정책연구 25-23

지능정보사회 플랫폼 자율규제 활성화

2025. 12

연구기관 : 정보통신정책연구원

방송미디어통신위원회
Korea Media and Communications Commission

정책연구 25-23

지능정보사회 플랫폼 자율규제 활성화

(Enhancing Self-Regulation for Platforms in the Intelligent Information Society)

김현수/이경은

2025. 12

연구기관 : **정보통신정책연구원**

방송미디어통신위원회
Korea Media and Communications Commission

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

제 출 문

방송미디어통신위원회 위원장 귀하

본 보고서를 『**지능정보사회 플랫폼 자율규제 활성화**』의 연구 결과보고서로 제출합니다.

2025년 12월

연 구 기 관 : 정보통신정책연구원
총괄책임자 : 김 현 수 선임연구위원
참여연구원 : 이 경 은 부 연 구 위 원

목 차

표 목 차

그 림 목 차

요 약 문

본 연구는 온라인 플랫폼을 주축으로 한 ICT 환경 변화 속에서 새롭게 대두되는 이용자 보호 이슈에 대한 자율규제 방안 마련 및 이행을 지원함으로써 협력적 생태계 구축에 기여하는 것을 목표로 하였다. 이를 위해 시장 환경 변화와 해외 정책동향 등을 참고하여 플랫폼 자율규제 개선방안을 제시하였다.

□ 해외의 플랫폼 자율규제 동향

본 보고서는 글로벌 디지털 플랫폼 환경에서 확산되는 허위조작정보, 불법 혐오 표현, 이용자 안전 위협 등에 대응하기 위한 해외 플랫폼 자율규제 동향을 유럽연합과 영국을 중심으로 분석한다. 디지털 플랫폼의 사회적 영향력이 확대됨에 따라 전통적인 사후 규제나 단일 법률 중심의 접근만으로는 복합적 온라인 위험에 효과적으로 대응하기 어렵다는 인식이 확산되고 있으며, 이에 따라 자율규제와 법적 규제를 결합한 새로운 거버넌스 모델이 주요 정책 대안으로 부상하고 있다.

먼저 유럽연합의 플랫폼 자율규제는 '법적 규제와 정합성을 전제로 한 보완적 자율규제'라는 특징을 가진다. EU는 허위조작정보와 불법 혐오 표현을 민주주의, 사회적 신뢰, 기본권 보호에 중대한 영향을 미치는 구조적 위험으로 인식하고, 이에 대응하기 위해 다자 참여형 행동강령을 중심으로 한 자율규제 체계를 구축해 왔다. 대표적으로 허위조작정보 행동강령(Code of Conduct on Disinformation)과 온라인 불법 혐오 표현 대응 행동강령(Code of Conduct on Countering Illegal Hate Speech Online)은 플랫폼 사업자의 자발적 참여를 전제로 하면서도, 시민사회, 팩트체커, 연구기관, 광고 산업 등 다양한 주체의 역할과 책임을 제도적으로 규정한다.

EU 허위조작정보 행동강령은 허위조작정보의 생산과 유통을 개별 콘텐츠 차원이 아니라 정보 생태계 전반의 구조적 문제로 접근한다. 이를 위해 허위조작정보의 광고 수익 차단, 정치 광고의 투명성 강화, 조작적 행위 억제, 이용자 역량 강화, 연구자 데이터 접근성

확대, 팩트체킹 네트워크 강화 등 다층적인 대응 영역을 설정한다. 특히, 허위조작정보 확산의 경제적 유인을 제거하기 위한 광고 생태계 규율과 정치·이슈 광고에 대한 공개성과 검증 가능성 확보는 EU 자율규제의 핵심 축으로 기능한다.

EU의 자율규제는 이행과 검증을 위한 거버넌스 구조를 중시한다. 투명성 센터를 통해 각 서명 기관의 이행 현황을 공개하고, 상설 태스크포스를 통해 기술적·사회적·정책적 변화에 따라 행동강령을 지속적으로 조정한다. 서비스 수준 지표(SLI)와 정성적 보고 요소(QRE)를 결합한 보고 체계는 자율규제 이행을 정기적으로 측정하여 비교 가능하게 하며, 자율규제가 선언적 규범에 머무르지 않도록 하는 핵심 장치로 작동한다.

2022년 이후 EU의 자율규제는 디지털서비스법(DSA)과의 제도적 연계를 통해 규범적 위상이 크게 강화되었다. 허위조작정보 행동강령과 혐오 표현 행동강령은 DSA 제45조에 따른 공식 행동강령으로 기능하며, 초대형 온라인 플랫폼(VLOPs)이 부담하는 시스템적 위험 관리 의무를 구체화하는 수단으로 활용된다. 이에 따라 자율규제의 성실한 이행 여부는 단순한 기업의 자발적 노력 차원을 넘어 법적 감독과 평가의 기준으로도 작동하게 되었다.

영국의 플랫폼 자율규제 동향은 EU와 유사한 문제의식을 공유하면서도 보다 명확한 법적 책임 구조와 위험 기반 접근을 강조하는 방향으로 이루어지고 있다. 영국은 기존의 자율규제 중심 접근이 대형 플랫폼 환경에서 실효성을 확보하지 못했다는 평가를 바탕으로, 온라인 안전법(Online Safety Act)을 제정하여 법적 규율의 틀을 정립하였다. 이 과정에서 자율규제는 법적 의무를 대체하는 수단이 아니라 법이 요구하는 의무를 이행하기 위한 실행 메커니즘으로 재정의된다.

영국의 온라인 안전법은 플랫폼에 대해 불법 콘텐츠뿐만 아니라 합법이지만 유해한 콘텐츠에 대해서도 위험 평가와 완화 의무를 부과하며, 특히 아동 보호를 핵심 정책 목표로 설정한다. 이러한 구조 하에서 플랫폼이 스스로 위험을 식별하고 관리하는 내부 절차, 가이드라인, 운영 기준이 작동되며, 규제 이행의 결과를 투명성 보고, 위험 평가 문서, 내부 거버넌스 구조를 통해 외부에 공개하도록 요구한다.

영국 규제기관인 Ofcom은 법 집행 기관이자 동시에 규제 이행을 평가하고 유도하는 핵심 주체로 기능한다. Ofcom이 제시하는 투명성 보고 가이드라인과 위험 평가 지침은 플랫폼의 자율적 조치를 표준화하고, 비교·감독 가능하게 만드는 역할을 수행한다. Ofcom이 제정한 투명성 가이드라인은 온라인 안전법 체계 하에서 플랫폼 사업자의 책임 이행을 외

부에서 검증 가능하도록 만드는 핵심 수단으로 기능한다. 이 가이드라인의 목적은 플랫폼이 불법 및 유해 콘텐츠에 어떻게 대응하고 있는지를 이용자, 규제기관, 사회 전반이 이해할 수 있도록 정보 공개의 범위와 방식을 표준화하는 데 있다. 투명성 가이드라인은 단순한 정보 공개를 넘어 플랫폼의 내부 의사결정 구조와 위험 관리 방식을 외부에 설명 가능하도록 요구한다는 점이 특징이다. 플랫폼은 콘텐츠 삭제 및 제한 조치, 알고리즘 추천 관리, 신고 및 이의 제기 처리, 위험 완화 조치의 효과 등에 대해 정량적·정성적 정보를 체계적으로 보고해야 한다. 특히 보고 내용은 Ofcom이 비교 분석할 수 있도록 공통된 형식과 지표를 기반으로 작성하도록 설계되어 있다.

아동 접근성 평가 가이드라인은 온라인 서비스가 아동에게 실제로 접근 가능하거나 아동이 이용할 개연성이 있는지를 사전에 평가하도록 하는 데 목적을 둔다. 이는 서비스가 명시적으로 아동 대상이 아닌 경우에도 아동의 이용 가능성이 존재한다면 보호 의무를 부담해야 한다는 위험 기반 접근을 전제로 한다. 플랫폼은 연령 제한 여부나 서비스의 명목상 대상만을 기준으로 삼지 않고, 이용자 구성, 서비스 기능, 콘텐츠 성격, 유입 경로, 알고리즘 추천 구조 등을 종합적으로 고려하여 아동 접근 가능성을 판단해야 한다. 아동 접근성 평가는 이후 아동 위험성 평가의 전제가 되는 절차로서 기능하며, 플랫폼이 아동 보호 조치를 회피하지 못하도록 하는 구조적 장치로 작동한다.

아동 위험성 평가 가이드라인은 아동에게 접근 가능한 서비스가 아동의 안전, 정신적·신체적 건강, 발달에 어떠한 위험을 초래할 수 있는지를 체계적으로 식별하고 완화하도록 하는 데 목적을 둔다. 이는 온라인 환경에서 아동이 노출될 수 있는 위험을 구조적·예방적으로 관리하기 위한 핵심 규범이다. 이 가이드라인은 콘텐츠의 불법성 여부를 중심으로 한 전통적 규제 방식에서 벗어나 아동에게 미치는 잠재적 영향 자체를 평가 대상으로 삼는다는 점이 특징이다. 플랫폼은 자해·자살 유도, 폭력적·성적 콘텐츠, 혐오 표현, 중독적 설계, 과도한 비교 및 평가를 유발하는 기능 등 다양한 위험 요소를 서비스 설계 단계부터 분석해야 한다.

종합적으로 볼 때, 유럽연합과 영국의 플랫폼 자율규제 동향은 공통적으로 자율성과 책임성의 결합을 지향한다. EU는 다자 참여형 행동강령과 법적 규제의 연계를 통해 자율규제를 제도화하고 있으며, 영국은 강력한 법적 틀 하에서 자율규제를 위험 관리 수단으로 활용한다. 두 사례 모두 자율규제를 규제의 '대안'이 아니라, 규제의 '구성 요소'로 재정의

하고 있다는 점에서 중요한 시사점을 제공한다. 이러한 해외 동향은 향후 국내 플랫폼 정책과 자율규제 설계에 있어 법·제도·거버넌스 차원의 종합적 접근이 필요함을 명확히 보여준다.

□ 플랫폼 자율규제 개선방안

디지털·AI 기술 혁신과 디지털 환경의 급진적 확산으로 디지털 정보 유통의 영향력이 지대해짐에 따라 표현의 자유를 확대시킨 이점이 있는 반면, 허위조작정보 등 불법유해정보의 대량 확산이라는 새로운 사회적 위험을 수반하였다. 디지털 정보의 비대칭성과 플랫폼 기반 확산 구조는 피해 회복의 곤란, 신속 대응의 어려움, 사회적 혼란 등의 문제를 야기한다. 생성형 AI 기술의 급속한 발전은 콘텐츠 다양화 등에 기여하지만, 사실과 허위의 경계가 모호해지는 정보 불확실성이 심화될 우려가 있다.

허위조작정보는 사회적 혼란을 야기하고 사회의 건전한 토론을 왜곡하여, 특히 선거를 앞두고 여론을 부적절하게 반영하거나 사회적 분열과 양극화를 심화시킬 가능성이 있다. 또한, 정부와 정책에 대한 신뢰를 훼손하며, 특히 코로나-19와 같은 정부 위기 상황에서 국가 보건정책에 대한 혼란을 야기할 수 있다.

허위조작정보 대응은 정보의 유형(권리침해정보·법령위반정보·유해정보)과 확산 구조를 통합적으로 고려하는 체계가 필요하다. 허위조작정보의 정의 방식에 따라 규율 대상이 달라지며, 특히 사회적 법익 영역(유해정보)에서의 직접 규제는 표현의 자유와 충돌할 소지가 크다. 권리침해정보는 플랫폼의 신속 삭제 의무와 피해자 중심 절차의 제도화가 핵심이고, 법령위반정보는 행정기관 요청에 따른 우선 대응체계를 갖추며, 유해정보는 사회적 합의에 기반한 자율규제와 플랫폼 설계 구조를 활용한 간접규제가 유용하다.

허위조작정보를 포함하여 불법유해정보에 대한 대응은 단순 삭제 중심을 넘어 플랫폼의 시스템 구조 자체를 재설계하는 접근이 중요해지고 있다. 이는 알고리즘 설계, 콘텐츠 수익화 구조, 계정 생성 방식 등 기능 전반에 대해 규범적 설계 요구가 필요하다는 의미다. 이를 위해 '규제화된 자율규제'(공동규제) 도입을 검토해야 한다. 공동규제의 행동강령 준수는 법 준수의 안전지대, 모범사례 활용, 평판 리스크 완화 등 인센티브가 있다. 이행 체계로는 플랫폼이 자체 이행평가보고서를 공개하고, 제3자 감사를 거쳐 정부·이해관계자가

함께 평가·개선하는 순환형 거버넌스를 구축한다. 정책효과 측정을 위해 측정가능한 목표와 표준화된 KPI 개발이 필요하며, 준수 기준과 근거자료를 감사기관에 제공해 실효성을 담보해야 한다.

