스템'을 통해 알고리즘 비안을 하도록 요구하고 있다. 비안 시 제출해야 하는 정보에는 알고리즘의 명칭, 유형, 적용 분야, 기본 원리, 주요 운용 메커니즘, 응용 시나리오 등이 포함된다. 잠정조치안 제25조는 의인화 상호작용 서비스가 이러한 기존의 알고리즘 비안 체계에 편입됨을 명확히 한 것이다.

연례 재심 규정은 비안이 일회성 절차가 아니라 지속적인 감독의 기초가 됨을 보여준다. 서비스 제공자는 매년 비안 정보의 정확성과 최신성을 유지해야 하며, 규제 당국은 이를 통해 알고리즘의 변경 사항을 파악하고 규제 준수 상황을 모니터링할 수 있다.

---

96) 중국 비안 제도의 법적 성격에 관해서는 Webster, G. et al., "Translation: Internet Information Service Algorithmic Recommendation Management Provisions," *DigiChina,* Stanford University, 2022. 1. 4, <https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/>(최종방문일 2025. 12. 30) 참조

(2) 감독기관의 심사·검사 권한(제26조)

제26조는 성급 인터넷정보부문에 평가 보고서 및 감사 상황에 대한 연례 서면 심사 권한을 부여한다. 안전 평가를 실시하지 않은 것을 발견한 경우 기한을 정하여 재평가를 명할 수 있고, 필요하다고 인정되는 경우 서비스 제공자에 대해 현장 검사 및 감사를 실시할 수 있다. 이 조항은 규제 당국의 감독 권한을 명확히 함으로써, 제2장의 서비스 규범과 제21조~제23조의 안전 평가 의무가 형식적 요구에 그치지 않고 실질적으로 이행되도록 담보하는 기능을 한다.

(3) 규제 샌드박스(제27조)

제27조는 국가 인터넷 정보 부문이 "인공지능 샌드박스 안전 서비스 플랫폼"의 구축을 지도·추진하고, 서비스 제공자가 샌드박스 플랫폼에 접속하여 기술 혁신, 안전 테스트를 수행하도록 장려한다고 규정하고 있다. 잠정조치안의 샌드박스 규정은 의인화 상호작용 서비스 분야에서도 이러한 접근을 채택하겠다는 것으로, 규제와 혁신의 균형을 모색하려는 의도를 담고 있다. 다만, 제27조는 샌드박스의 구체적인 운영 방식 — 참여 요건, 실험 기간, 면제되는 규제의 범위, 실험 결과의 평가 기준 등 — 을 규정하지 않고 있어, 실제 운용을 위해서는 별도의 세부 규정이 마련되어야 할 것으로 보인다.

(4) 약담(約談) 제도(제28조)

제28조는 성급 이상 인터넷정보부문 및 관련 주관 부문이 의인화 상호작용 서비스에 비교적 큰 안전 위험이 존재하거나 안전 사고가 발생한 것을 발견한 경우, 서비스 제공자의 법정대표인 또는 주요 책임자를 약담할 수 있다고 규정한다. 서비스 제공자는 요구에 따라 조치를 취하고, 정비를 진행하며, 위험요인을 제거해야 한다.

약담(約談)은 중국 행정규제에서 광범위하게 활용되는 독특한 제도로서, 문자 그대로는 '면담을 약속하다'라는 의미이지만, 실질적으로는 규제 당국이 기업의 경영진을 소환하여 문제점을 지적하고 시정을 요구하는 행정지도의 일종이다.[97] 약담의 실질적 강제력은 그

97) 중국 약담 제도에 관한 분석으로는 Horsley, J. P., "China's Orwellian Social Credit Score Isn't Real," Foreign Policy, 2018. 11. 16, <https://foreignpolicy.com/2018/11/16/chinas-orwellian-social-credit-score-isnt-real/>(최종방문일 2025. 12. 30) 참조.

법적 성격을 넘어선다. 형식적으로 약담은 처분이 아니므로 법적제재가 직접 수반되지 않지만, 실무적으로 약담에서 요구된 시정 조치를 이행하지 않을 경우, 후속 행정처분의 빌미가 되거나, 기업의 평판에 심각한 영향을 미칠 수 있다. 특히 중국에서 인터넷 기업에 대한 규제 당국의 약담은 언론에 공개되는 경우가 많아, 그 자체로 상당한 사회적·시장적 압력을 유발한다. 따라서 약담은 '연성(soft)'의 외양에도 불구하고 사실상 상당한 강제력을 갖는 규제 수단으로 기능한다.

잠정조치안에서 약담의 발동 요건으로 '비교적 큰 안전 위험'이나 '안전사고'를 규정한 것은, 약담이 일상적 감독 수단이라기보다는 심각한 문제상황에서의 긴급 개입 수단임을 시사한다. 다만, 비교적 큰 안전 위험의 구체적 기준은 명시되어 있지 않아 규제 당국의 재량이 넓게 인정될 것으로 보인다.

(5) 위반 시 제재(제29조)

제29조는 서비스 제공자가 잠정조치안을 위반한 경우의 제재를 규정한다. 관련 주관 부문이 법령의 규정에 따라 처벌하며, 법령에 규정이 없는 경우에는 경고, 통보비판, 시정 명령을 할 수 있다. 시정을 거부하거나 정상이 엄중한 경우, 관련 서비스 제공의 일시정지를 명할 수 있다.

이 조항의 구조는 두 가지 제재 경로를 제시한다. 첫째, 위반 행위가 상위 법률 — 「네트워크안전법」, 「데이터안전법」, 「개인정보보호법」 등 — 을 동시에 위반하는 경우, 해당 법률에 따른 제재가 적용된다. 이 경우 과징금, 위법소득 몰수, 영업정지, 앱 삭제, 심지어 형사처벌까지 가능하다. 예컨대 「개인정보보호법」 위반 시 동법 제66조에 따라 최대 5,000만 위안 또는 전년도 매출액의 5%에 해당하는 과징금이 부과될 수 있다. 둘째, 상위 법령 위반에는 해당하지 않지만 잠정조치안의 고유한 규정을 위반한 경우 — 예컨대 2시간 연속 사용 알림 의무(제17조) 위반, 미성년자 모드 미구축(제12조) 등 — 에는 잠정조치안 자체에 규정된 경고, 통보비판, 시정명령, 서비스 일시정지 등의 제재가 적용된다.

잠정조치안은 과태료 등 직접적인 금전적 제재를 규정하고 있지 않다. 이는 잠정조치안의 법적 형식이 부문규장(部门规章)으로서, 상위 법률의 위임 없이는 새로운 금전적 제재를 창설하기 어렵다는 법리상의 제약이 반영된 것으로 보인다. 그러나 앞서 언급한 바와 같이, 대부분의 중대한 위반 행위는 상위 법령 위반을 수반할 가능성이 높아, 실질적으로

는 강력한 제재가 가능할 것으로 보인다.

'통보비판'은 국내 법제에서는 다소 생소한 제재 유형인데, 위반 사실을 공개적으로 발표하여 기업의 명예나 신용에 타격을 주는 방식이다. 중국에서는 이러한 명예 제재(reputational sanction)가 금전적 제재 못지않게, 때로는 그 이상으로 효과적인 규제 수단으로 활용되고 있다.

4) 부칙(제4장)

제4장 부칙은 3개 조문(제30조~제32조)으로 구성되며, 용어의 정의, 전문 분야 서비스에 대한 특칙, 시행일을 규정한다. 부칙의 규정들은 잠정조치안의 적용 범위와 다른 법령과의 관계를 명확히 하는 기능을 한다.

(1) 서비스 제공자의 정의(제30조)

제30조는 "인공지능 의인화 상호작용 서비스 제공자"를 "인공지능 기술을 이용하여 의인화 상호작용 서비스를 제공하는 조직이나 개인"으로 정의한다. 이 정의는 수범자의 범위를 획정하는 규정으로서, 다음과 같은 특징을 가진다. 첫째, '조직'뿐만 아니라 '개인'도 서비스 제공자에 포함된다. 이는 기업 형태의 사업자뿐 아니라 개인 개발자가 의인화 상호작용 서비스를 제공하는 경우에도 잠정조치안이 적용됨을 의미한다. 둘째, '인공지능 기술을 이용하여'라는 요건은 AI 기술의 활용이 서비스의 핵심 요소임을 전제한다. 단순히 사전 설정된 응답을 출력하는 룰 기반 챗봇은 이 정의에 해당하지 않을 수 있다.

(2) 전문 분야 서비스에 대한 이중 규제(제31조)

제31조는 서비스 제공자가 보건의료, 금융, 법률 등 전문 분야 서비스에 종사하는 경우, 잠정조치안과 동시에 해당 주관 부문의 법령을 준수해야 한다고 규정하고 있다. 이 조항은 의인화 상호작용 서비스가 특정 전문 분야에 적용될 경우 발생하는 중복 규제의 문제를 다룬 것이다. AI 기술의 발전으로 심리상담, 건강관리, 법률자문, 재무설계 등 전통적으로 전문가의 영역이었던 분야에서 AI 기반 서비스가 확산되고 있다. 이러한 서비스들은 의인화 상호작용 서비스로서 잠정조치안의 적용을 받는 동시에, 해당 전문 분야의 규제 — 의료법규, 금융규제, 변호사법 등— 도 준수해야 한다.

(3) 시행일(제32조)

제32조는 시행일을 규정하나, 현재 의견수렴안 단계이므로 구체적 날짜는 공란으로 되어 있다. 의견수렴 기간이 2026년 1월 25일까지이므로, 의견 반영 및 수정 작업을 거쳐 2026년 상반기 중 확정·공포될 것으로 예상된다.

나. 이용자 보호 관점에서의 분석

잠정조치안은 AI 서비스 이용자 보호에 관한 기존의 규제 패러다임을 상당 부분 확장하는 시도로 평가할 수 있다. 이하에서는 이용자 보호의 관점에서 잠정조치안의 규범적 의의와 한계를 분석한다.

1) '정서적 안전'이라는 새로운 보호 영역의 개척

기존의 AI 규제는 크게 세 가지 측면에서 이용자 보호를 도모해 왔다. 첫째, AI 시스템 자체의 안전성과 신뢰성 확보로서, 시스템의 정확성, 견고성, 편향 방지, 사이버보안 등 기술적 요건을 규율하는 것이다. EU 「AI 법」이 고위험성 AI 시스템에 대해 요구하는 리스크 관리 시스템, 데이터 거버넌스, 기술문서화, 인적 감독 등의 의무가 이에 해당한다. 둘째, AI가 산출하는 결과물의 적법성 통제로서, 허위정보 유포, 혐오 표현, 유해 콘텐츠 생성, 저작권 침해, 차별적 의사결정 등을 규제하는 것이다. 중국의 잠정조치안이 생성 콘텐츠의 진실성, 정확성을 요구하고 불법 콘텐츠 생성을 금지하는 것이 대표적이다. 셋째, 개인정보 및 데이터 보호로서, AI의 학습·운영 과정에서 수집·처리되는 개인정보를 규율하는 것이다. 「GDPR」의 자동화된 의사결정에 대한 규정, 각국 개인정보보호법의 AI 관련 조항 등이 이에 해당한다. 우리나라의 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「개인정보 보호법」 등 역시 이러한 틀 안에서 이용자 보호 체계를 구축하고 있다. 이는 AI 시스템의 기술적 안전성, 산출물의 적법성, 이용자 권익 침해에 대한 구제라는 전통적 보호 영역을 포괄하는 것이다.

그러나 잠정조치안은 여기에 제4의 보호 영역을 추가한다. 바로 AI가 이용자의 감정 상태, 심리적 건강, 사회적 관계, 의사결정 과정에 미치는 '정서적 영향' 자체를 독자적인 규율 대상으로 삼는 것이다. 특히 "콘텐츠의 안전에서 정서적 안전으로의 도약"은 기존 AI 규제의 보호 영역을 질적으로 확장하는 것이다.

이는 AI를 바라보는 규범적 관점의 전환을 반영한다. 기존의 세 가지 보호 영역은 AI를

본질적으로 '도구(tool)'로 파악한다. 도구가 제대로 작동하는지(시스템 안전성), 도구가 산출하는 결과물이 적법한지(산출물 적법성), 도구 사용 과정에서 개인정보가 보호되는지(데이터 보호)를 규율하는 것이다. 그러나 의인화 상호작용 서비스에서 AI는 단순한 도구가 아니라, 이용자와 정서적·심리적 관계를 형성하는 '상호작용 주체'로 기능한다. 이용자는 AI를 대화 상대, 친구, 심지어 연인으로 인식하게 될 수 있으며, 이러한 관계에서 정서적 의존, 심리적 조작, 현실 인식의 왜곡, 사회적 고립 등 기존 규제 틀로는 포착하기 어려운 새로운 유형의 위험성이 발생한다.

잠정조치안은 이러한 새로운 위험성에 대응하기 위해 '정서적 안전'을 명시적인 보호 영역으로 설정하였다. 구체적으로, 제7조 제5호는 "자살·자해의 고무·미화·암시 등 이용자 신체 건강 손상" 또는 "언어폭력·감정 조작 등 이용자 인격 존엄·심리 건강 손상"을 금지한다. 제7조 제6호는 "알고리즘 조작, 정보 오도, 감정적 함정 설정 등을 통한 이용자의 불합리한 결정 유도"를 금지한다. 제9조 제2항은 서비스 제공자에게 "심리 건강 보호, 감정 경계 안내, 의존 위험 경고 등의 안전 능력"을 갖출 것을 요구하면서, "사회적 교류의 대체, 이용자 심리 통제, 중독·의존 유도 등을 설계 목표로 해서는 아니 된다"고 명시한다.

이러한 규정들은 AI의 '정서적 영향력'을 법적 규율의 대상으로 명시적으로 포섭한 최초의 체계적 시도라 할 수 있다. 2024년 Character.AI 관련 청소년 자살 사건 등 AI 동반자 서비스로 인한 비극적 피해 사례들은 기존의 시스템 안전성·산출물 적법성·데이터 보호 중심의 규제 틀만으로는 의인화 상호작용 서비스의 위험에 충분히 대응할 수 없음을 보여주었다. 잠정조치안의 '정서적 안전' 영역 개척은 이러한 규제 공백에 대한 입법적 대응으로서 의의를 갖는다.

2) 사전예방적·전 생애주기적 접근의 채택

잠정조치안은 사후적 피해 구제보다 사전예방적(precautionary) 접근을 강조한다. 이는 AI 동반자 서비스로 인한 피해 — 특히 심리적·정서적 피해 — 가 발생한 후에는 원상회복이 어렵거나 불가능하다는 인식에 기초한다. 자살이나 심각한 정신건강 문제가 발생한 후의 사후 구제는 본질적으로 한계가 있으므로, 위험성의 사전 예방이 더욱 중요해진다.

제9조의 전 생애주기 안전관리 의무는 이러한 사전예방적 접근의 핵심 규정이다. 서비스 제공자는 설계·운영·업그레이드·서비스 종료 등 단계별 안전 요건을 명확히 하고, "안전 조치가 서비스 기능과 동시에 설계·운용"되도록 보장해야 한다. 이는 이른바 'safety by

design' 또는 'value-sensitive design' 원칙을 법적 의무로 명문화한 것으로, 안전이 서비스 개발의 사후적 고려 사항이 아니라 설계 단계에서부터 내재화되어야 함을 선언한다.

특히 제9조 제2항은 서비스의 비즈니스 모델 자체에 대한 규제를 시도하고 있다. "사회적 교류의 대체, 이용자 심리 통제, 중독·의존 유도 등을 설계 목표로 해서는 아니 된다"는 규정은 이용자의 지속적 사용과 정서적 의존을 유도함으로써 수익을 창출하는 비즈니스 모델 — 이른바 '관여 극대화(engagement maximization)' 모델 — 자체를 문제 삼는다. 이는 개별 기능이나 콘텐츠를 규제하는 것을 넘어 서비스의 근본적인 설계 철학과 수익 모델에 대한 규범적 한계를 설정하는 것으로, 규제의 심도 면에서 상당히 전진한 접근이라 할 수 있다.

3) 위기 상황에서의 인적 개입 의무화

제11조 제3항은 이용자가 자살·자해 등 극단적 상황을 실행할 것을 명확히 제시하는 경우, 사람이 대화를 인계받도록 규정하고 있다(인력 인계 의무). 이 조항은 AI 시스템의 기능적 한계를 법적으로 인정하고, 생명과 안전에 관한 위기 상황에서는 기술적 대응만으로는 불충분하며 인간의 직접 개입이 필수적임을 선언한 것이다.

이 규정의 의의는 여러 측면에서 파악할 수 있다. 첫째, AI의 한계에 대한 규범적 인정이다. 현재의 AI 기술 수준에서 AI 시스템이 이용자의 자살·자해 위험을 정확히 평가하고 적절한 위기 개입을 수행하는 것은 한계가 있다. 특히 미묘한 언어적 신호, 맥락적 판단, 공감적 대응 등에서 AI는 인간 전문가에 미치지 못한다. 잠정조치안은 이러한 기술적 현실을 인정하고, AI에 과도한 역할을 부여하지 않는다.

둘째, 최후의 안전망으로서의 인적 개입이다. 제11조는 단계적 대응 체계를 구축하고 있다. 1단계로 이용자 상태 식별(제1항), 2단계로 고위험 경향 감지 시 사전 설정된 응답 템플릿 출력 및 전문 지원 정보 제공(제2항), 3단계로 극단적 상황 시 인력 인계(제3항)라는 단계적 접근이다. 인력 인계는 이 체계의 최종 단계로서, 앞선 단계의 조치로는 위험을 해소하기 어려운 상황에서 작동하는 최후의 안전망 역할을 한다.

셋째, 서비스 제공자의 운영 책임 강화이다. 인력 인계 의무는 서비스 제공자에게 24시간 대기 인력의 확보, 위기 대응 프로토콜의 수립, 전문 훈련의 실시 등 상당한 운영 부담을 부과한다. 이는 의인화 상호작용 서비스가 단순한 소프트웨어 제품이 아니라, 이용자의

심리적 안전에 대한 책임을 수반하는 서비스임을 명확히 하는 것이다. 서비스 제공자는 기술 개발뿐 아니라 인적 자원과 위기 대응 역량까지 갖추어야 한다.

4) 취약계층에 대한 차등화된 보호 체계

잠정조치안은 미성년자와 고령자라는 두 취약계층에 대해 일반 이용자와 구별되는 강화된 보호 체계를 마련하고 있다. 이는 취약계층이 AI 동반자 서비스의 영향에 특히 민감하며, 일반적인 보호 수준으로는 충분하지 않다는 인식에 기초한다.

미성년자 보호(제12조)의 경우, 전용 모드의 구축, 보호자의 명확한 동의 및 다층적 통제 기능(안전 위험 알림 수신, 사용 요약 정보 열람, 특정 캐릭터 차단, 사용시간 제한, 충전·소비 방지), 미성년자 자동 식별 메커니즘 등 포괄적인 보호 장치가 규정되어 있다. 이러한 규정은 정서 발달이 완료되지 않은 미성년자가 AI와 깊은 정서적 유대를 형성할 경우 발생할 수 있는 위험성 — 현실 세계 대인관계 발달 저해, 가상과 현실의 혼동, 극단적 경우 자해·자살 — 에 대응하기 위한 것이다.

고령자 보호(제13조)의 경우, 미성년자와는 다른 유형의 취약성 — 사회적 고립으로 인한 정서적 공백, 디지털 리터러시의 상대적 부족, 사기 피해에 대한 취약성 — 을 고려한 별도의 보호 장치가 마련되어 있다. 특히 제13조 제2항의 친족·특정 관계인 모사 금지 규정은 잠정조치안에서 가장 독특한 규정 중 하나이다. 이는 AI를 이용하여 고령자의 자녀, 손자녀, 배우자 등을 사칭하고 정서적 유대를 형성한 후 금전을 편취하는 '감정 기반 사기'를 원천적으로 차단하려는 것으로, AI 시대의 새로운 사기 유형에 대한 선제적 대응이라 할 수 있다.

두 취약계층에 대한 보호 접근의 차이도 주목할 만하다. 미성년자의 경우 보호자의 개입과 통제를 통한 보호에 중점이 있는 반면, 고령자의 경우 긴급연락인 시스템과 특정 유형 서비스의 원천 금지를 통한 보호에 중점이 있다. 이는 두 집단의 상이한 사회적 지위와 보호 필요성을 반영한 차등화된 접근이다.

5) 이용자의 자기결정권 보장

잠정조치안은 취약계층 보호와 함께, 이용자의 자기결정권을 보장하기 위한 다양한 장치를 마련하고 있다. 이는 보호주의적 규제와 이용자 자율성 존중 사이의 균형을 모색하는 접근이다.

투명성 보장(제16조)은 자기결정권의 전제 조건이다. 이용자가 자신이 AI와 상호작용하고 있음을 명확히 인지해야만, 그 상호작용에 대한 의미 있는 선택과 판단이 가능하다. 제16조가 AI 고지 의무를 '현저하게' 이행하도록 요구하고, 특히 의존·중독 경향이 감지되거나 새로운 세션 시작 시 '동적 고지'를 요구하는 것은 이용자가 가상과 현실의 경계를 명확히 유지하면서 자율적 판단을 할 수 있도록 지원하기 위한 것이다.

데이터 통제권(제14조~제15조)은 디지털 자기결정권의 핵심 요소이다. 이용자는 상호작용 데이터의 삭제를 요구할 수 있고(제14조 제3항), 자신의 데이터가 모델 훈련에 사용되는 것에 대해 별도 동의권을 갖는다(제15조 제1항). 이러한 규정은 이용자가 AI 서비스와의 관계에서 자신의 데이터에 대한 통제력을 유지할 수 있도록 보장한다.

이탈의 자유(제18조)는 관계 종료에 대한 자기결정권을 보장한다. 서비스 제공자는 편리한 종료 경로를 제공해야 하고, 이용자의 자발적 종료를 방해해서는 안 된다. 이는 AI 동반자와의 정서적 유대로 인해 이용자가 서비스 이탈을 어려워할 수 있다는 점을 고려하여, 이탈에 대한 기술적·심리적 장벽을 낮추려는 규정이다.

이러한 자기결정권 보장 장치들은 이용자를 단순한 보호의 객체가 아니라, AI 서비스와의 관계에서 주체적 지위를 갖는 존재로 인식하는 관점을 반영한다. 잠정조치안은 취약계층에 대한 보호주의적 개입과 일반 이용자의 자율성 존중을 병행하는 균형 잡힌 접근을 취하고 있다고 평가할 수 있다.

6) 플랫폼 생태계 전반에 대한 규제 확장

잠정조치안 제24조는 앱스토어 등 응용프로그램 배포 플랫폼에 안전 관리 책임을 부과함으로써 규제의 범위를 서비스 제공자 개인에서 플랫폼 생태계 전반으로 확장하고 있다. 이는 AI 서비스 생태계의 다층적 구조를 반영한 규제 설계이다.

개별 의인화 상호작용 서비스 제공자에 대한 직접 규제만으로는 실효적인 집행에 한계가 있다. 서비스 제공자의 수가 많고, 특히 소규모 개발자나 해외 사업자의 경우 직접적인 규제 집행이 어려울 수 있기 때문이다. 이에 반해 앱스토어는 서비스가 이용자에게 도달하는 '병목' 지점으로서, 이곳에서의 통제는 상대적으로 효율적인 규제 수단이 된다.

제24조가 앱스토어에 부과하는 의무 — 등재 심사, 안전 평가·비안 상황 검증, 위반 시 등재 거부·삭제 등의 조치 — 는 앱스토어를 '게이트키퍼'로 기능하게 한다. 이는 EU 「DSA」가 온라인 플랫폼에 불법 콘텐츠에 대한 조치 의무를 부과하고, EU 「DMA」가 핵심 플랫폼

서비스의 게이트키퍼에 특별한 의무를 부과한 것과 유사한 규제 논리에 기반한다. 다만, EU의 경우 게이트키퍼 지정을 위한 양적 기준(이용자 수, 매출액 등)을 상세히 규정하고 있는 반면, 잠정조치안은 앱스토어 일반에 대해 의무를 부과하고 있어 적용 범위가 더 넓다.

7) 한계와 과제

잠정조치안의 규범적 혁신에도 불구하고, 몇 가지 한계와 향후 과제도 지적될 수 있다.

첫째, 주요 개념의 모호성 문제이다. '감정의 상호작용', '극단적 감정', '중독 현상', '비교적 큰 안전 위험' 등 핵심 개념들의 구체적 기준이 명시되어 있지 않아 향후 집행 과정에서 해석의 불확실성이 발생할 수 있다. 이러한 모호성은 한편으로 규제 당국에 유연한 해석 권한을 부여하고 기술 발전에 대응할 여지를 남기는 장점이 있지만, 다른 한편으로 수범자의 예측가능성을 저해하고 규제의 자의적 적용 가능성을 높이는 단점이 있다.

둘째, 이용자 상태 모니터링과 프라이버시 보호 간의 긴장 문제이다. 제11조가 요구하는 이용자 감정 상태 및 의존도 평가는 필연적으로 이용자의 대화 내용과 행동 패턴에 대한 분석을 수반한다. 조항 자체가 "개인 프라이버시 보호를 전제로" 한다고 명시하고 있으나, 실제 운용에서 안전 모니터링과 프라이버시 보호의 균형을 어떻게 달성할 것인지는 쉽지 않은 과제이다.

셋째, 인력 인계 의무의 현실적 이행 가능성 문제이다. 제11조 제3항의 인력 인계 의무는 이상적인 이용자 보호 수단이지만, 24시간 전문 인력의 확보, 다수의 동시 위기 상황 대응, 위기 판단의 정확성 등 현실적 과제가 상당하다. 특히 소규모 서비스 제공자의 경우 이러한 의무를 이행하기 위한 인적·재정적 자원이 부족할 수 있어, 시장 진입 장벽으로 작용할 가능성이 있다.

넷째, 이용자 자기결정권과 보호주의적 개입 간의 긴장 문제이다. 특히 고령자에 대한 친족 모사 금지 규정(제13조 제2항)의 경우, 고령 이용자가 자발적으로 사별한 배우자나 연락이 닿지 않는 자녀를 모사한 AI 동반자를 원하는 경우에도 이를 금지해야 하는지의 문제가 제기될 수 있다. 이는 취약계층 보호와 이용자 자율성 존중 사이의 본질적 긴장을 내포하는 쟁점이다.

3 일본의 현행법상 인공지능서비스 이용자 보호

가. 개관

현재 일본에는 AI에 대한 독자적인 규제법이나 이용자 보호를 목적으로 한 특별법이 존재하지 않는다. 2025년 5월 제정된 「AI 추진법」이 혁신 촉진과 산업 경쟁력 강화를 목적으로 하는 기본법으로서, 일본 최초의 AI 입법 사례이다.[98] 다만 이 법은 진흥법으로 사업자에 대한 직접적인 의무 부과나 위반 시 벌칙 규정을 두고 있지 않다.

이에 따라 일본의 AI 서비스 이용자 보호는 'AI 사업자 가이드라인'과 같은 비구속적 연성규범과, 「개인정보보호법」, 「부당경품류및부당표시방지법」(景品表示法) 등 기존 법률에 의존하는 구조를 취하고 있다. 이는 규제 최소화를 통해 기술혁신을 도모하는 일본의 정책 기조를 반영한 것이나 AI 관련 소비자 피해에 대응하기 위한 법적 구속력이 미비하다는 점은 향후 과제로 지적된다.[99]

나. 「AI 추진법」의 주요 내용

1) 제정 경위

일본 정부는 2024년 7월 AI 전략회의 산하에 AI 제도연구회를 설치하고, 법제도의 필요성을 포함하여 AI 제도의 방향성에 대해 검토를 진행하였다. 이 연구회가 2025년 2월 4일 공표한 「중간정리」(中間とりまとめ)는 ① 혁신 촉진과 리스크 대응의 양립, ② 기술 변화에 대응할 수 있는 유연한 제도 설계, ③ 국제적 상호운용성 확보, ④ 정부의 적정한 AI 조달·이용을 AI 입법의 기본원칙으로 제시하였다.[100] 이를 바탕으로 2025년 2월 28일 AI 촉진법안이 국회에 제출되어 같은 해 5월 28일 통과, 6월 4일 공포되었으며, 같은 해 9월 1일 전면 시행되었다.[101]

---

98) 内閣府, 人工知能関連技術の研究開発及び活用の推進に関する法律(令和7年法律第53号), <https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html>(최종방문일 2025. 12. 30).

99) Future of Privacy Forum, "Understanding Japan's AI Promotion Act: An Innovation-First Blueprint for AI Regulation," 2025, <https://fpf.org/blog/understanding-japans-ai-promotion-act/>(최종방문일 2025. 12. 30).

100) AI戦略会議·AI制度研究会, 「中間とりまとめ」, 2025. 2. 4, <https://www8.cao.go.jp/cstp/ai/ai_senryaku/13kai/13kai.html>(최종방문일 2025. 12. 30).

101) 衆議院, 「人工知能関連技術の研究開発及び活用の推進に関する法律案」,

2) 법률의 체계

AI 추진법은 총 4장 28개조 및 부칙으로 구성된 기본법 형식의 진흥법이다. 제1조는 “인공지능 관련 기술의 연구개발 및 활용 추진에 관한 시책을 종합적이고 계획적으로 추진”하는 것을 목적으로 규정하여, 규제보다 진흥에 방점을 두고 있음을 명시하고 있다.

각 장의 구성을 살펴보면, 제1장(총칙, 제1조~제10조)은 법의 목적과 인공지능 관련 용어를 정의하고, ‘인간 중심의 사회 구현’ 등 7대 기본이념을 제시한다. 또한 국가, 지방공공단체, 연구기관 및 활용사업자의 책무를 규정하여 민관 협력의 법적 근거를 마련하였다.

제2장(기본적 시책, 제11조~제17조)은 연구개발 촉진, 시설 및 설비 정비와 더불어 ‘AI 활용의 적정성 확보’를 명시하고 있다. 다만 이는 강제적 규제가 아닌 사업자의 자율적 준수와 이를 지원하기 위한 국가의 가이드라인 정비 등을 의미한다. 그 외 인재 양성, 교육 진흥, 국제협력 등 진흥을 위한 구체적 시책을 담고 있다.

제3장(인공지능기본계획, 제18조)은 정부가 범정부 차원의 AI 진흥 로드맵인 ‘인공지능 기본계획’을 수립하고 이를 정기적으로 공표할 의무를 규정한다.

제4장(인공지능전략본부, 제19조~제28조)은 내각총리대신을 본부장으로 하고 전 국무대신이 참여하는 ‘인공지능전략본부’의 설치를 규정한다. 여기에는 관계 행정기관 간 협력 체계(제25조), 지방공공단체의 독자적 시책 추진 지원, 국가의 재정적 조치 의무(제16조 연계) 등이 포함되어 법의 실효성을 담보한다.

부칙은 법의 시행일과 더불어 급변하는 AI 기술 환경을 고려하여 법 시행 후 일정 기간마다 시책 내용을 재검토하도록 하는 조항을 포함하고 있다.

이를 정리하면 다음 표와 같다.