주요 세부과제로는 첫째, 허위조작정보 유포의 실질적 동기를 약화시키기 위해 수익화 인센티브를 차단하는 것이 핵심이다. 불법적 허위조작정보를 반복 유포하는 이용자에 대해 정보 삭제와 계정 정지·폐쇄를 엄정히 적용하고, 유해한 허위조작정보는 표현의 자유와의 균형을 위해 가시성에 직접 영향을 미치지 않는 수익화 차단 중심으로 대응한다. 콘텐츠 수익화 및 광고 수익 공유 프로그램의 자격요건과 콘텐츠 검토 절차를 강화하고, 광고주와 대행사에게 광고가 게재될 콘텐츠 유형에 대한 정보를 제공해 선택권을 보장한다.

둘째, 딥페이크 위험 완화는 AI 개발자부터 플랫폼까지 공급망 전 단계 대응이 필요하다. 예방 측면에서 AI 모델 개발자는 프롬프트 필터를 사용하여 특정 유형의 콘텐츠 생성 방지, 학습 데이터셋에서 유해 콘텐츠 제거, 유해 콘텐츠 생성을 자동으로 차단하는 출력 필터 사용 등을 시행할 수 있다. 임베딩 측면에서는 AI 모델 개발자와 플랫폼은 콘텐츠에 워터마크를 임베드하여 딥러닝 알고리즘을 사용하여 감지할 수 있도록 한다. 마지막으로 탐지 측면에서 플랫폼은 맥락 데이터가 첨부되지 않은 경우에도 콘텐츠 검토를 통해 허위조작정보를 식별할 수 있다.

셋째, 이용자 측면에서는 정보 제공과 역량 강화가 중요하다. 추천시스템의 투명성 및 통제권을 보장하면 허위조작정보에 접하는 것을 줄일 수 있다. 이용자가 콘텐츠의 출처, 편집이력, 진위 여부 등을 평가할 수 있도록 도구를 제공하고, 게시자의 AI 생성물 표시 반복 위반 시 제재 등을 통해 자발적으로 AI 생성물임을 표시하도록 유도한다. 팩트체커 콘텐츠의 통합과 가시성 제고로 팩트체크의 영향력을 강화한다. 권위 있는 정보의 가시성을 높이고 허위조작정보의 가시성을 낮추는 방향으로 추천시스템을 설계하고, 위기 상황에서는 이용자를 신뢰할 수 있는 공식 출처로 안내하는 기능을 제공한다.

다음으로 청소년의 SNS/숏폼 과의존과 관련하여 청소년은 도파민을 지속적으로 제공하는 중독성 피드에 특히 취약해 충동 통제가 어렵고, 이는 발달 저해와 장기적 피해로 이어질 수 있다. 또한, 정체성이 형성되는 시기에 개인화된 콘텐츠가 과도하게 제공되면 다양한 관점을 접할 기회가 제한된다는 비판도 제기된다. 이에 플랫폼은 각자의 특성·규모·목적·이용자 기반에 따라 위험 수준이 다르다는 점을 고려한 위험기반 접근을 취하되, 조치

가 아동의 권리를 부당하게 제한하지 않도록 균형을 유지해야 한다.

구체적으로는 과도한 사용을 유발하는 기능을 기본설정에서 비활성화해야 한다. 동영상 자동 재생과 라이브 스트리밍 호스팅을 끄고, 푸시 알림은 기본적으로 비활성화하되 연령에 맞춘 핵심 수면시간에는 항상 차단한다. '좋아요' 수 표시나 읽음 확인 등 참여 경쟁과 압박을 유도하는 기능도 기본적으로 비활성화해야 한다. 청소년이 이러한 기본설정을 변경하려는 경우에는 변경 시점에 경고를 제공해 잠재적 결과를 명확히 알리고, 주기적으로 안내하며 언제든 쉽게 기본설정으로 복귀할 수 있도록 해야 한다.

인터페이스 설계·운영 측면에서 무한 스크롤, 불필요한 반복 행동 요구, 인위적 알림, 희소성·긴급성을 강조하는 신호, 가상 보상 등 참여 유도 요소에 청소년이 과도하게 노출되지 않도록 해야 한다. 플랫폼은 정기적으로 그리고 인터페이스에 중대한 변경이 있을 때마다 위험 평가와 완화 조치를 시행하고, 제3자의 실효적 감사를 위해 관련 데이터와 기술 정보에 대한 접근과 협력을 제공해야 한다.

보호자 통제 기능은 기본설정 및 설계 조치를 보완하는 수단으로 활용하되, 기본설정 관리, 이용 시간 제한, 소통 계정 확인, 계정·지출 관리 등을 포함해야 한다. 이 과정에서 청소년의 프라이버시를 과도하게 침해하지 않도록, 모니터링 기능이 작동할 경우 청소년에게 실시간으로 명확히 알리는 등 적절한 보호 조치를 병행해야 한다.

제1장 서 론

플랫폼은 산업 융합·혁신, 신시장 창출로 스타트업·소상공인·창작자 등에게 성장 기회를 제공하고, 이용자 편의성 제고, 비용 절감 등 사회 후생을 증대시킨다. 반면, 플랫폼의 확산에 따른 이해관계자 갈등이나 불공정 행위 및 이용자 이익 저해행위의 논란, 서비스 장애로 인한 국민 불편을 야기한 바 있다. 따라서 플랫폼을 둘러싼 부작용을 해소하면서도 시장의 역동성과 혁신이 저해되지 않는 균형 잡힌 정책을 위해 자율규제 필요성이 제기되었다. 이에 2022년 8월 출범한 '플랫폼 자율규제 및 상생발전 촉진 기구'(이하 "플랫폼 자율기구")는 플랫폼 사업자, 중소 상공인 단체, 소비자 단체, 전문가 등 이해관계자 간의 협업 및 현안 논의를 통해 자율규제 방안을 도출하고, 이해관계자 간 소통 창구로서의 역할을 수행하고 있다.

플랫폼 자율기구는 2023년 5월 11일 "플랫폼 자율기구 자율규제 방안 발표회"를 개최하여 산하 4개 분과에서 마련한 자율규제 방안을 발표하였다. 갑을 분과는 '오픈마켓 분야 자율규제 방안', 이용자/소비자 분과는 '오픈마켓 소비자 집단피해 신속 대응 방안', 데이터/AI 분과는 '플랫폼 검색·추천 서비스 투명성 제고를 위한 자율규제 원칙', 혁신공유/거버넌스 분과는 '플랫폼의 사회 가치 제고를 위한 8대 원칙'을 발표하였다. 또한, 2024년 6월 플랫폼 자율기구의 법적 근거를 마련하고, 정부가 자율규제 활동의 지원 시책 마련 및 자율규제의 확산에 필요한 사업을 추진할 수 있도록 하기 위한 전기통신사업법 개정안이 국회에 제출되었다.

본 연구는 해외 주요국의 플랫폼 자율규제 동향 및 성과를 분석하여 국내 자율규제 정책방안에 대한 시사점을 도출하는 것을 목표로 하였다. 이를 위해 제2장에서는 유럽연합 및 영국의 플랫폼 자율규제 동향과 시사점을 분석한다. 제3장에서는 시장 환경 변화와 해외 정책동향 등을 참고하여 플랫폼 자율규제 개선방안을 제시한다. 마지막으로 제4장에서는 앞서 논의된 국내외 현황과 정책 제언을 정리한다.

제 2 장 해외의 플랫폼 자율규제 동향

제 1 절 유럽연합

1. 개요

유럽연합은 온라인 플랫폼의 사회적 영향력이 확대됨에 따라, 법적 규제와 자율규제를 결합한 혼합형 거버넌스 모델을 통해 온라인 위험에 대응해 오고 있다.[1] 특히 허위조작정보(disinformation)와 불법 혐오 표현(hate speech)은 민주적 여론 형성, 사회적 신뢰, 기본권 보호에 중대한 영향을 미치는 핵심 위험 요소로 인식되어 왔으며, 이에 대응하기 위한 대표적인 자율규제 수단으로 「온라인 허위조작정보 행동강령(Code of Conduct on Disinformation)」과 「온라인 불법 혐오 표현 대응 행동강령(Code of Conduct on Countering Illegal Hate Speech Online)」이 각각 제정·운영되고 있다.[2][3] 두 행동강령은 모두 플랫폼 사업자의 자발적 참여를 전제로 하되, 시민사회, 연구기관, 팩트체커, 광고 산업 등 다양한 이해관계자의 협력을 제도화함으로써 온라인 위험에 대한 공동 책임 구조를 형성하는 데 초점을 두고 있다.

특히 EU의 자율규제 체계는 단순한 비구속적 권고 수준을 넘어, 2022년 이후 디지털서비스법(DSA)과의 제도적 연계를 통해 규범적 위상이 강화되었다는 점에서 특징적이다. 허위조작정보 행동강령은 DSA상 시스템적 위험 관리 의무를 구체화하는 공식 행동강령으로 기능하도록 개편되었으며,[4] 혐오 표현 행동강령 역시 2023년 '행동강령+' 형태로 강화되어 DSA 제45조에 근거한 자발적 준수 메커니즘으로 자리매김하였다.[5] 이와 같은 구조 하에

---

1) European Commission(2020), European democracy action plan.

2) European Commission(2018), Code of conduct on disinformation.

3) European Commission(2016), Code of conduct on countering illegal hate speech online.

4) European Commission(2022), Strengthened code of practice on disinformation.

5) European Commission(2023), Strengthening the code of conduct on countering illegal

서 행동강령은 법적 의무를 대체하지는 않지만, 플랫폼이 취한 위험 완화 조치를 평가하는 중요한 준거로 활용되고 있으며, 성실한 이행은 DSA상 적절한 위험 관리 조치로 인정될 수 있다.[6] 결과적으로 EU의 플랫폼 자율규제는 법적 규제와의 정합성을 전제로 한 '보완적·증거 기반 자율규제' 모델로 진화하고 있으며, 이는 글로벌 플랫폼 규제 논의에서 중요한 참고 사례로 평가될 수 있다.

2. 허위조작정보 행동강령

EU의 「허위조작정보 행동강령(Code of Conduct on Disinformation)」은 온라인 환경에서 확산되는 허위조작정보(disinformation)에 체계적으로 대응하기 위한 자율규제 기반의 정책 수단으로서 플랫폼·광고업계·팩트체커·시민사회·연구기관 등 다양한 이해관계자의 공동 참여를 전제로 설계되었다. 허위조작정보 행동강령은 허위조작정보의 생산·유통·확산·수익화 전반을 포괄적으로 관리하는 것을 주요 목표로 하며, 특히 투명성 제고와 협력 강화, 사용자 보호 및 연구 기반 확충을 핵심 축으로 삼고 있다.

가. 주요영역

행동강령은 첫째, 허위조작정보의 수익창출 금지를 핵심 영역으로 설정하고 있다. 이는 허위 또는 조작된 정보 콘텐츠 인접 영역에 광고가 게재되는 것을 차단함으로써, 허위조작정보 유통이 경제적 이익으로 연결되는 구조를 근본적으로 차단하고자 하는 목적을 가진다. 아울러 광고주, 광고기술 기업, 플랫폼 간 협력을 강화하여 광고 생태계 전반에서 허위조작정보에 대한 공동 대응 체계를 구축하도록 요구하고 있다.

둘째, 투명한 정치광고 체계 구축을 주요 영역으로 규정하고 있다. 정치적 광고 및 이슈 광고에 대해 이용자가 이를 명확히 인식할 수 있도록 효율적인 라벨링 체계를 도입하고, 광고의 출처, 자금 흐름, 타깃팅 기준 등에 대한 투명성 의무를 강화함으로써 민주적 의사결정 과정의 왜곡을 방지하고자 한다.

셋째, 조작적 행동의 감소를 중요한 목표로 제시하고 있다. 이는 현재 확인된 조작 행위

---

hate speech online ("Code of Conduct+")

6) European Parliament Research Service(2023), The EU approach to tackling disinformation.

뿐만 아니라, 기술 발전과 사회적 환경 변화에 따라 새롭게 등장하는 다양한 조작적 행태를 포괄적으로 대상으로 삼는다. 이 과정에서 행동강령 서명 기관 간의 협력을 한층 강화하여, 조작적 정보 유통에 대한 공동 감시 및 대응 역량을 제고하도록 하고 있다.