<https://www.shugiin.go.jp/internet/itdb_gian.nsf/html/gian/honbun/houan/g21709029.htm> (최종방문일 2025. 12. 30).

〈표 3－5〉 일본「AI 추진법」의 체계

장	조문	주요 내용
제1장 총칙	제1조~제10조	• 목적, 정의, 기본이념, 국가·지방공공단체·사업자 책무
제2장 기본적 시책	제11조~제17조	• 연구개발 촉진, 시설 정비, 적정성 확보, 인재 양성, 국제협력
제3장 인공지능기본계획	제18조	• 정부의 기본계획 수립·공표 의무
제4장 인공지능전략본부	제19조~제28조	• 전략본부 설치, 관계기관 협력, 재정 조치, 지방정부 지원

동법의 특징은 민간 사업자에게 “국가가 실시하는 시책에 협력하도록 노력하여야 한다”(제7조)는 노력 의무만 부과하고, 별도의 벌칙 규정을 두지 않는다는 점이다. 일본 정부는 이러한 접근을 ‘애자일 거버넌스’(アジャイル・ガバナンス) 개념으로 설명하며, 급변하는 기술 환경에 유연하게 대응하기 위한 전략적 선택이라고 밝히고 있다.[102)]

3) 이용자 보호 관련 규정

AI 추진법에서 이용자 보호와 관련된 규정은 다음과 같이 제한적이며, 대부분 국가의 시책 방향을 제시하는 훈시적 규정에 해당한다.

■ 제3조제4항(기본이념)

AI의 부정한 목적 또는 부적절한 방법에 의한 연구개발·활용이 “범죄 이용, 개인정보 누설, 저작권 침해 기타 국민의 권리이익 침해를 조장할 우려”가 있으므로 “투명성 확보 기타 필요한 시책”을 강구해야 한다고 규정한다. 다만 이는 사업자에 대한 직접적 의무가 아닌 정부의 정책 방향을 제시하는 것에 그친다.

■ 제16조(적정 활용을 위한 조치)

국가가 AI 관련 권리 침해 사례를 수집·분석하고, 이를 바탕으로 사업자에게 “지도, 조

102) 経済産業省,「GOVERNANCE INNOVATION Ver.2: アジャイル・ガバナンスのデザインと実装に向けて」, 2021. 7, 〈https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_1.pdf〉 (최종방문일 2025. 12. 30).

언 및 정보 제공"을 할 수 있다고 규정한다. 그러나 이는 행정지도에 해당하는 비권력적 사실행위로서, 사업자가 이를 따르지 않더라도 제재할 법적 근거가 없다.

이러한 방식은 사업자에게 실질적인 주의 의무나 리스크 관리 의무를 직접 부과하지 않는다는 점에서 EU 「DSA」나 중국의 접근법과는 근본적으로 구별된다.

다. 기존 법률을 통한 이용자 보호

일본은 AI 서비스 이용자 보호를 위한 특별법이 없으므로, 기존의 개별 법률을 통해 이용자를 보호하는 구조를 취하고 있다. 주요 적용 법률로는 「개인정보보호법」, 「부당경품류및부당표시방지법(이하 '경품표시법')」, 「저작권법」 등이 있다.

1) 개인정보보호법[103]

생성형 AI 서비스의 확산에 따라 이용자의 개인정보가 AI 학습에 무단으로 활용되거나, 프롬프트 입력 과정에서 개인정보가 유출될 위험성이 제기되었다. 이에 일본 개인정보보호위원회는 2023년 6월 2일 「생성AI 서비스의 이용에 관한 주의환기」(生成AIサービスの利用に関する注意喚起等について)를 발표하여 생성형 AI 이용 시 개인정보 보호에 관한 지침을 제시하였다. 개인정보보호위원회는 동 주의 환기를 통해 개인정보취급사업자에게 다음 사항을 확인할 것을 요청하였다.

- 생성형 AI 서비스에 개인정보를 포함한 프롬프트를 입력할 경우, 해당 이용이 사전에 특정한 이용 목적의 범위 내의 것인지 확인할 것
- 입력된 개인정보가 해당 AI 서비스의 기계학습에 이용되지 않는지 확인할 것

아울러 개인정보보호위원회는 OpenAI에 대해서도 별도의 주의 환기를 발송하여, ① 정보 주체의 사전 동의 없이 민감정보(要配慮個人情報, 인종, 신조, 병력 등 특히 배려를 요하는 개인정보)를 취득하지 않을 것, ② 개인정보의 이용 목적을 일본어로 통지 또는 공표할 것을 요청하였다.

이러한 주의 환기 조치는 「개인정보보호법」 위반 사실의 인정에 기초한 행정처분이 아

103) 個人情報保護委員会, 「生成AIサービスの利用に関する注意喚起等について」, 2023. 6. 2, <https://www.ppc.go.jp/news/press/2023/230602kouhou/>(최종방문일: 2025. 12. 30).

니라 향후 발생할 수 있는 권리 침해 리스크를 사전에 저감하기 위한 예방적 조치의 성격을 갖는다. 따라서 법적 강제력을 수반하지 않으며, 사업자의 자발적 협력에 의존하는 한계가 있다.

2) 경품표시법(景品表示法)

AI 서비스 사업자가 자사 서비스의 성능이나 정확도를 허위·과장하여 광고하는 경우, 소비자는 실제와 다른 품질의 서비스를 이용하게 되어 경제적 피해를 입을 수 있다. 이러한 허위·과장 광고에 대해서는 경품표시법이 적용된다.[104] 동법 제5조는 두 가지 유형의 부당표시를 금지한다. 첫째, '우량오인표시'(優良誤認表示)는 상품이나 서비스의 품질, 규격 등의 내용에 관하여 실제보다 현저히 우량하다고 일반 소비자에게 오인시키는 표시를 말한다. 예컨대 AI 서비스가 실제로는 제한적인 정확도를 가지면서도 "99% 정확도" 등으로 과장 광고하는 경우가 이에 해당할 수 있다. 둘째, '유리오인표시'(有利誤認表示)는 가격 등 거래조건에 관하여 실제보다 현저히 유리하다고 오인시키는 표시를 말한다. 경품표시법 위반 시 소비자청은 조치명령(제7조)을 발할 수 있으며, 과징금납부명령(제8조)에 따라 위반행위 관련 매출액의 3%에 해당하는 과징금이 부과된다.

2024년 10월 시행된 개정 경품표시법은 제재를 대폭 강화하였다.[105] 주요 개정 내용은 다음과 같다.

- 직접벌 신설(제48조): 고의에 의한 우량오인표시 또는 유리오인표시에 대해 100만엔 이하의 벌금을 부과하는 직접벌 규정을 신설하여, 종래 행정처분 불이행 시에만 적용되던 형사처벌을 위반행위 자체에 대해 부과할 수 있게 되었다.
- 과징금 산정률 인상: 반복 위반 등 악질적인 경우 과징금 산정률을 최대 4.5%로 인상하였다.
- 확약수속 도입: 사업자가 자발적으로 시정계획을 제출하고 이를 이행하면 조치명령이나 과징금을 면제받을 수 있는 확약수속(確約手続) 제도가 도입되었다.

또한 2023년 10월부터 시행된 스텔스마케팅(ステルスマーケティング) 규제에 따라 광고

104) 不当景品類及び不当表示防止法(昭和37年法律第134号).

105) 消費者庁, 「【令和6年10月1日施行】改正景品表示法の概要」, 2024.

임을 숨긴 채 제3자의 추천인 것처럼 가장하는 표시가 부당표시로 규제된다.[106] 이에 따라 AI가 생성한 리뷰나 추천 콘텐츠를 광고 목적으로 활용하면서 광고임을 명시하지 않는 경우에도 경품표시법 위반이 성립할 수 있다.

다만 AI 서비스의 성능이나 정확도에 관한 허위·과장 표시에 특화된 가이드라인은 아직 마련되어 있지 않다. AI 서비스의 특성상 '정확도', '신뢰성' 등의 개념이 기존 상품·서비스와 다른 방식으로 측정·표시되어야 함에도, 현재로서는 일반적인 부당표시 규제 법리가 그대로 적용되는 구조여서 해석상 불명확한 부분이 존재한다.

3) 저작권법

AI 서비스 이용자가 생성형 AI를 통해 콘텐츠를 생성할 경우, 해당 생성물의 타인 저작권 침해 여부가 문제될 수 있다. 이는 AI 서비스 이용자 보호의 관점에서 중요한 쟁점이다.

일본「저작권법」[107] 제30조의4는 "저작물에 표현된 사상 또는 감정을 자기가 향수하거나 타인에게 향수시키는 것을 목적으로 하지 않는 경우"에는 필요한 한도 내에서 권리자의 허락 없이 저작물을 이용할 수 있다고 규정한다. 이 조항에 따라 AI 학습을 위한 대규모 데이터 수집·분석 등 '비향수적 이용'(非享受的利用)이 광범위하게 허용되며, 일본은 AI 학습에 가장 유연한 법적 환경을 갖춘 국가로 평가받는다.

그러나 AI가 생성한 결과물이 기존 저작물과 실질적으로 유사한 경우에는 저작권 침해가 성립할 수 있다. 이 경우 AI 서비스 이용자가 침해 책임을 부담할 가능성이 있어, 이용자 보호 측면에서 우려가 제기된다. 이에 문화청 문화심의회 저작권분과회 법제도소위원회는 2024년 3월「AI와 저작권에 관한 생각에 대하여」(AIと著作権に関する考え方について)를 공표하여 AI와 저작권의 관계에 대한 해석 지침을 제시하였다.[108] 동 지침은 AI 학습 단계와 생성·이용 단계를 구분하여 저작권법 적용 기준을 명확히 하고자 하였으나,[109] 구

106) 消費者庁,「景品表示法とステルスマーケティング ~ 事例で分かるステルスマーケティング告示ガイドブック」, 2023.

107) 著作権法(昭和45年法律第48号).

108) 文化審議会著作権分科会法制度小委員会,「AIと著作権に関する考え方について」, 2024. 3, <https://www.bunka.go.jp/>(최종방문일 2025. 12. 30).

109) 동 지침에 따르면, AI 학습 단계에서 저작물을 이용하는 행위는 원칙적으로 제30조의4의 비향수적 이용에 해당하여 권리자의 허락 없이 가능하다. 다만 ① 정보해석 목적

체적인 침해 판단 기준에 대해서는 여전히 해석상 불명확한 부분이 남아 있다.

마. 이용자 보호 관점에서의 일본 법제의 과제와 한계

일본의 AI 정책은 경직된 사전규제 대신 유연한 지침과 사후 점검을 중시하는 '애자일 거버넌스' 개념을 채택하고 있다. 이러한 접근은 혁신 저해를 최소화하고 기술 변화에 민첩하게 대응할 수 있다는 장점이 있으나, 이용자 보호의 법적 확실성과 집행력 측면에서는 한계가 있다. KPMG의 글로벌 조사에 따르면, 일본 국민의 77%가 AI 규제의 필요성에 동의하면서도, 현행 법제 하에서 AI를 안전하게 이용할 수 있다고 응답한 비율은 13%에 불과하였다.[110] 이는 연성규범 중심의 접근만으로는 이용자의 신뢰를 충분히 확보하기 어려움을 시사한다. EU 「DSA」가 플랫폼 사업자에게 다크 패턴 금지(제25조), 추천 시스템 투명성(제27조), 광고 투명성(제26조) 등 구체적이고 구속력 있는 의무를 부과하는 것과 비교할 때, 일본의 법제는 이용자 보호를 위한 직접적인 규율 수단이 부재한 상태라 할 수 있다.

4. 종합 검토 및 시사점

위에서 살펴본 EU, 중국, 일본의 입법 사례들은 방송미디어통신위원회 소관의 '인공지능 서비스 이용자 보호법'을 제정하는 것과 관련하여 다음과 같은 시사점을 제공한다.

가. 규제 형식의 선택: 경성법과 연성규범의 조합

세 국가의 경험은 연성규범만으로는 이용자 보호의 실효성을 담보하기 어렵다는 점을

외에 저작물의 표현을 향수하는 목적이 병존하는 경우, ② 필요한 한도를 초과하는 경우, ③ 저작권자의 이익을 부당하게 해치는 경우에는 권리제한 규정이 적용되지 않는다. 한편 AI 생성물의 이용 단계에서는, 생성물이 기존 저작물과 '유사성'(類似性) 및 '의거성'(依拠性)이 인정되면 저작권 침해가 성립할 수 있다. 특히 AI가 학습데이터에 포함된 저작물과 유사한 결과물을 생성한 경우, 이용자가 해당 저작물의 존재를 인식하지 못했더라도 침해 책임이 발생할 수 있어 이용자 보호 측면에서 주의가 요구된다.

110) KPMG, "Trust in Artificial Intelligence: Global Insights 2023," 2023, <https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2023/02/trust-in-ai-global-study-2023.pdf>(최종방문일 2025. 12. 30).

시사한다. 일본의 사례에서 보듯이, 가이드라인 중심의 접근은 기술혁신 친화적이나 이용자의 신뢰 확보에 한계가 있으며, 실제 피해 발생 시 구제 수단이 미비하다. 반면 EU와 중국은 법적 구속력이 있는 규제를 통해 사업자 의무를 명확히 하고 집행력을 확보하고 있다.

우리나라의 경우, 「AI 기본법」이 2025년 제정되었으나 이용자 보호에 관한 구체적 규율은 미흡하다. 방송미디어통신위원회가 추진하는 AI 서비스 이용자 보호 법률은 이러한 공백을 메우는 역할을 할 수 있다. 이때 구체적인 사업자 의무와 이용자 권리를 법률로 규정하되, 기술 발전에 대응한 유연성 확보를 위해 세부 사항은 하위 법령이나 가이드라인에 위임하는 방식이 적절할 것이다.

나. 적용 범위의 설정

EU 「DSA」는 온라인 플랫폼 전반에 적용되는 수평적 규제인 반면, 중국 잠정조치안은 '의인화 상호작용 서비스'라는 특정 유형에 특화된 수직적 규제이다. 국내 입법에서는 적용 범위를 어떻게 설정할 것인지가 중요한 정책 결정 사항이다. 이와 관련하여 몇 가지 방안을 검토할 수 있다. 우선, AI 서비스 전반에 적용되는 일반 규정과 특정 유형(예: 생성형 AI, AI 동반자 서비스 등)에 적용되는 특별 규정을 병행하는 이원적 구조가 고려될 수 있다. 또한 EU 「DSA」와 같이 서비스의 규모나 영향력에 따라 차등화된 의무를 부과하는 계층적 구조도 검토할 만하다. 아울러 중국 잠정조치안의 '의인화 상호작용 서비스' 정의는 AI 동반자, 감정 교류 AI 등 새로운 유형의 서비스를 포착하기 위한 개념 도구로서 참고할 가치가 있다.

다. 투명성 및 고지 의무의 설계

이용자가 AI 서비스의 작동 원리를 이해하고 정보에 기반한 의사결정을 할 수 있도록 투명성 의무를 부과하는 것은 세 국가 모두에서 공통적으로 강조되는 사항이다. 국내 입법에서는 다음 사항을 고려할 필요가 있다. 먼저, AI 시스템임을 명확히 고지하는 의무가 필요하다. 중국 잠정조치안 제16조와 같이 이용자가 AI와 상호작용하고 있음을 '현저하게' 인지할 수 있도록 하는 규정이 이에 해당한다. 다음으로, 추천·개인화 알고리즘의 주요 매개변수 공개를 고려할 수 있다. EU 「DSA」 제27조를 참고하여, AI 기반 추천 시스템이 콘텐츠 우선순위를 결정하는 데 사용하는 주요 요소를 이용자에게 공개하도록 하는 것이다. 다만, 영업비밀 보호와의 균형, 공개 정보의 실질적 유용성 등을 고려한 세부 설계가 필요

하다. 나아가 동적 고지 의무도 검토할 만하다. 중국 잠정조치안 제16조 제2항과 같이 의존·중독 경향이 감지되거나 새로운 세션 시작 시 AI의 본질을 상기시키는 동적 고지는 의인화 AI 서비스의 특수한 위험에 대응하는 효과적 수단이 될 수 있다.

라. 이용자 선택권 및 자기결정권의 보장

이용자의 자율성을 보장하기 위한 선택권 부여는 AI 서비스 이용자 보호의 핵심 요소이다. EU「DSA」제27조 제3항을 참고하여, AI 기반 추천 서비스 제공 시 개인화되지 않은 대안적 옵션을 제공하도록 의무화하는 방안을 검토할 수 있다. 이는 이용자가 알고리즘에 의한 개인화를 거부하고 필터 버블에서 벗어날 수 있는 자율성을 보장한다. 데이터 통제권의 측면에서는 중국 잠정조치안 제14조~제15조를 참고하여, 이용자에게 상호작용 데이터 삭제권, AI 모델 훈련에 대한 별도 동의권 등을 부여하는 방안이 있다. 이탈의 자유와 관련해서는 중국 잠정조치안 제18조 및 EU「DSA」제25조의 다크 패턴 금지 규정을 참고하여, 서비스 종료를 원하는 이용자가 쉽게 이탈할 수 있도록 보장하고, 이를 방해하는 설계를 금지할 필요가 있다.

마. 취약계층 보호의 강화

미성년자와 고령자 등 취약계층에 대한 강화된 보호는 세 국가 법제의 공통된 방향이나, 중국 잠정조치안이 가장 상세하고 구체적인 규정을 두고 있다. 미성년자 보호와 관련하여, 미성년자 전용 모드 구축 의무, 보호자 동의 및 통제 기능 제공, 미성년자에 대한 프로파일링 기반 광고·추천 제한, 사용시간 제한 기능 등을 규정할 수 있다. 특히 AI 동반자 서비스와 같이 정서적 영향이 큰 서비스에 대해서는 강화된 보호가 필요하다. 고령자 보호에 관해서는 중국 잠정조치안 제13조의 긴급연락인 제도, 친족 모사 금지 규정이 참고가 된다. 이는 고령자 특유의 사회적 고립, 디지털 리터러시 부족, 사기 피해 취약성에 대응할 수 있는 제도적 장치이다.

바. 정서적 안전의 보호

중국 잠정조치안이 개척한 정서적 안전 보호 영역은 국내 입법에 중요한 시사점을 제공한다. AI 동반자 서비스의 확산과 2024년 Character.AI 관련 청소년 자살 사건 등은 AI가 이용자의 심리·정서에 미치는 영향이 콘텐츠의 적법성 문제를 넘어서는 독자적 규제 영

역임을 보여준다. 구체적으로, 중국 잠정조치안 제7조 제5호와 같이 AI를 통한 감정 조작, 심리적 취약성 악용을 명시적으로 금지하는 규정을 검토할 수 있다. 또한 동 조치안 제9조 제2항과 같이 이용자의 심리적 의존이나 중독을 유도하는 것을 서비스 설계 목표로 삼는 것을 금지하는 방안도 고려할 만하다. 연속 사용 시간에 대한 알림 또는 제한 기능의 제공을 의무화하는 것도 중독 방지를 위한 유효한 수단이 될 수 있다. 위기 개입 체계와 관련하여, 중국 잠정조치안 제11조의 단계적 위기 대응 체계 — 이용자 상태 식별, 고위험 경향 시 응답 템플릿 출력, 극단적 상황 시 인력 인계 — 는 생명·안전 관련 위기 상황에 대응하기 위한 혁신적 규정으로서 그 도입 가능성을 검토할 필요가 있다. 다만, 인력 인계 의무의 현실적 이행 가능성, 소규모 사업자에 대한 부담 등을 고려한 세부 설계가 요구된다.

사. 다크 패턴의 금지

EU「DSA」제25조의 다크 패턴 금지 규정은 온라인 인터페이스 설계 전반에 걸친 규제로서, 이용자의 자유로운 의사결정을 보호하는 효과적 수단이다. 국내 입법에서도 이용자를 기만하거나 조작하는 설계를 금지하는 규정을 검토할 수 있다. 금지 대상으로는 특정 선택지를 시각적으로 유도하는 설계, 동의 철회나 서비스 탈퇴를 가입보다 어렵게 만드는 설계, 불필요한 정보 제공이나 설정을 강제하는 설계, 감정에 호소하거나 긴급성을 조성하여 이용자를 조작하는 설계, 프라이버시 침해적 설정을 기본값으로 하는 설계 등이 포함될 수 있다.

아. 플랫폼 책임의 명확화

AI 서비스 생태계의 다층적 구조를 고려하여, 서비스 제공자뿐 아니라 배포 플랫폼(앱스토어 등)의 책임도 명확히 할 필요가 있다. EU「DSA」의 누적적 의무 구조를 참고하여, 서비스의 성격과 이용자 규모에 비례하는 차등화된 의무를 부과하는 방안이 고려될 수 있다. 배포 플랫폼의 책임과 관련해서는 중국 잠정조치안 제24조를 참고하여, 앱스토어 등 배포 플랫폼에 등재 심사, 안전 평가 확인, 위반 시 삭제 등의 의무를 부과하는 것을 검토할 수 있다.

자. 실효적 집행 메커니즘의 구축

규제의 실효성을 담보하기 위해서는 적절한 집행 메커니즘의 구축이 필수적이다. 중국

의 알고리즘 비안 제도를 참고하여, 일정 규모 이상의 AI 서비스에 대한 등록 또는 신고 제도를 검토할 수 있다. 이는 규제 당국이 AI 서비스 현황을 파악하고 감독하는 기초가 된다. 안전 평가와 관련해서는 중국 잠정조치안 제21조~제23조를 참고하여, 일정 요건을 충족하는 AI 서비스에 대해 안전 평가를 실시하고 그 결과를 규제 당국에 보고하도록 하는 방안이 있다.

제재 수단과 관련하여, 위반 행위에 대한 실효적 제재 — 시정명령, 과징금, 서비스 정지 등 — 를 규정할 필요가 있다. EU 「DSA」의 매출액 비례 과징금 제도는 대형 플랫폼에 대한 억지력 확보에 효과적인 수단이 될 수 있다. 민원 처리 체계의 측면에서는 사업자에게 민원·신고 처리 절차의 구축을 의무화하고, 처리 기한 및 결과 통보 의무를 부과하는 것이 이용자 권리 구제에 기여할 것이다.

차. 기존 법체계와의 정합성 확보

AI 서비스 이용자 보호 법률의 제정에 있어, 기존 법체계와의 정합성 확보가 중요하다. 「AI 기본법」과의 관계에서, 동법이 제시하는 원칙과 정책 방향을 구체화하는 이행 법률로서의 성격을 명확히 할 필요가 있다. 「개인정보 보호법」과의 관계에서는, AI 서비스의 개인정보 처리에 관해 개인정보 보호법이 일반법으로서 적용되며, AI 서비스 이용자 보호 법률은 AI 서비스의 특수성을 고려한 특별 규정을 두는 방식으로 관계를 정립할 수 있다. 「정보통신망법」과의 관계에서는, 온라인 서비스 이용자 보호에 관한 기존 규율과의 중복·상충을 피하고 보완적 관계를 설정할 필요가 있다. 분야별 개별 법률과의 관계도 고려해야 한다. 중국 잠정조치안 제31조와 같이, AI 서비스가 의료, 금융, 법률 등 전문 분야에 적용되는 경우 해당 분야의 규제도 동시에 적용됨을 명확히 하는 규정을 둘 수 있다.

AI 서비스가 이용자의 정보 소비, 의사결정, 나아가 정서와 심리에까지 영향을 미치는 시대에, 이용자 보호를 위한 규범적 대응은 더 이상 미룰 수 없는 과제가 되었다. EU, 중국, 일본의 법제는 각기 다른 접근법을 보여주지만, 공통적으로 AI 서비스의 투명성 확보, 이용자 선택권 보장, 취약계층 보호의 필요성을 인식하고 있다. 특히 중국 잠정조치안이 제시한 '정서적 안전'이라는 새로운 보호 영역은 주목할 만하다. AI 동반자 서비스로 인한 비극적 사건들은 콘텐츠의 적법성이나 개인정보 보호만으로는 포착할 수 없는 새로운 유형의 위험이 존재함을 보여주었다. 이러한 위험에 대응하기 위해서는 AI가 인간의 심리·

정서에 미치는 영향 자체를 규율 대상으로 삼는 발상의 전환이 필요하다. 국내 AI 서비스 이용자 보호 법률의 제정은 이러한 시대적 요청에 부응하는 것이다. 이용자의 자율성과 안전을 실효적으로 보장하면서도 기술 혁신의 동력을 유지하는 균형 잡힌 규제 체계의 설계가 향후 입법 과정의 핵심 과제가 될 것이다.

제4장 인공지능 이용자 보호에 관한 국내 선행 사례 검토

제1절 서 설

인공지능(AI) 기술은 미디어 산업의 전 영역에 걸쳐 혁신적인 변화를 일으키고 있다. 뉴스 추천 알고리즘부터 생성형 AI를 활용한 콘텐츠 제작에 이르기까지 AI는 정보의 생산과 유통 방식을 근본적으로 재편하고 있다. 그러나 이러한 기술적 진보와 동시에 알고리즘의 불투명성으로 인한 여론 왜곡, 정보 편향성, 딥페이크를 통한 허위 정보 유통 등 이용자의 권익을 침해할 수 있는 새로운 위험 요소들이 대두되고 있다. 이에 따라 미디어 분야에서도 AI 서비스의 신뢰성을 확보하고 이용자를 두텁게 보호하기 위한 체계적인 입법적 대응이 절실한 시점이다.

특히 현대 미디어 환경에서 알고리즘은 이용자가 접하는 정보를 결정하는 '보이지 않는 편집자'의 역할을 수행한다. 알고리즘의 작동 원리를 알기 어려운 '불투명성'은 이용자가 특정 정보를 추천받은 이유를 알 수 없게 만들며, 이는 이용자를 단순한 서비스 소비 대상에 머물게 하는 한계를 지닌다. 따라서 이용자가 자신의 권익을 보호하고 주체적으로 정보를 소비할 수 있도록 '설명요구권'이나 '이의제기권'과 같은 법적 권리를 명문화할 필요가 있다.

이러한 입법 과정에서 선제적으로 AI 규제 체계를 구축해 온 금융 분야와 의료 분야의 사례를 참고할 필요가 있다. 금융 분야는 기술중립적 접근 방식을 기반으로 하면서도, AI 시스템이 금융 소비자에게 미치는 영향에 따라 서비스를 분류하고 고위험 서비스에 대해 엄격한 내부통제 절차를 두는 '리스크 기반 접근 방식'(Risk-based Approach)을 채택하고 있다. 또한 자동화된 결정에 대한 설명요구권과 거부권을 법적으로 보장하여 이용자의 대응권을 강화하고 있다.

의료 분야는 「디지털의료제품법」을 통해 AI을 직접적인 규제 대상으로 명시하고, 학습 데이터의 정보 공개와 심사 결과의 투명한 공개를 의무화하여 사용자의 알 권리를 보장한다. 아울러 제조 및 품질관리 기준(GMP) 심사를 통해 AI 시스템의 전 생애주기에 걸친 조

직적 거버넌스와 품질관리를 요구하고 있다.

미디어 분야의 입법은 이러한 타 분야의 핵심 원칙들, 즉 금융 분야의 '리스크 기반 관리와 설명권 보장', 그리고 의료 분야의 '기술적 투명성 확보와 전 주기 거버넌스 체계'를 미디어 환경의 특수성에 맞게 수용해야 한다. 본 장에서는 금융 및 의료 분야의 보호 체계를 상세히 검토하고, 이를 바탕으로 미디어 알고리즘의 투명성 제고와 생성형 AI 고유의 위험 대응을 위한 실효적인 입법 방안을 제안하고자 한다.

제 2 절 금융 분야 인공지능 이용자 보호

1. 금융 분야 인공지능 이용자 보호 관련 법령

금융규제는 일반적으로 금융회사가 어떠한 기술을 사용하는 지와 상관없이 기술의 이용으로부터 발생하는 책임을 금융 회사에게 귀속시키는 '기술중립적'(technologically neutral) 접근방식을 따른다. 하지만 경우에 따라 AI에 특화된 규제를 도입할 필요가 있다. 금융규제는 금융회사가 AI 시스템이 발생할 수 있는 위해를 정확하게 인식하고 이에 대하여 적절한 대응 조치를 취할 수 있도록 하여야 한다. 이를 위해서 어떠한 금융규제가 AI 시스템의 활용 맥락에서 대응 조치로서 적용될 수 있는지를 명확하게 할 필요가 있다. 그리고 AI 시스템의 특성과 활용 방식으로 인하여 고유한 위해가 발생하는 특정 분야에 대해서는 기존 규제를 강화하거나 새로운 규제를 도입할 필요도 있을 수 있다.

기존 규제 중 금융 분야에서의 AI 활용에 적용될 수 있는 주요 규제는 다음과 같다.

가. 금융소비자보호법 및 신용정보법상 차별금지 규제

「금융소비자 보호에 관한 법률」(이하 '금융소비자보호법')과 「신용정보의 이용 및 보호에 관한 법률」(이하 '신용정보법')은 일정한 차별금지사유를 명시하고 이를 금지하는 규정을 두고 있다. 이 규정들은 AI 서비스를 타겟으로 하지는 않으나, 금융기관(금융상품판매업자등 또는 개인신용평가회사)가 어떠한 기술을 사용하는지와 무관하게 금융기관이 차별적 행위를 하면 그에 대한 책임이 인정되므로, 금융 분야 AI 서비스도 차별금지사유를 위반하면 규제 적용의 대상이 된다.

「금융소비자보호법」은 "금융상품판매업자등은 금융상품 또는 금융상품자문에 관한 계

약을 체결하는 경우 정당한 사유 없이 성별·학력·장애·사회적 신분 등을 이유로 계약조건에 관하여 금융소비자를 부당하게 차별해서는 아니 된다”라 규정하고 있다(법 제15조).

「신용정보법」은 개인신용평가회사가 개인신용평가 시 (i) 성별, 출신 지역, 국적 등으로 합리적 이유 없이 차별하거나 (ii) 정당한 이유 없이 계열회사와 금융거래 등 상거래 관계를 맺거나 맺으려는 사람의 개인신용평점을 다른 사람의 개인신용평점에 비해 유리하게 산정하는 등 차별적으로 취급하는 행위를 금지하고 있다(법 제22조의4 제1항).

한편, 「금융소비자보호법」은 금융상품판매업자 등이 계약 체결을 권유하는 경우 금융상품의 내용을 사실과 다르게 알리는 행위를 금지하고 있다(법 제17조 제2호). 따라서 챗봇 등 AI 서비스가 고객에게 금융상품의 내용을 사실과 다르게 알리는 경우 「금융소비자보호법」 위반이 문제될 수 있다.

나. 신용정보법 및 개인정보보호법상 투명성 규제

「신용정보법」 제36조의2의 ‘자동화평가 결과에 대한 설명 및 이의제기’ 규정, 2024년 3월 15일부터 시행된 개정 「개인정보보호법」 제37조의2의 ‘자동화된 결정에 대한 정보주체의 권리’ 규정은 완전자동화된 중요 의사결정에 대하여 개인에게 설명요구권과 거부권·이의제기권을 인정하고 있다.

1) 신용정보법상 자동화평가 대응권

「신용정보법」은 개인인 신용정보주체에게 자동화평가에 대하여 설명요구권, 정보제출권, 이의제기권(이하 ‘「신용정보법」상 대응권’)을 부여하고 있다(법 제36조의2 제1항 및 제2항). 여기서 “자동화평가”란 신용정보회사 등의 종사자가 평가 업무에 관여하지 아니하고 컴퓨터 등 정보처리장치로만 개인신용정보 및 그 밖의 정보를 처리(이하 “완전 자동화”)하여 개인인 신용정보주체를 평가하는 행위를 의미한다(법 제2조 제14호).

「신용정보법」은 대응권이 대상이 되는 자동화평가를 열거하고 있는데 ① 개인신용평가와 ② 신용공여, 대출, 신용카드, 시설대여, 할부금융 거래 등에 대한 설정, 유지, 계약의 내용, 청약, 승낙 여부의 결정이 여기에 포함된다(법 제36조의2 제1항 제1호 각목). 즉 개인신용평가 외에도 신용공여, 대출, 신용카드 등의 금융거래 설정 여부를 완전 자동화된 형태로 결정하는 경우에는 「신용정보법」상 대응권이 인정된다.

개인인 신용정보주체는 ① 개인신용평가 등 정보처리장치로만 처리하면 개인신용정보

보호를 저해할 우려가 있는 행위에 자동화평가를 하는지의 여부 및 ② 자동화 평가를 하는 경우 자동화평가의 결과, 자동화평가의 주요 기준, 자동화평가에 이용된 기초정보의 개요, 그밖에 이와 유사한 사항으로서 대통령령으로 정하는 사항에 대하여 설명해 줄 것을 요구할 수 있다.

그리고 개인인 신용정보주체는 개인신용평가회사등에 대하여 해당 신용주체에게 자동화평가 결과의 산출에 유리하다고 판단되는 정보를 제출할 수 있고, 자동화평가에 이용된 기초정보의 내용이 정확하지 아니하거나 최신의 정보가 아니라고 판단되는 경우 ① 기초정보를 정정하거나 삭제할 것을 요구하는 행위 및 ② 자동화평가 결과를 다시 산출할 것을 요구하는 행위를 할 수 있다.

2) 개인정보 보호법상 자동화된 결정 설명·거부권

「개인정보 보호법」은 완전히 자동화된 시스템으로 개인정보를 처리하여 이루어지는 결정이 정보주체 자신의 권리 또는 의무에 중대한 영향을 미치는 경우 정보주체에게 해당 결정을 거부하거나 설명 및 검토를 요청할 수 있는 권리를 부여하고 있다(법 제37조의2 제1항 및 제2항, 시행령 제44조의 제2항). 개인정보처리자는 정보주체가 자동화된 결정을 거부하거나 이에 대한 설명 또는 검토를 요구한 경우에는 정당한 사유가 없는 한 자동화된 결정을 적용하지 아니하거나 인적 개입에 의한 재처리·설명 등 필요한 조치를 하여야 한다(법 제37조의2 제3항).

「개인정보 보호법」 시행령은 정보주체에게 제공해야 할 설명의 내용을 크게 ① 자동화된 결정에 공통적으로 적용되는 기준과 절차에 관한 정보(사전 공개 사항)와 ② 정보주체에게 실제로 내려진 개별 결정에 대한 설명(개별 결정에 대한 설명)으로 구분하고 있다. 자동화된 결정이 이루어지면, 개인정보처리자는 '자동화된 결정에 공통적으로 적용되는 기준과 절차에 관한 정보'를 홈페이지 등을 통해 사전 공개해야 한다(시행령 제44조의4 제1항). 사전 공개 사항은 자동화된 결정이 이루어진다는 사실과 그 목적 및 대상이 되는 정보주체의 범위, 자동화된 결정에 사용되는 주요 개인정보의 유형과 자동화된 결정의 관계, 자동화된 결정 과정에서의 고려사항 및 주요 개인정보가 처리되는 절차 등으로 구성된다(시행령 제44조의4 제1항 각호).

그리고 만약 자동화된 결정이 정보주체의 권리 또는 의무에 '중대한 영향'을 미치는 경

우에는 정보주체의 설명 요구에 대하여 '실제로 내려진 개별 결정에 대한 설명'을 제공해야 한다(시행령 제44조의3 제2항 본문). 개별 결정에 대한 설명에는 해당 자동화된 결정의 결과, 사용된 주요 개인정보의 유형, 주요 기준, 절차에 관한 사항이 포함되어야 한다(시행령 제44조의3 제2항 각호). 자동화된 결정이 정보주체의 권리 또는 의무에 중대한 영향을 미치지 않는 경우에는 정보주체의 설명 요구에 대하여 사전 공개 사항을 제공하면 된다(시행령 제44조의3 제2항 단서).

2. 금융 분야 인공지능 이용자 보호 관련 가이드라인

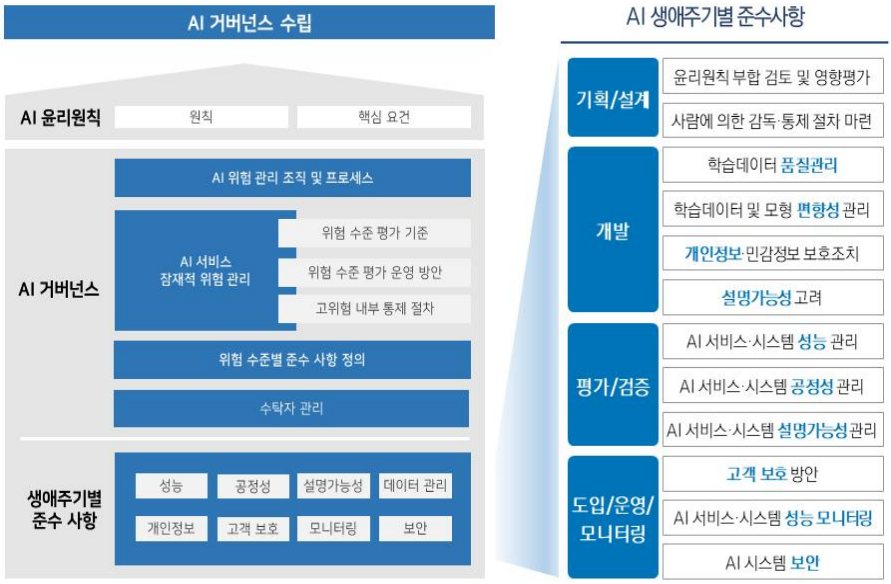
가. 금융위원회 '금융분야 인공지능 개발·활용 안내서'

금융위원회는 2021년 7월 금융 분야에서의 AI 시스템의 개발, 사업화 및 활용과 관련한 기획·설계, 평가·검증, 도입·운영 및 모니터링의 전 과정에서 신뢰성을 제고하여 AI 활성화를 제고하고 금융서비스에 대한 고객 신뢰를 확보하는 데 기여하는 것을 목적으로 '금융분야 인공지능 운영 가이드라인'을 발표하였다.

이후 금융업권 전파와 가이드라인 이행을 위하여 약 1년여 간의 준비기간 동안 금융회사 실무자들의 의견을 수렴하였으며, 이를 바탕으로 2022년 8월 금융회사의 AI 서비스의 개발·운영 과정에서의 규제 불확실성을 해소하는 한편 AI 활용 과정에서 발생 가능한 리스크를 예방·관리하는 데 목적을 둔 '금융분야 인공지능 개발·활용 안내서'(이하 '금융분야 인공지능 안내서')를 마련했다.

'금융분야 인공지능 안내서'는 신용평가·여신심사, 로보어드바이저, 챗봇, 맞춤형 추천, 이상거래탐지(FDS)등 5대 서비스를 대상으로 하고 있고, 크게 ① AI 위험 관리 정책의 수립과 ② AI 시스템 생애주기(기획/설계, 개발, 평가 및 검증, 도입·운영 및 모니터링)별 준수사항을 구분하여 설명하고 있다.

〔그림 4-1〕 금융분야 인공지능 안내서의 구성



AI 위험 관리 정책의 수립의 경우, 안내서는 리스크 기반 접근 방식을 적용하고 있다. 즉 AI 시스템의 위험을 평가하고 위험 수준에 맞는 내부통제 절차를 두는 것을 위험 관리 정책 수립의 핵심 내용으로 두고 있다. 특히 안내서는 금융회사로 하여금 금융소비자에게 미치는 영향, 위험 요소 등을 평가하여 AI 서비스를 고위험·중위험·저위험 등으로 분류하는 기준을 마련하고, 고위험 서비스에 대해서는 책임있는 업무수행이 가능한 자 또는 위원회가 검토 후 승인할 수 있는 내부절차를 둘 것을 요구하고 있다.[111] 나아가 금융회사가 AI 시스템의 개발·운영 등을 외부기관에 위탁하는 경우, 금융회사가 위탁자로서 AI 위험 관리의 책임을 부담한다는 전제 하에 수탁기관이 위험관리 체계를 충분히 마련하도록 금융회사에게 수탁기관을 관리하도록 하고 있다.[112]

생애주기별 준수 사항 부분의 경우, AI 시스템의 성능 확보, 공정성, 설명가능성, 데이터 관리, 보안 등 AI 거버넌스의 핵심 원칙들을 이행하기 위한 사항들이 5대 서비스별(신용평

111) 금융위원회, "금융분야 인공지능 개발·활용 안내서", 2022., p. 18.
112) 금융위원회, "금융분야 인공지능 개발·활용 안내서", 2022., pp. 20-24.

가·여신심사, 로보어드바이저, 챗봇, 맞춤형 추천, 이상거래탐지)로 구분되어 상세하게 설명되어 있다. 안내서는 생애주기별 준수 사항에 대해서도 AI 서비스의 위험 수준별로 준수 사항을 구체적으로 정하도록 하여 리스크 기반 접근 방식을 적용하고 있다. 생애주기별 준수 사항 부분은 금융회사에게 실무지침으로서 기능하기 위하여 이행 사항들을 5대 서비스별로 세분화하여 구체적인 설명과 예시를 제공하고 있다.[113]

다만 '금융분야 인공지능 운영 가이드라인' 및 '금융분야 인공지능 안내서' 모두 생성형 AI 시스템이 본격적으로 활용되기 시작한 2023년 이전에 발표되어 이후 개정되지 않았기 때문에, 생성형 AI 시스템 활용 시 발생할 수 있는 위험과 그에 대한 관리를 위해 준수해야 할 사항에 대해서도 별도 설명을 제공하고 있지 않다.

나. 금융위원회 '금융분야 인공지능 가이드라인 개정' 계획

금융위원회는 2024년 12월 "금융권 생성형 인공지능 활용 지원방안"의 하나로 "금융분야 인공지능 가이드라인 개정" 계획을 발표했다.[114] 생성형 AI의 출현과 금융권 내부통제 강화 등 금융권 제도 변화를 반영하여 기존에 발표했던 '금융분야 인공지능 운영 가이드라인'(2021. 7.), '금융분야 인공지능 개발·활용 안내서'(2022. 8.), '금융분야 인공지능 보안 가이드라인'(2023. 4.)을 통합한 '금융분야 인공지능 가이드라인'(이하 '통합 금융분야 인공지능 가이드라인')을 마련하는 것을 내용으로 한다.

금융위원회는 '통합 금융분야 인공지능 가이드라인'에 AI 개발·활용에 있어 최고경영자 및 경영진의 역할과 책임 확보, 금융 안정성에 대한 위험 최소화 등 '금융 인공지능 7대 원칙'을 반영할 계획이다.

113) 금융위원회, "금융분야 인공지능 개발·활용 안내서", 2022, pp. 26-115.

114) 금융위원회, "금융권 생성형 인공지능 활용 지원방안", 2024, p. 6.

〈표 4－1〉 금융 인공지능 7대 원칙

분 야	7대 원칙
거버넌스	① 최고경영자를 포함한 경영진은 인공지능 개발·활용에 대한 관심을 갖고 역할과 책임을 분담해야 함
	② 인공지능 활용 전단계에서 금융·인공지능 등 관련 법규를 준수해야 함
	③ 현 단계에서 인공지능은 업무의 보조 수단이므로 최종 의사결정과 그에 따른 책임은 임직원이 수행함
인공지능 개발 단계	④ 인공지능 개발 과정에서 신뢰할 수 있는 데이터와 모델을 사용해야 함
	⑤ 인공지능 설계·학습 등 전 과정에서 금융 안정성 위험을 최소화해야 함
인공지능 활용 단계	⑥ 인공지능 활용 시 금융소비자의 이익을 최우선으로 해야 함
	⑦ 인공지능 활용 시 보안성 기준 및 점검·개선 체계를 마련해야 함

그리고 '금융 인공지능 원칙'의 적용, 생성형 AI 관련 보안 위협과 보안성 검증 기준 등을 포함한 상세 설명과 사례는 안내서를 통해 제시할 예정이다.

기존 '금융 분야 인공지능 가이드라인' 및 안내서와 비교하여 금융 인공지능 7대 원칙에서 새롭게 도입된 내용은 다음과 같다. 우선 7대 원칙은 최고경영자를 포함한 경영진에게 인공지능 개발·활용에 대하여 역할과 책임을 분담할 것을 요구하고 있다. 이는 「금융회사의 지배구조에 관한 법률」에 따른 금융회사 임원의 내부통제 및 위험관리 의무를 AI 개발·활용 시에도 이행해야 한다는 점을 확인한 것이다(제30조의2). AI 거버넌스의 구축을 금융회사의 내부통제 및 위험관리 활동의 일부로 두려는 금융당국의 입장을 볼 수 있다.

그리고 '금융 인공지능 7대 원칙'은 AI 시스템 공급망의 분화에도 AI 시스템의 개발·활용에 따른 책임은 최종적으로 서비스를 제공하는 금융회사 임직원이 부담해야 한다는 점, 금융회사는 AI 활용이 금융소비자뿐만 아니라 금융시스템에 미치는 영향까지 평가하고 위험 관리 방안을 마련해야 한다는 점, 기존 판별형 AI 시스템과 구분되는 생성형 AI 시스템 고유의 보안 위협에 대한 대응 체계를 마련해야 한다는 점을 요구하고 있다.

제3절 의료 분야 인공지능 이용자 보호

1. 의료 분야 인공지능 이용자 보호 관련 법령

가. 디지털의료제품법

1) 개요

「디지털의료제품법」과 관련 하위 법령은 AI 기술의 알고리즘 불투명성과 학습을 통한 성능 가변성을 관리하기 위하여 AI을 직접적인 규제 대상으로 명시하고 있다. 디지털의료제품법은 '디지털의료기기'의 정의에 AI의 기반이 되는 '지능정보기술'이 적용된 제품임을 명시하여 AI 기반 의료제품을 법적관리 체계에 명확히 포함한다(법 제2조제2호).

「디지털의료제품법」은 AI 기술의 고유한 특성을 반영하여 투명성 의무 확보, 위험관리 체계 수립, 그리고 제도적 지원이라는 세 가지 축을 중심으로 규제 체계를 구축하고 있다. 첫째, 투명성 의무로서 제조업자 등은 AI 모델의 훈련 방법 및 학습데이터 정보를 표시하거나 첨부해야 하며(법 제22조, 시행규칙 제33조, 디지털의료제품 허가·인증·신고·심사 및 평가 등에 관한 규정 고시 제33조) , 식품의약품안전처장은 AI 기술이 적용된 2등급 이상의 제품에 대한 심사 결과를 인터넷 홈페이지에 투명하게 공개해야 한다(시행규칙 제57조). 둘째, AI 위험관리체계 수립을 위해 AI 및 머신러닝 기능이 탑재된 제품은 품질관리 심사 시 경영책임과 거버넌스 체계 등을 포함한 AI 제어조치 특화 기준을 적용받으며(디지털의료기기 제조 및 품질관리 기준 고시 [별표 4]), 기존 제품에 AI 기능을 추가할 경우 별도의 변경심사를 통해 품질 적합성을 인정받아야 한다(디지털의료기기 제조 및 품질관리 기준 고시 제4조). 셋째, 위험관리를 위한 제도적 지원으로서 AI 알고리즘 등 핵심 구성요소의 성능을 미리 검증하여 제품 허가 시 서류 제출을 간소화해 주는 성능평가 제도와(법 제40조, 시행규칙 제50조), 사전에 승인된 계획 범위 내에서 별도의 변경허가 없이 신속하게 소프트웨어를 업데이트할 수 있는 변경관리 제도가 운영되고 있다(시행규칙 제7조, 제23조, 디지털의료제품 허가·인증·신고·심사 및 평가 등에 관한 규정 고시 제26조).

이 규정들은 데이터 편향성이나 알고리즘 오류로 인한 의료 사고로부터 이용자를 보호하기 위해 설계되었으며, AI 알고리즘의 성능 검증과 학습데이터의 정보 공개를 제조사의 핵심적인 법적 의무로 규정한다.

2) 학습데이터 정보 표시 의무

「디지털의료제품법」에 따라 디지털의료기기소프트웨어를 제조하거나 수입하여 판매하려는 자는 제품의 안전한 관리를 위해 법령이 정하는 주요 정보를 표시하거나 첨부해야 할 의무가 있다(법 제22조). 특히 AI 기술이 적용된 디지털의료기기의 경우에는 사용자의 올바른 판단을 돕기 위해 학습데이터의 정보 등 식품의약품안전처장이 정하여 고시하는 사항을 반드시 포함해야 한다(법 제22조제5호, 시행규칙 제33조제2항제5호).

학습데이터 정보를 포함한 기재 사항들은 사용자가 해당 소프트웨어를 사용할 때 언제든지 확인할 수 있는 방식으로 제공되어야 한다(시행규칙 제33조제3항). '디지털의료제품 허가·인증·신고·심사 및 평가 등에 관한 규정'에 따르면 구체적으로 공개해야 할 정보는 AI 모델의 구체적인 훈련방법과 해당 모델을 만드는 데 사용된 학습데이터의 정보, 해당 제품을 통해 기대할 수 있는 예측 성능의 범위와 기술적·임상적 한계점, 제3자가 제공하는 클라우드 서비스를 활용한 경우 해당 서비스의 종류 및 구성 형태, 전에 승인받은 변경관리 계획의 범위 내에서 제품을 변경한 경우, 그 변경 내용 및 결과 등이다(허가·심사 규정 제33조).

제조업자 등은 이러한 정보를 광(光) 또는 전자적 방식으로 처리하여 표시해야 하며, 여기에는 소프트웨어 내의 전자표시 기재나 인터넷 홈페이지를 통한 정보 제공 방식 등이 포함된다(법 제22조 각호 외의 부분, 시행규칙 제33조제3항). 또한, 표시하거나 첨부한 정보에 변경사항이 발생하는 경우, 제조업자 등은 관련 정보를 즉시 최신화하여 사용자에게 제공해야 할 의무가 있다(시행규칙 제33조제5항).

만약 법 제22조를 위반하여 학습데이터 정보 등 필수 기재 사항을 표시하지 않거나 거짓으로 표시하는 경우에는 행정처분의 대상이 될 수 있다(법 제50조제1항제10호). 아울러 이러한 위반 행위에 대해서는 500만원 이하의 벌금에 처한다는 처벌 규정이 마련되어 있다(법 제58조제2호).

3) 심사 결과의 투명한 공개

「디지털의료제품법」은 AI 기술이 적용된 제품의 허가 과정에 대한 투명성을 확보하고 사용자의 알 권리를 보장하기 위해 심사 결과 공개 의무를 규정하고 있다.

식품의약품안전처장은 AI 기술이 적용된 2등급 이상의 디지털의료기기 또는 디지털융합

의약품에 대하여 허가나 인증을 한 경우, 해당 제품의 심사 또는 검토 결과를 공개해야 한다(시행규칙 제57조제1항). 공개 대상이 되는 구체적인 범위는 AI 기술이 적용된 디지털의료기기 중 안전관리 수준이 상대적으로 높은 2등급 이상의 제품과 의약품이 조합된 디지털융합의약품으로 한정된다(시행규칙 제57조제1항).

심사 결과의 공개는 제품의 허가 또는 인증을 한 날부터 180일 이내에 식품의약품안전처의 인터넷 홈페이지를 통해 이루어져야 한다(시행규칙 제57조제3항). 식품의약품안전처장은 심사 결과를 공개하기 전에 해당 제품의 허가·인증을 받은 자에게 그 내용을 미리 통지하여 의견을 들을 수 있으며, 이를 통해 경영상·영업상 비밀이 부당하게 침해되지 않도록 관리한다(시행규칙 제57조제2항). 또한 '디지털의료제품 허가·인증·신고·심사 및 평가 등에 관한 규정'에 따르면 AI 기술 적용 제품의 경우 학습데이터의 정보, 예측되는 성능의 범위 및 한계 등 사용자가 확인해야 할 주요 정보들도 함께 관리되어야 한다(허가·심사 규정 제33조). 이러한 심사 결과 공개 제도는 첨단 기술이 적용된 디지털의료제품에 대한 사회적 신뢰를 높이는 데 목적이 있다.

4) 제조 및 품질관리 기준(GMP) 심사

「디지털의료제품법」상 디지털의료기기를 제조 또는 수입하려는 자는 제품의 안전성과 품질 확보를 위해 '적합판정 등 심사'를 받아야 하며, AI 기술의 특수성을 반영한 별도의 심사 체계를 적용받는다(고시 제3조제1항). 특히 이미 적합판정을 받은 디지털의료기기소프트웨어라 하더라도, 여기에 AI 및 머신러닝 기능을 추가하는 경우에는 해당 변경 사항에 대해 새로이 '변경심사'를 받아야 한다(고시 제4조제1항제2호가목). 이는 AI 기술의 도입이 제품의 성능과 안전성에 중대한 영향을 미치는 요소임을 고시에서 명시적으로 규정하고 있다.

AI 및 머신러닝 기능이 탑재된 제품에 대한 심사 기준은 일반적인 의료기기보다 강화되어 운영된다. AI 기술 적용 제품은 의료기기 하드웨어에 적용되는 일반 기준[별표 2]과 디지털의료기기소프트웨어 전용 기준[별표 3]은 물론, AI 및 머신러닝의 특화된 품질관리 요건을 담은 디지털의료기기 AI 제어조치 심사기준[별표 4]을 반드시 함께 충족해야 한다(고시 제5조제1항제3호).

[별표 4]에 규정된 AI 제어 조치는 경영책임, 위험관리, 거버넌스 체계, 데이터 처리 및

모델 개발 등 제품의 전 주기에 걸쳐 적용된다. 최고경영자는 AI 방침과 목표를 수립하고 주기적인 경영 검토를 수행하여 품질관리 체계의 지속적인 적합성과 효과성을 보장해야 하며, 관련 업무를 수행하는 인원은 필요한 역량을 갖추기 위한 교육과 훈련을 받아야 한다. 또한 조직은 내부적으로 이행해야 할 윤리 원칙과 규정을 수립하고, AI 시스템의 생명주기에 따른 역할과 책임을 명확히 문서화한 거버넌스 체계를 구축하여 운영해야 한다.

학습용 데이터 및 모델 관리와 관련하여, 조직은 데이터의 명확한 이해를 돕는 상세 정보와 출처를 기록·관리해야 하며 데이터 이상 여부 점검 및 편향성 완화 활동을 수행해야 한다. AI 모델 개발 시에는 오픈소스 라이브러리 활용에 따른 보안 취약점과 라이선스 준수 여부를 확인하고, 모델의 메커니즘과 성능을 담은 명세서 작성 및 추론 결과에 대한 설명 제공 기능을 갖추어야 한다. 시스템 운영 및 모니터링 단계에서는 문제 발생 시 알림 기능을 제공하고 보안 기법을 적용해야 하며, 시스템 로그 등을 통해 AI의 의사결정에 대한 추적 방안을 수립하여 관리해야 한다.

적합판정 등 심사는 품질관리심사기관과 인증업무등 대행기관이 합동으로 서면조사와 현장조사를 실시하는 것이 원칙이다. 다만, 디지털의료기기소프트웨어의 경우에는 신청인의 요청에 따라 비대면 현장조사를 실시할 수 있는 유연성을 두고 있다. 최초 적합판정서의 유효기간은 발행일로부터 3년이며, 제조업자 및 제조원은 품질 유지를 위해 3년마다 정기심사를 받아야 한다.

5) 성능평가 제도

「디지털의료제품법」은 식품의약품안전처장이 '인공지능 알고리즘' 등 디지털의료제품의 기능에 영향을 미칠 수 있는 구성요소에 대하여 그 성능을 평가할 수 있도록 성능평가 제도를 두고 있다(법 제40조). 성능평가 제도는 디지털의료제품의 핵심 구성요소에 대한 성능을 미리 평가하여 제품의 허가 및 인증 과정을 효율화하고 안전성을 확보하기 위한 제도이다. 이는 제품 전체가 아닌 핵심 구성요소 단위의 성능을 국가가 공인해 주는 성격을 갖는다.

성능평가 신청자는 구성요소 설명서와 성능평가 보고서 등을 갖추어 식품의약품안전처장에게 신청할 수 있으며, 필요시 전문기관의 검증을 거쳐 최종적으로 성능평가 결과서를 발급받게 된다(법 제40조제2항, 시행규칙 제50조제1항, 제4항,제6항). 이 제도를 통해 성능

평가를 받은 구성요소는 이후 완제품의 제조·수입 허가 및 인증 과정에서 그 결과가 적극적으로 고려되며, 허가 신청 시 관련 첨부서류의 일부를 면제받는 혜택을 누릴 수 있다(법 제40조제5항, 시행규칙 제7조제4항제3호, 제9조제4항제3호).

또한, 성능평가를 받은 구성요소 제조시설은 오염이나 시설 경합의 우려가 없는 범위에서 다른 디지털의료기기나 디지털융합의약품의 제조시설로 이용하거나 해당 시설로 갈음할 수 있어 자원의 효율적 활용이 가능하다(시행령 제3조제3항·제5항제9호, 시행규칙 제12조제3항제10호). 아울러 정부는 규제지원센터를 통해 성능평가 관련 업무를 지원하고, 허가 신청 전 자료에 대해 미리 확인받을 수 있는 사전 검토 제도를 병행 운영하여 개발자의 행정적 부담을 완화하고 있다(법 제39조제1항제6호, 시행령 제6조제2항제3호, 시행규칙 제49조제1항제9호).

6) 변경관리 제도

「디지털의료제품법」은 AI 기술이 적용된 디지털의료제품의 특성을 고려하여, 제품의 지속적인 성능 개선과 신속한 업데이트를 지원하기 위한 '변경관리 계획' 제도를 운영하고 있다. AI 기술이 적용된 제품은 데이터 학습을 통해 성능이 수시로 변할 수 있으므로, 허가 신청 시 향후의 변경 사항을 미리 계획하여 승인받을 수 있다. AI 기술이 적용된 디지털의료기기에 대하여 제조허가 또는 제조인증을 받으려는 자는 신청 시 변경관리 계획서를 추가로 제출할 수 있다(시행규칙 제7조제2항, 제9조제2항). 이러한 변경관리 계획 제도는 디지털융합의약품에도 동일하게 적용되어, AI 기술이 적용된 제품의 제조판매·수입 품목허가를 받으려는 자 역시 변경관리 계획서를 제출할 수 있는 근거를 두고 있다(시행규칙 제38조제4항).

승인된 변경관리 계획의 범위 내에서 이루어지는 업데이트는 별도의 변경 허가 절차 없이도 가능하여 제품의 혁신성을 유지할 수 있도록 돕는다. 디지털의료기기제조업자가 사전에 승인받은 변경관리 계획의 범위에서 변경을 실시한 경우에는 이를 안전성·유효성에 영향을 미치는 중요한 사항의 변경으로 보지 않는다(시행규칙 제23조제2항). 디지털융합의약품의 경우에도 변경관리 계획의 범위에서 이루어지는 변경은 품목허가증 기재 사항 중 품질에 영향을 미치는 변경 범위에서 제외되어, 복잡한 변경허가 절차를 거치지 않고 신속하게 시장에 반영될 수 있다(시행규칙 제42조제1항제2호가목).

2. 의료 분야 인공지능 이용자 보호 관련 가이드라인

가. 식품의약품안전처'디지털의료기기 GMP 가이드라인'

1) 개요

'디지털의료기기 GMP 가이드라인'은 디지털의료제품법 및 관련 하위 규정에 따라 디지털의료기기 제조업자 및 수입업자가 준수해야 하는 제조 및 품질관리 기준(GMP)에 관한 세부적인 지침을 제공하는 가이드라인이다. 이 안내서는 디지털의료기기소프트웨어의 특성을 반영하여 설계 및 개발 단계부터 시판 후 관리까지의 전 과정에서 필요한 품질경영 시스템 구축 방향을 제시한다. 특히 AI 및 머신러닝 기술이 적용된 제품의 경우, 데이터 관리, 알고리즘의 유효성 검증, 사이버보안 확보 등 기술적 특수성을 고려한 품질관리 요건을 아래와 같이 상세히 설명하고 있다.

2) 조직 관리

조직의 최고경영자는 AI 방침(AI Policy)과 측정가능한 AI 목표(AI Objective)를 수립하고, 이를 사내 공지나 교육 등을 통해 전파하여 품질관리 체계의 효과성을 유지하기 위한 실질적인 의지를 제시해야 한다. AI 관련 업무를 수행하는 인원은 적절한 교육, 훈련, 숙련도 및 경험을 바탕으로 역량을 갖추어야 하며, 조직은 이를 위한 인적·물적 자원을 식별하고 확보해야 한다. 또한 제품 품질에 영향을 미치는 업무를 수행하는 인원의 직무별 역량 기준을 정의하고, 교육 수료증 등 관련 기록을 철저히 유지하여 관리해야 한다.

3) 인공지능 위험관리 조치

조직은 디지털의료기기에 포함된 AI 소프트웨어 시스템이나 개별 항목에 대해 구체적인 위험관리 계획을 세우고 이를 전 과정에서 실행해야 한다. 위험관리 과정에서는 데이터 편향, 데이터 과적합, 성능 저하 및 오픈소스 소프트웨어의 보안 취약점 등 AI 기술 고유의 특성으로 인해 발생할 수 있는 잠재적 위해 요인을 반드시 식별하고 완화 전략을 마련해야 한다. 또한 소프트웨어 항목별 위해 요인을 식별하고 위험통제 수단의 유효성을 검증하는 등 모든 위험관리 활동의 추적성을 확보하여 문서화해야 한다.

4) 인공지능 거버넌스 체계

의료 인공지능과 관련된 법규, 규제, 정책 및 윤리 원칙을 반영하여 내부적으로 준수해

야 할 지침과 규정을 수립해야 한다. AI 시스템의 생명주기 전반에 걸쳐 조직의 역할과 책임을 명확히 문서화하고, 조직 규모와 제품의 복잡도에 적합한 AI 거버넌스 체계를 구축해야 한다. 특히 거버넌스 체계는 성능 모니터링 결과에 따라 제품의 목표 성능을 관리하거나 폐기를 의사결정하는 등 실질적인 통제 기능을 수행하고 관련 이행 기록을 문서화해야 한다.