넷째, 사용자 권한 강화를 주요 영역으로 설정하고 있다. 이를 위해 이용자가 허위조작정보를 식별하고 대응할 수 있도록 더욱 다양하고 우수한 사용자 역량 강화 도구를 제공하며, 동시에 신뢰할 수 있는 정보 및 맥락에 대한 접근성을 확대함으로써 이용자의 정보 판단 능력을 실질적으로 제고하는 것을 목표로 한다.

다섯째, EU 전역에 걸친 팩트체킹 체계 강화를 주요 영역으로 포함하고 있다. 이는 사실 확인 기관의 작업 결과가 플랫폼 전반에서 일관되게 활용될 수 있도록 하는 한편, 사실 확인 기관의 독립성과 지속 가능성을 확보하기 위해 공정한 재정적 기여가 이루어지도록 요구하는 내용을 포함한다.

여섯째, 연구를 위한 데이터 접근성 확대를 주요 영역으로 설정하고 있다. 플랫폼이 보유한 데이터에 대해 연구자가 보다 쉽게 접근할 수 있도록 접근성과 이용 편의성을 개선하고, 허위조작정보 연구 및 정책 평가를 위한 연구 활동을 적극적으로 지원하도록 규정하고 있다.

〔그림 2－1〕 허위조작정보 행동강령의 주요 영역

나. 목적 달성 및 향후 증빙을 위한 이행 체계

행동강령은 이러한 목적을 실질적으로 달성하고 향후 이행 성과를 검증하기 위해 구체적인 이행 및 관리 체계를 함께 제시하고 있다. 우선, 행동강령 이행 상황을 투명하게 공개하기 위한 투명성 센터(Transparency Center)의 개설을 요구하고 있다. 해당 센터는 각 서명 기관의 이행 개요를 체계적으로 정리하여 공개 접근이 가능하도록 운영된다.

또한, 상설 태스크포스(Permanent Task Force)의 설치를 통해 행동강령의 지속적인 발전과 조정을 도모하도록 하고 있다. 해당 태스크포스에는 모든 서명 기관을 비롯하여 ERGA(현재 EBMS), EDMO, EEAS, 그리고 유럽연합 집행위원회 의장이 참여하여 다자 간 협력 구조를 형성한다.

아울러, 강력한 모니터링 체계를 구축하여 정성적·정량적 정보를 종합적으로 보고하도록 하고 있으며, 해당 체계는 EU 차원뿐만 아니라 회원국 수준에서도 설치되어 다국어 환경을 지원하도록 설계되어 있다.

다. 서명 기관 구성

행동강령에는 다양한 유형의 주체가 서명 기관으로 참여하고 있다. 주요 온라인 플랫폼 및 검색 엔진으로는 Google(Google 광고, 검색 및 YouTube), Meta(Facebook, Instagram, Messenger 및 WhatsApp), LinkedIn, Microsoft Ads 및 Bing, TikTok, 그리고 무역 단체인 DOT Europe 등이 포함된다.

이와 함께 트위치(Twitch), 비메오(Vimeo), 세즈남, 더 브라이트 앱 등 소규모·전문 온라인 플랫폼, 광고 업계 단체 및 기업, 다수의 유럽 내 팩트체킹 기관, 시민사회 및 연구 기관, 그리고 허위조작정보 탐지·분석을 위한 기술 솔루션 제공 업체들이 폭넓게 참여하고 있다. 이러한 다층적 참여 구조는 허위조작정보 문제를 단일 주체가 아닌 생태계 차원에서 대응하고자 하는 EU의 정책적 접근을 반영한다.

라. 2022년 행동강령

EU 허위조작정보 행동강령은 2018년 주요 온라인 플랫폼이 자율적으로 서명한 Code of Practice로 최초 제정되었다. 이후 허위조작정보의 영향력이 확대되고 기존 자율규제의 한계가 지적됨에 따라, 2022년 6월 16일 개정·강화된 버전이 발표되었다.

강화된 행동강령은 EU 디지털서비스법(DSA) 체계 내에서 공식적인 행동강령(Code of Conduct)으로 인정받을 수 있도록 구조화되었으며, 허위조작정보 대응의 실효성을 제고하는 것을 핵심 목표로 하였다. 이를 위해 투명성 강화, 플랫폼 데이터 접근성 확대, 허위조작정보 광고 수익 차단, 팩트체킹 확산, 연구자 접근 보장 등의 요소가 종합적으로 반영되었다. 개정된 주요사항의 특징을 정리하면 다음과 같다.

1) 법적 연계성의 강화: 디지털서비스법과의 연동

2022년 개정된 EU 허위조작정보 행동강령은 기존의 자율규제적 성격을 넘어, EU 디지털서비스법과의 법적 연계성을 명확히 강화한 것이 가장 큰 특징이다. DSA 제35조에 따르면, 초대형 온라인 플랫폼(Very Large Online Platforms, VLOPs)은 허위조작정보로 인해 발생할 수 있는 시스템적 위험(systemic risks)을 식별하고 이를 완화하기 위한 조치를 이행할 의무를 부담한다.

개정된 행동강령은 이러한 DSA상의 위험 완화 의무를 구체적으로 이행하기 위한 공식적인 행동강령(Code of Conduct)으로 기능하도록 설계되었다. 이에 따라 본 행동강령은 단순한 자율적 권고 수준을 넘어, DSA가 요구하는 위험관리 프레임워크의 핵심 구성 요소로 작동하게 되었다. 특히 2023년 이후에는 허위조작정보 대응과 관련한 플랫폼의 조치가 자율규제의 영역에 머무르지 않고, 법적 감독과 평가의 대상이 되는 구조로 전환되었다는 점에서 제도적 의미가 크다.

2) 서명기관의 확대 및 다양화

개정된 행동강령은 서명 기관의 범위와 구성 면에서도 2018년 최초 제정 당시와 비교해 현저한 확장을 보인다. 초기에는 주로 대형 온라인 플랫폼 중심의 참여 구조를 가졌다면, 개정안에서는 플랫폼, 광고 산업, 팩트체킹 기관, 연구기관, 시민사회 단체를 포괄하는 다자 간 협력 체계로 발전하였다.

플랫폼 사업자 측면에서는 Google, Meta, TikTok, Microsoft, X(구 Twitter), Twitch, Vimeo 등 주요 글로벌 및 전문 플랫폼이 참여하고 있다.

광고 생태계에서는 TradeDesk, GroupM, Magnite 등 광고기술 및 미디어 대행 관련 기업들이 참여하여, 허위조작정보와 광고 수익 간의 연결고리를 차단하기 위한 책임을 공동으로 부담하고 있다.

또한 AFP, EU DisinfoLab, EDMO 등 팩트체킹 기관과 시민사회 단체의 참여를 통해, 허위조작정보 대응이 기술적·상업적 차원을 넘어 공공성과 민주주의 보호라는 관점에서 다층적으로 이루어지도록 설계되었다.

3) 핵심 약속 및 세부 조치의 체계화

개정된 행동강령은 총 43개의 핵심 약속(Commitments)과 128개의 세부 조치를 제시하며, 허위조작정보 대응을 위한 구체적 이행 수단을 명문화하고 있다.

우선, 광고 배치 재검토를 통한 수익창출 차단을 핵심 과제로 설정하고 있다. 이는 허위조작정보 확산에 기여하는 웹사이트나 계정으로 광고 수익이 유입되지 않도록 광고 공급망의 투명성을 확보하고, 광고 배치 기준을 재검토하는 조치를 포함한다.

다음으로, 정치 광고의 투명성 제고를 주요 약속으로 규정하고 있다. 정치 및 이슈 광고와 관련하여 광고주 정보, 타게팅 기준, 광고 노출 범위 등을 공개함으로써, 이용자가 정치적 메시지의 배경과 의도를 명확히 인식할 수 있도록 하는 것이 목적이다.

또한, 서비스 무결성 보장을 위해 허위조작정보 확산에 기여하는 조작적 행위를 줄이기 위한 조치를 강화하고 있다. 이는 봇, 가짜 계정, 인위적 증폭 행위 등 다양한 형태의 조작적 행동을 포괄적으로 관리하는 것을 의미한다.

이와 함께 이용자 역량 강화를 중요한 축으로 설정하여, 미디어 리터러시 도구 제공, 허위조작정보 식별을 위한 경고 라벨 도입 등 이용자가 스스로 정보를 판단할 수 있는 환경을 조성하도록 하고 있다.

팩트체킹 및 연구 협력 확대 역시 주요 약속에 포함된다. 각국 언어를 기반으로 한 팩트체킹 네트워크를 지원하고, 연구자가 허위조작정보의 확산 양상과 영향을 분석할 수 있도록 플랫폼 데이터 접근 API를 개선하는 조치가 이에 해당한다.

아울러, 투명성 센터 및 상설 태스크포스 운영을 통해 행동강령의 이행 현황을 시민 누구나 쉽게 확인할 수 있도록 하고, 기술적·사회적·시장적·입법적 환경 변화에 따라 약속사항을 지속적으로 검토·조정하는 구조를 마련하였다.

마지막으로, 보고 및 모니터링 체계의 강화가 중요한 특징으로 제시된다. EU 차원 및 회원국 차원에서 행동강령 이행 현황을 체계적으로 측정하기 위해 서비스 수준 지표(Service Level Indicators, SLIs)와 정성적 보고 요소(Qualitative Reporting Elements, QREs)를 포함

한 강력한 보고 체계를 도입하였다. 이에 따라 2023년 초 서명 기관들은 행동강령 이행에 관한 첫 기준 보고서를 제출하였으며, DSA에 정의된 초대형 온라인 플랫폼(VLOPs) 및 초대형 온라인 검색 엔진(VLOPSE)은 6개월마다, 그 외 서명 기관은 매년 정기적으로 보고할 의무를 부담하게 되었다.

〈표 2－1〉 허위조작정보 행동강령의 DSA로의 편입 과정

연도	주요 이행내용	비고
2022.6	강화된 행동강령 채택 (EU집행위, DisinfoCode.eu 공개)	34개 서명기관 참여
2023.2	첫 번째 기준(베이스라인) 보고서 제출 (플랫폼별 KPI 결과)	투명성 지표 도입
2023.8	DSA 발효(VLOP 지정) - Google, Meta, TikTok, X 등	행동강령을 DSA 제45조에 따른 의무 준수 수단으로 간주
2024.2	EU집행위, 투명성센터(Transparency Centre) 정식 운영	disinfocode.eu/transparency
2024~2025	Generative AI 확산 대응 조항 추가 논의 (딥페이크, AI정보 조작 등)	행동강령 수정안 발표('24.10)[7] DSA로의 공식 통합 승인('25.2) DSA-AI Act 연계 시행('25.7)

마. 행동강령의 구체적 내용

1) 서문: 규약의 성격과 기본 원칙

허위조작정보 행동강령은 2018년 제정된 기존 행동강령의 성과를 토대로, 유럽연합 집행위원회의 정책 지침에 따라 허위조작정보 대응의 실효성을 강화하기 위해 개정·보완된 규범이다. 해당 규약은 선언적 성격을 넘어, 허위조작정보 확산에 대한 구조적이고 지속적인 대응을 목적으로 하는 실천 규범으로서의 성격을 가진다.

서명기관들은 유럽 정보공간에서 확산되는 허위조작정보, 오정보, 정보조작, 외국의 정보 개입 등 다양한 정보 위협에 대응하기 위한 각자의 역할과 책임을 인식하고 있다. 유럽

7) DSA 발효에 맞추어 전문에 기술적 수정 및 최소한의 변경만 이루어졌으며 본문은 2022년 버전에서 크게 변경되지 않았음.

집행위원회는 대규모 허위조작정보 노출을 유럽 사회의 중대한 과제로 규정하며, 개방된 민주사회 유지를 위해 시민들이 정보에 기반하여 자유롭고 공정한 정치적 의사 형성을 할 수 있는 공론장이 필수적임을 강조하고 있다.

행동강령은 한편, 허위조작정보 대응 조치가 기본권을 전적으로 존중하는 범위 내에서 이루어져야 함을 분명히 한다. 서명기관들은 표현의 자유, 정보 접근권, 사생활 보호 등 기본권의 중요성을 인식하는 동시에, 합법적 콘텐츠의 확산을 보장하면서 허위조작정보의 영향력을 제한하기 위한 조치 간의 균형을 유지해야 한다. 모든 약속과 이행 조치는 EU 기본권 헌장에 따라 해석·적용되며, 기본권 간 충돌이 발생할 경우 비례성 원칙에 따라 조정되어야 한다. 또한, 서명기관들은 허위조작정보 확산이 온라인과 오프라인 전반에서 다양한 행위자에 의해 촉진된다는 점을 인식하고, 플랫폼·광고업계·팩트체커·시민사회·연구기관 등 생태계 전반의 공동 책임을 인정한다. 또한 각 기관은 서비스의 특성, 규모, 가용 자원에 따라 대응 방식과 이행 수준이 달라질 수 있음을 전제로 규약에 참여한다.