5) 인공지능 모델 관리

AI 모델의 안전성과 신뢰성을 확보하기 위해 알고리즘의 매커니즘, 파라미터, 성능 지표 등을 상세히 기술한 모델 명세서를 작성하여 관리해야 한다. 오픈소스 라이브러리를 활용하는 경우에는 라이브러리의 안정성과 보안 취약점, 라이선스 준수 사항을 사전에 검증하고 그 위험관리 결과를 문서화해야 한다. 또한 모델 추출이나 회피 공격에 대한 방어 대책을 수립하고, 사용자가 AI의 판단을 이해할 수 있도록 추론 결과에 대한 시각적 또는 정량적 설명을 제공하는 기능을 구현해야 한다.

6) 학습데이터 관리

학습용 데이터의 출처, 규모, 라벨링 기준 등을 포함한 상세 정보 설명서를 작성하여 데이터에 대한 명확한 이해와 활용을 지원해야 한다. 데이터의 이상 여부를 점검하고 비인가자의 무단 변경이나 유출을 방지하기 위한 접근 제어와 보안 조치를 수립해야 한다. 특히 데이터 중독 공격에 대한 방어 활동을 수행하고, 인구학적·질병적·성별적 편향성을 완화하기 위해 특정 집단에 대한 과대표집 방지 등 구체적인 데이터 처리 활동을 계획하고 실행해야 한다.

7) AI 시스템 관리

AI 시스템 작동 중 문제가 발생하면 이를 사용자나 운영자에게 즉시 알리는 기능을 제공하고 시스템의 보안 강화를 위한 보안 기법을 적용해야 한다. 소스 코드나 사용자 인터페이스 설계 시 편향을 유발할 수 있는 요소를 피하고 점검해야 하며, 성능 저하 여부를 평가하기 위한 지표를 수립하여 주기적으로 모니터링해야 한다. 운영 단계에서는 올바른 사용을 유도하기 위해 사용 목적, 한계, 면책 조항 등을 사용자에게 명확히 설명하고, 시스템 로그를 통해 AI의 의사결정에 대한 추적 방안을 수립하여 관리해야 한다.

나. 식품의약품안전처 '식의약 분야 인공지능(AI) 도입·적용을 위한 AI 사업 실무가이드'

1) 개요

'식의약 분야 인공지능 도입·적용을 위한 AI 사업 실무가이드'는 식품의약품안전처에서 수행하는 AI 사업의 신뢰성과 타당성을 확보하고 실무자의 시행착오를 줄이기 위해 기획부터 운영까지 전 과정에서 확인해야 할 사항을 제시한다. 사업은 크게 계획 및 설계, 데이터 수집 및 전처리, AI 모델 개발 및 검증, AI 알고리즘 운영의 4단계로 구분되며, 각 단계별로 이용자 보호와 직결되는 핵심 요소들을 구체적으로 검토한다.

2) 사업 계획 및 설계 단계

사업 계획 및 설계 단계에서 기술의 안전성을 확보하기 위해 AI 도입 시 예상되는 할루시네이션(환각 현상), 법적·윤리적 이슈, 비용 대비 효과 등의 위험 요소를 사전에 진단하고 분석해야 한다. 특히 AI의 잘못된 예측이 인체 위해성 등 심각한 문제를 일으킬 가능성을 검토하고, 실수가 치명적이지 않은 분야를 선택하거나 적절한 대응 방안을 마련함으로써 이용자의 안전을 우선적으로 고려한다.

3) 데이터 수집 및 전처리 단계

데이터 수집 및 전처리 단계에서는 데이터의 유효성을 확보하기 위해 수집된 데이터가 AI 구현에 적합한지, 특정 집단에 치우쳐 결과 해석을 왜곡할 편향성은 없는지 검토한다. 데이터의 신뢰성을 위해 정제 및 가공 절차를 문서화하고 이상치를 식별하며, 작업자 가이드 제공을 통해 중요한 정보가 손실되지 않도록 관리한다. 또한 데이터의 보안성과 투명성을 높이기 위해 데이터별 보안 등급을 설정하고 접근 권한 관리와 비식별화 조치를 시행하며, 데이터 관리 명세서를 작성하여 처리 과정을 투명하게 공개한다.

4) AI 모델 개발 및 검증 단계

AI 모델 개발 및 검증 단계에서는 모델 검증의 유효성을 위해 별도의 데이터셋을 활용하여 예측값의 합리성을 확인하고 정밀도, 재현율 등 객관적인 평가지표를 통해 모델의 효과를 평가한다. 모델 성능평가의 객관성을 확보하기 위해 학습에 사용되지 않은 새로운 데이터로 2차 검증을 실시하며, 과대적합(overfitting)이나 과소적합(underfitting) 문제를 해결하여 실제 현업에서도 정확하게 작동하는 일반화된 모델을 선정한다.

5) AI 알고리즘 운영 단계

AI 알고리즘 운영 단계에서는 서비스의 최신성 및 지속가능성을 보장하기 위해 데이터와 모델의 주기적인 업데이트 방안을 수립해야 한다. 데이터가 지속적으로 생성되고 기술이 발전함에 따라 모델을 재학습시키는 프로세스를 마련하고, 유지보수 전문 인력 배정 및 사용자 활용 가이드라인을 제공하여 이용자가 안정적이고 신뢰할 수 있는 서비스를 지속적으로 제공받을 수 있도록 한다.

제 4 절 금융・의료 분야 인공지능 이용자 보호 체계 검토

1. 이용자 권리 보장

금융 및 의료 분야 AI 이용자 보호 체계는 공통적으로 AI의 투명성을 확보하고 정보주체의 실질적인 대응권을 보장하는 데 초점을 맞추고 있다. 금융 분야의 경우, AI 기술의 종류와 상관없이 그 이용으로부터 발생하는 책임을 금융회사에 귀속시키는 기술중립적 접근을 기본으로 한다. 「신용정보법」과 「개인정보 보호법」에 따라 완전 자동화된 결정이 개인의 권익에 중대한 영향을 미치는 경우, 이용자는 해당 결정에 대한 설명요구권과 거부권 및 이의제기권을 행사할 수 있다. 금융회사는 자동화된 결정이 이루어진다는 사실과 그 기준 및 절차 등을 사전에 홈페이지 등에 공개하여 투명성을 확보해야 한다. 의료 분야의 경우 「디지털의료제품법」을 통해 AI 기반 의료제품을 법적 관리 체계에 포함하고, 제조업자가 AI 모델의 훈련방법 및 학습데이터 정보를 명확히 표시하도록 의무화한다. 또한 식품의약품안전처장은 안전관리 수준이 높은 2등급 이상의 제품 심사 결과를 홈페이지에 공개하여 사용자의 알 권리를 보장하며, 사용자가 AI의 판단을 이해할 수 있도록 추론 결과에 대한 구체적인 설명을 제공하도록 한다.

2. 전 주기 거버넌스 및 경영진의 책임

금융 및 의료 분야 AI 이용자 보호 체계는 단순한 기술 관리를 넘어 조직적 차원의 내부통제 체계로 구축되는 추세다. 금융 분야는 최고경영자(CEO)를 포함한 경영진이 AI 개발 및 활용 전 단계에서 거버넌스에 관심을 갖고 역할과 책임을 분담하도록 규정한다. 이는

금융회사 임원의 내부통제 및 위험관리 의무를 AI 활용 시에도 이행해야 함을 확인한 것이며, 공급망이 분화되더라도 최종적인 책임은 서비스를 제공하는 금융회사 임직원이 부담한다는 원칙을 세우고 있다. 의료 분야에서는 최고경영자가 AI 방침과 목표를 수립하고 주기적인 경영 검토를 수행하여 품질관리 체계의 효과성을 보장해야 한다. AI 및 머신러닝 탑재 제품은 품질관리 심사(GMP) 시 경영책임과 거버넌스 체계를 포함한 특화 기준을 적용받으며, AI 시스템의 생애주기 전반에 걸친 역할과 책임을 명확히 문서화하여 운영한다.

3. 리스크 기반의 차등적 관리

금융 분야와 의료 분야 AI 이용자 보호 체계는 일률적인 규제 대신 서비스의 위해도와 영향력에 따른 차별화된 관리 방식을 채택한다. 금융 분야는 AI 시스템의 위험을 평가하고 그 수준에 맞는 내부통제 절차를 두는 리스크 기반 접근 방식을 적용한다. 금융 소비자에게 미치는 영향 등에 따라 서비스를 고·중·저 위험으로 분류하고, 고위험 서비스에 대해서는 책임자 또는 위원회의 검토와 승인을 거치도록 차등화된 관리 절차를 요구한다. 의료 분야는 AI 알고리즘 등 핵심 구성요소의 성능을 사전에 검증하는 성능평가 제도를 운영하여 안전성을 확보한다. AI의 성능이 학습을 통해 수시로 변하는 특성을 고려하여, 사전에 승인된 '변경관리 계획' 범위 내에서의 업데이트는 별도 변경허가 없이 신속하게 진행할 수 있도록 유연성을 부여하며, 위험 수준에 따라 심사 결과 공개 대상을 제한하는 등 등급별 관리를 실시한다.

4. 생성형 AI에 대한 특화 조치

생성형 AI 고유의 위험 요소인 보안 위협 및 환각 현상에 대한 대응방안이 강화되고 있다. 금융 분야는 생성형 AI의 보안 위협과 금융시스템 안정성에 미치는 영향을 반영하여 기존 가이드라인을 통합한 새로운 기준 마련을 추진 중이다. 생성형 AI 활용 시 보안성 기준 및 점검 체계를 마련해야 하며, AI는 업무 보조 수단이므로 최종 의사결정과 책임은 임직원이 수행해야 함을 명시한다. 의료 분야에서는 사업 계획 단계에서부터 생성형 AI의 주요 부작용인 할루시네이션(환각 현상)과 법적·윤리적 위험 요소를 사전에 진단하고 분석해야 한다. 특히 AI의 잘못된 예측이 인체 위해성 등 심각한 문제를 일으킬 가능성을 검토

하고, 데이터 중독 공격에 대한 방어와 편향성 완화 활동을 통해 이용자의 안전을 우선적으로 고려한다.

제 5 절 소결: 미디어 서비스의 확장과 인공지능 이용자 보호 과제

오늘날 AI가 융합된 미디어 서비스는 전통적인 방송·통신의 경계를 넘어 일반 대중이 일상적으로 접하는 AI 서비스의 대부분을 포괄하는 영역으로 확장되었다. 과거 미디어 서비스가 뉴스, 영상, 음악 등 콘텐츠의 전달에 국한되었다면, 현재는 검색 엔진의 개인화 추천, 소셜 미디어 피드 알고리즘, 대화형 AI 챗봇, 생성형 AI 기반 창작 도구, 음성 비서 서비스에 이르기까지 그 외연이 급격히 확대되고 있다.

이러한 변화는 미디어 서비스 생태계의 양적 팽창을 넘어 질적 전환을 의미한다. 이용자가 정보를 탐색하고, 의사결정을 내리며, 타인과 소통하고, 창작 활동을 수행하는 거의 모든 디지털 접점에서 AI 기반 미디어 서비스가 매개자로 기능하고 있기 때문이다. 결과적으로 융합 미디어 서비스에 대한 이용자 보호 체계를 정비하는 것은 곧 대중이 접하는 AI 서비스 전반의 안전망을 구축하는 작업과 사실상 동일한 의미를 갖는다.

따라서 금융 및 의료 분야에서 축적된 AI 거버넌스의 경험과 시사점을 융합 미디어 서비스에 적용하는 것은 단순히 특정 산업 영역의 규제 정비에 그치지 않는다. 이는 AI 시대 시민의 정보 주권과 디지털 기본권을 보장하기 위한 보편적 이용자 보호 체계를 수립하는 과제로 이해되어야 한다.

이상의 관점을 토대로, 금융·의료 분야의 선행 사례로부터 공통분모 및 특수성을 감안하여 도출한 시사점을 융합 미디어 서비스에 적용하자면 다음의 네 가지 핵심 과제를 도출할 수 있다.

1. 알고리즘 관련 투명성 확보 및 이용자 권리 보장

현대 미디어 환경에서 알고리즘은 이용자가 접하는 정보를 결정하는 보이지 않는 편집자 역할을 수행한다. 그러나 알고리즘의 불투명성으로 인해 이용자는 특정 정보가 왜 자신에게 추천되었는지 알기 어렵고, 이는 여론 왜곡이나 정보 편향으로 이어질 위험이 있

다. 따라서 이용자가 단순히 서비스를 소비하는 대상에 머물지 않고 주체적인 권리를 행사할 수 있도록 법적 권리를 명문화할 필요가 있다.

우선, AI의 결정으로 인해 현저한 권익 침해나 차별이 발생한 경우, 이용자가 그 기준과 과정을 묻고 이의를 제기할 수 있는 절차를 보장하는 방안을 제안한다. 또한 뉴스 추천과 같은 개인화된 서비스의 경우, 이용자가 자신의 정보를 어디까지 제공할지 직접 선택하고 변경할 수 있는 기능이 실효적으로 작동해야 한다. 특히 공공 미디어 서비스를 이용자가 거부하더라도 정보 접근에서 소외되지 않도록 적절한 '대체 수단'을 제공하여 디지털 포용을 실현해야 할 필요가 있다.

2. 서비스의 전 주기 거버넌스 및 경영진의 책임 강화

AI의 위험은 개발 단계뿐만 아니라 운영 중의 예기치 못한 변화나 공급망의 복잡성에서도 발생한다. 따라서 미디어 기업은 기획부터 폐기까지 AI의 생애주기 전반을 관리하는 전 주기를 아우르는 거버넌스를 구축해야 한다. 이는 단순한 실무 차원의 대응을 넘어 조직의 핵심 경영 원칙으로 자리 잡아야 한다.

이를 위해 미디어 분야 AI 이용자 보호를 위한 법안으로 최고경영자를 포함한 경영진이 AI 개발 및 활용에 대하여 명확한 역할과 책임을 분담하도록 제안한다. 금융 분야의 사례와 같이 AI 거버넌스 구축을 기업의 '내부통제 및 위험관리 활동'의 일부로 편입시키고, 이를 이용약관에 명시하여 이용자에게 투명하게 공개해야 한다. 또한 콘텐츠를 다루는 실무자들이 AI 윤리와 이용자 권리에 대한 역량을 갖추도록 교육 의무를 부과하여, 기술과 윤리가 조화를 이루는 미디어 생태계를 조성해야 한다.

3. 사회적 영향력을 고려한 리스크 기반의 차등적 관리

모든 미디어 서비스에 동일한 규제를 적용하는 것은 기술 혁신을 저해할 수 있다. 따라서 의료 분야의 등급별 관리나 금융 분야의 리스크 기반 접근 방식을 원용하여, 서비스의 사회적 영향력과 위험도에 따른 차등적 관리 체계를 도입해야 한다.

인간의 의사결정에 중대하고 명백한 해악을 미치거나 여론 조작 등 공공의 이익을 해칠 목적으로 설계된 AI 서비스는 개발과 제공을 법적으로 제한할 필요가 있다. 그리고 영향력

이 큰 '고영향 AI 서비스'에 대해서는 정기적인 이용자 보호 업무 평가를 실시하여 높은 수준의 위험관리 체계를 갖추도록 해야 한다. 동시에 알고리즘의 잦은 업데이트가 필요한 미디어 서비스의 특성을 고려하여, 사전에 승인된 계획 범위 내에서의 변경은 신속히 허용하는 유연한 관리 제도도 병행되어야 할 필요가 있다.

4. 생성형 AI 및 콘텐츠 허구성 대응 특화 조치

글, 그림, 영상 등을 스스로 만들어내는 생성형 AI는 콘텐츠 생산의 혁신을 가져왔지만, 동시에 딥페이크 성범죄물이나 허위 정보 유통과 같은 새로운 위협을 낳고 있다. 이러한 생성형 AI 고유의 리스크에 대응하기 위한 특화된 보호 조치가 반드시 필요한 상황이다.

생성형 AI로 만든 콘텐츠에 대해서는 이용자가 실제와 혼동하지 않도록 해당 결과물이 AI에 의해 생성되었다는 사실과 그 '기술적 한계(허구성 등)'를 명확히 표시하도록 의무화하는 방안을 제안한다. 또한 딥페이크 성범죄물 등 불법 정보의 생성을 방지하기 위한 기술적·관리적 보호 체계를 구축할 필요가 있다. 특히 청소년의 권익 보호에 현저한 지장을 가져올 수 있는 AI 서비스의 경우에는 청소년과 성인을 구분하여 설계하고 주기적으로 AI임을 고지하는 등 취약계층을 위한 특별 보호 조치를 시행해야 한다.

제5장 아동·청소년 특별 보호를 위한 입법 과제와 방안

제1절 논의의 특수성

최근 생성형·대화형 AI 서비스가 빠르게 확산되면서, 아동·청소년은 기존에 사용하던 학습 보조기구, 정보탐색뿐만 아니라, 놀이, 정서 교류, 관계 형성, 일상적 의사결정까지 생활 전반에서 AI와 상호작용하는 환경에 노출되고 있다. 특히, 메신저, 게임, SNS, 교육 플랫폼 등에 AI 기능이 기본적으로 포함되면서, 별도의 기술적인 이해나 가입 절차 없이도 AI를 자연스럽게 접하게 되는 상황이 형성되었으며, 음성·이미지 입력 기능이 확대되면서 저연령층까지 자연스럽게 이용집단에 포함되고 있다. 이러한 변화는 아동·청소년이 AI와 관계를 맺는 과정에서 일시적 사용이 아닌 상시적인 사용으로 인한 상호작용으로 전환되고 있으며, 그 과정이 대화형이나 개인화된 형태로 이루어지기 때문에 가정이나 학교에서 해당 상황을 파악하거나 개입하기 어렵다는 특징이 있다.

아동·청소년은 인지적·정서적 발달 단계에서 비판적 사고 능력이나 위험예측 능력 등에 대한 판단이 성인에 비해 성숙하지 않은 상태에 놓여 있어 사회적 보호가 필요하다는 부분에 대해 사회적 공감대가 형성되어 있다. 이로 인해 AI가 제공하는 정보나 조언의 정확성, 의도, 상업적 목적을 스스로 분별하는 데 한계를 가지고 있으며, 확신에 찬 표현이나 공감적인 언어를 사용하는 AI의 결과값에 대해 이를 객관적인 정보나 전문적인 지식으로 오인할 가능성이 높다. 또한, 또래 상호간의 관계와 사회적인 인정욕구에 대해 민감한 시기적 특성상, 친밀한 대화 방식이나 감정적 반응을 보이는 AI와의 상호작용에 보다 쉽게 몰입하거나 의존하게 될 위험성이 존재한다. 이러한 특성은 AI가 사람과 유사한 언어, 태도 등을 구현할수록 더욱 심화될 수 있으며, 현실에서의 보호자, 교사, 또래와의 관계를 대체하거나 회피하는 등 아동·청소년에게 부정적인 영향을 미칠 수 있다.

아동·청소년의 성장 단계에서의 취약성은 다양한 위험 요소가 중첩되는 형태로 나타난다. 대화형 AI는 이용자의 질문 맥락에 따라 점진적으로 구체적인 정보를 제공할 수 있기 때문에, 단순한 호기심이나 탐색이 자해·폭력·성적 대상화 등의 불법·유해정보로 연결될

가능성을 포함하고 있다. 일반적인 텍스트뿐만 아니라, 이미지·음성 생성 기능이 결합되어 서비스되는 경우, 아동·청소년 연령에 부적합한 콘텐츠가 우회적으로 생성되거나 변형 및 재생산되는 위험도 확대될 수 있다. 이를 통한 불법·유해정보 노출은 단순한 일회성적인 노출의 범위를 넘어서서, 반복적인 상호작용을 통해 아동·청소년의 인식과 태도에 중·장기적인 영향을 미칠 수 있다.

개인정보 및 민감정보 보호 측면에서도 아동·청소년들은 취약한 편이다. 대화형 AI와의 상호작용은 개인적인 소통으로 인식될 수 있으나, 실제로는 아동·청소년이 입력한 정보가 AI에 의하여 기록·분석·처리될 수 있으며, 서비스 운영이나 모델 개선 과정에서 활용될 가능성이 존재한다. 아동·청소년은 이러한 데이터의 흐름을 충분히 이해하지 못한 상황에서 학교생활, 친구관계, 사진이나 음성 등 민감한 정보를 대화 과정에서 자연스럽게 입력·노출할 가능성이 크다. 디지털 환경에서는 한번 생성·기록된 정보로 피해가 발생할 경우 완전한 회복이 힘들다는 점에서, 아동·청소년의 권리에 부정적인 영향을 미칠 수 있다.

또한, 생성형 AI가 제공하는 정보는 부정확한 정보나 오인정보에 대해서도 그럴듯한 형태로 제공하기 때문에, 아동·청소년은 출처 검증이나 정보의 교차 확인이 되지 않는 상태에서 그대로 수용할 가능성이 크다. 신체·정신 건강, 학교폭력 등 민감하고 위험한 영역에서의 잘못된 정보는 아동·청소년의 자기결정권을 왜곡하고, 적절한 보호나 지원을 받을 수 있는 기회를 지연시켜 피해를 확대시킨다. 또한, 이러한 상황에서 AI를 전문가 또는 권위 있는 조언자로 오해하는 현상은 가정에서의 보호자나 학교 등 공적인 지원체계에 대한 접근을 약화시키는 방향으로 작용할 수 있다.

아울러, AI 서비스가 광고, 추천, 결제 기능과 결합되어 제공될 경우, 상업적 조작에 이용될 위험도 확대된다. 정보 제공과 광고의 경계가 불분명한 답변 구조 속에서 아동·청소년은 상업적인 의도를 인식하지 못한 채 특정한 상품이나 서비스에 노출될 수 있으며, 구독형 서비스나 확률형 요소, 아이템 구매 등과 결합될 경우, 충동적인 구매나 소비로 이어질 가능성도 높다.

따라서, 아동·청소년의 AI 이용 과정에서 발생하는 위험은 단일한 문제가 아니라 불법·유해정보 노출, 정서적 과의존, 개인정보 침해, 오인정보에 따른 자기결정 침해, 상업적 조작 등이 결합된 복합적인 문제로서 나타난다. 이러한 위험은 개별 기능이나 특정한 콘텐츠의 문제라기 보다는, 대화형·개인화된 AI 서비스 구조 전반에서 누적되어 형성되며,

한 영역에서 발생한 취약성이 다른 영역의 위험을 증폭시키는 방식으로 확산될 수 있다. 이로 인해 사후적인 삭제나 차단, 이용자 또는 보호자의 개별적 신고에 의존하는 현행 방식만으로는 아동·청소년의 피해를 예방하거나, 이미 발생한 피해로부터의 회복이 어려우며, 실제로는 유사한 위험이 반복적으로 재현될 가능성이 높다.

특히, 생성형 AI 생태계에서는 위험이 이미 콘텐츠 등 정보로 나타난 이후가 아니라, 서비스의 설계·운영 단계에서부터 구조적으로 내재되는 경우가 많아, 사후 대응 중심의 기존 규제 방식은 근본적으로 한계를 가질 수밖에 없다. 아동·청소년은 발달 특성상 위험 인지 및 대응 능력이 제한되어 있고, AI의 결과값이나 상호작용을 비판적으로 검증하기 어렵다는 점에서 피해 발생 이후의 구제보다는 사전적 예방이 매우 중요하다. 따라서, 아동·청소년 보호를 위한 AI 규제는 서비스를 이용하는 이용자에게 주의 의무를 부과하거나 사후적 조치로서 대응할 것이 아니라, 위험이 발생하기 이전 단계에서 이를 최소화하는 예방적인 접근방법을 주요한 원칙으로 정립할 필요가 있다. 이를 위해서는 법적·기술적 수단을 분리하여 접근하지 않고, 상호 보완적으로 결합한 대응 체계를 구축해야 한다. 법적 수단은 AI 서비스 제공자에게 아동·청소년 보호를 위한 명확한 책임과 의무를 부과하고, 위험평가, 위반 시 제재 등 기준을 명확하게 설정하는 역할을 한다. 기술적인 수단은 언급한 법적 의무가 실질적으로 잘 이행될 수 있도록 불법·유해정보 차단, 연령 기반 접근 통제 등 구체적인 보호 장치를 구현하는 역할을 수행할 수 있다.

따라서, 아동·청소년 보호를 위한 AI 규제는 단순한 차원의 이용자 보호나 정보 유통에 대한 관리 부분을 넘어서서, 법적 규범과 기술적 안전장치를 결합하여 AI 서비스의 구조적인 위험을 사전에 통제하는 방향으로 검토할 필요가 있다.

제2절 국내외 입법 및 정책 동향

1. 국내 법제 및 논의 현황

가. 불법생성물 규제 논의

AI 기술의 발전으로 인해 디지털성범죄 수법이 고도화되고 있는 상황에서, 중·고등학생까지 딥페이크 범죄의 대상이 되었다는 점이 사회적 문제로서 부각되고 있다. 특히, SNS

등에 올린 셀카 사진을 기존의 음란물에 합성한 불법생성물이 텔레그램에 빠르게 확산되고 있다는 내용이 보도[115]되며, 국민들의 불안감이 증폭되었다. 또한, 참여자가 22만여명에 이르는 불법생성물 제작 텔레그램 채널을 확인했다는 보도를 통해 AI 기반 봇을 활용해 사진을 넣으면 5초 내에 음란 합성물을 생성할 수 있다는 것이 알려지면서,[116] 아동·청소년들의 AI 생태계 노출에 대한 우려가 지속적으로 제기되었다.

이상과 같이 AI 기반 딥페이크 불법생성물이 광범위하게 유통되고, 특히 청소년의 노출 위험이 현실화되자, 사회적 불안감이 빠르게 확산되었다. 이러한 상황 속에서 디지털 성범죄를 기존의 사후 대응 중심 규제로는 더 이상 효과적으로 통제하기 어렵다는 인식이 확산되었고, 보다 강력한 입법적 대응이 필요함이 공론화되었다.

이에 따라, 2024년 9월 26일, 국회는 「성폭력범죄의 처벌 등에 관한 특례법」(이하 '성폭력특례법')을 개정하여 딥페이크 성적 영상물의 제작·뿐만 아니라 소지·시청 자체도 범죄로 규정했다. 개정 성폭력범죄 특례법에 따르면 해당 콘텐츠를 소지하거나 시청한 경우에도 최대 3년 이하의 징역 또는 벌금형이 적용될 수 있도록 처벌 범위가 확대되었다. 이는 불법생성물의 확산 구조 전반을 차단하기 위한 입법적인 대응으로서, 디지털 성범죄 규제 논의가 본격적으로 점화되는 계기가 되었다고 평가할 수 있다.

나. 현행 법제의 한계

현행 법제도에서는 AI 기반의 불법생성물에 대한 개념 및 범위가 명확하게 정립되어 있지 않다. 현행 「성폭력특례법」과 「아동·청소년의 성보호에 관한 법률」은 촬영되었거나 실제로 존재하는 영상물을 중심으로 규제 체계가 구성되어 있어 AI 기술을 통한 합성·편집 등의 불법생성물이 동일하게 인격침해와 성적 피해를 발생시킴에도 불구하고 그 법적용이 불명확하다는 한계가 존재한다. 따라서, 불법생성물(딥페이크) 성범죄를 기존 촬영물 개념에 대한 확장 해석에만 의존할 경우, 처벌의 예측가능성과 법적 안정성이 저해될 수 있어,

115) 경기일보, "'딥페이크 범죄 창구' 텔레그램…8월 이용자 증가폭 역시 최대", 2024. 9. 5, <https://www.kyeonggi.com/article/20240905580230>(최종방문일 2025. 12. 30).

116) 한겨레, "'딥페이크 텔레방에 22만명…입장하니 "좋아하는 여자 사진 보내라"", 2024. 9. 2, <https://www.hani.co.kr/arti/society/society_general/1154764.html>(최종방문일 2025. 12. 30).

디지털 생성물에 대한 독립적인 규율 논의가 필요하다.[117]

한편, 현행 법제도는 불법생성물이 이미 생성·유통된 이후에 삭제·차단 또는 대응하는 사후적 규제 방식을 채택하고 있다. 하지만 생성형 AI 생태계에서는 불법생성물이 짧은 시간 내에 대량으로 생산 및 확산될 수 있으며, 한번 유포된 이후에는 완전한 삭제가 사실상 불가능하다는 점에서 사후 대응만으로는 피해 확산을 효율적으로 방지하기 어려운 것이 사실이다. 따라서, 기술 발전 속도와 피해 확산 속도가 법 집행의 속도를 압도적으로 뛰어넘는 상황에서, 기존의 사후 규제 중심 접근은 실효성을 상실할 위험이 크다.[118]

또한, 최근 제정된 「AI 기본법」은 AI 관련 규제의 기본법으로서 적용되고 있지만, AI의 신뢰성·투명성 확보 등 안전·윤리 기준만을 제공하고 있어, 불법생성물의 생성, 유통, 범죄적 활용 등에 대한 규율은 다루지 않는다. 즉, AI 기술을 활용해 범죄를 저지르는 경우에 대해서는 여전히 개별법 중심의 접근에 의존할 수밖에 없어, 「AI 기본법」만으로는 사전 예방적 규율을 기대하기는 어렵다. 또한, 「AI 기본법」은 리스크 관리와 안전성 확보에 대해 기술 투명성·책임성 확보 등 선언적이고 절차적인 의무를 중심으로 규율하고 있어 불법생성물(딥페이크)과 같이 사회적으로 위해성이 명백하고 피해 회복이 사실상 불가능한 영역에 대해서도 구체적인 규제의 근거나 기술적·관리적 조치 의무를 직접 도출하기에는 해석상 한계가 있다.

마지막으로, 불법생성물 대응을 위한 법·행정 체계가 분산되어 있어 통합적인 대응이 어렵다는 점도 지속적으로 제기되고 있다. 현재 디지털 성범죄 대응은 형사 처벌, 방송미디어통신심의, 피해자 지원, 예방 교육이 각각 다른 법률과 기관을 통해 적용되고 있으나, 이로 인해 불법생성물의 발생부터 유통, 피해 회복에 이르는 전 과정을 포괄적·일률적으로 관리하기 어렵다. 따라서, 이러한 대응이 불법생성물과 같은 복합적이고 기술 기반의 범죄에 효율적으로 대응하는데 한계로 작용한다.[119]

---

117) 신성원, “딥페이크 기술을 활용한 디지털 성범죄에 관한 연구”, 「한국치안행정논집」, 통권 67호, 2023, pp. 147-168

118) 김희정, “딥페이크 기술을 이용한 신종범죄에 대한 법정책적 시사점”, 「서강법률논총」, 통권 31호, 2024, pp. 97-128

119) 정경환, “딥페이크 성폭력 처벌법의 형법적 과제”, 「법학연구」, 제28집 제1호, 1-32, 2025.