2) 약속 및 이행 관련 규정

행동강령은 서명기관의 약속이 형식적 선언에 그치지 않도록, 가입·변경·이행·보고에 관한 구체적인 절차를 규정하고 있다. 각 서명기관은 자사 및 자회사의 제품, 활동, 서비스와 관련된 약속과 조치에 가입하며, 가입 문서에 이행 대상 약속과 적용 범위를 명확히 기재해야 한다. 특정 약속이 자사 서비스와 관련성이 없다고 판단될 경우, 그 사유를 가입 문서에 명시하도록 하고 있다.

서명기관들은 규약의 목적 달성을 위해 약속과 조치의 관련성·적절성·실행 가능성을 정기적으로 검토하며, 필요 시 새로운 약속에 추가 가입하거나 더 이상 관련성이 없는 조치에서 탈퇴할 수 있다. 약속의 갱신이 필요한 경우, 개정 초안을 상설 태스크포스에 공유하고 논의를 거쳐 공식적으로 반영한다.

행동강령에 따른 모든 조치는 디지털서비스법(DSA)의 규제 목적과 원칙을 보완하는 것으로 이해되며, 특히 초대형 온라인 플랫폼(VLOPs) 및 초대형 온라인 검색엔진(VLOSEs)의 허위조작정보 관련 체계적 위험을 다루는 공동규제 체계와 조화를 이룬다. DSA에 따라 VLOPSE 제공자는 시스템적 위험을 평가·완화하기 위한 합리적이고 비례적인 조치를 취해야 하며, 본 규약의 관련 약속에 가입하는 행위는 DSA상 위험 완화 조치로 간주된다.

VLOPSE로 분류되지 않는 서비스 제공기관의 참여 역시 장려되며, 이 경우 서비스의 규모와 성격, 가용 자원에 비례한 수준에서 약속을 이행할 수 있다. 해당 기관들은 질적 보고요소(QRE)와 서비스 수준지표(SLI)를 통해 단계적 이행 계획을 제시할 수 있으며, 태스크포스는 이러한 조정된 조치가 규약의 목적에 부합하는지를 검토한다.

모든 서명기관은 규약 서명 후 6개월 이내에 약속 이행을 개시하고, 7개월 이내에 유럽집행위원회에 기준선(베이스라인) 보고서를 제출해야 한다. 이 보고서는 이후 보고가 점진적으로 구체화·정교화되는 것을 전제로 한 최초 기준 자료로 기능한다. VLOPSE로 지정된 서명기관은 규약이 DSA 제45조에 따른 행동강령으로 전환된 이후 감사를 받게 되며, 약속 이행 여부를 입증하기 위한 관련 기준과 자료를 감사기관에 제공해야 한다.

행동강령은 EU 및 회원국의 기존 법률을 대체하거나 우선하지 않으며, 자발적 약속과 법적 의무가 중복될 경우에는 DSA 및 기타 EU 법률이 항상 우선 적용된다. 서명기관은 법적 의무 이행 과정에서 규약이 요구하는 추가적 조치와 보고 요건을 함께 고려해야 한다. 동일한 원칙은 정치 광고 투명성 및 표적화 규정(TTPA)에도 적용되며, 해당 규정이 본 행동강령보다 우선적인 효력을 가진다.

행동강령은 유럽경제지역(EEA) 내에서 제공되는 서비스에 한해 적용되며, 서명기관은 태스크포스에 통보함으로써 규약에서 탈퇴할 수 있다. 다만, 탈퇴는 다른 서명기관의 규약 효력에는 영향을 미치지 않는다. 추가 서명기관은 언제든지 규약에 참여할 수 있으며, 이 경우 준수 의사가 있는 약속과 조치를 태스크포스에 제출해야 한다.

3) 주요영역① : 광고배치 재검토

광고배치 재검토 영역은 허위조작정보 유포자가 광고를 통해 경제적 이익을 획득하지 못하도록 차단하는 것을 핵심 목표로 한다. 본 영역은 허위조작정보 확산의 구조적 동인으로 작용해 온 광고 수익 모델을 직접적으로 겨냥함으로써, 허위조작정보 유통을 억제하기 위한 실질적이고 예방적인 대응 수단으로 기능한다. 이를 위해 서명기관들은 허위조작정보 콘텐츠 인접 영역에 광고가 게재되는 것을 방지하고, 허위조작정보를 포함하거나 허위조작정보 확산에 기여하는 광고의 유통을 차단하기 위한 강화된 조치를 공동으로 이행할 것을 약속하고 있다. 아울러 본 규약은 플랫폼, 광고업계, 광고 기술 기업 등 관련 주체 간 협력을 제도화함으로써 보다 강력한 공동 대응을 가능하게 한다.

서명기관들은 2021년 유럽집행위원회가 발표한 「허위조작정보 행동강령 강화 지침」(EU Guidance on Strengthening the Code of Practice on Disinformation)에서 제시된 우선순위를 재확인하고 있다. 해당 지침은 허위조작정보 유포자의 주요 수익원이 광고 배치라는 점을 명확히 지적하며, 광고 수익 차단이 허위조작정보 대응의 핵심 요소임을 강조하였다. 이에 따라 서명기관들은 허위조작정보 확산을 억제하기 위해 광고 배치에 대한 감시와 통제를 체계적으로 강화해야 한다는 공통된 인식을 형성하고, 이를 규약상의 구체적 약속과 이행 조치로 전환하였다.

〈표 2－2〉 광고 배치 재검토 주요 조치

목적	구분	핵심 내용	주요조치
허위조작정보 수익 차단	약속 1	허위조작정보 유포에 대한 광고·수익화를 차단하기 위해 콘텐츠 수익화 기준, 광고 게재 통제, 정책 집행 및 보고 체계를 전반적으로 강화	■ 허위조작정보 인접 광고 차단 및 반복 위반 출처 제재 ■ 콘텐츠 수익화·광고 수익 공유 심사 강화 ■ 광고 게재 위치에 대한 투명성 제공 ■ 독립 제3자 감사 및 DSA 감사 체계와 연계 ■ 브랜드 안전성 도구 및 출처 평가·팩트체킹 정보 활용 확대
허위조작정보 광고 대응	약속 2	광고 메시지 자체를 통해 허위조작정보가 유포되는 것을 방지하기 위해 광고 시스템 오용을 차단	■ 허위조작정보 광고를 규율하는 맞춤형 광고 정책 도입 ■ 허위조작정보 콘텐츠·출처 탐지 도구 및 파트너십 구축 ■ 광고 검증·검토 시스템 강화 ■ 위반 광고 삭제 시 광고주 통지 및 이의제기 절차 명확화
협력 강화	약속 3	광고 생태계 전반에서 허위조작정보 대응을 위한 공동 대응 체계 구축	■ 플랫폼·광고 공급망·팩트체커·출처 평가 기관 간 정보 공유 ■ 태스크포스·업계 포럼을 통한 허위조작정보 동향 및 TTP 교환 ■ 제3자 기관과 협력하여 허위조작정보 수익화 및 광고 확산 차단

4) 주요영역② : 정치 광고 투명성

정치 광고 투명성 영역은 정치 광고가 여론 형성과 민주적 의사결정에 미치는 영향을 고려하여, 서명기관들이 보다 강화된 투명성 조치를 마련하도록 하는 데 목적이 있다.

서명기관들은 이용자가 정치 광고를 쉽게 식별할 수 있도록 효율적인 라벨링 체계를 도입·적용해야 한다. 이러한 조치는 정치 광고와 일반 콘텐츠를 명확히 구분함으로써, 이용자가 정치적 정보의 성격과 출처를 보다 분명히 인식할 수 있도록 한다.

또한 서명기관들은 정치 광고에 대한 공개성과 책임성을 강화하기 위해, 효율적이고 검색 가능한 정치 광고 라이브러리를 구축할 것을 약속한다. 이를 통해 정치 광고의 집행 현황에 대한 사회적 감시와 검증이 가능해지며, 온라인 공론장의 투명성과 신뢰성이 제고된다.

〈표 2－3〉 정치 광고 투명성 주요 조치

목적	구분	핵심 내용	주요 조치
정치·이슈 광고에 대한 공통 기준 확립	약속 4	정치 및 이슈 광고에 대한 공통 정의를 채택하여 규제 및 이행의 기준을 통일	■ 정치 광고 정의를 규정 2024/900과 정합적으로 적용 ■ 규정상 합의 부재 시 태스크포스와 협력해 실무적 정의 마련
	약속 5	정치·이슈 광고에 대해 서비스 전반에 일관된 정책을 적용하고 허용·금지 범위를 명확화	■ 정치·이슈 광고 관련 정책·지침 공개 ■ 라벨링·투명성·검증 원칙을 관련 광고 전반에 적용
정치·이슈 광고의 명확한 식별	약속 6	사용자가 정치 또는 이슈 광고임을 명확히 인식할 수 있도록 표시 체계 강화	■ 공통 라벨·마크 모범사례 개발 및 서비스별 적용 ■ 라벨에 스폰서 신원 등 핵심 정보 포함 또는 접근 보장 ■ 라벨 인식·이해도 제고를 위한 연구 수행 및 반영 ■ 광고 공유 시에도 라벨 유지 ■ 메시징 서비스 내 정치 광고 라벨 가시성 제고 노력
정치·이슈 광고 검증 강화	약속 7	정치·이슈 광고주 및 대리 광고 서비스 제공자에 대한 비례적 신원 검증 체계 구축	■ 광고 게재 전 광고주·스폰서 신원 확인 및 재확인 ■ 검증 절차의 적시성·비례성 확보 및 관련 도구 보고 ■ 검증 회피 시 계정 정지 등 제재 조치

목적	구분	핵심 내용	주요 조치
			■ 정치 광고 미표시 광고에 대한 사용자 신고 기능 제공 ■ 광고주에게 정치·이슈 광고 해당 여부 선언 요구
정치·이슈 광고에 대한 사용자 투명성	약속 8	사용자가 접하는 정치·이슈 광고에 대한 기본 투명성 정보 제공	■ 후원자, 게재 기간, 지출, 수신자 집계 정보 등 최소 투명성 의무 합의 ■ 광고에서 광고 저장소로의 직접 링크 제공
	약속 9	사용자가 특정 정치·이슈 광고를 보게 된 이유를 명확히 설명	■ 광고 노출 사유를 쉽게 확인할 수 있는 수단 제공 ■ 타겟팅 기준 및 사용 도구를 평이한 언어로 설명
정치·이슈 광고 데이터 접근성 확보	약속 10	정치·이슈 광고 저장소의 최신성·완전성·품질 보장	■ 모든 정치·이슈 광고에 대한 검색 가능한 저장소 구축·운영 ■ 광고주, 지출, 노출, 타겟팅 기준, 지역·수신자 정보 제공 ■ 광고 데이터 최소 5년간 공개
	약속 11	사용자 및 연구자를 위한 광고 데이터 API 제공	■ 실시간에 가까운 맞춤형 검색 기능 제공 ■ 광고 저장소와 동등 이상 수준의 데이터 제공 ■ API 접근성·가용성 보장 ■ 연구자 수요에 따른 기능 개선
시민사회 및 감시 기능 강화	약속 12	정치·이슈 광고에 대한 사회적 감독과 정책 개선을 건설적으로 지원	■ 회원국 전반의 정치 광고 연구·모니터링·보고 ■ 선거 기간 감시를 위한 도구·대시보드 제공 ■ 정책·이행 과정의 문제를 타 서명 기관에 경고
지속적 협력 및 위험 대응	약속 13	정치·이슈 광고 관련 허위조작 정보 위험에 대한 지속적 모니터링과 공동 대응	■ 태스크포스를 통한 신규·진화 위험 식별 및 대응 논의 ■ 선거 기간 광고 금지 기간의 효과·영향 평가 ■ 회원국 내 독립적 감시 체계 충분성 정기 검토

5) 주요영역③ : 서비스 무결성 보장

서비스 무결성 보장 영역은 허위조작정보 확산을 억제하기 위해 온라인 서비스의 설계와 운영 전반에서 무결성을 확보하는 것을 목적으로, 개별 콘텐츠 대응을 넘어서 허위조작정보가 확산되는 구조적 환경 자체를 개선하는 데 중점을 둔다.