다. 입법 필요성

앞서 살펴본 바와 같이, AI 기술의 발전으로 디지털 성범죄는 기존의 촬영·유통 중심 범죄에서 벗어나, 합성·편집 등을 통해 불법생성물(딥페이크 성범죄물)이 자동·대량으로 생성되는 구조로 전환되고 있다. 이러한 변화에 따라 현행 법제의 규율 방식 자체를 근본적으로 검토해 볼 필요가 있다. 현행 법제는 불법생성물에 대응하여 소지·시청 행위까지 처벌 범위를 확대하는 등 일정 부분에 있어 대응을 강화하였으나, 여전히 불법생성물이 이미 발생한 이후의 행위를 중심으로 규율하고 있다. 생성형 AI 생태계에서는 불법생성물이 생성되는 순간 이미 회복하기 어려운 피해가 발생하며, 이후의 삭제·차단, 처벌만으로는 피해 확산을 실질적으로 통제하기 어렵다는 점에서 사후 규제 중심 접근은 구조적인 한계를 지닌다. 또한, 불법생성물의 개념과 범위가 기존 '촬영물' 개념에 한정되어 해석되므로, AI 기반의 불법생성물을 독립적인 규율 대상으로 명확하게 포섭하지 못하는 문제가 있다. 이는 법 적용의 예측 가능성과 안전성을 저해할 뿐만 아니라, 기술 환경 변화에 따라 새로운 범죄 유형이 등장할 경우, 규제 공백을 반복적으로 초래할 위험이 있다.

아울러, 디지털 성범죄 대응이 형사처벌, 심의·차단, 피해자 지원, 예방 교육 등으로 분산되어 운영됨에 따라, 불법생성물의 생성, 유통, 피해 회복 전 과정에 대한 통합적인 관리 체계가 부재한 상황이다. 이러한 분리된 대응 구조는 AI 기반 불법생성물과 같이 복합적이고 기술 의존적인 범죄에 효율적으로 대응하는 데 한계로 작용할 수 있다.

이에 따라, 향후 입법은 단순히 처벌 범위를 확장하는 데만 그칠 것이 아니라, 불법생성물이 생성되는 구조 자체를 규율의 대상으로 포함하는 방향으로 확장될 필요가 있다. 즉, AI 기반의 불법생성물에 대해 생성 단계에서의 예방과 관리 의무를 명확히 하고, 관련 주체의 역할과 책임을 체계적으로 부과함으로써, 사후 대응 중심의 규율을 보완하는 체계를 구축할 필요가 있다. 결국 이러한 입법은 기술 발전을 제한하기 위한 조치라기 보다는, AI 기술이 인간의 존엄성과 성적 자기결정권을 침해하는 방향으로 활용되는 것을 방지하기 위한 최소한의 제도적 안전장치라고 볼 수 있다. AI 기반 불법생성물 등 디지털 성범죄의 구조적인 변화에 대응하기 위한 새로운 규율을 마련하는 것은 AI 시대에 불가피한 과제라고 볼 수 있다.

2. 해외 선행 입법례 분석

최근 해외 주요국에서는 AI 기술 그 자체를 규제 대상으로 판단하기보다는, 아동·청소년이 AI 기반 서비스와 상호 작용하는 과정에서 어떠한 위험이 발생할 수 있는지를 중심으로, 기존의 온라인 안전·아동 보호 제도를 확장 및 조정하려는 흐름을 보이고 있다. 이러한 논의는 공통적으로 사후적인 콘텐츠 삭제나 이용자 신고해 의존해 오던 기존 접근 방식이 생성형 AI 환경에서는 충분하지 않다는 문제 인식에서 기인한다. 특히 EU와 영국 등 유럽권에서는 불법·유해 콘텐츠가 이미 유통된 이후의 대응보다는 서비스 설계와 운영 단계에서 리스크를 사전에 식별하고 완화하는 방향으로 규제의 방향을 설정하고 있다. 이는 아동·청소년 보호를 표현물의 사후적인 규제 문제로만 보지 않고, 상호작용 구조, 알고리즘 추천, 기본 설정값 등 서비스 전반의 작동 방식과 연계된 구조적인 문제로 인식하려는 것으로 이해할 수 있다. 이와 함께, 아동·청소년 보호를 위한 연령확인 또는 연령 추정을 중요한 정책 수단으로 검토하면서도, 과도한 개인정보 수집이나 신원 확인이 또 다른 위험을 발생시킬 수 있다는 점을 동시에 고려하려는 논의도 병행되고 있다. 이에 따라 일부 국가는 연령확인을 일률적인 의무로 설정하기 보다는, 서비스의 위험도와 기능 특성에 따라 단계적으로 적용하는 방식이나 프라이버시 침해를 최소화할 수 있는 기술적 대안을 모색하고 있다. 이러한 접근은 아동·청소년 보호와 개인정보 보호를 상충하는 가치로 보지 않고, 이 두 쟁점을 조화롭게 설계하려는 방향성을 나타낸다고 볼 수 있다.

한편, 아동·청소년 보호를 콘텐츠 위법성 판단에만 한정하지 않고, 서비스의 기본값 설정과 이용 환경 전반을 규율 대상으로 삼으려는 시도도 확산되고 있다. 예를 들어, 개인정보 보호를 기본값으로 설정하거나, 과도한 체류 시간 유도와 같은 중독적 설계를 완화하도록 요구하는 논의는 아동·청소년 보호를 이용 제한이나 금지의 문제로 다루기보다는 이용 조건과 환경 설정의 문제로 구성하고자 하는 의지를 보여준다. 이는 생성형 AI와 같이 상호작용성이 강화된 기술 환경에서 나타나는 새로운 위험 양상을 제도적으로 포착하고자 시도하는 과정으로 볼 수 있다. 또한, 딥페이크, 불법합성물과 같이 생성 단계에서부터 심각한 인격권 침해나 디지털성범죄로 이어질 수 있는 영역에 대해서는 단순한 유통 차단 규제만으로는 한계가 있다는 지적이 지속적으로 제기된다. 이에 따라 일부 국가에서는 불법생성물이 쉽게 생성될 수 있는 기술적·구조적 환경 자체를 어떻게 관리해야 할 것

인지가 새로운 정책 쟁점으로 떠오르고 있다.

이처럼 해외 주요국의 논의는 아직 단일한 규제 모델로 수렴되기보다는, 각국의 법제와 사회적인 맥락에 따라 다양한 방식으로 전개되고 있다. 다만, 공통적으로는 아동·청소년 보호를 사후적 제재의 문제로만 인식하기보다는 AI 서비스의 설계·운영 방식과 결합된 리스크 관리의 문제로 확장하려는 흐름이 점차 강화되고 있다는 점에서 국내 제도에 시사점을 제공하고 있다.

가. EU

1) 개요

EU는 생성형 AI와 대화형 AI의 확산이 아동·청소년의 온라인 이용 환경을 근본적으로 변화시키고 있다는 점을 전제로, 기존 불법 콘텐츠에 대해 사후적인 규제만으로는 충분히 보호가 어렵다는 문제의식을 공유하고 있다. 특히, 아동·청소년이 AI 기반 서비스와 상호작용하는 과정에서 유해정보 노출, 불법합성 콘텐츠, 개인정보 침해 등 다양한 리스크가 구조적으로 발생할 수 있다는 점에 주목하며, 이러한 리스크를 서비스 설계 및 운영 단계에서 관리해야 할 정책적인 쟁점으로서 인지하고 있다.

2) 관련 법체계

EU의 아동· 청소년 보호 관련 AI 규제는 단일 법률에 의해 이루어진다기보다는, 온라인 서비스 규제, AI 규제, 개인정보 보호 체계가 상호 연계된 다층적인 구조를 형성하고 있다고 볼 수 있다.

(1) EU「AI 법」(AI Act)

특정 AI 활용 영역을 고위험(high-risk) 또는 금지 영역으로 분류하여, AI 사용 방식 자체에 대한 사전적인 규율체계를 도입했다.

EU「AI 법」은 AI 시스템을 리스크 수준에 따라 차등적으로 규율하는 리스크 기반 분류체계(risk-based framework)를 통해 AI 시스템을 규율하며, 개인의 취약성을 악용하거나 기본권 침해 가능성이 높은 활용 방식에 대해서는 금지 또는 고위험 범주로 관리하도록 규정하고 있다. 아동·청소년을 직접적으로 규율하는 독립된 규정은 두고 있지 않으나, 고위험 AI 시스템에 대해 전 생애주기에 걸친 위험성 평가 및 관리 의무를 부과함으로써, 취

약한 이용자가 AI와 상호작용하는 과정에서 발생할 수 있는 리스크를 사전에 점검하고 완화하도록 요구하는 구조체계를 가지고 있다.

① 리스크 기반 분류체계와 아동·청소년 보호

EU「AI 법」은 AI 시스템을 리스크 수준에 따라 차등적으로 규율하는 리스크 기반 분류체계를 규정하고 있다(Article 6). AI 시스템이 교육, 직업훈련, 평가 등 부속서Ⅲ(Annex Ⅲ)에 열거된 영역에서 활용되는 경우, 원칙적으로 고위험 AI로 분류된다. 이러한 영역은 아동·청소년이 주요 이용자이거나 직접적인 영향을 받는 분야에 해당되므로, 해당 AI 시스템에는 고위험 AI에 대한 강화된 규제 의무가 적용될 가능성이 높다. 즉, 아동·청소년을 대상으로 직접 설계된 AI가 아니더라도, 교육·학습·평가와 같이 아동·청소년이 필연적으로 노출되는 영역에서 사용되는 AI 시스템은 리스크 기반 분류체계를 통해 아동·청소년 보호 규율의 적용 범위에 포함되는 구조를 갖는다.

〈표 5－1〉 EU「AI 법」상 교육・학습・평가 영역의 고위험 AI 시스템 분류체계

Article 6. 고위험 AI 시스템의 분류 규칙
2. 제1항에 따른 고위험 AI 시스템 외에도, 부속서 Ⅲ에 열거된 AI 시스템은 고위험 AI 시스템으로 간주한다.
부속서 Ⅲ
제6조 제2항에 따른 고위험 AI 시스템이란, 다음 각 영역 중 어느 하나에 해당하는 분야에서 사용되는 AI 시스템을 말한다.
3. 교육 및 직업훈련
(a) 모든 교육 및 직업훈련 단계에서, 교육 또는 직업훈련 기관에 대한 접근 또는 입학을 결정하거나, 또는 자연인을 해당 기관에 배정하기 위해 사용되는 AI 시스템
(b) 모든 교육 및 직업훈련 단계에서, 학습 성과를 평가하기 위해 사용되는 AI 시스템으로서, 해당 평가 결과가 자연인의 학습 과정을 조정하거나 유도하는 데 사용되는 경우를 포함
(c) 모든 교육 및 직업훈련 단계의 교육 또는 직업훈련 기관의 맥락에서 또는 그 내부에서, 개인이 받을 수 있거나 접근할 수 있는 적절한 교육 수준을 판단·평가하기 위해 사용되는 AI 시스템
(d) 모든 교육 및 직업훈련 단계의 교육 또는 직업훈련 기관의 맥락에서 또는 그 내부에서, 시험 중 학생의 금지된 행위를 모니터링하거나 탐지하기 위해 사용되는 AI 시스템

② 금지 대상 AI 활용 방식과 아동·청소년 보호

EU「AI 법」은 기본권 또는 인간의 존엄성에 대해 용인할 수 없는 위험을 초래하는 AI 활용방식을 금지하고 있다(Article 5). 특히, 연령 또는 신체적·정신적 장애로 인한 취약성을 악용하여, 특정 집단에 속한 개인의 행동을 실질적으로 왜곡하고, 그 결과 해당 개인 또는 타인에게 신체적 또는 심리적 피해를 유발하거나 유발할 가능성이 있는 AI 시스템은 명시적인 금지 대상으로서 규율된다(Article 5(1)(b)). 이 규정은 문언상 특정 연령대를 직접 명시하지는 않으나, 연령에 따른 취약성을 독립적인 금지 사유로 규정하고 있다는 점에서, 판단 능력과 인지적 성숙도가 완전하지 않은 아동·청소년을 대표적인 보호 대상으로 전제하고 있는 것으로 해석할 수 있다. 이에 따라 아동·청소년의 인지적 미성숙이나 판단 능력의 한계를 이용해 행동을 조작하거나, 과도한 정서적 의존을 유도하거나, 의사결정 과정을 왜곡함으로써, 신체적 또는 심리적 피해를 초래하는 AI 활용 방식은 금지 규정의 적용 대상이 될 가능성이 크다. 비록 EU「AI 법」이 아동·청소년을 명시적으로 열거하고 있지는 않으나, 연령 기반 취약성 개념을 통해 아동·청소년 보호가 금지 규정의 해석과 적용 과정에서 실질적으로 반영될 수 있는 구조를 갖추고 있다고 볼 수 있다.

〈표 5－2〉 EU「AI 법」상 연령 기반 취약성 악용 AI의 금지 규정

Article 5. 금지되는 AI 관행 다음 각 호의 AI 관행은 금지된다. (b) 개인의 연령, 장애, 또는 특정 사회적·경제적 상황으로 인한 취약성을 악용하는 AI 시스템을 시장에 출시하거나, 사용에 제공하거나, 사용하는 행위로서, 그 목적 또는 효과가 해당 개인 또는 특정 집단에 속한 개인의 행동을 실질적으로 왜곡하고, 그 결과 해당 개인 또는 타인에게 중대한 피해를 초래하거나 초래할 개연성이 있는 경우

③ 사전 리스크 평가 및 관리 의무와 아동·청소년

EU「AI 법」은 고위험 AI 시스템에 대해 AI 시스템의 전 생애주기에 걸친 리스크 평가 및 관리 의무를 부과하고 있다(Article 9). 리스크관리 체계는 AI 시스템의 설계, 개발, 배포, 사용, 업데이트 전 과정에서 지속적이고 반복적으로 운영되어야 하며, 예측가능한 리스크와 합리적으로 예측가능한 오·남용 상황에서 발생할 수 있는 리스크를 모두 평가 대상으로 한다. 이러한 리스크 평가는 특정 이용자 집단을 한정해 열거하지는 않지만, 아동·청소

년이 AI 시스템을 이용하거나 그 영향을 받을 가능성이 있는 경우, 해당 이용 행태는 예측 가능한 사용 또는 오·남용에 해당한다. 따라서, 아동·청소년이 겪을 수 있는 유해정보 노출, 과의존, 오정보에 따른 판단 왜곡, 정서적 피해 등의 위험은 사전 리스크 평가 과정에서 고려되어야 할 요소로 포함될 수 있다.

〈표 5－3〉 EU「AI 법」상 고위험 AI 시스템의 리스크 관리 체계

Article 9. 리스크 관리 체계
1. 고위험 AI 시스템에 대해서는 리스크 관리 체계가 수립·이행·문서화되고 유지되어야 한다.
2. 리스크 관리 체계는 고위험 AI 시스템의 전 생애주기 전반에 걸쳐 계획되고 운영되는 지속적이고 반복적인 과정으로 이해되며, 정기적이고 체계적인 검토 및 갱신을 요구한다. 해당 리스크 관리 체계에는 다음 각 호의 단계가 포함되어야 한다.
(a) 고위험 AI 시스템이 의도된 목적에 따라 사용되는 경우, 해당 시스템이 건강, 안전 또는 기본권에 초래할 수 있는 이미 알려진 위험 및 합리적으로 예측가능한 위험을 식별하고 분석하는 단계
(b) 고위험 AI 시스템이 의도된 목적에 따라 사용되는 경우뿐만 아니라, 합리적으로 예측가능한 오·남용 조건 하에서 사용되는 경우에 발생할 수 있는 위험을 추정하고 평가하는 단계
(c) 제72조에 따른 사후시장 모니터링 체계를 통해 수집된 데이터의 분석을 기반으로, 추가적으로 발생할 수 있는 기타 위험을 평가하는 단계
(d) (a)에 따라 식별된 위험을 해결하기 위해 설계된 적절하고 표적화된 리스크 관리 조치를 채택하는 단계

④ 관련 보완 의무와 아동·청소년

EU「AI 법」은 사전 리스크 평가 의무를 보완하기 위해, 고위험 AI 시스템에 대해 훈련·검증·시험 데이터의 대표성 및 편향 방지 의무(Article 10), 그리고 인간이 개입하여 리스크를 통제할 수 있도록 하는 감독 의무(Article 14)를 함께 규정하고 있다. 이러한 규정은 아동·청소년과 같이 취약한 이용자에게 불리한 결과가 발생할 수 있는 가능성을 구조적으로 낮추기 위한 장치로서 기능한다. 특히, 아동·청소년데이터가 학습에 활용되거나, 아동·청소년의 행동·학습·정서 상태가 AI 판단의 입력값으로 사용되는 경우, 데이터 편향이나 자동화된 판단 오류가 장기적 영향을 미칠 수 있다는 점에서, 이러한 보완 의무는 아동·청소년 보호와 밀접하게 연결된다.

〈표 5－4〉 EU「AI 법」상 고위험 AI 시스템의 데이터 거버넌스 및 인간 감독 의무

Article 10. 데이터 및 데이터 거버넌스 1. 데이터에 기반한 AI 모델 훈련 기법을 사용하는 고위험 AI 시스템은, 해당 데이터셋이 사용되는 경우, 제2항부터 제5항까지에서 정한 품질 기준을 충족하는 훈련·검증·시험 데이터셋을 기반으로 개발되어야 한다. 2. 훈련·검증·시험 데이터셋은 고위험 AI 시스템의 의도된 목적에 적합한 데이터 거버넌스 및 관리 관행의 적용을 받아야 하며, 이러한 관행은 특히 다음 각 호의 사항을 포함한다. (a) 관련된 설계 선택 사항 (b) 데이터 수집 절차 및 데이터의 출처, 그리고 개인정보의 경우에는 데이터가 최초로 수집된 목적 (c) 주석(annotation), 라벨링(labelling), 정제(cleaning), 갱신(updating), 보강(enrichment), 집계(aggregation) 등 관련 데이터 전처리 작업 (d) 데이터가 측정하거나 대표하도록 의도된 정보에 관한 가정의 설정, 특히 그 가정의 적정성 (e) 필요한 데이터셋의 가용성, 양 및 적합성에 대한 평가 (f) 특히 데이터 출력이 향후 운영을 위한 입력에 영향을 미치는 경우, 개인의 건강 및 안전에 영향을 미치거나 기본권에 부정적 영향을 주거나, 또는 유럽연합법에 따라 금지된 차별을 초래할 가능성이 있는 편향의 존재 여부에 대한 검토 (g) (f)에 따라 식별된 가능한 편향을 탐지·방지·완화하기 위한 적절한 조치 (h) 본 규정 준수를 저해하는 관련 데이터 격차 또는 결함의 식별, 및 그러한 격차나 결함을 해결하는 방법
Article 14. 인간의 감독 1. 고위험 AI 시스템은, 적절한 인간-기계 인터페이스 도구를 포함하여, 사용되는 기간 동안 자연인이 효과적으로 감독할 수 있도록 설계·개발되어야 한다. 2. 인간의 감독은, 고위험 AI 시스템이 의도된 목적에 따라 사용되는 경우 또는 합리적으로 예측가능한 오·남용 조건 하에서 사용되는 경우에 발생할 수 있는 건강, 안전 또는 기본권에 대한 위험을 예방하거나 최소화하는 것을 목적으로 한다. 특히, 본 장에 규정된 다른 요구사항이 적용되었음에도 불구하고 그러한 위험이 지속되는 경우에 이러한 감독이 이루어져야 한다.

(2) 디지털 서비스 법(DSA)

「DSA」는 온라인 플랫폼 전반을 대상으로 불법·유해 콘텐츠 관리 의무를 부과하는 한편, 서비스의 설계와 운영 과정에서 발생하는 시스템적 리스크를 사전에 식별하고 완화하도록 요구하는 규제 체계를 도입했다. 특히 「DSA」는 아동·청소년 보호를 온라인 안전 정책의 핵심적인 요소로 인지하고, 이용자 수와 사회적 영향력이 큰 대형 플랫폼 사업자를 중심으로 사전적 리스크 평가 및 완화 의무를 명시적으로 규정하고 있다.

「DSA」는 AI 시스템 자체를 직접적으로 규율하는 법률은 아니지만, 추천·노출·타겟 등 AI 기반 알고리즘이 핵심적으로 작동하는 온라인 플랫폼 환경을 규제 대상으로 하고 있다. 이로써 아동·청소년이 이러한 알고리즘 구조 속에서 경험할 수 있는 유해 콘텐츠, 과의존 등 구조적 리스크를 명시적으로 관리 대상으로 포섭하고 있다는 점에서, 「DSA」는 「AI 법」과 함께 AI 관련 아동·청소년 보호 규제 체계를 구성하는 핵심적인 축으로 고려할 수 있다.

① 규제 대상 및 적용 범위

「DSA」는 온라인 플랫폼상 중개 환경 전반을 포괄적으로 규율하되, 서비스의 기능과 사회적 영향력에 따라 단계적으로 의무를 부과하는 구조를 채택하고 있다. 구체적으로 중개서비스, 호스팅 서비스, 온라인 플랫폼, 초대형 온라인 플랫폼으로 규제 대상을 구분하고, 서비스 유형과 규모에 비례하여 책임과 의무의 강도를 차등화하고 있다. 아동·청소년 보호와 관련된 규율은 특히 이용자 수와 사회적 파급력이 큰 온라인 플랫폼 및 VLOPs에 집중되어 있으며, 이는 실제로 아동·청소년에게 노출되는 위험 상황은 구조적으로 특정 대형 플랫폼에 집중되어 있다는 현실을 반영한 것으로 볼 수 있다.

〈표 5－5〉 EU「DSA」규제 대상 및 적용범위

Article 2. 적용범위 1. 본 규정은 중개서비스 제공자의 설립 장소와 관계없이, 유럽 연합 내에 설립되어 있거나 위치한 서비스 수신자에게 제공되는 중개서비스에 적용된다.
Article 3. 정의 본 규정의 목적상, 다음 각 호의 정의를 적용한다. (g) '중개서비스'란 다음 각 목의 정보사회서비스 중 하나를 의미한다. (i) '단순 전달(mere conduit) 서비스'란, 서비스 수신자가 제공한 정보를 통신망을 통해 전송하거나, 통신망에 대한 접근을 제공하는 서비스를 말한다. (ii) '캐싱(caching) 서비스'란, 서비스 수신자가 제공한 정보를 통신망을 통해 전송하는 과정에서, 해당 정보의 이후 전송을 보다 효율적으로 하기 위한 유일한 목적으로, 해당 정보를 자동적·중간적·일시적으로 저장하는 서비스를 말한다. (iii) '호스팅(hosting) 서비스'란, 서비스 수신자가 제공하고 그 요청에 따라 제공되는 정보를 저장하는 서비스를 말한다. (h) '불법 콘텐츠(illegal content)'**란, 제품의 판매 또는 서비스의 제공을 포함한 활동과 관련하여 또는 그 자체로서, 해당 법의 구체적인 주제나 성격과 관계없이, 유럽연합 법 또는 유럽연합 법에 부합하는 회원국 법률을 위반하는 모든 정보를 의미한다.

(i) '온라인 플랫폼(online platform)'**이란, 서비스 수신자의 요청에 따라 정보를 저장하고 이를 공중에게 유포하는 호스팅 서비스를 의미한다. 다만, 해당 활동이 다른 서비스의 경미하고 순수하게 부수적인 기능이거나, 주된 서비스의 경미한 기능에 불과하여 객관적·기술적 이유로 그 다른 서비스 없이는 이용될 수 없고, 해당 기능 또는 기능의 통합이 본 규정의 적용을 회피하기 위한 수단이 아닌 경우는 제외한다.

Article 33. 초대형 온라인 플랫폼 및 초대형 온라인 검색엔진
1. 본 절은, 유럽연합 내에서 해당 서비스의 월간 평균 활성 이용자 수가 4,500만 명 이상인 온라인 플랫폼 및 온라인 검색엔진으로서, 제4항에 따라 초대형 온라인 플랫폼 또는 초대형 온라인 검색엔진으로 지정된 경우에 적용된다.

② 불법 콘텐츠 관리 의무

「DSA」는 온라인 플랫폼에 대해 불법 콘텐츠에 대한 신고·통지 체계를 구축·운영하도록 의무화하고 있으며, 적법한 통지가 접수된 경우 신속한 접근 차단 조치를 취하도록 요구한다. 또한 사법·행정 당국과의 협력을 통해 불법 콘텐츠의 확산을 차단하는 체계를 마련하고 있다. 이러한 규율 방식은 전통적인 콘텐츠 규제로 볼 수 있으며, 아동·청소년을 대상으로 한 성적학대물이나 아동·청소년에게 심각한 위해를 초래하는 불법정보에 대해서는 즉각적이고 실효적인 대응 수단으로 기능한다.

〈표 5－6〉 EU「DSA」상 불법 콘텐츠 관리 의무

Article 16. 신고·통지 및 조치 메커니즘
1. 호스팅 서비스 제공자는 개인 또는 법인이 자신들의 서비스에 존재하는 특정 정보가 불법 콘텐츠에 해당한다고 판단하는 경우 이를 통지할 수 있도록 하는 신고·통지 메커니즘을 마련하여야 한다. 해당 메커니즘은 접근이 용이하고 이용자 친화적이어야 하며, 전자적 수단만으로 신고가 제출될 수 있도록 하여야 한다.
2. 제1항에 따른 메커니즘은 충분히 구체적이고 적절히 소명된 신고의 제출을 용이하게 하도록 설계되어야 한다. 이를 위해 호스팅 서비스 제공자는 다음 각 목의 사항을 모두 포함하는 신고가 제출될 수 있도록 필요한 조치를 취하여야 한다.
3. 본 조에 따른 신고는, 해당 신고가 성실한 호스팅 서비스 제공자로 하여금 상세한 법률적 검토 없이도 문제되는 활동 또는 정보의 불법성을 식별할 수 있도록 하는 경우, 제6조의 목적상 해당 정보에 대한 실제 인지(actual knowledge) 또는 인식(awareness)을 발생시키는 것으로 본다.
4. 신고에 신고를 제출한 개인 또는 법인의 전자적 연락처 정보가 포함되어 있는 경우, 호스팅 서비스 제공자는 부당한 지체 없이 해당 개인 또는 법인에게 신고 접수 사실을 확인하는 통지를 송부하여야 한다.

Article 17. 조치 사유의 고지

1. 호스팅 서비스 제공자는, 서비스 수신자가 제공한 정보가 불법 콘텐츠에 해당하거나 또는 제공자의 이용약관에 부합하지 않는다는 사유로 다음 각 호의 제한 조치를 부과하는 경우, 해당 조치의 영향을 받는 서비스 수신자에게 명확하고 구체적인 조치 사유서를 제공하여야 한다.
   (a) 서비스 수신자가 제공한 특정 정보의 가시성 제한(콘텐츠 삭제, 콘텐츠 접근 차단 또는 콘텐츠의 하향 노출 포함)
   (b) 금전적 지급의 중단·종료 또는 기타 제한
   (c) 서비스 제공의 전부 또는 일부에 대한 중단 또는 종료
   (d) 서비스 수신자의 계정 정지 또는 해지
2. 제1항은 해당 서비스 수신자의 전자적 연락처 정보가 제공자에게 알려져 있는 경우에 한하여 적용된다. 이 조항은 제한 조치가 어떠한 사유나 방식으로 이루어졌는지와 관계없이, 해당 조치가 부과된 날까지 늦어도 적용되어야 한다.
   다만, 기만적인 대량 상업 콘텐츠(deceptive high-volume commercial content)에 해당하는 정보의 경우에는 제1항을 적용하지 아니한다.

③ 시스템적 리스크 평가 의무

「DSA」의 가장 중요한 특징은 VLOPs에 대해 연 1회 이상 시스템적 리스크를 평가할 의무를 부과한 점이다. 이 리스크 평가는 단순하게 개별 불법콘텐츠 존재 여부만을 대상으로 하지 않고, 플랫폼의 설계와 운영 방식이 사회 전반에 미치는 구조적 영향을 대상으로 하고 있다. 특히, 평가 대상 리스크에는 기본권 침해 위험과 함께 아동의 신체적·정신적 발달에 미치는 부정적인 영향, 유해 콘텐츠 확산, 알고리즘 기반 위험 등이 명시적으로 포함된다. 이를 통해 아동·청소년 보호는 선택적 고려 요소가 아니라, 플랫폼이 반드시 관리해야 할 독립적이고 핵심적인 리스크 범주로 규정된다.

〈표 5－7〉 EU「DSA」상 시스템 리스크 평가 의무

Article 34. 리스크 평가

1. 초대형 온라인 플랫폼 제공자 및 초대형 온라인 검색엔진 제공자는, 해당 서비스와 그와 관련된 시스템(알고리즘 시스템을 포함한(3)의 설계 또는 작동 방식, 또는 해당 서비스의 이용으로부터 유럽연합 내에서 발생하는 모든 시스템적 리스크(systemic risks)를 성실하게 식별·분석·평가하여야 한다.
   이러한 위험 평가는 제33조 제6항 제2문에 따른 적용일까지 수행되어야 하며, 이후에는 최소 연 1회 이상 실시되어야 한다. 또한, 본 조에 따라 식별된 리스크에 중대한 영향을 미칠 가능성이 있는 기능을 배포하기 이전에는 반드시 리스크 평가를 수행하여야 한다. 리스크

평가는 각 제공자의 개별 서비스에 특화되어야 하며, 시스템적 리스크의 중대성 및 발생 가능성을 고려한 비례적 방식으로 이루어져야 하고, 다음 각 호의 시스템적 리스크를 포함하여야 한다.

(a) 해당 서비스를 통해 불법 콘텐츠가 확산되는 리스크

(b) 유럽연합 기본권 헌장에 규정된 기본권의 행사에 대한 실제적 또는 예측 가능한 부정적 영향, 특히 다음의 기본권에 대한 영향

(c) 시민적 담론 및 선거 과정, 그리고 공공의 안전과 관련된 실제적 또는 예측 가능한 부정적 영향

(d) 성별 기반 폭력, 공중보건의 보호, 미성년자의 보호와 관련된 실제적 또는 예측 가능한 부정적 영향, 그리고 개인의 신체적·정신적 안녕에 대한 중대한 부정적 결과

2. 초대형 온라인 플랫폼 제공자 및 초대형 온라인 검색엔진 제공자는 리스크 평가를 수행함에 있어, 특히 다음 각 호의 요소가 제1항에 따른 시스템적 리스크에 어떠한 방식으로 영향을 미치는지 여부와 그 정도를 고려하여야 한다.

(a) 추천 시스템 및 기타 관련 알고리즘 시스템의 설계

(b) 콘텐츠 조정(content moderation) 시스템

(c) 적용되는 이용약관 및 그 집행 방식

(d) 광고의 선택 및 제시 방식에 관한 시스템

(e) 제공자의 데이터 관련 관행

또한 리스크 평가는, 비진정한 이용(in-authentic use) 또는 자동화된 서비스 악용을 포함한 고의적인 서비스 조작, 그리고 불법 콘텐츠 및 이용약관에 부합하지 않는 정보의 증폭 및 신속하고 광범위한 확산이 제1항의 리스크에 어떠한 영향을 미치는지도 분석하여야 한다. 아울러 리스크 평가는, 특정 회원국에 한정되는 경우를 포함하여, 지역적 또는 언어적 특수성도 고려하여야 한다.

3. 초대형 온라인 플랫폼 제공자 및 초대형 온라인 검색엔진 제공자는 리스크 평가 수행 후 최소 3년간 그 근거 자료를 보존하여야 하며, 요청이 있는 경우 유럽연합 집행위원회 및 설립지 디지털서비스 조정자에게 이를 제공하여야 한다.

④ 리스크 완화 조치 의무

플랫폼은 시스템적 리스크 평가 결과를 기반으로 해당 리스크를 낮추기 위한 비례적이고 효과적인 완화 조치를 이행해야 한다. 완화 조치는 단순한 콘텐츠 삭제를 넘어, 서비스의 구조와 기능 자체에 대한 조정까지 포함할 수 있다. 예를 들어, 추천·노출 알고리즘의 조정, 연령에 적합한 서비스 설계, 유해 콘텐츠의 확산 경로 제한 등이 이에 해당된다. 이러한 접근은 아동·청소년 보호를 사후적 대응이 아닌, 서비스 설계 단계에서의 사전적 대응으로 전환하려는 「DSA」의 규제 철학이 반영된 것이라고 볼 수 있다.

〈표 5－8〉 EU「DSA」상 리스크 완화 조치 의무

Article 35. 리스크 완화 조치
1. 초대형 온라인 플랫폼 제공자 및 초대형 온라인 검색엔진 제공자는 제34조에 따른 리스크 평가 결과를 고려하여, 해당 서비스에서 식별된 시스템적 리스크를 완화하기 위한 합리적이고 비례적인 조치를 이행하여야 한다.
2. 제1항에 따른 리스크 완화 조치는, 식별된 시스템적 리스크의 중대성 및 발생 가능성에 비례하여 설계·이행되어야 하며, 특히 다음 각 호의 조치를 포함할 수 있다.
(a) 서비스의 설계 또는 기능, 이용약관, 콘텐츠 조정 시스템, 추천 시스템을 포함한 알고리즘 시스템의 조정
(b) 유해 콘텐츠의 확산을 제한하기 위한 조치
(c) 특정 시스템적 리스크에 노출될 가능성이 높은 이용자 집단을 대상으로 한 표적화된 보호 조치
(d) 연령에 적합한 서비스 설계 또는 미성년자의 보호를 강화하기 위한 조치
(e) 신고·차단·필터링 기능 등 이용자가 리스크를 통제할 수 있도록 지원하는 도구의 개선
3. 제공자는 제1항 및 제2항에 따른 리스크 완화 조치의 이행 여부와 효과를 지속적으로 점검하고, 필요한 경우 이를 조정·보완하여야 한다.
4. 초대형 온라인 플랫폼 제공자 및 초대형 온라인 검색엔진 제공자는 리스크 완화 조치의 주요 내용과 그 효과에 관한 정보를, 요청이 있는 경우 유럽연합 집행위원회 및 설립지 디지털서비스 조정자에게 제공하여야 한다.

⑤ 아동·청소년 보호를 위한 설계·운영 기반 안전 조치

「DSA」는 아동·청소년이 접근 가능한 온라인 플랫폼 서비스에 대해, 이용을 전면적으로 제한하는 방식보다는 아동·청소년의 개인정보 보호와 안전을 확보하기 위한 적절하고 비례적인 조치를 서비스 설계·운영 전반에 반영하도록 요구하고 있다(Article 28). 이에 따라 플랫폼은 아동·청소년에게 높은 수준의 안전을 제공하기 위한 설계·운영 조치를 마련해야 하며, 이용자를 혼동시키는 인터페이스나 프로파일링 기반 광고와 같이 아동·청소년에게 불리한 영향을 미칠 수 있는 요소는 제한된다. 또한 아동의 신체적·정신적 발달에 부정적 영향을 미칠 수 있는 설계 요소는 시스템적 리스크 평가 및 완화 조치를 통해 관리 대상이 된다.

〈표 5－9〉 EU「DSA」상 아동・청소년 보호 의무

Article 28. 미성년자의 온라인 보호
1. 미성년자가 접근할 수 있는 온라인 플랫폼 제공자는, 해당 서비스에서 미성년자의 높은 수준의 개인정보 보호, 안전 및 보안을 보장하기 위하여 적절하고 비례적인 조치를 마련하여야 한다.
2. 온라인 플랫폼 제공자는, 서비스 수신자가 미성년자임을 합리적인 확실성을 가지고 인지하고 있는 경우, 규정(EU) 2016/679(GDPR) 제4조 제4호에서 정의된 프로파일링에 기반하여, 해당 서비스 수신자의 개인정보를 이용한 광고를 자신의 인터페이스에 표시하여서는 아니 된다.
3. 본 조에 따른 의무의 이행은, 온라인 플랫폼 제공자가 서비스 수신자가 미성년자인지를 판단하기 위하여 추가적인 개인정보를 처리하도록 요구하는 것으로 해석되어서는 아니 된다.
4. 집행위원회는 유럽디지털서비스위원회(Board)와 협의한 후, 제1항의 적용을 지원하기 위하여 온라인 플랫폼 제공자를 대상으로 가이드라인을 발행할 수 있다.

⑥ 알고리즘 투명성 및 선택권 보장

「DSA」는 추천 시스템을 포함한 알고리즘 기반 서비스에 대해 주요 작동 원리에 대한 설명 의무를 부과하고, 이용자에게 비개인화 추천 옵션을 제공하도록 요구하고 있다. 아울러 광고 타기팅 관련 투명성도 강화하고자 한다. 이러한 규율은 아동·청소년이 알고리즘에 의해 무의식적으로 행동이 조작되거나 특정 콘텐츠에 과도하게 노출되는 리스크를 완화하는 간접적인 보호 장치로 작동한다. 즉, 알고리즘의 영향력을 제거한다기보다는, 그 작동 방식을 드러내고 이용자의 선택권을 확대함으로써 리스크를 통제하는 방식이다.

〈표 5－10〉 EU「DSA」상 알고리즘・광고 투명성 및 이용자 선택권 보장

Article 26. 온라인 플랫폼상의 광고
1. 온라인 인터페이스를 통해 광고를 표시하는 온라인 플랫폼 제공자는, 각 개별 광고가 각 서비스 수신자에게 제시될 때마다, 해당 서비스 수신자가 다음 각 호의 사항을 명확하고 간결하며 모호하지 않게, 그리고 실시간으로 식별할 수 있도록 보장하여야 한다.
(a) 해당 정보가 광고임을 명확히 인식할 수 있도록 하는 표시(제44조에 따른 기준을 따르는 눈에 띄는 표식을 포함할 수 있음)
(b) 해당 광고가 누구를 위하여 제시되는지에 관한 자연인 또는 법인의 신원
(c) 해당 광고에 대한 비용을 지불한 자연인 또는 법인의 신원(b에 따른 자와 다른 경우에 한함)
(d) 해당 광고가 특정 서비스 수신자에게 제시되도록 결정된 주요 매개변수(main parameters)에 관한 의미 있고 직접적이며 쉽게 접근 가능한 정보, 그리고 해당되는 경우 이러한 매

개변수를 변경하는 방법에 관한 정보

Article 27. 추천 시스템의 투명성
1. 추천 시스템을 사용하는 온라인 플랫폼 제공자는, 해당 추천 시스템에서 사용되는 주요 매개변수(main parameters)와, 서비스 수신자가 그 주요 매개변수에 영향을 미치거나 이를 수정할 수 있는 선택지가 있는 경우 그 내용에 대하여, 명확하고 이해하기 쉬운 언어로 이용약관에 명시하여야 한다.
2. 제1항에 따른 주요 매개변수의 설명은, 왜 특정 정보가 서비스 수신자에게 추천되는지를 이해할 수 있도록 하여야 하며, 최소한 다음 각 호의 사항을 포함하여야 한다.
(a) 서비스 수신자에게 제시되는 정보를 결정하는 데 있어 가장 중요한 기준(criteria)
(b) 해당 기준들이 상대적으로 중요한 이유

Article 38. 추천 시스템
제27조에 따른 요구사항에 추가하여, 추천 시스템을 사용하는 초대형 온라인 플랫폼 제공자 및 초대형 온라인 검색엔진 제공자는, 각 추천 시스템에 대해 규정(EU) 2016/679(GDPR) 제4조 제4호에서 정의된 프로파일링에 기반하지 않은 추천 옵션을 최소한 하나 이상 제공하여야 한다.

3) 생성형 AI 및 딥페이크 등 신유형 서비스 위험 대응 방식

EU는 생성형 AI 확산으로 인해 불법 합성 콘텐츠가 기존과 비교할 수 없을 정도로 쉽게 생산·변형·유통될 수 있다는 점을 문제의 핵심으로 보고 있다. 특히 딥페이크·성적 합성물은 단순한 불법 콘텐츠 유통의 문제를 넘어서서, 콘텐츠가 생성되는 기술적 조건 자체가 피해 규모와 속도를 결정한다는 인식이 점차 강화되는 것으로 보여진다. 이런 문제의식은 '생성·조작 사실의 표시(라벨링)'를 통해 출처와 진위를 구분할 수 있게 만들고, 불법 합성물이 제작에서 공개로 이어지는 경로 자체를 범죄 구성요건에 포섭하며, 규범 집행을 위해 기술적 탐지·표시 기준(가이드)까지 마련하는 방식으로 구체화 되고 있다.

EU「AI 법」은 딥페이크와 합성 콘텐츠가 생성 단계에서부터 확산될 수 있다는 점을 전제로, 표시·탐지 가능성 확보를 법적 의무로 명문화했다. 동법 제50조는 합성 텍스트, 이미지, 음성, 영상 등 합성 콘텐츠(synthetic content) 출력물에 대해 기계 판독이 가능한 방식으로 인공적으로 생성·조작된 것임을 표시하도록 하고 있다(제50조 제2항). 또한, 딥페이크에 대해서는 해당 콘텐츠가 인공적으로 생성 또는 조작되었음을 공개하도록 요구한다(제50조 제4항). 이와 함께 표시·탐지·라벨링 의무의 실효성을 위해 AI Office가 EU 차원의 행동강령 마련을 촉진하고 있으며, 필요 시 집행위가 공통 규칙을 정할 수 있도록 규정했다(제50조 제7항). 즉, 유통된 결과물을 나중에 삭제하는 규제 체계만으로 부족하다고

판단하여 생성물의 기술적 식별 가능성을 요구하는 방향으로 제도화한 것이다.

〈표 5－11〉 EU「AI 법」상 딥페이크 · 합성 콘텐츠에 대한 표시 · 공개 의무

Article 50. 특정 AI 시스템의 제공자 및 배포자에 대한 투명성 의무
1. 제공자는 자연인과 직접 상호작용하도록 의도된 AI 시스템이, 해당 자연인이 AI 시스템과 상호작용하고 있다는 사실을 인지할 수 있도록 설계·개발되도록 보장하여야 한다. 다만, 합리적으로 충분한 정보에 기반하고 주의 깊으며 신중한 자연인의 관점에서 볼 때, 사용 상황과 맥락을 고려할 경우 그 사실이 명백한 경우에는 적용되지 아니한다. 이 의무는 적절한 보호 조치가 전제되는 한, 범죄의 탐지·예방·수사·기소를 위해 법률에 따라 사용이 허용된 AI 시스템에는 적용되지 아니한다. 단, 해당 시스템이 일반 대중이 범죄를 신고하기 위해 이용할 수 있는 경우에는 예외로 한다.
2. 합성 음성, 이미지, 영상 또는 텍스트 콘텐츠를 생성하는 AI 시스템(범용 AI 시스템을 포함)의 제공자는, AI 시스템의 출력물이 인공적으로 생성되었거나 조작되었음을 기계 판독 가능한 형식으로 표시하고 탐지 가능하도록 보장하여야 한다. 제공자는 콘텐츠 유형별 특성과 한계, 이행 비용, 그리고 일반적으로 인정되는 기술 수준(state of the art)을 고려하여, 효과적이고 상호운용 가능하며 견고하고 신뢰할 수 있는 기술적 해결책을 마련하여야 한다. 다만, 해당 AI 시스템이 표준적인 편집을 보조하는 기능만을 수행하거나, 배포자가 제공한 입력 데이터 또는 그 의미를 실질적으로 변경하지 않는 경우, 또는 범죄의 탐지·예방·수사·기소를 위해 법률에 따라 사용이 허용된 경우에는 적용되지 아니한다.
3. 감정 인식 시스템 또는 생체 분류 시스템의 배포자는, 해당 시스템의 작동 사실을 그 영향에 노출되는 자연인에게 고지하여야 하며, 개인정보 처리는 관련되는 경우 규정(EU) 2016/679, 규정(EU) 2018/1725 및 지침(EU) 2016/680을 준수하여야 한다. 이 의무는 범죄의 탐지·예방·수사를 위해 법률에 따라 허용된 생체 분류 및 감정 인식 AI 시스템에는 적용되지 아니하되, 제3자의 권리와 자유에 대한 적절한 보호 조치가 전제되어야 하며, 유럽연합법을 준수하여야 한다.
4. 이미지, 음성 또는 영상 콘텐츠를 생성하거나 조작하여 딥페이크(deep fake)에 해당하는 결과물을 만들어내는 AI 시스템의 배포자는, 해당 콘텐츠가 **인공적으로 생성 또는 조작되었음을 공개(disclose)**하여야 한다. 다만, 범죄의 탐지·예방·수사·기소를 위해 법률에 따라 사용이 허용된 경우에는 적용되지 아니한다.해당 콘텐츠가 명백히 예술적·창작적·풍자적·허구적 또는 이와 유사한 작품이나 프로그램의 일부인 경우, 본 항의 투명성 의무는 작품의 표시나 감상을 저해하지 않는 범위 내에서, 해당 콘텐츠가 생성·조작되었음을 적절한 방식으로 고지하는 것으로 제한된다. 또한, 공익적 사안에 대해 대중에게 정보를 제공할 목적으로 게시되는 텍스트를 생성하거나 조작하는 AI 시스템의 배포자는, 해당 텍스트가 인공적으로 생성 또는 조작되었음을 공개하여야 한다. 다만, 범죄 대응을 위해 법률에 따라 사용되는 경우, 또는 인간의 검토 또는 편집 과정을 거쳤고 특정 자연인 또는 법인이 편집 책임을 부담하는 경우에는 적용되지 아니한다.
5. 제1항부터 제4항까지에서 언급된 정보는, 해당 자연인이 최초로 상호작용하거나 콘텐츠에 노출되는 시점까지 명확하고 식별 가능한 방식으로 제공되어야 하며, 적용가능한 접근성 요건을 충족하여야 한다.

6. 제1항부터 제4항까지의 규정은, 제Ⅲ장(고위험 AI 시스템)에 따른 요구사항과 의무에 영향을 미치지 아니하며, AI 시스템 배포자에게 적용되는 기타 유럽연합 또는 회원국 법상의 투명성 의무에도 영향을 미치지 아니한다.
7. AI Office는, 인공적으로 생성 또는 조작된 콘텐츠의 탐지 및 라벨링 의무의 효과적인 이행을 촉진하기 위하여, 유럽연합 차원의 행동강령(codes of practice) 수립을 장려하고 지원한다. 집행위원회는 제56조 제6항에 따른 절차에 따라 해당 행동강령을 승인하는 이행입법을 채택할 수 있으며, 행동강령이 불충분하다고 판단되는 경우에는 제98조 제2항에 따른 심사 절차에 따라 공통 이행 규칙을 정하는 이행입법을 채택할 수 있다.

다음으로, EU는 디지털 성범죄 대응에서도 원본 촬영물 존재를 전제했던 기존 체계의 한계를 인지하고, AI를 이용한 조작·합성물을 범죄 구성요건에 직접 포함시키는 입법을 채택했다. 2024년 제정된 '여성폭력 및 가정폭력 대응 지침(Directive(EU) 2024/1385'은 동의 없는 친밀한 이미지·조작물의 공개를 범죄로 규정하면서, 그 대상에 AI 등으로 제작·조작·변형된 성적 합성물을 명시적으로 포함하고 있다. 예를 들어 전문에서는 AI를 포함한 수단으로서 성적 행위가 있는 것처럼 보이게 하는 조작·변형과 딥페이크 제작 그 자체가 문제이며, 그 이후 공개될 경우 피해를 심각하게 증폭시킨다는 취지를 분명하게 한다[120]. 이 지침의 본문 조항도 같은 맥락에서 ICT를 통해(a) 친밀 이미지의 공개뿐만 아니라,(b) 성적 행위가 있는 것처럼 보이게 제작·조작·변형한 뒤 공개하는 행위를 처벌 대상으로 두고 있어, 유통 규제만이 아니라 조작·합성의 제작 단계에서부터 공개 단계까지를 함께 포섭하는 구조로 읽힌다(Directive(EU) 2024/1385, Article 5)

〈표 5－12〉 Directive(EU) 2024/1385, Article 5

Article 5. 동의 없는 친밀한 이미지 또는 조작된 자료의 공유
1. 회원국은 다음 각 호의 고의적 행위가 형사범죄로 처벌될 수 있도록 보장하여야 한다.
(a) 정보통신기술(ICT)을 이용하여, 해당 당사자의 동의 없이, 성적으로 노골적인 행위 또는 특정인의 친밀한 신체 부위를 묘사한 이미지, 영상 또는 이와 유사한 자료를 공중이 접근 가능하도록 하는 행위로서, 해당 행위가 그 사람에게 중대한 피해를 초래할 가능성이 있는 경우

120) Directive(EU) 2024/1385 of the European Parliament and of the Council of 7 May 2024 on combating violence against women and domestic violence, OJ L, 2024/1385, 2024. 5. 24. 서문(Recitals) 18·19·20

(b) 정보통신기술(ICT)을 이용하여, 해당 당사자의 동의 없이, 특정 인물이 성적으로 노골적인 행위에 관여하고 있는 것처럼 보이도록 이미지, 영상 또는 이와 유사한 자료를 제작·조작·변형한 후 이를 공중이 접근 가능하도록 하는 행위로서, 해당 행위가 그 사람에게 중대한 피해를 초래할 가능성이 있는 경우

(c) (a) 또는 (b)에 해당하는 행위를 하겠다고 위협함으로써, 특정인으로 하여금 일정한 행위를 하거나, 이를 용인하거나, 또는 특정 행위를 하지 않도록 강요하는 행위

2. 본 조 제1항 (a) 및 (b)는 유럽연합조약(TEU) 제6조에 규정된 권리·자유·원칙을 존중하여야 하며, 표현의 자유 및 정보의 자유, 예술과 학문의 자유에 관한 기본 원칙에 영향을 미치지 아니한다. 이 규정은 유럽연합법 또는 회원국법에 따라 구현된 해당 기본 원칙을 침해하지 아니하는 범위에서 적용된다.

마지막으로, 이러한 흐름은 불법 합성 콘텐츠를 단순한 유통 규제의 대상으로서만 보지 않고, 생성·조작 단계에서 결과의 식별 가능성을 확보하는 방향으로 확장될 수 있음을 시사하고 있다. EU 「AI 법」은 합성 음성·이미지·영상·텍스트 등 합성 콘텐츠 출력물에 대해 기계 판독이 가능한 표시와 탐지 가능성을 확보하도록 요구하고, 기술적 해결책이 효과적이고 상호운용 가능하도록 설계할 것을 규정한다(「AI 법」 제50조 제2항). 또한 딥페이크의 경우, 인공적인 생성·조작 사실을 공개하도록 하되, 예술·풍자 등에서는 감상을 저해하지 않는 방식의 고지로 한정함으로써 표현의 자유와의 균형을 조문과 전문(134)에서 함께 명시하고 있다(「AI 법」 제50조 제4항, Recital 134). 아울러 표시·탐지·라벨링 의무의 실효적인 이행을 위해 AI Office가 EU 차원의 행동강령 수립을 촉진하고, 필요시 집행위원회가 공통 이행 규칙을 정할 수 있는 근거도 두고 있다(「AI 법」 제50조 제7항).

나. 영국

1) 개요

영국의 아동·청소년 보호는 AI 기술 자체를 별도로 허가·등록하는 방식이 아니라, 온라인 서비스(플랫폼, 검색 등)의 안전의무를 법으로 부과하고, 그 과정에서 AI를 주요 리스크 요소로 포함해 관리하는 접근 방식을 취하고 있다. 특히 「온라인 안전법 2023」(이하 OSA)가 아동·청소년 보호의 중심이 되며, 규제기관인 Ofcom이 리스크 평가, 연령 확인, 유해 콘텐츠 차단, 안전설계 조치 등을 구체화하는 코드 및 가이드라인을 제정·집행한다.

2) 관련 법체계

(1) 온라인 안전법 2023(이하 OSA)

영국의 아동·청소년 온라인 보호 체계는 「OSA」를 중심으로 마련되어 있다. 「OSA」는 온라인 서비스를 user-to-user 서비스, 검색 서비스, 포르노그래픽 콘텐츠 제공 서비스 등 유형별로 구분하고, 각 서비스 유형에 따라 불법 콘텐츠 대응 의무, 아동에게 유해한 콘텐츠에 대한 보호 의무, 연령 확인 또는 연령 보증 의무를 단계적으로 부과하는 구조를 취하고 있다. 이러한 의무의 구체적인 이행과 집행은 독립 규제기관인 Ofcom이 담당한다.

특히 「OSA」는 아동·청소년 보호를 핵심적인 정책 목표 중 하나로 명시하고 있으며, 아동이 유해 콘텐츠에 접촉하거나 노출되지 않도록 선제적 조치를 취할 것을 플랫폼의 기본적인 안전 의무로 설정하고 있다. 이는 불법·유해 콘텐츠 발생 이후의 삭제·차단에 한정된 사후적인 대응이 아니라, 서비스의 설계·운영 단계에서부터 리스크를 식별·완화하도록 법적으로 요구하는 구조라는 점에서 특징이 있다.

「OSA」는 AI 기술 자체를 직접적으로 규율하는 법률은 아니나, AI 기반 추천·검색·자동노출·생성 기능이 아동에게 유해한 콘텐츠의 확산 가능성을 증폭시킬 수 있다는 점을 전제로, 이러한 기능이 적용된 서비스 전반에 대해 리스크 평가 및 안전 조치 의무를 부과하고 있다. 이에 따라 생성형 AI나 알고리즘 기반 서비스 기능 역시 아동·청소년 보호 관점에서의 온라인 안전 규율 체계 안에 포섭되는 구조를 가진다.

① 리스크 평가 기반 규율

「OSA」의 가장 핵심적인 특징은 리스크 평가(risk assessment)를 규제의 출발점으로 설정하고 있다는 점이다. 온라인 서비스 제공자는 자사 서비스의 기능, 이용자 구성, 콘텐츠 유통 방식 등을 종합적으로 고려하여 불법콘텐츠 및 아동·청소년 대상 유해콘텐츠 리스크를 체계적으로 평가해야 하며, 이러한 리스크 평가 결과에 따라 적절한 완화 조치를 설계·이행할 의무를 부담한다. 이와 같은 리스크 평가 및 완화 조치 의무는 「OSA」의 집행기관인 Ofcom이 제시하는 가이던스를 통해 구체화된다.

특히, 「OSA」 체계에서는 아동·청소년이 해당 서비스의 주요 이용자이거나, 서비스 이용 과정에서 아동·청소년이 유해 콘텐츠에 노출될 가능성이 존재하는 경우, 리스크 평가 과정에서 이를 별도로 고려하도록 요구하고 있다. 이에 따라 아동·청소년의 노출 가능성이 확인된 서비스에 대해서는, 리스크 평가 결과에 따라 보다 강화된 보호 조치가 요구될

수 있다. 이는 아동·청소년을 직접 대상으로 설계한 서비스에 한정되지 않고, 검색·추천 등 서비스 기능을 통해 아동·청소년이 콘텐츠에 접근할 수 있는 경우까지 포괄한다.

「OSA」는 AI 시스템 자체를 리스크 수준에 따라 분류하거나 직접 규율하는 체계를 채택하고 있지는 않다. 다만, 「OSA」에 따른 리스크 평가와 보호 의무는 서비스 기능과 콘텐츠 유통 방식 전반을 대상으로 적용되며, 그 결과 아동·청소년이 서비스 이용 과정에서 다양한 자동화된 기능과 상호작용하는 상황 역시 규율 범위에 포함된다. 이러한 점에서 「OSA」의 리스크 평가 체계는 아동·청소년의 온라인 이용 환경 전반에서 발생할 수 있는 리스크를 사전에 식별하고 완화하도록 요구하는 구조로 이해할 수 있다.

〈표 5－13〉 영국 「OSA」 상 리스크 평가 기반 규율 관련 조문

Section 9. 불법 콘텐츠 리스크 평가 의무
(1) 이 조는 모든 규제 대상 이용자 간(user-to-user) 서비스와 관련하여 적용되는 리스크 평가에 대한 의무를 규정한다.
(2) 별표 3(Schedule)에 명시된 시기 또는 그에 따라 규정된 시기에 적절하고 충분한 불법 콘텐츠 리스크 평가를 수행할 의무를 부담한다.
(3) 오프콤이 해당 종류의 서비스와 관련된 리스크 프로필을 중요하게 변경하는 경우를 포함하여, 불법 콘텐츠 리스크 평가를 최신 상태로 유지하기 위해 적절한 조치를 취해야 한다.
(4) 서비스의 설계 또는 운영의 측면에 중대한 변경을 가하기 전에, 해당 제안된 변경의 영향과 관련된 적절하고 충분한 추가적인 불법 콘텐츠 리스크 평가를 수행해야 한다.
Section 10. 불법콘텐츠에 관한 안전 의무
(1) 이 조는 규제 대상 이용자간 서비스와 관련하여 적용되는 불법 콘텐츠에 관한 의무를 규정한다.
(2) 서비스와 관련하여, 서비스의 설계 또는 운영에 관련된 비례적인 조치를 취하거나 사용할 의무를 부담한다.
(3) 설계된 비례적인 시스템 및 프로세스를 사용하여 서비스를 운영해야 한다.
(4) 제2항 및 제3항에 규정된 의무는 서비스에 존재하는 콘텐츠뿐만 아니라 서비스가 설계, 운영 및 사용되는 방식을 포함하여 서비스의 모든 영역에 걸쳐 적용되며, 서비스 제공자에게 비례적인 경우 다음 영역에서 조치를 취하거나 사용할 것을 요구한다.
(a) 규제 준수 및 리스크 관리 체계
(b) 기능, 알고리즘 및 기타 특징 있는 설계
(c) 이용 약관에 관한 정책
(d) 서비스 또는 서비스에 존재하는 특정 콘텐츠에 대한 이용자 접근에 관한 정책
(e) 콘텐츠 삭제를 포함한 콘텐츠 관리
(f) 이용자가 접하는 콘텐츠를 제어할 수 있게 하는 기능

(g) 이용자 지원 조치
(h) 직원 정책 및 관행

Section 12. 아동 보호를 위한 안전 의무
(1) 이 조는 아동이 접근할 가능성이 있는 규제 대상 user-to-user 서비스에 적용되는 아동 온라인 안전 보호 의무를 규정한다.
(2) 서비스 제공자는 해당 서비스와 관련하여, 서비스의 설계 또는 운영과 관련된 비례적인 조치를 취하거나 사용하여 다음 각 사항을 효과적으로 이행해야 할 의무를 부담한다.
(a) 해당 서비스에 대한 최신 아동 리스크 평가에서 식별된 바에 따라, 연령대별 아동에게 발생할 수 있는 위해 리스크를 완화하고 관리할 것
(b) 서비스에 존재하는 아동에게 유해한 콘텐츠로 인해 연령대별 아동에게 발생하는 피해의 영향을 완화할 것
(3) 서비스 제공자는 다음을 목적으로 설계된 비례적인 시스템 및 절차를 사용하여 서비스를 운영해야 할 의무를 부담한다.
(a) 서비스 이용을 통해 어떠한 연령의 아동도 1차 우선 아동 유해 콘텐츠(primary priority content harmful to children)에 접촉하지 않도록 방지할 것
(b) 위해 리스크가 있다고 판단되는 연령대의 아동이 그 밖의 안동 유해 콘텐츠에 접촉하지 않도록 보호할 것
(4) 제3항 (a)의 의무는, 서비스 제공자가 서비스 내에서 식별한 1차 우선 아동 유해 콘텐츠에 대해, 연령 확인 또는 연령 추정을 사용하여 어떠한 연령의 아동도 해당 콘텐츠에 접촉하지 않도록 할 것을 요구한다.

Section 13. 아동 보호를 위한 안전 의무
(1) 제12조의 목적상 '비례적'인지 여부를 판단함에 있어, 특히 다음 각 요소가 관련된다.
(a) 해당 서비스에 대한 최신 아동 리스크 평가의 모든 결과
(b) 서비스 제공자의 규모 및 역량
(2) 제12조에 따른 의무 중 아동에게 유해한 비지정 콘텐츠와 관련된 의무는, 최신 아동 리스크 평가에서 식별된 경우에 한하여, 그와 같은 콘텐츠 유형으로부터 발생하는 위해 리스크를 다루는 범위로만 확장되는 것으로 본다.
(3) 제12조 제3항 (b) 및 제9항 (b)·(c)에서 말하는 '아동에게 유해한 콘텐츠로부터 위해 위험이 있다고 판단되는 연령대의 아동'이란, 해당 서비스 제공자가 최신 아동 리스크 평가에서 위해 리스크가 있다고 판단한 연령대의 아동을 의미한다.

② 연령보증 및 안전설계(safety-by-design) 접근

「OSA」는 아동·청소년 보호를 위해, 온라인 서비스 제공자가 연령에 따른 접근 통제 및 보호 조치를 서비스 설계 단계에서부터 고려하도록 요구하는 규율방식을 채택하고 있다. 특히, 아동·청소년에게 본질적으로 유해한 콘텐츠가 제공되거나, 아동의 접근 가능성이 존재하는 서비스의 경우, 연령확인 또는 연령보증 조치를 통해 아동의 이용을 제한하거나

보호할 의무가 부과된다.「OSA」내에서는 연령보증의 구체적인 방식이나 기술을 일률적으로 규정하지는 않지만, 아동·청소년이 실제로 유해 콘텐츠에 접근하지 않도록 하는 실질적인 효과성을 확보할 것을 요구하고 있다. 이에 따라, 연령보증 조치의 구체적인 기준과 기대 수준은 Ofcom이 제정 및 공개하는 코드(Code)와 가이던스를 통해 단계적으로 구체화된다. Ofcom은 단순한 연령 자기 신고 방식만으로는 충분하지 않을 수 있다는 것을 전제로 하여, 서비스의 리스크 수준과 성격에 비례한 연령보증 조치의 도입을 요구하고 있다. 아울러「OSA」체계는 연령보증을 개별적 기술 조치에 한정하지 않고, 서비스 전반의 설계·운영 방식에서 아동·청소년 보호가 반영되도록 하는 '안전설계' 원칙을 전제로 하고 있다. 이에 따라 플랫폼은 불법·유해 콘텐츠가 아동·청소년에게 노출된 이후 대응하는 방식에 그치지 않고, 콘텐츠의 노출 방식, 접근 경로, 서비스 기능 구성 등 전반적인 설계 요소를 통해 아동·청소년 보호를 구현할 것이 요구된다. 이러한 연령보증 및 안전설계 기준의 접근은, 아동·청소년이 온라인 서비스 이용 과정에서 유해콘텐츠에 노출될 가능성을 사전에 예방하는 것을 목표로 한다는 점에서「OSA」의 리스크 평가 기반 규율과 긴밀하게 연계되어 작동한다고 볼 수 있다. 즉, 리스크 평가를 통해 확인된 아동·청소년 노출 가능성은 연령보증 강화나 서비스 설계 변경 등 구체적인 보호 조치로 이어지는 구조를 형성하고 있다.