행동강령은 가짜 계정, 봇 기반 확산, 사칭, 악의적 딥페이크 등 허위조작정보 확산에 활용되는 조작적 행위를 주요 대응 대상으로 규정하고, 이러한 행위를 줄이기 위한 탐지 및 완화 조치를 강화하도록 요구한다. 서명기관들은 서비스 운영 과정에서 이러한 조작적 행위의 발생과 영향을 최소화하기 위한 조치를 이행할 것을 약속한다.

서명기관들은 악의적 행위자들이 사용하는 전술·기법·절차(TTPs)가 지속적으로 변화한다는 점을 인식하고, 관련 목록을 정기적으로 검토·갱신하며 확인된 행위와 관행을 포괄하는 명확한 정책을 시행하기로 하였다. 이 과정에서 서명기관 간 협력을 강화하여 정보 공유와 공동 대응의 효과성을 높이고자 한다.

〈표 2－4〉 서비스 무결성 보장 주요 조치

목적	구분	핵심 내용	주요 조치
허용되지 않는 조작적 행위에 대한 공통 기준 확립	약속 14	허위조작정보 확산에 활용되는 허용되지 않는 조작적 행위와 관행에 대해 서비스 간 공통된 이해를 확립하고, 이를 반영한 정책을 강화	■ 가짜 계정·봇 증폭, 사칭, 해킹·유출, 악의적 딥페이크, 가짜 참여 구매 등 주요 조작 행위 정의 ■ AMITT 프레임워크 등 최신 TTP 증거를 반영하여 정책 정기 검토
			■ 허용되지 않는 조작 행위에 대한 명확한 공개 정책 유지 및 투명성 센터 보고 ■ TTP 대응 조치의 효과성 지표 및 가짜·비진정 계정 침투율 측정 지표 개발 ■ 태스크포스를 통해 공통 TTP 용어·목록 합의 및 연례 갱신 ■ 대응 정책의 공통 기준 요소, 목표 및 벤치마크 마련

목적	구분	핵심 내용	주요 조치
인공지능 기반 조작 대응 및 투명성 확보	약속 15	AI 생성·조작 콘텐츠(딥페이크 등)에 대한 금지된 조작 관행 대응 및 투명성 확보	■ AI 생성·조작 콘텐츠에 대한 금지 행위 대응 정책 수립 또는 점검(사전 탐지, 사용자 경고 등) ■ 콘텐츠 탐지·제재에 사용되는 알고리즘의 신뢰성 확보 및 기본권 존중 ■ AI 시스템이 사용자 행동을 부당하게 왜곡하는 금지된 조작 관행을 구성하지 않도록 정책 마련
협력 및 정보 공유 강화	약속 16	플랫폼 간 협력을 통해 외국의 정보 개입 및 조작적 영향력 행사에 공동 대응	■ 정보 조작, 외부 간섭 및 서비스 내 관련 사건 정보 사전 공유 ■ 허위조작정보 주체의 플랫폼 간 이동(전술적 마이그레이션)에 대한 정보 공유 및 공동 대응

6) 주요영역④ : 이용자 역량 강화

이용자 역량 강화 영역은 허위 및 오인시키는 콘텐츠에 대응하기 위해, 이용자가 정보를 식별하고 대응할 수 있는 권한과 수단을 제공하는 것을 목적으로 한다. 이는 허위조작정보 대응을 이용자 참여 중심으로 보완하는 접근이다.

이를 위해, 신뢰할 수 있는 콘텐츠의 가시성과 검색 가능성을 개선하고, 허위조작정보를 식별·신고할 수 있는 도구를 제공하는 조치를 포함한다. 이를 통해 이용자는 오인시키는 콘텐츠를 보다 쉽게 인지하고 대응할 수 있다.

서명기관들은 콘텐츠 진위 검증 도구의 개발, 미디어 리터러시 강화 노력, 정보 추천 방식에 대한 이용자 선호 조정 기능 제공 등을 통해 이용자의 자율성과 통제권을 확대할 것을 약속한다. 이러한 조치는 이용자의 장기적 정보 판단 역량을 강화하고 온라인 정보 환경의 신뢰성을 높이는 데 기여한다.

〈표 2－5〉 이용자 역량 강화 주요 조치

목적	구분	핵심 내용	주요 조치
미디어 리터러시 및 비판적 사고 강화	약속 17	미디어 리터러시 및 비판적 사고 역량을 강화하여 허위조작정보에 대한 이용자의 인식·판단 능력을 제고	■ 콘텐츠 맥락 제공, 평가 가이드 등 사용자 역량 강화 도구 개발·운영 ■ EU 전역 인식 제고 캠페인 및 TTP 대응 교육 추진 ■ 취약 계층 참여 확대 ■ 미디어 리터러시 전문가·EDMO·ERGA 등과 협력
안전한 서비스 설계 및 책임 있는 추천	약속 18	서비스 설계와 운영 단계에서 허위조작정보의 확산 위험을 완화	■ 허위조작정보 확산을 억제하도록 설계된 추천 시스템 운영 ■ 사전 테스트 등 안전 설계 접근 적용 ■ 허위·오도 정보에 대한 비례적 정책(금지·순위 하락·추천 중단) 시행 및 반복 위반 제재 ■ 안전 설계 관련 연구 수행 및 결과 공유
	약속 19	추천 시스템의 작동 원리와 사용자 선택권에 대한 투명성 확보	■ 추천 시스템 주요 매개변수에 대한 명확한 설명 제공 ■ 이용자가 추천 옵션을 선택·수정할 수 있는 기능 제공
콘텐츠 출처·진위성 판단 지원	약속 20	이용자가 콘텐츠의 출처·편집 이력·진위성을 평가할 수 있도록 지원	■ 콘텐츠 출처 확인 및 진위 검증 기술 솔루션 개발 ■ C2PA 등 글로벌 출처 추적 표준 및 이니셔티브 참여·지원
허위조작정보 식별 및 경고 강화	약속 21	팩트체킹 및 경고 도구를 통해 허위조작정보 식별 역량을 강화	■ EU 전 회원국 언어로 팩트체크 라벨·경고·정보 패널 제공 ■ 팩트체커와 협력한 라벨·알림 기능 개발 ■ 경고·업데이트 효과성에 대한 연구 및 테스트 수행
정보 신뢰성 평가 지원	약속 22	사용자가 사회적 쟁점 정보의 신뢰성을 평가할 수 있도록 지원	■ 독립 제3자의 신뢰성 지표 및 신뢰 마크 접근 제공 ■ 추천 시스템에 신뢰성 신호 포함 여부 선택 옵션 제공 ■ 신뢰성 지표의 투명성·중립성·감사 가능성 보장 ■ 위기 상황 시 권위 있는 정보로 안내하는 기능 제공

목적	구분	핵심 내용	주요 조치
유해 허위조작정보 신고 기능 제공	약속 23	이용자가 허위·오도 정보를 신고할 수 있는 수단 제공	■ 모든 회원국 언어에서 사용자 친화적 신고 기능 제공 ■ 대량 신고 등 악용 방지 장치 마련
투명한 항소 절차 보장	약속 24	콘텐츠·계정 제재에 대한 이용자 권리 보호	■ 제재 사유·근거·이의제기 절차에 대한 명확한 통지 ■ 신속하고 공정한 항소 처리 및 정당한 경우 조치 철회
메신저 환경에서의 허위조작정보 대응	약속 25	사적 메시징 환경에서도 허위조작정보 식별을 지원	■ 암호화 수준을 유지하면서 신뢰 정보 접근 지원 기능 개발 ■ 정부·팩트체커·시민사회와 협력한 정보 제공 ■ 전달 제한, 사실 확인 라벨 등 확산 억제 기능 도입

7) 주요영역⑤ : 연구커뮤니티 역량 강화

연구커뮤니티 역량 강화 영역은 허위조작정보 대응 전략의 실효성을 높이기 위해, 연구자들이 플랫폼 데이터에 안정적으로 접근할 수 있는 환경을 구축하는 것을 목적으로 한다. 이는 허위조작정보 확산과 대응 효과를 증거 기반으로 분석하기 위한 핵심 요소이다.

행동강령은 비개인적·익명화·집계 처리되거나 공개된 데이터에 대해 자동화된 접근을 보장하는 방향을 제시하며, 개인정보 보호 원칙을 전제로 연구 효율성을 높이고자 한다. 또한 추가 검토가 필요한 데이터 접근에 대해서도 절차를 간소화하기 위한 거버넌스 구조 구축을 추진함으로써, 플랫폼과 연구자 간 협력을 제도적으로 강화하고자 한다.

〈표 2－6〉 연구 커뮤니티 역량 강화 주요 조치

목적	구분	핵심 내용	주요 조치
허위조작정보 연구를 위한 데이터 접근 확대	약속 26	허위조작정보 연구를 목적으로 비개인적·익명화·집계·공개 데이터를 자동화된 방식으로 안정적으로 제공	■ API 등 기술적 수단을 통한 실시간·준실시간 데이터 접근 제공 ■ 참여도·노출 등 허위조작정보 연구 관련 데이터 공개 ■ 남용 방지를 위한 안전장치(API 정책 등) 마련

목적	구분	핵심 내용	주요 조치
			■ 기계 판독 가능한 접근 제공 및 과도하지 않은 신청 절차 운영 ■ 접근 장애 발생 시 신속한 복구 절차 마련
검증된 연구자를 위한 데이터 접근 거버넌스 구축	약속 27	검증된 연구자에게 허위조작정보 연구에 필요한 데이터 접근을 제공하기 위한 독립적 제3자 기구 설립	■ 연구자·연구계획 심사를 담당하는 독립 기구 개발 및 협력 ■ 해당 기구에 대한 공동 재정 지원 ■ 기구 설립 후 정의된 프로토콜에 따라 데이터 공유 협력 ■ 기구 설립 전까지 허가된 연구자 대상 시범 데이터 공유 프로그램 운영
연구자와의 협력 및 연구 지원	약속 28	플랫폼 관련 허위조작정보에 대한 선의의 공익 연구를 적극 지원	■ 연구 지원을 위한 전담 인력 확보 및 연구자와의 상시 소통 체계 구축 ■ 제공 가능한 데이터 유형에 대한 투명한 공개 ■ 선의의 허위조작정보 연구를 방해하거나 제재하지 않을 것 명시 ■ EDMO 지원 하에 허위조작정보 연구 자금 마련 및 독립적 배분 절차 보장
연구 투명성 및 결과 공유	약속 29	허위조작정보 연구의 투명성과 재현성을 제고하고 연구 결과를 공유	■ 투명한 방법론과 윤리 기준에 따른 연구 수행 ■ 데이터셋·연구 결과·방법론을 태스크포스 및 대중과 공유 ■ 허위조작정보 대응 조치(라벨·경고 등)의 효과성 연구 수행 ■ 허위조작정보 포함 광고의 투명성 제고를 위한 광고 저장소 모델 개발

8) 주요영역⑥ : 팩트체킹 커뮤니티 역량 강화

팩트체킹 커뮤니티 역량 강화 영역은 허위조작정보 대응의 효과성을 높이기 위해, 플랫폼과 사실 확인 기관 간 협력 체계를 강화하는 것을 목적으로 한다. 이는 팩트체킹이 EU 전반의 정보 환경에서 일관되게 활용될 수 있도록 하기 위한 조치이다.

행동강령은 플랫폼들이 모든 EU 회원국 및 언어로 사실 확인의 범위를 확대하고, 서비스 전반에서 팩트체킹 결과를 보다 일관되게 활용할 것을 약속하도록 한다. 이를 통해 이용자는 정보의 사실적 정확성에 대한 판단을 보다 쉽게 할 수 있다.

본 섹션은 팩트체킹 커뮤니티가 플랫폼 데이터에 안정적으로 접근할 수 있는 체계를 구축하고, 협력을 확대함으로써 허위조작정보 퇴치에 있어 지속적이고 실질적인 역할을 수행할 수 있도록 지원한다.