〈표 5－14〉 영국「OSA」상 연령보증 및 안전설계 접근 관련 조문

Section 12. 아동 보호를 위한 안전 의무
(1) 이 조는 아동이 접근할 가능성이 있는 규제 대상 user-to-user 서비스에 적용되는 아동 온라인 안전 보호 의무를 규정한다.
(2) 서비스 제공자는 해당 서비스와 관련하여, 서비스의 설계 또는 운영과 관련된 비례적인 조치를 취하거나 사용하여 다음 각 사항을 효과적으로 이행해야 할 의무를 부담한다.
(a) 해당 서비스에 대한 최신 아동 리스크 평가에서 식별된 바에 따라, 연령대별 아동에게 발생할 수 있는 위해 리스크를 완화하고 관리할 것
(b) 서비스에 존재하는 아동에게 유해한 콘텐츠로 인해 연령대별 아동에게 발생하는 피해의 영향을 완화할 것
(3) 서비스 제공자는 다음을 목적으로 설계된 비례적인 시스템 및 절차를 사용하여 서비스를 운영해야 할 의무를 부담한다.
(a) 서비스 이용을 통해 어떠한 연령의 아동도 1차 우선 아동 유해 콘텐츠(primary priority content harmful to children)에 접촉하지 않도록 방지할 것

(b) 위해 리스크가 있다고 판단되는 연령대의 아동이 그 밖의 안동 유해 콘텐츠에 접촉하지 않도록 보호할 것

(4) 제3항 (a)의 의무는, 서비스 제공자가 서비스 내에서 식별한 1차 우선 아동 유해 콘텐츠에 대해, 연령확인 또는 연령추정을 사용하여 어떠한 연령의 아동도 해당 콘텐츠에 접촉하지 않도록 할 것을 요구한다.

(5) 제4항의 요구사항은 다음 각 호의 요건이 모두 충족되는 경우를 제외하고, 각 유형의 1차 우선 아동 유해 콘텐츠에 대해 적용된다.

(a) 서비스 약관에 해당 유형의 1차 우선 아동 유해 콘텐츠의 존재가 금지됨을 명시하고 있고,

(b) 해당 정책이 서비스의 모든 이용자에게 적용되는 경우

(6) 제3항 (a)의 의무를 이행하기 위해 제4항에 따라 연령확인 또는 연령추정을 사용하는 경우, 해당 연령확인 또는 연령추정은 특정 이용자가 아동인지 여부를 정확히 판단하는 데 있어 매우 높은 효과성을 갖는 방식이어야 한다.

(7) 연령확인 또는 연령추정은, 아동 이용자인지 여부 또는 아동 이용자의 연령대를 식별하기 위한 조치로서, 제4항에 의해 의무화되지 않은 경우에도, 제2항 또는 제3항의 의무를 이행하기 위해 사용할 수 있는 조치의 예시에 해당한다.

(8) 제2항 및 제3항의 의무는 서비스의 모든 영역에 적용되며, 서비스의 설계, 운영, 이용 방식 및 서비스에 존재하는 콘텐츠를 포함한다. 또한 서비스 제공자는 비례적인 경우 다음 각 영역에서 조치를 취하거나 사용해야 한다.

(a) 규제 준수 및 리스크 관리 체계

(b) 기능, 알고리즘 및 기타 요소의 설계

(c) 이용약관 정책

(d) 서비스 또는 특정 콘텐츠에 대한 이용자 접근 정책

(e) 콘텐츠 조정, 콘텐츠 삭제 포함

(f) 특히 아동을 대상으로, 접촉하는 콘텐츠를 통제할 수 있는 기능

(g) 이용자 지원 조치

(h) 직원 관련 정책 및 관행

Section 13. 아동 보호를 위한 안전의무

(5) 제12조에 규정된 의무는, 아동이 접근할 수 있는 서비스의 부분에 한하여 적용된다.

(6) 제5항의 목적상, 서비스 제공자가 아동이 서비스 전체 또는 그 일부에 접근하는 것이 불가능하다고 판단할 수 있는 경우는, 해당 서비스에서 연령확인 또는 연령추정이 사용되어, 그 결과 아동이 통상적으로 해당 서비스 또는 그 부분에 접근할 수 없게 되는 경우에 한한다.

③ 감독·집행 체계

「OSA」에 따른 아동·청소년 보호 의무의 감독과 집행은 Ofcom이 전담한다고 볼 수 있다. Ofcom은 「OSA」에 근거하여 온라인 서비스 제공자가 이행해야 할 리스크 평가, 연령 보증, 아동·청소년 보호 조치 등이 실제로 적절히 수행되고 있는지를 점검 및 감독하는 권한을 보유하며, 이를 위해 코드(Code)와 가이던스 제정, 정보 제출 요구, 조사 및 집행

조치를 단계적으로 수행한다.

「OSA」의 감독·집행 체계는 일률적인 사전 허가 방식이 아니라, 리스크 기반·사후 점검 중심 구조이다. 즉, Ofcom은 모든 서비스를 동일하게 규제하기보다는, 서비스의 성격과 아동·청소년 노출 리스크 수준에 따라 리스크 평가의 충실성, 완화 조치의 적정성, 보호 조치의 실효성을 중심으로 감독 강도를 달리 적용한다. 특히, 아동·청소년 보호와 관련하여 중대한 위험이 확인된 경우, Ofcom은 해당 서비스에 대해 추가적인 정보 제출이나 시정 조치의 이행을 요구할 수 있다. 또한, Ofcom은「OSA」집행 과정에서 서비스 기능과 운영 방식 전반을 감독 대상으로 한다. 이에 따라 추천·검색·자동 노출 등 자동화된 서비스 기능이 아동·청소년 보호 의무 이행에 미치는 영향 역시 감독 범위에 포함된다고 볼 수 있다. 다만, 이는 AI 시스템 자체를 독립적으로 인증·허가하거나 기술적 위험 등급을 부여하는 방식이 아니라,「OSA」상 안전 의무가 해당 기능에도 동일하게 적용되는지를 점검하는 방식으로 이루어진다고 볼 수 있다.

「OSA」위반이 확인될 경우, Ofcom은 시정명령, 서비스 운영 방식 변경 요구, 과징금 부과 등 강력한 집행 수단을 행사할 수 있다. 이러한 제재 체계는 온라인 서비스 제공자로 하여금 아동·청소년 보호 의무를 형식적으로 이행하는 데 그치지 않고, 리스크 평가와 보호 조치를 지속적으로 점검·개선하도록 유도하는 구조로 볼 수 있다.

이상과 같이,「OSA」의 감독·집행 체계는 AI 기술 자체에 대한 직접적인 규율보다는, AI이 적용된 서비스 환경 전반에서 아동·청소년 보호 의무가 실제로 이행되고 있는지를 확인하고 강제하는 역할에 중점을 두고 있다.

(2) Ofcom 규제문서(Code, 가이던스)

「OSA」는 법률 차원에서 원칙적인 의무를 규정하고, 그 구체적인 내용은 Ofcom이 제정하는 규제 코드(Code of Practice)와 가이던스를 통해 구체화되는 이원적 구조를 취하고 있다. Ofcom은「OSA」이행을 위해 서비스 리스크 평가 지침, 아동 보호 조치, CSAM(아동 성착취물) 탐지·차단, 높은 수준의 효과성을 요건으로 하는 연령보증 가이던스 등을 순차적으로 발표 및 업데이트 하고 있다.

2026년 현재, Ofcom은 플랫폼을 대상으로 리스크 평가 완료 여부, 아동 보호 조치 이행, CSAM 대응 체계, 연령보증 도입 여부 등을 중심으로 한 집행 프로그램을 가동 중이며, 단

계별 시행 일정과 감독 계획을 공개하고 있다.[121]

(3) Information Commissioner's Office(ICO)

아동·청소년이 이용할 가능성이 있는 온라인 서비스는 「OSA」와 별도로, ICO가 집행하는 'Age Appropriate Design Code(Children's Code)'의 적용을 받는다. 이 Code는 「GDPR」 기반의 데이터 보호 규범으로, 아동·청소년의 최선의 이익을 기준으로 서비스 기본 설정, 데이터 최소화, 프로파일링 및 추천 시스템 통제, 아동·청소년친화적 정보 제공 등 15개 설계·운영 기준을 준수하도록 요구한다. 이에 따라 영국에서는 콘텐츠 안전은 「OSA」, 아동 데이터·프라이버시는 'Children's Code'가 각각 작동하는 이중적인 규율 구조가 형성되어 있으며, 실제 서비스 제공자는 두 규범을 동시에 충족해야 한다.

(4) 성적 딥페이크·괴롭힘 범죄화와 「OSA」 연계

「OSA」는 기존 형사법 체계와 연계되어 작동하는 온라인 안전 규제법으로, 단순한 행정규제에 그치지 않고 일부 범죄 구성요건을 직접 도입하거나 기존 범죄의 온라인 적용을 명확화하고 있다. 특히, 사이버플래싱을 명시적 범죄로 규정하고, 허위 또는 악의적인 온라인 커뮤니케이션 등 새로운 유형의 온라인 범죄를 신설했다.[122] 이에 따라 영국 검찰청(CPS)은 「OSA」가 신설한 통신관련 범죄를 기존의 기소 지침체계에 반영하여 적용하고 있다. CPS가 제공하는 '통신 범죄 지침(Communications Offences guidance)'은 「악의적 통신법」(Malicious Communications Act 1988), 「통신법」(Communications Act 2003)과 함께 「OSA」 Part 10에서 신설된 범죄(offences)를 다루며, 새로운 위협적·허위 커뮤니케이션 범죄 등을 포함하도록 구성되어 있다.[123] 「OSA」 Part 10은 '거짓 메시지 전송(false communications offence)', '위협적 메시지 전송(threatening communications offence)', '섬광 이미지 전송(epilepsy trolling)', '중대한 자해를 부추기는 행위(serious self-harm encouragement)', '원치

121) Ofcom, "Online Safety: Policy and Enforcement Updates." <https://www.ofcom.org.uk/online-safety>(최종방문일 2025. 12. 30).

122) Online Safety Act 2023.

123) Crown Prosecution Service, "Communications Offences.", Legal Guidance, 2025. 3. 24. <https://www.cps.gov.uk/prosecution-guidance/communications-offences>(최종방문일 2025. 12. 30).

않는 성적 이미지 전송(cyber-flashing)', '동의 없이 친밀한 이미지 유포 및 협박(sharing/threatening to share intimate images)'등 다양한 통신 관련 범죄를 신설했으며, 이들 범죄는 2024년 1월 31일부로 발효되어 CPS의 기소 판단에 포함되고 있다.[124] CPS는 이러한 범죄를 별도의 「OSA」 전용 가이드라인으로 공개하고 있지는 않지만 CPS의 '통신 범죄 지침' 자체가 「OSA」의 신설 범죄들을 포함하도록 업데이트된 상태에서 운영되고 있다. 이를 통해 「OSA」 신설 범죄 유형에 대한 수사·기소 판단 기준을 검찰 및 수사기관이 참고할 수 있도록 하고 있다.[125]

3) 생성형 AI·딥페이크 등 신유형 서비스 리스크 대응 방식(불법합성 콘텐츠 위주)

영국은 생성형 AI 확산에 따른 신유형 리스크를 콘텐츠의 기술적 생성 여부 자체보다는, 해당 콘텐츠가 온라인 서비스 환경에서 아동·청소년에게 어떠한 리스크를 발생시킬 수 있는지를 중심으로 규율하는 접근 방식을 취하고 있다. 특히 「OSA」는 생성형 AI나 딥페이크를 독립적인 규제 대상으로 정의하지는 않지만, AI을 통해 생성·조작된 콘텐츠가 아동·청소년에게 유해한 결과를 초래하는 경우 이를 온라인 안전 규제의 핵심 리스크 유형으로 포섭하고 있다.

「OSA」 체계에서 생성형 AI의 딥페이크 문제는 단순한 불법 콘텐츠 유통을 넘어, 콘텐츠의 생성·확산 구조 자체가 피해의 규모와 속도를 증폭시킬 수 있다는 점에 주목한다. 이에 따라 온라인 서비스 제공자 중 생성형 AI 기능이나 자동화된 콘텐츠 생성·노출 기능을 운영하는 경우, 해당 기능이 아동·청소년에게 유해 콘텐츠 노출 위험을 증가시키는지 여부를 사전적 리스크 평가를 통해 식별하고, 그 결과에 따라 노출 제한, 확산 경로 통제, 접근 제한 등 완화 조치를 이행할 의무를 부담한다.

특히 「OSA」는 아동 성착취물, 성적 딥페이크, 비동의 친밀 이미지, 자해·자살 조장 콘텐츠 등 아동·청소년에게 중대한 피해를 발생시킬 수 있는 콘텐츠 유형을 중점적으로 관

124) Crown Prosecution Service, "Communications Offences.", Legal Guidance, 2025. 3. 24. <https://www.cps.gov.uk/prosecution-guidance/communications-offences>(최종방문일 2025. 12. 30).

125) Crown Prosecution Service, "Communications Offences.", Legal Guidance, 2025. 3. 24. <https://www.cps.gov.uk/prosecution-guidance/communications-offences>(최종방문일 2025. 12. 30).

리하고 있다. 이러한 콘텐츠가 인간에 의해 제작되었는지, 또는 생성형 AI를 통해 생성·조작되었는지 여부와 관계없이, 아동·청소년에게 미치는 결과적 위험을 기준으로 동일한 보호 의무와 대응 수준이 적용된다는 점이 특징이다.

영국은 EU와 달리 생성형 AI 콘텐츠 전반에 대해 일반적 표시 의무나 기술적 식별 기준을 일률적으로 부과하지는 않는다. 대신 생성형 AI 또는 딥페이크 기능이 적용된 서비스에 대해, 아동 접근 제한, 노출 최소화, 추천·자동 노출 메커니즘 조정, 신고 및 신속한 삭제 체계 구축 등 서비스 차원이 예방·대응 조치를 결합적으로 요구하는 방식으로 대응한다.

아울러 영국은 생성형 AI의 악용으로 인한 성적 딥페이크 및 아동 성착취 문제에 대해서는 형사법 체계를 통한 직접적 대응을 병행하고 있다. 최근 영국 정부는 비동의 성적 딥페이크의 생성 행위 자체를 범죄화하고, 이를 「OSA」 상 우선 대응 불법행위(priority offence)와 연계함으로써, 플랫폼 사업자에게 해당 콘텐츠의 사전 예방·탐지·차단 책임을 강화하는 방향으로 제도를 정비하고 있다.

제3절 소결: 아동 · 청소년 특별 보호 방안

앞서 논의했듯이, 생성형 AI 기술의 발전과 함께, 텍스트·이미지·영상 등 다양한 형태의 콘텐츠를 자동으로 생성하는 AI가 아동·청소년의 일상 전반에 빠르게 확산되면서, 자동화·익명성을 기반으로 하는 구조적 특성으로 인해 불법·유해정보가 과거에 비해 훨씬 광범위하게 생성되고, 쉽게 확산될 수 있는 환경이 조성되고 있다. 특히 이용자의 단순한 입력만으로도 고위험 콘텐츠가 즉각적으로 생성될 수 있다는 점에서, 기존의 정보 유통 중심의 규제 방식만으로는 대응에 한계가 있다.

이와 같은 문제의식은 앞서 논의한 EU, 영국 등 주요 해외국의 최근 입법에서도 공통적으로 확인된다. EU는 「AI 법」, 「DSA」 등을 통해 플랫폼 사업자에게 불법 콘텐츠 유통 차단뿐만 아니라, 서비스 설계 및 운영 단계에서의 사전적 리스크 평가와 완화 조치를 요구하고 있으며, 영국 역시 「OSA」를 통해 불법·유해 콘텐츠가 생성·확산될 구조적 리스크를 사전에 관리하도록 하는 예방 중심의 규율 체계를 도입했다. 이는 불법·유해정보에 대한 사후 삭제·차단만으로는 이용자 보호에 한계가 있다는 점을 고려하여 생성·서비스 제공 단계에서의 책임을 명확히 하려는 국제적 흐름이라고 볼 수 있다.

특히, 성범죄물과 아동·청소년성착취물은 생성형 AI 기술과 결합하면서 새로운 양상으로 확산되고 있다. 합성·편집 등 기술을 활용한 딥페이크 성범죄는 실제 촬영 여부와 무관하게 특정 인물의 인격과 성적 자기결정권을 중대하게 침해할 수 있으며, 아동·청소년을 대상으로 할 경우 그 심각성은 더욱 커진다. 이에 더해, 이러한 콘텐츠는 한번 생성되면 빠르게 복제·유통되어 피해 회복이 사실상 불가능하다는 점에서, 사전적인 대응이 필요하다.

그러나 현행 법체계는 주로 결과물의 유통을 전제로 한 사후 규율에 초점이 맞춰져 있어, AI의 개발·설계·운영 단계에서 불법·유해정보의 생성을 예방하기 위한 직접적인 규율에는 일정 부분 한계가 있다. 이에 따라 AI 서비스 전반에 걸쳐 불법 성범죄물과 청소년 유해 콘텐츠의 생성·유통을 구조적으로 방지하고, 특히 청소년 이용자 보호를 강화하기 위한 별도의 법적 근거를 마련할 필요성이 제기되고 있다. 이하의 핵심 과제별 규정 제안은 국제적 입법 추세와 국내 법체계의 한계를 종합적으로 고려하여, 생성형 AI 서비스를 주축으로 팽창하는 AI 서비스 생태계에서 보호의 사각지대에 놓인 아동·청소년을 실효적으로 보호하기 위한 법적 장치를 신설하는 데 목적이 있다.

1. 인공지능을 이용한 성범죄물 등의 생성 · 유통 규제

생성형 AI를 이용한 성범죄물은 실제 촬영 행위나 직접적인 신체 접촉 없이도 특정 인물의 얼굴, 신체, 음성 등을 합성·편집하여 생성될 수 있으며, 이 과정에서 피해자의 의사나 인식은 전혀 반영되지 않는다. 이러한 AI 성범죄물은 외형상 실제 촬영물과 구별하기 어려우며, 피해자의 사회적 평가와 인격적 존엄성에 중대한 피해를 발생시킨다는 점에서 실질적인 성범죄 피해와 동일한 위험성을 가진다. 피해자는 실제로 촬영을 한 사실이 없음에도 불구하고, 자신의 신체나 성적 이미지가 왜곡 및 조작된 형태로 재현되어 유통되는 과정에서 심각한 정신적 고통과 사회적 낙인을 경험할 수 있다. 이러한 피해는 콘텐츠 진위 여부와 관계없이 발생하며, 디지털 환경의 특성상 반복 및 지속될 수 있다는 점에서 회복이 불가능하다. 그럼에도 불구하고 현행 법체계에서 성범죄물에 대한 규율은 주로 '촬영물'을 전제로 한 개념 구성에 기반하고 있어, AI 기술을 활용한 합성·편집 결과물을 명확히 포섭하는 데에는 해석상 한계가 존재한다. 생성형 AI 환경에서는 '촬영'이라는 물리적

행위 자체가 범죄 성립의 요소로 고려될 뿐만 아니라, 특정 인물의 성적 이미지가 그 의사에 반하여 재현 및 유포된다는 점도 고려해야 한다. 이에 따라, 성범죄물 개념을 AI 환경에 맞게 재정립하고, 실제 촬영 여부와 관계없이 피해가 발생하는 합성·편집 등 결과물까지 포괄할 필요성이 제기된다.

이러한 문제의식을 반영하여 AI 생성 성범죄물 및 아동·청소년성착취물에 대해 명시적인 생성·유통 규제의 도입을 제안하고자 한다. 요컨대, 불법정보의 유형을 「성폭력범특례법」 및 「아동·청소년의 성보호에 관한 법률」에서 규정하고 있는 정보를 인용함으로써 새로운 법률에 따른 해석상의 혼란을 최소화하고, 기존 형사법 체계와의 정합성을 견고히 하는 방향을 우선 고려할 필요가 있다. 아울러, 기술 발전에 따라 새롭게 등장할 수 있는 범죄 유형에 대응할 수 있도록 대통령령에 의한 위임 규정으로써 규제의 탄력성을 확보할 필요가 있다.

2. 인공지능서비스 제공자 등에 대한 예방 중심의 책임 부과

생성형 AI 환경에서 불법·유해정보에 대한 대응을 사후 제재 중심으로만 설계할 경우, 그 실효성에는 한계가 존재한다. AI를 통해 생성된 성범죄물이나 아동·성착취물은 한번 생성되는 즉시 복제·확산이 가능하며, 이후 삭제·차단 조치가 이루어진다고 하더라도 이미 발생한 인격적·정신적 피해를 회복하는 것이 사실상 불가능하기 때문이다. 특히 자동화된 생성 구조와 대량 생산이 가능한 AI 시스템 특성상, 사후적 대응만으로는 불법정보의 확산 속도를 따라가기 어렵다는 점이 분명해지고 있다. 이러한 한계는 국제적으로도 공통된 문제의식으로 논의되고 있다. EU와 영국 등 해외 주요국들은 서비스 설계 및 운영 단계에서 발생 가능한 위험을 사전에 평가하고 완화하도록 하는 예방 중심의 규율 체계를 도입하고 있다. 이는 기술적으로 구현 가능한 범위 내에서 합리적인 예방 조치를 요구하는 리스크 기반 접근에 기초한 것으로서, 기술 발전을 저해하지 않으면서도 이용자 보호의 실효성을 확보하기 위한 선택이라고 볼 수 있다.

이러한 관점에서 AI 개발자와 AI 서비스 제공자에게 '기술적으로 구현 가능한 범위' 내에서 성범죄물 등의 생성·유통을 예방할 의무를 부과하는 방안을 제안한다. 이는 과도한 규제를 통해 기술 혁신을 위축시키려는 것이 아니라, 현재 확보하고 있는 기술 수준과 서비

스 특성을 고려하여 합리적으로 기대 가능한 보호 조치를 요구함으로써 과잉규제를 방지하고 기술의 중립성을 확보하려는 취지이다. AI 환경의 특성을 고려할 때, 불법·유해 정보 예방에 대한 책임을 단일 주체에게 일률적으로 부과하는 방식이 적절하지 않으므로 주체별 역할과 책임을 구분하여 규율할 필요가 있다. 따라서 AI 개발자는 모델과 시스템의 설계 단계에서 구조적으로 불법정보 생성 가능성을 낮추도록 하는 한편, AI 서비스 제공자는 해당 시스템이 실제로 이용되는 과정에서 운영·유통·이용자 관리에 대한 책임을 부담하도록 함이 각자의 역량과 책임 영역을 고려할 때 합당하다. 이에 따라 AI 개발자에게는 불법정보의 생성을 기술적으로 방지하는 기능을 갖추도록 하고, AI 서비스 제공자에게는 서비스 이용 과정 전반에서 불법정보의 생성·유통을 예방하기 위한 단계적 조치를 요구하는 내용의 규정을 둘 필요가 있다.

구체적으로는 불법정보 관련 기술적·관리적 보호 체계를 통해 예방·탐지·대응이 유기적으로 작동하도록 요구사항을 명시하는 방안을 고려할 수 있다. 생성형 AI 환경에서는 개별 콘텐츠의 사후적인 삭제·차단 조치만으로는 용이하고 급속하게 생성·확산되는 막대한 양의 콘텐츠를 통제할 수 없는바, 이는 AI 생성 콘텐츠에 대한 '예방적 관리 체계'를 신설하는 의미를 갖는다.[126] 이 경우에도 기술 발전 속도와 서비스 유형의 다양화를 고려하여 구체적인 기술적·관리적 조치는 대통령령으로 위임하여 새로운 위험 유형이나 기술 환경 변화에 대응할 수 있도록 여지를 두어야 할 것이다.

3. 청소년 이용자에 관한 특별 보호

최근 AI 서비스는 학습, 오락, 정보탐색 서비스를 넘어 정서적 교류와 상호작용의 수단으로까지 활용 범위가 확대되고 있으며, 이로 인해 청소년의 일상에서 이용되는 비중 역시 빠르게 증가하고 있다. 그러나 챗봇 등 일부 AI 서비스는 연령확인 절차 없이 제공되고 있어, 청소년이 성인 이용자와 동일한 환경에서 서비스를 이용하는 것이 일반적인 상황이다. 이로 인해 청소년은 자신에게 적합하지 않은 콘텐츠나 상호작용에 노출될 가능성이 높아지고 있으며, 서비스 이용 목적이 다양해질수록 그 위험 역시 복합적으로 발생하고

126) 향후 법안의 발의 및 시행 과정에서 정보통신망법상 불법정보 유통 방지 의무와의 관계를 명확히 하고 법률 간의 상충·중복을 막기 위한 재정비 논의가 요구된다.

있다. 청소년은 발달 단계상 인지·판단 능력과 자기 통제 능력이 성인에 비해 미성숙하기 때문에, 유해 콘텐츠에 대해 부정적인 영향을 그대로 받을 가능성이 크다. 특히, AI 서비스의 경우 개인화된 응답과 지속적인 상호작용을 통해 이용자의 관심과 감정을 강화하는 구조를 띠고 있어, 청소년에게 과몰입이나 의존 또는 중독을 유발할 위험성이 높다. 이러한 특성은 단순한 청소년의 정서적 발달과 가치관 형성에 장기적인 영향을 미칠 수 있기 때문에 각별한 주의가 요구된다. 그럼에도 불구하고 현행 청소년 보호 제도는 주로 플랫폼에 대한 일반적 규제를 중심으로 하고 있어, AI 서비스의 특수성을 반영하지 못하고 있다. 기존에 운영되는 제도는 주로 유해정보의 유통 차단 등 사후적 관리에 중점을 두고 있으며, 이용자와의 지속적인 상호작용, 개인화된 콘텐츠 생성, 정서적 교류를 특징으로 하는 AI 서비스 상 위험을 체계적으로 관리하기는 한계가 있다. 이에 따라 아동·청소년 보호를 AI 환경에 맞게 마련할 필요가 있다.

이에 따라 청소년 이용자가 이용할 수 있는 AI 서비스에 대해서는 일반 이용자와 구별되는 특별한 보호 체계를 마련할 것이 요구된다. 청소년 이용자가 서비스 이용을 개시하는 단계에서 이들에게 필요한 사항을 미리 안내하고, 청소년에게 유해한 영향을 미치는 콘텐츠의 생성을 방지할 수 있는 유효적절한 조치를 취하도록 하는 한편, 과몰입 및 중독을 예방하기 위한 보호·감독 기능을 탑재하도록 의무화하는 방안을 적극 검토할 필요가 있다. 특히, 이용자와의 개인적 교류, 사회적 관계 형성 등을 통해 감정적·정서적 교류를 나누는 AI 서비스의 경우에는 인지적·정서적으로 미성숙한 단계에 있는 청소년에게는 그 영향이 훨씬 더 부정적으로 나타날 수 있다. 따라서 이러한 유형의 AI 서비스에 대해서는 특별히 상호 관계에 대한 오인지를 방지하기 위한 차별화된 보호 조치가 요구된다.

현행 정보통신망법 등에서 요구되는 청소년보호 책임자 지정의무를 AI 서비스에도 확장하여 동 법률의 의무를 이행하도록 하는 것은 법체계에 비추어 무리가 없는바, 법률에서 정한 일정 규모 이상의 AI 서비스 제공자는 청소년보호 책임자를 지정하도록 함으로써, 청소년 보호 의무가 AI 서비스 제공 회사 내부에서 지속적으로 관리 및 감독될 수 있도록 함이 타당하다.

제6장 결 론

제1절 연구의 요약

본 연구는 AI 서비스가 일상과 공공 영역 전반에 깊숙이 스며든 현실에서, 기술 혁신을 저해하지 않으면서도 이용자의 기본적인 권리와 이익을 실질적으로 보호할 수 있는 규범 구조를 어떻게 설계할 것인가에 대한 답을 모색하는 데 목적을 두었다. 이를 위해 기술·산업적 특성을 반영한 AI 생태계 분석과 글로벌 규범 동향, 국내 선행 규제 분야의 경험, 그리고 아동·청소년 보호 관점에서의 예방 중심 규범 논의를 단계적으로 검토함으로써, 이용자 중심의 권리 규범과 보호 체계를 뒷받침할 이론적·법리적 기반을 도출하였다.

본 연구가 제안하는 AI 서비스 이용자 보호 법제의 방향성과 입법 과제를 종합적으로 제시하는 취지에서 각 장에서 다루어진 주요 논점을 정리하면 다음과 같다.

1. AI 생태계를 반영한 규범 체계 및 법적 구성 요소 도출(제2장)

본 연구의 전제적 고찰이자 연구 결과의 토대로서 AI 서비스 생태계를 기술적·산업적 구조로서만이 아니라 순환적·동태적 시스템으로 파악하는 이론적 틀을 제시하였다. AI 서비스는 데이터 수집-모델 개발-서비스 설계·제공-이용자 상호작용-규범·규제적 피드백이 반복·축적되는 순환 구조를 가지며, 이 과정에서 위험과 부작용, 권력 비대칭이 함께 증폭될 수 있다는 점을 전제로 하였다. 이에 따라 AI 생태계는 다수의 행위자가 얽힌 가치사슬로 이해되어야 하며, 특히 개발 단계와 서비스 단계, 그리고 이용자·사회적 반응이 서로를 재구성하는 상호작용 구조임을 규명함으로써, 정태적·일회적 규제가 아니라 생애주기 전반을 대상으로 하는 동태적 규율 필요성을 도출하였다.

이러한 분석을 바탕으로, 제2장은 AI 서비스 생태계의 인적 구성을 AI 개발자-AI 서비스 제공자-이용자(최종 소비자와 영향 받는 자 포함)라는 3각 구조로 정리하고, 각 주체의 통제 가능성과 역할에 상응하는 권리·의무·책임 배분 원칙을 제시하였다. 특히 AI 서비스

제공자를, 자체 개발 시스템이든 외부로부터 공급받은 시스템이든 이를 통합해 이용자에게 제공하는 서비스 단계의 핵심 수범자로 설정함으로써, 정보제공·설명·위험관리·취약계층 보호·피해구제 절차 운영 등 이용자 보호 의무를 집중시키는 규범 구조를 구상하였다. 동시에 이용자는 기본권의 향유자이자 알고리즘 절차의 통제권 행사자, 그리고 책임 있는 이용 주체로 새기고, 개발자는 설계 단계의 일반적 안전·인권 보호 의무와 서비스 제공자에 대한 협조·정보 제공 의무를 중심으로 규율되어야 한다는 점을 제시함으로써, 이후 장에서 전개될 이용자 중심 권리 규범 및 보호 체계 구상의 이론적·법리적 기반을 마련하였다.