〈표 2－7〉 팩트체킹 커뮤니티 역량 강화 주요 조치

목적	구분	핵심 내용	주요 조치
사실 확인 커뮤니티와의 지속가능한 협력 구축	약속 30	팩트체커에 대한 지원과 협력을 투명·개방·차별 없이, 재정적으로 지속 가능한 방식으로 제도화	■ 모든 회원국을 포괄하는 플랫폼-팩트체커 협정 체결 ■ 윤리·전문성 기준 충족 및 팩트체커 독립성 보장 ■ 공정한 재정적 기여 제공(개별·다자·대표기구 협정 방식) ■ 팩트체커 간 국경 간 협력 지원 ■ EDMO 및 유럽 팩트체킹 대표기구와 협의
플랫폼 서비스 내 팩트체킹의 일관된 활용	약속 31	모든 회원국·언어에서 플랫폼 서비스 전반에 팩트체킹 결과를 통합·전시·일관되게 활용	■ UGC 전반에 팩트체킹 결과 통합 및 게시 ■ 라벨·정보 패널·정책 집행 등 신속한 반영 메커니즘 운영 ■ EDMO와 협력한 팩트체킹 콘텐츠 공동 저장소 구축·운영 ■ 저장소 설립 자금 공동 부담 및 연례 재평가 ■ 플랫폼·언어 간 활용을 위한 기술적 해결책 개발
팩트체커의 정보 접근성 강화	약속 32	팩트체킹의 품질과 영향력 제고를 위해 관련 정보에 대한 신속·자동화된 접근 제공	■ 팩트체크된 콘텐츠의 노출·상호작용·조치 결과 제공 ■ 영향력 정보 접근을 위한 자동화된 인터페이스 제공 ■ 플랫폼의 팩트체크 활용·출처 표시·피드백 방식 일관성 확보 ■ 서명기관-팩트체커 간 정기적 정보 교환
사실 확인 기관의 윤리·독립성 보장	약속 33	팩트체킹 기관이 엄격한 윤리 기준과 투명성에 기반해 독립적으로 운영되도록 보장	■ IFCN 원칙 강령 또는 향후 EU 윤리 강령 준수 ■ 검증된 팩트체킹 기관 기준 충족 요구

9) 주요영역⑦ : 투명성 센터 설치

투명성 센터 설치 영역은 강령에 따른 약속과 조치가 실제로 어떻게 이행되고 있는지를 대중이 확인할 수 있도록, 정확하고 시의적절한 정보를 체계적으로 제공하는 것을 목적으로 한다.

〈표 2－8〉 투명성 센터 설치 주요 조치

구분	핵심 내용	주요 조치
약속 34	본 규약 이행의 투명성과 책임성을 확보하기 위해 공개 접근이 가능한 공동 투명성 센터 웹사이트를 구축·운영	■ 규약 서명 후 6개월 이내 공동 투명성 센터 구축 및 공개 ■ 웹사이트 구축·운영·유지 관리를 위한 적절한 재정 지원 제공 ■ 각 서명 기관의 서비스 범위에 따라 정보 제공에 기여 ■ 태스크포스 내에서 운영 및 자금 조달 방식 합의·기록·연례 검토 ■ 태스크포스 논의 결과에 따른 조정 사항을 합리적 기간 내 이행
약속 35	투명성 센터에 규약 이행과 관련된 모든 정보를 서비스별로 이해하기 쉽고 검색 가능하게 제공	■ 각 약속·조치별로 적용되는 서비스 약관 및 정책 공개 ■ 서비스별 정책 이행·집행 정보 및 지역·언어 적용 범위 제공 ■ 규약 이행 약속 평가 보고서 보관·공개 ■ 위기 상황 시 관련 완화 조치 정보 공개 ■ 사용자 친화적·최신 기술 기반 검색 기능 제공 ■ SLI, QRE, 구조적 지표(SI)를 표준화된 방식으로 제공(회원국별 세부 포함)
약속 36	투명성 센터 내 정보의 적시성·완전성 확보	■ 정책 및 이행 조치 변경 사항을 발표 또는 시행 후 50일 이내 반영 ■ 정기 보고 주기에 맞춰 SLI·QRE·구조적 지표 업데이트 ■ 상설 태스크포스의 최신 결정 사항을 반영하여 지속적 갱신

10) 주요영역⑧ : 상설 태스크포스 운영

이 영역은 기술적·사회적·시장적·입법적 환경 변화에 대응하여 허위조작정보 행동강령의 약속과 조치를 지속적으로 발전시키기 위한 거버넌스 체계를 규정한다. 태스크포스는 서명 기관 대표를 중심으로, 유럽 시청각 미디어 서비스 규제 기관 그룹(ERGA, 향후 유럽 미디어 서비스 위원회), 유럽 디지털 미디어 관측소(EDMO), 유럽 대외행동청(EEAS) 대표가 참여하는 다자 협력 구조로 구성되며, 유럽 집행위원회가 의장을 맡는다.

〈표 2－9〉 상설 태스크포스 운영 주요 조치

구분	핵심 내용	주요 조치
약속 37	행동강령의 이행·조정·발전을 위해 상설 태스크포스에 참여하고 공동 거버넌스를 운영	■ 서명 기관, EDMO, ERGA, EEAS 대표로 태스크포스 구성 ■ 유럽 집행위원회가 의장 역할 수행 ■ 필요 시 관련 전문가를 참관인으로 초청 ■ 의사 결정은 합의(consensus)에 의해 수행 ■ 모든 서명 기관은 태스크포스 업무에 참여·기여 ■ 소규모·신생 서비스는 자원·위험 프로필에 비례해 기여하거나 공동 대표 가능 ■ 최소 6개월마다 전체 회의 개최, 필요 시 사안별 소그룹 회의 운영 ■ 선거·위기 상황 대응을 위한 위험 평가 방법론 및 신속 대응 체계 마련 ■ 특별 상황에서의 공동 대응 및 조정 수행 ■ 통합 보고 양식, 보고 방법론 및 데이터 공개 방식에 합의 ■ 모니터링 지표(SLI·QRE·구조적 지표)의 품질·효과성 평가 및 조정 ■ 악의적 행위자의 TTP 목록 합의·공개·갱신 및 대응 기준 설정 ■ 안전 설계, 사후 표시, 팩트체킹 저장소, 출처 도구 관련 연구·증거 논의 ■ 팩트체커 협약 이행 관련 정량 정보 제공 기준 논의 ■ 기술·사회·시장·입법 변화 및 신규 서명 기관 수용에 따른 약속·조치 업데이트 검토 및 수행 ■ 소규모·신생 서비스에 적용된 조치의 적절성·일관성 검토 ■ 규약 이행 촉진 및 신규 서명 기관 통합

구분	핵심 내용	주요 조치
		■ 태스크포스 운영 규칙을 핸드북으로 정의하고 합의(제3자 전문가 참여 포함) ■ 특정 사안을 전담하는 소그룹 구성 가능 ■ 최소 연 1회 이상 이해관계자·전문가 회의 개최를 통해 운영 현황 공유 및 의견 수렴 ■ 약속·조치 변경 필요 시 태스크포스에 통보 후 논의 ■ 변경 사항을 가입 문서에 반영하고 차기 보고서에서 보고

11) 주요영역⑨ : 보고 및 모니터링 강화

이 섹션은 허위조작정보 행동강령의 이행 여부를 체계적으로 점검하고, 그 효과성을 지속적으로 평가하기 위한 강력한 모니터링 및 보고 체계를 구축하는 것을 목적으로 한다.

이를 위해 행동강령은 EU 전체 및 회원국 차원에서 규약 이행 현황을 정기적으로 측정할 수 있도록 서비스 수준 지표(Service Level Indicators, SLIs)와 정성적 보고 요소(Qualitative Reporting Elements, QREs)를 결합한 보고 체계를 도입한다.

서명기관들은 2023년 초 규약 이행에 관한 첫 기준선 보고서를 제출하였으며, 이후 디지털서비스법에 따라 초대형 온라인 플랫폼(VLOPs) 및 초대형 온라인 검색 엔진(VLOPSE)은 6개월마다, 그 외 서명기관은 매년 1회 정기 보고를 수행한다. 다만, 규약에 따른 보고 의무는 서명기관의 규모, 서비스의 성격 및 가용 자원을 고려하여 차등 적용되며, VLOPs 및 VLOPSE로 분류되지 않는 플랫폼 제공자에게 불균형적인 부담이 발생하지 않도록 설계되었다.

〈표 2－10〉 보고 및 모니터링 강화 주요 조치

구분	핵심 내용	주요 조치
약속 38	규약 이행을 위해 충분한 재정적·인적 자원을 투입하고, 이를 뒷받침할 내부 절차를 마련	(추가적인 조치 내용은 따로 기술되지 않음)
약속 39	규약 서명일로부터 6개월의 이행 기간 종료 후 1개월 이내에 기준선(베이스라인) 보고서를 유럽 집	(추가적인 조치 내용은 따로 기술되지 않음)

구분	핵심 내용	주요 조치
	행위원회에 제출	
약속 40	서비스 수준 지표(SLIs) 및 정성적 보고 요소(QREs)에 기반한 정기 보고 제공	■ VLOPSE 제공자: 서비스·회원국 수준에서 6개월마다 QRE·SLI 보고 ■ 기타 서명기관: 서비스·회원국 수준에서 연 1회 QRE·SLI 보고 ■ QRE·SLI를 투명성 센터에 정기 업데이트 ■ 태스크포스 합의에 따른 공통 보고 서식 개발 ■ 유럽위원회·ERGA·EDMO 피드백을 반영해 보고·모니터링 체계 개선 ■ 유럽위원회의 합리적 요청에 따라 추가 정보·데이터 제공
약속 41	규약의 효과성을 평가하기 위한 구조적 지표를 공동 개발하고, 9개월 이내 첫 지표 세트와 초기 측정 결과를 공개	■ 서명 후 1개월 이내 구조적 지표 워킹 그룹 구성 ■ 데이터 포인트 공유 및 측정 방법론 제안 ■ 분기별 태스크포스 진행 보고 및 전문가 협의 ■ 6개월 이내 실행 가능한 지표 제안 제출 ■ 9개월 이내 구조적 지표 공개 및 데이터 제공 ■ 첫 종합 보고서와 함께 측정 결과·방법론 공개 ■ 측정 결과를 보고 주기에 맞춰 갱신하고 투명성 센터에 게시
약속 42	선거·위기 등 특별한 상황에서 유럽위원회 요청 시 신속 대응 체계에 따라 특별 보고 제공	(추가적인 조치 내용은 따로 기술되지 않음)
약속 43	태스크포스에서 합의된 통합 보고 양식 및 개선된 보고·데이터 공개 방법론에 따라 보고 수행	(추가적인 조치 내용은 따로 기술되지 않음)

3. 혐오 표현 행동강령

가. 혐오 표현 행동강령의 배경 및 목적

EU 온라인 불법 혐오 표현 대응 행동강령(Code of Conduct on Countering Illegal Hate Speech Online)은 온라인 환경에서 급속히 확산되는 불법 혐오 발언에 체계적으로 대응하기 위해 마련된 자발적 규범으로, 최초 채택 시기는 2016년 5월 31일이다. 이 행동강령은 초기 단계에서 EU 집행위원회와 주요 온라인 플랫폼 사업자가 공동으로 참여하여 마련되었으며, 당시 구글(유튜브), 페이스북, X(구 트위터), 마이크로소프트가 주요 주체로 참여하

였다.

행동강령은 이후 지속적인 확장을 거쳐 2018년부터 2022년 사이 다수의 온라인 서비스 제공자가 추가로 참여하였다. 이러한 참여 기업들은 행동강령 문서상 "서명 기관"(Signatories)으로 지칭되며, Dailymotion, Facebook, Instagram, Jeuxvideo.com, LinkedIn, Microsoft에서 호스팅하는 소비자 서비스, Snapchat, Rakuten Viber, TikTok, Twitch 등이 추가 서명 기관으로 포함되었다.

행동강령은 불법 혐오 발언 콘텐츠에 대한 신속한 검토 및 제거를 촉진하는 데 실질적으로 기여해 왔으며, 플랫폼 사업자, 시민사회단체, EU 회원국 당국 간의 신뢰와 협력을 증진하는 역할을 수행해 왔다. 특히, 서명 기관들은 EU 기본권 헌장에 명시된 표현의 자유, 인간 존엄성, 차별 금지 등 기본권을 존중해야 할 공동의 책임을 공유하고, 불법적 온라인 혐오 발언에 대응하려는 EU 집행위원회와 회원국의 정책적 노력을 지지해 왔다.

나. 디지털서비스법과의 관계

온라인 불법 혐오 표현 대응 행동강령은 2023년 개정을 통해 '행동강령+' 형태로 확대·강화되었으며, 디지털서비스법(DSA) 제45조에 근거한 자발적 준수 메커니즘으로 발전하였다. 이 과정에서 서명 기관들은 온라인 불법 혐오 발언 퇴치의 효율성을 제고하기 위해 기존 행동강령을 DSA의 규율 체계와 정합적으로 강화할 필요성이 있음을 인정하였다.