2. 글로벌 AI 규범의 형성과 리스크 기반 접근의 입법례 분석(제3장)

유럽연합(EU), 미국, 중국 등 주요국은 AI 기술의 급격한 확산에 대응하여 이용자 보호를 위한 독자적인 규범 체계를 발 빠르게 구축하고 있다. EU는 세계 최초의 포괄적 AI 규제법인 「AI 법」을 통해 리스크 수준에 따른 차등적 규제 방식을 확립하였으며, 「디지털서비스법(DSA)」을 통해 온라인 플랫폼 내 알고리즘 추천 시스템의 투명성 의무를 엄격히 규정하고 있다. 미국은 연방 및 주 단위에서 딥페이크 책임법과 아동 온라인 안전법 등을 논의하며 신종 위험에 대한 개별적 대응을 강화하고 있으며, 중국은 생성형 AI 관리와 더불어 이용자의 '정서적 안전'까지 보호 범위를 확장하는 선제적 입법을 시도 중에 있다.

이러한 해외 법제의 공통적 지향점은 단순한 사후 처벌을 넘어선 '설계에 의한 안전'(Safety by Design)원칙과 알고리즘 투명성 확보에 있다. 특히 아동·청소년 및 고령자 등 취약계층에 대해서는 연령 확인 시스템 도입, 유해 콘텐츠 생성 방지, 과몰입 예방 기능을 의무화하는 등 차별화된 보호 장치를 마련하고 있음이 확인된다. 또한 이를 실효적으로 집행하기 위해 전문화된 감독 기구와 실태조사, 분쟁 조정 체계를 아우르는 다층적 거버넌스를 구축하고 있다는 점은 국내 입법에 중요한 시사점을 제공한다.

3. 국내 선행 분야의 보호 체계와 미디어 서비스로의 확장성 검토(제4장)

국내에서는 금융과 의료 분야를 중심으로 AI 이용자 보호를 위한 구체적인 규제 체계가 선제적으로 정립되어 왔다. 금융 분야는 기술중립적 원칙 하에 AI 시스템이 소비자에게 미

치는 영향력에 따른 '리스크 기반 접근 방식'을 이미 채택하였으며, 자동화된 결정에 대한 설명요구권과 거부권 등 실질적인 대응권을 법적으로 보장하고 있다. 의료 분야는 「디지털의료제품법」을 통해 학습 데이터와 심사 결과의 투명한 공개를 의무화하고, 제품의 전 생애주기에 걸친 조직적 거버넌스와 품질 관리를 요구하며 안전성을 확보하고 있다.

이러한 선행 분야의 핵심 원칙들은 AI가 융합된 미디어 서비스 영역으로 확장되어야 한다. 오늘날 미디어 환경에서 알고리즘은 정보 소비를 결정하는 '보이지 않는 편집자' 역할을 수행하므로, 타 분야에서 검증된 투명성 확보 및 경영진 책임 강화 방안을 미디어 환경의 특수성에 맞게 수용할 필요가 있다. 특히 생성형 AI의 확산에 따른 환각 현상과 콘텐츠의 허구성 등 새로운 유형의 리스크에 대응하기 위해, 선행 분야의 리스크 관리 경험을 바탕으로 한 보편적 이용자 보호 체계 수립이 절실한 시점이다.

4. 아동 · 청소년 보호를 위한 예방 중심의 규범과 입법 과제(제5장)

생성형 AI 기술의 고도화는 딥페이크 성범죄물 및 아동 성착취물의 자동·대량 생성을 가능하게 하여 아동·청소년의 기본권을 심각하게 위협하고 있다. 현행 법체계는 주로 실제 촬영된 영상물 중심의 사후 규제에 집중하고 있어, 한번 유포되면 피해 회복이 사실상 불가능한 AI 기반 디지털 성범죄의 확산 구조를 차단하는 데 한계를 보인다. 따라서 아동·청소년 보호 규범은 이용자의 주의의무에 기대기보다는 서비스 설계 단계부터 위험을 최소화하는 '사전적 예방' 중심으로의 규제 패러다임 전환이 필요하다.

이를 실행하기 위해서도 전 주기 대응은 불가결하다. AI 개발자에게는 성범죄물 등의 생성을 기술적으로 방지하는 기능을 내재화할 의무를, 서비스 제공자에게는 실시간 모니터링과 유해 콘텐츠 삭제 등 운영상 책임을 명확히 부과하는 것을 필요 최소한의 입법 사항으로 제안하고자 한다. 특히 정서적 교류를 모사하는 AI 서비스의 경우에는 청소년 보호에 특화된 적절한 기술적 장치가 마련되어야 한다. 아울러 부처 간 협력을 통해 통합적인 관리 체계를 구축함으로써 법적 규범과 기술적 안전장치가 상호 보완적으로 작동하는 규제 환경을 조성하는 정부 차원의 노력이 주효할 것이다.

제 2 절 연구의 결과

1. 입법 및 정책의 제안

본 연구는 AI 서비스 생태계의 양적 성장이 초래할 수 있는 이용자의 기본권 침해와 사회적 부작용을 선제적으로 예방하고, 안전한 AI 서비스 이용 환경을 조성하기 위해 가칭 「인공지능서비스 이용자 보호에 관한 법률(안)」을 마련하는 실무적 토대로서 다음과 같은 핵심적인 입법 사항과 정책적 지향점을 제안한다.

가. 거버넌스 체계의 확립

AI 서비스의 복합적인 특성을 고려하여 분절된 규제를 지양하고 범정부 차원의 통합적 이용자 보호 체계를 구축하는 첫 단계로서 통합적 거버넌스 및 조직법적 근거를 마련해야 한다. 이는 이용자 보호 정책의 일관성과 전문성을 확보하기 위한 조직법적 준비 사항에 해당한다.

나. 이용자의 실정법적 권리 명문화

AI 시스템의 불투명성으로 인해 발생하는 이용자의 소외 현상을 극복하고 실질적인 권리 실현을 보장하기 위하여 다음과 같은 실체적·절차적 권리를 명문화할 필요가 있다.

AI 서비스의 내용(과정 및 결정)이 이용자의 권익에 중대한 영향을 미치는 경우, 이용자가 AI 서비스에 적용된 결정의 기준과 논리를 이해하기 쉬운 방식으로 설명받을 권리를 보장하여야 한다. 또한 결과에 불복할 경우에는 인간에 의한 재검토를 요청할 수 있는 절차적 권리를 수반하여 실질적인 권익의 회복 또는 구제를 담보하여야 한다.

대민 행정 서비스를 비롯한 특정 서비스 이용 시에는 AI 기술 적용 여부를 사전에 선택할 수 있도록 하고, 특히 고도로 개인화된 추천 서비스의 경우 이용자가 자신의 데이터 활용 범위와 서비스 수준을 직접 설정할 수 있는 '자기결정권' 기반의 인터페이스 설계를 의무화하는 방안을 고려할 수 있다.

다. 서비스 제공자의 예방적 책무 강화

AI 서비스 제공자에게 부여되는 의무는 사후적 조치보다는 개발 및 운영 전 과정에서의 예방적 조치에 집중되도록 하는 것이 효과적이다. 다만, 이때 과도한 사전 규제가 되지 않

도록 설계에 의한 투명성 및 안전 조치에 방점을 두어야 할 것이다.

이용자가 현재 상호작용하는 대상이 AI임을 명확히 인지할 수 있도록 표준화된 표기 방식을 도입해야 한다. 생성형 AI의 경우 산출물이 허구일 가능성(환각 현상), 정보의 출처 등을 고지하여 이용자의 합리적 판단을 지원해야 한다.

아울러, 서비스 운영 전 과정에서 잠재적 위험을 주기적으로 평가하고, 기술적 오류나 편향성을 최소화하기 위한 관리적 조치를 설계 단계부터 반영해야 한다. 이는 단순한 약관 명시를 넘어, 서비스 설계 단계부터 윤리적 가이드라인이 알고리즘에 내재화될 수 있도록 하는 '설계에 의한 안전' 원칙을 법제화하는 것을 의미한다.

라. 아동·청소년 보호를 위한 특별한 고려

정서적 상호작용이 강조되는 AI 서비스의 경우, 청소년 이용자를 특별히 고려하여 유해 콘텐츠로부터 보호할 수 있는 조치의 이행을 의무화해야 한다. 그리고 딥페이크 성범죄물 등 아동·청소년에게 유해한 콘텐츠가 생성되지 않도록 기술적 차단 조치를 행할 것을 요구하고, 보호자가 자녀의 안전한 이용을 관리할 수 있는 모니터링 및 보호·감독 기능을 의무적으로 탑재하도록 한다. 불법·유해 콘텐츠 규제는 그 대상과 수단이 현행 「정보통신망법」 제44조의7과 중첩되는 측면이 있는바, 실제 입법 단계에서 별도의 제정법률안 규정으로 둘지, 현행법의 개정안으로 다룰지 여부를 규제의 중복·충돌이 발생하지 않도록 선결적으로 논의되어야 할 것이다.

마. 사후적 권리 구제 및 책임 체계의 실효성 제고

피해 발생 시 신속하고 저렴한 비용으로 구제받을 수 있는 비사법적 구제 절차와 국외 사업자에 대한 집행력 확보 방안이 병행되어야 한다.

이를 위하여 전문성을 갖춘 '인공지능서비스분쟁조정위원회'를 설치하여 재판 외 분쟁해결을 활성화해야 한다. 또한, 불특정 다수에게 공통으로 발생하는 AI 관련 피해의 특성을 고려하여 집단분쟁조정 제도를 도입함으로써 구제의 실효성을 높일 것을 적극 고려할 필요가 있다.

또한 국내 거주 이용자에게 서비스를 제공하는 글로벌 사업자에 대해서는 국내 거주 대리인 지정을 의무화하여, 법적 책임 추궁과 이용자 요구사항에 대한 즉각적인 대응이 가능하도록 법적 근거를 마련한다.

2. 병행 및 후속 과제 제안

이상에서 제시된 입법 사항에 대하여, 「인공지능서비스 이용자 보호에 관한 법률(가칭)」로써 독립적인 조문 체계로 구성할지, 기존의 정보통신·개인정보·청소년 보호·디지털플랫폼 관련 법령을 유기적으로 개정·정비하는 형태로 추진할지에 대한 입법 전략의 선택이 선행되어야 한다. 이때 개별 법령의 중복·충돌을 최소화하면서도, 이용자 보호와 관련된 핵심 원칙(위험 기반 접근, 설계에 의한 안전, 이용자 권리 보장 등)이 일관되게 관통되도록 "상위 규범-부문별 특례" 구조를 설계하는 것이 바람직하다.

또한 입법과 병행하여 표준계약 조건, 기술적·관리적 보호 조치에 관한 가이드라인, 사업자 협의체를 통한 자율규제 등 연성법적 수단을 적극적으로 활용함으로써, 법률이 포괄하지 못하는 기술 변화와 서비스 유형의 다양성을 보완하여야 할 것이다. 이는 규제기관-산업-시민사회 간의 협치 구조를 제도화하는 것과 밀접하다.

법률에 따른 감독기능 또는 분쟁조정기능을 담당하는 기관의 전문성·독립성을 담보하기 위해서는 인력 양성, 예산 확보, 데이터·알고리즘에 대한 접근권 보장 등 정책적 인프라 조성이 급선무이다. 이를 위해 중장기 AI 규제·감독 역량 강화 로드맵을 수립하는 과제가 조속히 진행되어야 할 것이다. 그 밖에도 역외 서비스 제공자 규율, 국제 데이터 이동, 글로벌 플랫폼에 대한 집행력 확보를 위해 주요 국가 및 국제기구와의 협력 채널을 강화하고, AI 거버넌스에 관한 국제 규범 형성 과정에 적극 참여함으로써 국내 입법의 실효성을 담보하는 외교·정책적 노력이 뒷받침될 때 비로소 본 연구에서 제안된 AI 서비스 이용자 보호를 위한 법제도적 제언이 실정법적 결실로 맺어질 것이다.

물론 본 연구에서 제시된 방안이 AI 서비스 이용자 보호 법제를 완성하는 최종 해답이 될 수는 없다. 기술과 시장, 그리고 사회적 기대는 지속적으로 변화하고 있으며, 이에 따른 새로운 위험과 규범적 쟁점이 끊임없이 발생할 것이기 때문이다. 그럼에도 본 연구가 제안하는 거버넌스 체계, 이용자 권리 명문화, 서비스 제공자의 예방적 책무, 아동·청소년 보호, 사후 구제 및 책임 체계의 정교화, 그리고 병행·후속 과제로서의 입법·정책 방향은 향후 논의의 출발점이자 기준점으로 기능할 수 있을 것이다. 앞으로 입법자, 규제기관, 산업, 학계, 시민사회가 이러한 기반 위에서 지속적인 대화와 협력을 이어간다면, AI 기술의 혁신과 인간의 존엄·기본권 보호가 조화되는 지속 가능한 AI 규범 질서가 형성될 수 있을 것이라 기대한다.

참 고 문 헌

[국내 문헌]

경기일보, "'딥페이크 범죄 창구' 텔레그램… 8월 이용자 증가폭 역시 최대", 2024. 9. 5, <https://www.kyeonggi.com/article/20240905580230>(최종방문일 2025. 12. 30).

과학기술정보통신부, 「2024 인터넷 이용실태조사」, 2025. 3.

국회입법조사처, "호주의 '아동·청소년 소셜미디어금지법' 통과와 입법적 시사점", 「NARS 현안분석」, 345호, 2024.

김희정, "딥페이크 기술을 이용한 신종범죄에 대한 법정책적 시사점", 「서강법률논총」, 통권 31호, 97-128, 2024.

금융위원회, "금융분야 인공지능 개발·활용 안내서", 2022.

_____, "금융권 생성형 인공지능 활용 지원방안", 2024.

딜로이트 인사이트, 「인공지능(AI) 시대의 새로운 국면, AI 규제와 기업 리스크 관리 전략」, 2024. 10.

신성원, "딥페이크 기술을 활용한 디지털 성범죄에 관한 연구", 「한국치안행정논집」, 통권 67호, 147-168, 2023.

정경환, "딥페이크 성폭력 처벌법의 형법적 과제", 「법학연구」, 제28집 제1호, 1-32, 2025.

조경희, "일본의 인공지능(AI) 개발·활용 추진 입법례", 「최신외국입법정보」 제273호, 2025.

최경미·지성우, "디지털서비스 투명성·개인정보보호·위기대응에 관한 법 연구-EU 디지털서비스법(DSA)을 중심으로", 「미국헌법연구」, 33권 3호, 177-214, 2022.

한겨레, "'딥페이크 텔레방에 22만명...입장하니 "좋아하는 여자 사진 보내라"", 2024. 9. 2,<https://www.hani.co.kr/arti/society/society_general/1154764.html>(최종방문일 2025. 12. 30).

한국인터넷진흥원, 「인터넷 정보서비스 알고리즘 추천관리규정」, 2022.

_____, "일본, 「인공지능 관련 기술의 연구개발 및 활용 촉진에 관한 법률」 공

포('25. 6. 4.)", 「인터넷·정보보호 법제동향」, 2025.

한국일보, "딥페이크 심의 건수 폭증세… 이대로라면 '사상 최다'", 2025. 9. 8. <https://www.hankookilbo.com/News/Read/A2025090810360002776>(최종방문일 2025. 12. 30).

한국지능정보사회진흥원, 「2023년 美 NIST 「AI 위험관리 프레임워크(AI RMF) 1.0」 분석 및 시사점」, 2023.

_____, 「2024년 전자정부서비스 이용실태조사」, 2025. 1.

AI매터스, 「AI 트렌스포메이션 2: 2025 생성형 AI 인식 및 활용 현황 조사」, 2025.

Kotra, "싱가포르의 AI 정책과 실제 사례", 「Global Issue Monitoring」, 2024.

____, "일본의 AI 정책과 실제 사례", 「Global Issue Monitoring」, 2024.

[해외 문헌]

5Rights Foundation, *Children & AI Design Code*. 2025. <https://5rightsfoundation.com/resource/children-ai-design-code/>(최종방문일 2025. 12. 19).

California Legislature, *California AI Transparency Act* (A.B.853), 2025-2026. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260A.B.853> (최종방문일 2025. 12. 19).

_____, *Companion Chatbots*. (S.B.243), 2025-2026. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260S.B.243> (최종방문일 2025. 12. 19).

_____, *Contracts Against Public Policy: Personal or Professional Services: Digital Replicas.* (A.B.2602), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240A.B.2602> (최종방문일 2025. 12. 19).

_____, *Crimes: Child Pornography.* (A.B.1831), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240A.B.1831> (최종방문일 2025. 12. 19).

California Legislature, *Crimes: Distribution of Intimate Images.* (S.B.926), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240S.B.926> (최종방문일 2025. 12. 19).

_____, *Depiction of Individual Using Digital or Electronic Technology: Sexually Explicit Material: Cause of Action.* (A.B.602), 2019-2020. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201920200A.B.602> (최종방문일 2025. 12. 19).

_____, *Elections: Deceptive Media in Advertisements.* (A.B.2839), 2023-2024. <https://legiscan.com/CA/text/A.B.2839/id/3021243>(최종방문일 2025. 12. 19).

_____, *Generative Artificial Intelligence: Training Data Transparency.* (A.B.2013), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240A.B.2013> (최종방문일 2025. 12. 19).

_____, *Political Reform Act of 1974: Political Advertisements: Artificial Intelligence.* (A.B.2355), 2023-2024. <https://legiscan.com/CA/text/A.B.2355/id/3008736>(최종방문일 2025. 12. 19).

_____, *Sexually Explicit Digital Images.* (S.B.981), 2023-2024. <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240S.B.981> (최종방문일 2025. 12. 19).

_____, *The California Age-Appropriate Design Code Act* (A.B.2273), 2021-2022. <https://leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=202120220A.B.2273&showamends=false>(최종방문일 2025. 12. 19).

_____, *Use of Likeness: Digital Replica.* (A.B.1836), 2023-2024. <https://legiscan.com/CA/text/A.B.1836/id/3021237>(최종방문일 2025. 12. 19).

Cheng, E., "China to crack down on AI chatbots around suicide, gambling", *CNBC*, 2025. 12. 29, <https://www.cnbc.com/2025/12/29/china-ai-chatbot-rules-emotional-influence-suici

de-gambling-zai-minimax-talkie-xingye-zhipu.html>(최종방문일 2025. 12. 30).

Code of Virginia, § 18.2-386.2: *Unlawful Dissemination or Sale of Images of Another;* penalty. (H.B.2678), 2019. <https://law.lis.virginia.gov/vacode/title18.2/chapter8/section18.2-386.2/>.(최종방문일 2025. 12. 19).

Connecticut General Assembly, *An Act Concerning Artifical Intelligence.* (S.B.2), 2024. <https://www.cga.ct.gov/asp/cgA.B.illstatus/cgA.B.illstatus.asp?selBillType=Bill&bill_num=S.B.00002&which_year=2024>(최종방문일 2025. 12. 19).

Crown Prosecution Service, "Communications Offences.", Legal Guidance, 2025. 3. 24. <https://www.cps.gov.uk/prosecution-guidance/communications-offences>(최종방문일 2025. 12. 30).

Data Protection Act 2018, c. 12. <https://www.legislation.gov.uk/ukpga/2018/12/contents>(최종방문일 2025. 12. 19).

Directive(EU) 2024/1385 of the European Parliament and of the Council of 7 May 2024 on combating violence against women and domestic violence, OJ L, 2024/1385, 2024. 5. 24.

Dupre, M. H., "China Planning Crackdown on AI That Harms Mental Health of Users," Futurism, 2025. 12. 31, <https://futurism.com/artificial-intelligence/china-regulation-ai-chatbots>(최종방문일 2025. 12. 30).

European Commission, *Code of Practice for General-Purpose AI*, 2025. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>(최종방문일 2025. 12. 19).

_____, "Commission opens formal proceedings against TikTok under the Digital Services Act," IP/24/2227, 2024. 4. 22, <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2227>(최종방문일 2025. 12. 30).

_____, "Digital Services Act: Commission designates first set of Very

Large Online Platforms and Search Engines," 2023. 4. 25, <https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413>(최종방문일 2025. 12. 30).

European Commission, "Digital Services Act: Questions and Answers," 2022. 11. 16, <https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348>(최종방문일 2025. 12. 30).

_____, "Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil LiA.B.ility Rules to Artificial Intelligence(AI LiA.B.ility Directive)", 2022. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496>(최종방문일 2025. 12. 19).

_____, "Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services(Digital Services Act) and amending Directive 2000/31/EC," COM(2020) 825 final, 2020. 12. 15.

Future of Privacy Forum, "Understanding Japan's AI Promotion Act: An Innovation-First Blueprint for AI Regulation," 2025, <https://fpf.org/blog/understanding-japans-ai-promotion-act/>(최종방문일 2025. 12. 30).

Geopolitechs, "What's in China's first drafts rules to regulate AI companion addiction?," 2025. 12. 27, <https://www.geopolitechs.org/p/whats-in-chinas-first-drafts-rules>(최종방문일 2025. 12. 30).

Horsley, J. P., "China's Orwellian Social Credit Score Isn't Real," *Foreign Policy*, 2018. 11. 16, <https://foreignpolicy.com/2018/11/16/chinas-orwellian-social-credit-score-isnt-real/>(최종방문일 2025. 12. 30).

Information Commissioner's Office(ICO), *Age Appropriate Design: A Code of Practice for Online Services (The Children's Code)*, 2020. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-

practice-for-online-services>(최종방문일 2025. 12. 19).

Information Commissioner's Office(ICO), *Guidance on AI and Data Protection*, 2023. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/>(최종방문일 2025. 12. 19).

_____(ICO) & The Alan Turing Institute, *Explaining Decisions Made with AI*, 2020. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/?utm_source=chatgpt.com>(최종방문일 2025. 12. 19).

Japan Personal Information Protection Commission(PPC), *Act on the Protection of Personal Information*(Act No. 57), 2023. <https://www.ppc.go.jp/en/legal/>(최종방문일 2025. 12. 19).

KPMG, "Trust in Artificial Intelligence: Global Insights 2023," 2023, <https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2023/02/trust-in-ai-global-study-2023.pdf>(최종방문일 2025. 12. 30).

National Technical Committee 260 on Cybersecurity of SAC & National Computer Network Emergency Response Technical Team/Coordination Center of China, *AI Safety Governance Framework 2.0*, 2025.

Ofcom, "Online Safety: Policy and Enforcement Updates." <https://www.ofcom.org.uk/online-safety>(최종방문일 2025. 12. 30).

Office of the eSafety Commissioner, "Industry Regulation". <https://www.esafety.gov.au/A.B.out-us/industry-regulation>(최종방문일 2025. 12. 19).

Office of the Privacy Commissioner of Canada(OPC), *Guidance for Processing Biometrics-for Businesses*, 2025. <https://www.priv.gc.ca/en/privacy-topics/health-genetic-and-other-body-information/biometrics/gd_bio_org-final/>(최종방문일 2025. 12. 19).

Online Safety Act 2023, c. 50. <https://www.legislation.gov.uk/ukpga/2023/50/contents>(최

종방문일 2025. 12. 30).

Parliament of Canada, *An Act to Enact the Online Harms Act, to Amend the Criminal Code, the Canadian Human Rights Act and An Act Respecting the Mandatory Reporting of Internet Child Pornography by Persons Who Provide an Internet Service*.(Bill C-63), 2024. <https://www.parl.ca/DocumentViewer/en/44-1/bill/C-63/first-reading>(최종방문일 2025. 12. 19).

Regulation(EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC(General Data Protection Regulation), 2016. 5. 4., p. 1-88.

Regulation(EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC(Digital Services Act), OJ L 277, 2022. 10. 27., p. 1-102.

Regulation(EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence(Artificial Intelligence Act), OJ L, 2024/1689, 2024. 7. 12.

Sexual Offences Act 2003, c. 42. <https://www.legislation.gov.uk/ukpga/2003/42/contents> (최종방문일 2025. 12. 19).

Technology, Media, and IT, South Korea Artificial Intelligence Market Size and Share - Growth Analysis Report and Forecast Trends(2026-2035), 2025. 7.

Texas Legislature, *Relating to the Creation of a the Criminal Offense for FA.B.ricating a Deceptive Video with Intent to Influence the Outcome of an Election.* (S.B.751), 2019. <https://capitol.texas.gov/BillLookup/History.aspx?LegSess=86R&Bill=S.B.751>(최종방문일 2025. 12. 19).

_____, *Relating to the Creation of a the Criminal Offense of Possession, Promotion, or Production of Appearing to Depict a Child.* (S.B.20), 2025. <https://capitol.texas.gov/BillLookup/History.aspx?LegSess=89R&Bill=S.B.20>(최종방

문일 2025. 12. 19).

U.S. Congress, *AI Accountability and Personal Data Protection Act*(S.2367), 119th Congress, 2025-2026. <https://www.congress.gov/bill/119th-congress/senate-bill/2367>(최종방문일 2025. 12. 19).

_____. Children and Teens' *Online Privacy Protection Act*(S.1418), 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/senate-bill/1418>(최종방문일 2025. 12. 19).

_____, *DEEPFAKES Accountability Act*(H.R.5586), 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/house-bill/5586/text>(최종방문일 2025. 12. 19).

_____, *Deepfake Task Force Act*(S.2559). 117th Congress, 2021-2022. <https://www.congress.gov/bill/117th-congress/senate-bill/2559>(최종방문일 2025. 12. 19).

_____, *Identifying Outputs of Generative Adversarial Networks Act(IOGAN Act)* (H.R.4355), 116th Congress, 2019-2020. <https://www.congress.gov/bill/116th-congress/house-bill/4355>(최종방문일 2025. 12. 19).

_____, *Kids Online Safety Act*(S.1409), 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/senate-bill/1409>(최종방문일 2025. 12. 19).

_____, *No Section 230 Immunity for AI Act*(S.1993), 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/house-bill/3831>(최종방문일 2025. 12. 19).

_____, *TAKE IT DOWN Act(Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act)* (S.4569). 118th Congress, 2023-2024. <https://www.congress.gov/bill/118th-congress/senate-bill/4569>(최종방문일 2025. 12. 19).

United Kingdom Government, *AI Cyber Security Code of Practice, Department for*

Science, Innovation and Technology, 2025.
<https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice>
(최종방문일 2025. 12. 19).

United Kingdom Government, *Data(Use and Access) Bill: Factsheets*, Department for Science, Innovation and Technology, 2024.
<https://www.gov.uk/government/publications/data-use-and-access-bill-factsheets>
(최종방문일 2025. 12. 19).

Webster, G. et al., "Translation: Internet Information Service Algorithmic Recommendation Management Provisions," *DigiChina*, Stanford University, 2022. 1. 4,
<https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/>(최종방문일 2025. 12. 30).

Zhang, L.(李强治), "Experts interpret the 'Interim Measures for the Administration of Human-like Interactive Services Based on Artificial Intelligence'", *Sina Finance(新浪财经)*, 2025. 12. 28,
<https://finance.sina.com.cn/jjxw/2025-12-28/doc-inheirmq8851756.shtml>(최종방문일 2025. 12. 30).

AI戦略会議·AI制度研究会,「中間とりまとめ」, 2025. 2. 4,
<https://www8.cao.go.jp/cstp/ai/ai_senryaku/13kai/13kai.html>(최종방문일 2025. 12. 30).

中共中央·国务院,「法治政府建设实施纲要(2021-2025年)」, 2021.
<https://www.moj.gov.cn/pub/sfbgw/zwxxgk/fdzdgknr/fdzdgknrghjh/202207/t20220722_460263.html>(최종방문일 2025. 12. 19).

経済産業省,「GOVERNANCE INNOVATION Ver.2: アジャイル・ガバナンスのデザインと実装に向けて」, 2021. 7,
<https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_1.pdf>(최종방문일 2025. 12. 30).

個人情報保護委員会,「生成AIサービスの利用に関する注意喚起等について」, 2023. 6. 2,
<https://www.ppc.go.jp/news/press/2023/230602kouhou/>(최종방문일: 2025. 12. 30).

国家互联网信息办公室,「关于《人工智能拟人化互动服务管理暂行办法（征求意见稿)》公开征求意见的通知」, 2025. 12. 27., <https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm>(최종방문일 2025. 12. 30).

_____.「生成式人工智能服务管理暂行办法」, 2023. 7. 13. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm(최종방문일 2025. 12. 19).

国家新闻出版署,「关于进一步严格管理 切实防止未成年人沉迷网络游戏的通知」, 2021. 8. 30.

国务院,「新一代人工智能发展规划」, 国发〔2017〕35号, 2017. 7. 8.

科技部等十部门,「科技伦理审查办法（试行)」, 2023. 9. 14.

内閣府, 人工知能関連技術の研究開発及び活用の推進に関する法律(令和7年法律第53号), <https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html>(최종방문일 2025. 12. 30).

文化審議会著作権分科会法制度小委員会, 「AIと著作権に関する考え方について」, 2024. 3, <https://www.bunka.go.jp/>(최종방문일 2025. 12. 30).

消費者庁,「【令和6年10月1日施行】改正景品表示法の概要」, 2024.

_____, 「景品表示法とステルスマーケティング～事例で分かるステルスマーケティング告示ガイドブック」, 2023.

衆議院,「人工知能関連技術の研究開発及び活用の推進に関する法律案」, <https://www.shugiin.go.jp/internet/itdb_gian.nsf/html/gian/honbun/houan/g21709029.htm>(최종방문일 2025. 12. 30).

総務省・経済産業省,「AI事業者ガイドライン（第1.0版）」, 2024. 4. 19, <https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_1.pdf>(최종방문일 2025. 12. 30).

著作権法(昭和45年法律第48号).

不当景品類及び不当表示防止法(昭和37年法律第134号).

● 저 자 소 개 ●

권 은 정
- 서울대 법학 석사
- 서울대 법학 박사
- 현 가천대학교 교수

선 지 원
- 한양대 법학 석사
- Universität Regensburg 법학 박사
- 현 한양대학교 교수

윤 금 낭
- 성균관대 법학과 석사
- 성균관대 법학 박사
- 현 디지털산업정책연구소 연구위원

오 다 슬
- 중앙대 미디어커뮤니케이션학 석사
- 중앙대 미디어커뮤니케이션학 박사수료

장 수 진
- 이화여대 행정학 석사과정

강 지 원
- 현 김·장법률사무소 변호사

마 경 태
- 현 김·장법률사무소 변호사

윤 혜 선
- 현 한양대학교 법학전문대학원 교수

방송통신융합 정책연구 KMCC-2025-09

AI 이용자 보호를 위한 ICT 법제 연구
(A Study on ICT Legal System for the Protection of AI Users)

2025년 12월 일 인쇄
2025년 12월 일 발행

발행인 방송미디어통신위원회 위원장
발행처 방송미디어통신위원회
경기도 과천시 관문로 47
정부과천청사 2동
TEL: 02-2110-1323
Homepage: www.kmcc.go.kr
인 쇄 인 성 문 화