행동강령+는 DSA 제45조에 따른 자발적 행동강령으로 기능하며, DSA의 준수 및 효과적인 집행을 지원하는 프레임워크를 구축하는 것을 주요 목적으로 한다. 이를 위해 최근 온라인 환경에서 나타나는 새로운 도전과 위협에 대응하기 위한 추가적 조치들이 포함되었다.

행동강령은 DSA에 따라 서명 당사자에게 직접 부과되는 법적 의무에 영향을 미치지 않으며, 법적 효력 측면에서는 DSA가 명확한 우선권을 가진다. 행동강령에 따른 약속 사항이 DSA상의 법적 의무와 중첩되는 경우에도 행동강령은 법적 의무를 대체하거나 완화하는 수단이 아니라, 특히, 투명성 조항을 중심으로 의무 이행에 관한 구체적이고 비례적인 정보를 수집·제공하는 보조적 수단으로 작동하도록 설계되었다.

DSA 제35조 제1항 (h)에 따라, 대규모 온라인 플랫폼(VLOPs) 또는 대규모 온라인 검색엔진(VLOSEs)이 본 자발적 행동강령의 약속을 성실히 준수하는 경우 이는 해당 사업자가

취한 적절한 위험 완화 조치로 간주될 수 있다. 이에 따라 행동강령+는 법적 강제력은 없으나, DSA 위험 관리 체계 내에서 실질적인 규제 의미를 갖는다.

서명 기관들은 각 서비스가 직면한 위험 수준, 서비스의 규모와 성격, 혐오 발언 발생 가능성, 가용 자원 등을 종합적으로 고려하여 합리적이고 비례적이며 효과적인 방식으로 행동강령 상의 약속을 이행하도록 요구된다. 이러한 이행 방식은 일률적 기준이 아니라 위험 기반 접근에 따라 차등적으로 적용된다. 유럽 집행위원회는 서명 기관들의 이러한 자발적 노력을 환영하며, 향후 행동강령+의 이행 과정에 대해 지속적인 지원과 모니터링을 수행할 것임을 명확히 하고 있다.

다. 주요 약속사항

1) 이용 약관 관련 약속

온라인 불법 혐오 표현 대응 행동강령의 서명 기관들은 불법적 혐오 발언에 효과적으로 대응하기 위한 이용 약관을 마련할 것을 약속한다. 이를 위해 서명 기관들은 자사 서비스 내에서 불법적 혐오 발언이 금지된다는 점을 이용자에게 명확히 고지해야 하며, 해당 금지 사항이 단순한 선언에 그치지 않도록 구체적인 정책 내용과 위반 시 적용되는 조치에 대한 명확한 정보를 이용 약관에 포함해야 한다.

또한 서명 기관들은 디지털서비스법(DSA) 제14조 제2항에 따라, 불법적 혐오 발언에 대한 대응 방식과 관련하여 이용 약관에 중대한 변경 사항이 발생하는 경우 이를 이용자에게 사전에 또는 적시에 공지할 책임을 부담한다. 이는 이용자가 플랫폼의 규칙과 집행 기준을 예측 가능하게 이해할 수 있도록 보장하기 위한 조치로서, 플랫폼 운영의 투명성과 책임성을 강화하는 기능을 수행한다.

2) 콘텐츠 검토 및 삭제 또는 접근 차단

서명 기관들은 불법적 혐오 발언 콘텐츠를 효과적으로 검토하고, 필요 시 이를 삭제하거나 접근을 차단할 수 있는 기능적·조직적 역량을 확보할 것을 약속한다. 이를 위해 서명 기관들은 디지털서비스법 제16조 및 제22조 제1항에 따라, EU 내 모든 사용자뿐만 아니라 신뢰할 수 있는 신고자(trusted flaggers)를 포함한 주체가 불법적 혐오 발언 콘텐츠를 신고할 수 있는 신고 및 조치 메커니즘을 구축해야 한다.

유효한 통지를 접수한 경우 서명 기관들은 해당 신고를 적시에, 성실하게, 임의적이지

않으며 객관적인 방식으로 검토해야 한다. 검토 결과 해당 콘텐츠가 서명 기관의 정책을 위반하거나 관련 법률 및 DSA 제16조에 따라 불법적 혐오 발언으로 신고된 경우에는 지체 없이 해당 콘텐츠를 삭제하거나 접근을 차단하는 조치를 취해야 한다.

서명 기관들은 모니터링 신고자(monitoring reporters)로부터 접수된 통지의 대다수, 즉 최소 50%를 24시간 이내에 검토할 것을 약속하며, 가능할 경우 이 비율을 초과하여 통지의 3분의 2 이상, 즉 최소 67%를 24시간 이내에 검토하는 것을 목표로 한다. 이와 같은 목표 달성을 위해 서명 기관들은 지정된 신고 절차 등 적합한 내부 프로세스를 자율적으로 설계·운영할 수 있다.

서명 기관들은 신고 조치에 소요되는 시간이 콘텐츠의 양, 심각성, 확산 가능성, 법적·사실적 복잡성 등 상황적 요인에 따라 달라질 수 있음을 인정하며, 이러한 요소를 고려한 합리적 판단이 필요함을 명시한다.

3) 투명성, 책임성 및 모니터링 관련 약속

서명 기관들은 앞서 제시된 통지 검토 약속의 이행 여부를 체계적으로 평가하기로 합의하다. 해당 평가는 행동강령 부속서 1에 명시된 방법론에 따라 수행되며, 단발성 점검이 아니라 시간 경과에 따른 추세를 지속적으로 모니터링하는 방식으로 이루어진다.

서명 기관들은 모니터링 활동의 결과와 함께, 불법 혐오 발언 대응과 관련된 추가 정보를 제공할 것을 약속한다. 이와 같은 추가 정보의 범위와 내용은 행동강령 부속서 2에 구체적으로 규정되어 있으며, 이를 통해 행동강령 이행에 대한 외부 검증 가능성과 책임성을 제고하는 것을 목표로 한다.

4) 협력 관련 약속

서명 기관들은 불법적 혐오 발언의 확산 위험에 효과적으로 대응하고 이를 사전에 예방하기 위해, 업계 내부 협력과 다중 이해관계자 협력의 중요성을 강조하다. 이를 위해 서명 기관들은 모니터링 리포터와의 기존 파트너십을 지속적으로 이행하고 강화할 것을 약속한다.

서명 기관들은 모범 사례, 전문성, 기술적 도구의 효과적인 공유와 확산을 촉진하기 위해 정기적인 교류 포럼에 참여하기로 한다. 해당 포럼은 유럽 집행위원회 주관 하에 개최되며, 서명 기관과 모니터링 리포터 간의 연례 회의를 포함한다. 이 과정에서 서명 기관의 의견이 반영되며, 서명 기관 간 모범 사례 교환은 연례 회의의 별도 세션으로 구성된다.

더불어 서명 기관과 모니터링 리포터 간의 체계적인 커뮤니케이션을 통해 불법 혐오 발언 관련 동향과 최신 발전 상황에 대한 정보 교환이 이루어진다.

서명 기관들은 공유 온라인 지식 허브 구축을 지원한다. 해당 지식 허브는 유럽 집행위원회가 주관하며, 개인정보 보호와 정보 보안을 위한 적절한 안전장치를 전제로 운영된다. 이 지식 허브의 주요 목적은 불법 혐오 발언으로 주장되는 콘텐츠의 합법성 평가에 참고할 수 있는 판례 및 법률 자료를 수집하고, 과거 보고서 분석에 활용 가능한 자료를 축적하는 데 있다. 서명 기관과 모니터링 담당자는 신기술이 제기하는 도전과 기회에 관한 정보를 포함한 모범 사례 자료를 통해 이 지식 허브에 기여할 수 있다.

플랫폼에서 불법적 혐오 발언 유포가 현저히 증가할 실질적 위험이 발생하는 경우, 서명 기관과 유럽 집행위원회는 교환 포럼의 틀 아래에서 공동 합의에 따라 회의를 개최할 수 있다. 이러한 회의는 유럽 집행위원회와 관련 서명 기관만이 공동으로 개시할 수 있으며, 서명 기관들의 참여 범위는 불법적 혐오 발언 유포의 임박성 및 중대성에 따라 달라질 수 있다. 해당 논의에 모니터링 리포터를 참여시킬지 여부는 서명 기관들과 유럽 집행위원회가 공동으로 결정한다.

서명 기관들이 이미 디지털서비스법 제36조에 따라 유럽 집행위원회와 공식적인 회의 절차를 진행 중인 경우에는, 앞서 언급한 별도의 회의는 개최되지 않는다.

서명 기관들은 개인정보 보호 법률을 준수하고 보안 및 기본권에 대한 위협을 고려하는 범위 내에서, 크로스 플랫폼 환경에서의 불법적 혐오 발언 확산을 방지하기 위해 관련 팀 간에 운영 정보와 동향을 교환할 것을 약속한다.

5) 인식 제고 및 대안적 서사 구축 노력

서명 기관들은 혐오적 발언과 편견에 대한 사회적 인식 제고의 중요성을 인정하다. 이에 따라 서명 기관들은 온라인 안전과 디지털 예절을 증진하기 위해, 대안적 서사 구축, 새로운 아이디어, 다양한 이니셔티브를 지원하기 위한 도구와 접근법을 지속적으로 제공할 것을 약속한다.

온라인 예절과 비판적 사고를 장려하는 교육 프로그램을 지원하며, 이러한 노력은 유럽 집행위원회가 자금을 지원하고 촉진하는 관련 이니셔티브를 보완하는 역할을 수행한다. 대표적인 예로는 현재 진행 중인 '시민, 평등, 권리 및 가치 프로그램(Citizens, Equality,

Rights and Values Programme)'이 있다.

서명 기관들은 모니터링 리포터와 관련 시민사회단체를 대상으로, 교육 및 인식 제고 캠페인을 효과적으로 전달하기 위한 모범 사례, 활용 방법, 자원 및 지원 정보 제공에 최선을 다할 것을 약속한다.

〈표 2－11〉 행동강령+의 핵심 이행 약속 요약

- 불법적 혐오 표현을 금지하는 이용 약관을 마련하고, 관련 정책 및 위반 시 조치를 명확히 고지함
- 불법적 혐오 표현 대응과 관련한 이용 약관의 중대한 변경 사항을 이용자에게 사전 또는 적시에 공지함
- EU 내 모든 사용자 및 신뢰할 수 있는 신고자가 불법적 혐오 표현을 신고할 수 있는 신고·조치 메커니즘을 구축함
- 유효한 통지에 대해 적시적·객관적·비임의적으로 검토하고, 위반 콘텐츠는 신속히 삭제 또는 접근 차단함
- 모니터링 신고자로부터 접수된 통지의 최소 50%를 24시간 이내 검토하고, 67% 달성을 목표로 노력함
- 통지 검토 약속의 이행 여부를 정기적으로 평가하고, 표준화된 방법론에 따라 결과를 모니터링함
- 모니터링 결과 및 추가 정보를 제공하여 불법 혐오 표현 대응의 투명성과 책임성을 제고함
- 불법적 혐오 표현 확산 예방을 위해 모니터링 리포터 및 시민사회와의 협력을 지속·강화함
- 정기적 교류 포럼, 연례 회의, 모범 사례 공유 등을 통해 업계 내 협력 체계를 구축함
- 실질적 위험 발생 시 유럽 집행위원회와 협의하여 공동 대응 논의를 진행함
- 혐오 표현과 편견에 대한 인식 제고를 위해 대안적 서사 및 온라인 예절 증진 이니셔티브를 지원함
- 미디어 리터러시 및 비판적 사고 강화를 위한 교육·캠페인을 지원하고, 시민사회 대상 교육·자원 제공에 협력함

라. 부속서 1 : 모니터링 활동 및 방법론

1) 소개 및 정의

부속서 1은 모니터링 활동을 규정하여 절차의 명확성, 디지털서비스법과의 일관성, 강화된 투명성을 보장한다. 서명 기관의 모니터링 활동 결과는 개별적으로가 아니라 종합적으로 검토되며, 이 과정에서 서명 기관의 콘텐츠 검토 정책, 사전 예방적 노력, 집행 조치에 기반한 불법적 혐오 발언 대응 조치에 관한 추가 정보가 함께 고려된다. 이러한 추가 정보는 부속서 2에 제시된 지침 질문에 따라 제공된다.

모니터링 활동은 행동강령상 서약 사항의 이행 현황을 평가하기 위한 절차로서, 모니터링 기간의 시작과 함께 개시되어 최종 보고서 발간 시점에 종료된다. 모니터링 기간은 연간 최대 6주, 즉 30영업일로 설정되며, 이 기간 동안 모니터링 리포터가 불법 혐오 발언 콘텐츠를 신고한다.

신뢰할 수 있는 신고자(Trusted Flaggers)는 DSA 제22조 제2항에 따라 불법 콘텐츠 신고에 관한 전문성과 역량을 인정받은 기관을 의미하며, 관련 정보는 DSA 제22조 제5항에 따른 공개 데이터베이스에 게시된다.

모니터링 리포터는 EU 회원국 내에서 불법 혐오 발언에 관한 전문성을 보유한 비영리 기관 또는 공공 기관이어야 하며, 유럽 집행위원회와 서명 기관의 승인을 받아 모니터링 활동에 참여한다. 모니터링 리포터는 필요에 따라 신뢰할 수 있는 신고자일 수 있다. 모니터링 리포터 목록에 새로운 기관을 추가하는 절차는 기밀로 유지되며, 위원회는 각 모니터링 기간 시작 최소 10주 전에 제안된 목록을 서명 기관에 공유한다. 서명 기관은 이후 2주 동안 구두 또는 서면으로 의견을 제출할 수 있으며, 해당 기간 내 의견이나 이의 제기가 없는 경우 제안된 목록은 승인된 것으로 간주된다. 의견 불일치가 발생하는 경우에는 1주일 이내에 타협점을 도출해야 하며, 타협이 이루어지지 않을 경우 서명 당사자가 최종 결정권을 가지되 배제 사유를 제시해야 한다.

제3자 조정자는 유럽 집행위원회가 선정한 기관으로서 모니터링 과정에서 행정적 역할을 수행한다. 모든 제3자 조정자는 위원회의 통보 이후 서명 당사자들의 서면 승인을 받아야 하며, 2주 이내에 반응이 없는 경우 승인된 것으로 간주된다. 제3자 조정자 참여로 발생하는 비용은 위원회가 부담하며, 조정자의 참여 시 업무 범위는 위원회와 서명 기관 간 합의를 통해 정해진다.

분쟁 사례는 신고된 콘텐츠의 불법성 여부 또는 삭제·접근 차단 조치의 적정성에 대해 서명 기관과 모니터링 리포터 간 의견 차이가 발생하는 경우를 의미한다. 오류 사례는 동일 콘텐츠의 중복 신고, 잘못된 계정이나 페이지 신고, URL 또는 스크린샷 오류, 서명 기관이 통지를 수령했다는 증거가 없는 경우 또는 DSA 제16조나 제22조에 따라 통지가 유효하지 않은 경우를 포함한다.

2) 모니터링 수행방법

모니터링 활동은 표준화되고 주기적인 방식으로 매년 실시된다. 모든 서명 기관은 예외 없이 모니터링 활동의 적용을 수락해야 한다. 다만 신규로 행동강령에 가입하는 서명 기관의 경우, 한 차례에 한해 시범 감사를 받을 수 있으며, 해당 기간의 결과는 시험 목적에 한하여 활용된다. 시범적 모니터링 결과의 공개 여부는 신규 서명 기관의 재량에 따른다.

3) 모니터링 리포터의 역할과 책임

모니터링 리포터는 모니터링 결과의 기록을 담당하며, 신고하는 각 콘텐츠에 대해 고유한 URL과 스크린샷을 제공해야 한다. 또한 불법 혐오 발언으로 주장되는 콘텐츠의 특정 부분, 예를 들어 동영상의 특정 구간, 게시물의 문구 또는 해시태그 등을 명확히 식별해야 한다.

오류 사례는 최종 모니터링 결과에 포함되나 불법 혐오 발언으로는 간주되지 않는다. 모니터링 리포터가 단일 모니터링 기간 동안 특정 서명 기관에 제출한 통지 중 오류 사례의 비율이 33%를 초과하는 경우, 위원회는 해당 상황에 대해 조치를 취할 수 있다. 이 경우 위원회는 해당 리포터의 설명을 청취하고 서명 기관과 협의하며, 합리적인 설명과 실질적인 개선 약속이 제시되지 않을 경우 향후 모니터링 활동 참여가 중단될 수 있다.

4) 불법 혐오 발언 신고 절차

모니터링 리포터는 모니터링 기간 동안 서명 기관의 서비스에서 발견된 불법 혐오 발언을 지정된 신고 절차를 통해 신고한다. 이는 DSA에 따른 신고 절차가 실질적으로 검증되고, 모니터링 결과의 무결성을 확보하기 위함이다.

본 규정은 서명 기관에게 모니터링 리포터 전용의 지정 신고 절차를 제공할 법적 의무를 부과하지는 않는다. 지정된 절차가 존재하지 않는 경우 모니터링 리포터는 일반 사용

자에게 제공되는 신고 절차를 이용해야 한다. 지정된 신고 절차는 서명 기관이 위원회에 통보하며, 위원회는 해당 정보를 모니터링 리포터에게 전달한다.

모니터링 리포터는 지정된 신고 절차 외 또는 지정 절차 이용이 불가능한 경우 일반 신고 절차를 통해 신고할 수 있으며, 이러한 사례는 모니터링 활동 보고서에서 별도의 범주로 표시된다. 다만 모니터링 리포터는 지정된 신고 절차의 사용을 적극적으로 권장받으며, 신고 절차에 관한 정보는 사전 교육을 통해 제공된다.

5) 분쟁 사례 처리

모니터링 수행 과정에서 모니터링 리포터와 서명 기관 간 분쟁 사례가 발생하는 경우, 서명 기관은 분쟁 해결을 위한 최선의 방안을 결정한다. 이 과정에는 추가적인 법적 검토 또는 모니터링 리포터와의 개별 접촉이 포함될 수 있다. 모니터링 리포터는 해당 절차에 적극 협조하고 신속히 응답해야 한다.

위원회 또는 제3자 조정자는 개별 콘텐츠의 잠재적 불법성 판단에 직접 관여하지 않는다. 본 규약의 기준에 따라 불법으로 간주되지 않는 콘텐츠는 모니터링 활동 대상에서 제외된다.

이견이 지속되는 경우 해당 분쟁 사례는 최종 보고서에서 별도의 범주로 표시되어 불법 혐오 발언 혐의와 혼동되지 않도록 한다. 분쟁 사례는 예정된 연례 회의에서 추가로 논의될 수 있다. 모니터링 리포터가 두 차례의 모니터링 기간 동안 전체 제출 사례 중 70% 이상의 미해결 분쟁 사례를 제출하는 경우 위원회는 해당 리포터의 의견을 청취하고 협의한 후 향후 모니터링 활동 참여를 중단할 수 있다.

6) 데이터 통합 및 보고서 공개

모니터링 활동에 앞서 유럽 집행위원회는 서명 기관의 의견을 반영하여 모니터링 리포터를 대상으로 통지 작성 및 데이터 수집에 관한 교육을 제공한다. 모니터링 기간 종료 후 서명 기관들은 위원회가 조정하는 데이터 통합 기간을 거치며, 이 기간 동안 잠정 결과의 분석·검증과 분쟁 사례 논의가 이루어진다.

특정 모니터링 기간 동안 단일 서명 기관에서 오류 사례와 분쟁 사례의 합계가 전체 통지 건수의 75%를 초과하는 경우, 해당 결과는 신뢰성이 낮은 것으로 간주되어 최종 결과 발표에서 제외될 수 있다.

데이터 통합이 완료되면 위원회는 모니터링 기간 종료 후 3개월 이내에 최종 결과를 발표하며, 공개 최소 2주 전에 서명 기관들과 사전 사본을 공유한다. 최종 결과 보고서에는 전체 통지 건수, 참여 모니터링 리포터 수, 평균 검토 및 삭제 비율, 검토 소요 시간, 분쟁 사례 비율 등 핵심 지표가 포함된다.

서명 기관과 위원회는 행동강령 이행 여부를 평가함에 있어 모니터링 결과를 단일 수치가 아닌 과거 성과, 표본 규모, 부속서 2의 정성 정보, 협력 수준 등을 종합적으로 고려한다. 또한 서명 기관들은 필요 시 다수결에 따라 본 부속서의 방법론을 검토할 수 있으며, 해당 검토는 디지털서비스법 제45조를 침해하지 않는 범위에서 위원회와 공동으로 수행된다.

〈표 2－12〉 온라인 혐오 표현 모니터링 활동 및 보고 절차 요약

• 모니터링 활동은 매년 표준화된 방식으로 실시 • 모니터링 리포터는 최대 6주(30영업일) 동안 불법 혐오 발언 콘텐츠를 신고함 • 서명 기관은 접수된 통지를 검토하고 삭제 또는 접근 차단 조치를 수행 • 모니터링 종료 후 위원회 주관 하에 데이터 통합 및 결과 검증이 이루어짐 • 오류 및 분쟁 사례가 과도한 경우 결과 공개에서 배제될 수 있음 • 최종 결과 보고서는 종료 후 3개월 이내 공개되며, 핵심 성과 지표가 포함됨

마. 부속서 2 : 핵심 성과 지표 및 추가 정보

부속서 2는 온라인 불법 혐오 발언 대응 행동강령의 서명 기관들이 콘텐츠 관리 정책의 일환으로 취한 조치에 대해, 추가적이며 의미 있고 이해하기 쉬운 정보를 제공하도록 규정한다.

서명 기관들은 불법적 혐오 발언의 해결을 위해 수행한 조치와 그 효과를 투명하게 설명할 책임을 지는데, 제공해야 할 정보에는 콘텐츠 관리 담당자에 대한 교육 및 지원 조치, 정보의 가용성·가시성·접근성에 영향을 미치는 운영 조치, 불법 혐오 발언과 관련하여 감지된 주요 추세 등이 포함될 수 있다. 이러한 정보는 각 모니터링 활동 결과와 함께 요약 문서 형태로 제출되어야 하며, 요약 문서는 서명 기관이 작성하되 부속서2에 명시된 체계적인 기준에 따라 구성된다.

제공 정보는 불법 혐오 발언 콘텐츠 및 또는 서명 기관의 자체 정책을 위반한 콘텐츠와

관련하여 제출될 수 있으며, 서명 기관은 제공 정보의 범위를 명확히 명시해야 한다. 서명 기관은 최대 4페이지 이내의 서면 진술서를 통해, 본 부속서 항목 중 해당 플랫폼과 직접적으로 관련된 사항만을 선택하여 정보를 제공한다. 이 과정에서는 플랫폼 간 구조적 차이가 충분히 고려되어야 하며, 관련 보고 기간은 직전 연도 또는 직전 6개월로 설정된다. 디지털서비스법 제15조, 제24조 및 제42조에 따라 사용된 가장 최근의 보고 기간과 일치시킬 것이 권고된다.

1) 불법적 혐오 발언 대응 조건 및 규정

서명 기관은 불법적 혐오 발언을 다루는 이용 약관에 대한 링크를 제공해야 한다. 또한 지난 모니터링 이후 이용 약관에서 불법적 혐오 발언의 정의 또는 대응 방식과 관련하여 중대한 변경 사항이 있었는지 여부를 설명해야 하며, 해당 변경 사항이 사용자에게 어떤 방식으로 전달되었는지를 함께 제시해야 한다.

2) 혐오 발언 콘텐츠 검토 및 삭제 또는 접근 차단 가능성

서명 기관은 혐오 발언 콘텐츠 신고를 허용하는지 여부와 그 방법에 관한 정보를 제공해야 하며, 이는 언어별 및 회원국별로 명확히 구분되어 제시되어야 한다. 또한 사용자가 혐오 발언 신고를 제출할 수 있는 경로에 대한 정보도 제공해야 하며, 해당 정보는 접근 경로(모바일 애플리케이션, 데스크톱 애플리케이션, 브라우저)와 신고 근거(디지털서비스법 제16조 또는 이용 약관)에 따라 구분된다.

서명 기관은 신고 근거별로 구분된 혐오 발언 신고의 검토 절차에 관한 정보를 제공해야 하며, 플랫폼 내에서 혐오 발언을 탐지하고 검토하기 위해 취하는 다양한 조치에 대한 추가 세부사항을 제시해야 한다. 여기에는 선제적 탐지 조치, 콘텐츠 노출 범위, 검토 절차, 이의 제기 처리 방식 등에 관한 정보가 포함될 수 있다.

가능한 범위 내에서 서명 기관은 자동화 도구를 통해 탐지된 혐오 발언 콘텐츠의 비율, 자동화와 인간 검토 간의 역할 분담, 이의 제기된 콘텐츠 중 조치 또는 복원된 비율, 신고 처리 시간(24시간, 48시간, 72시간 기준) 등에 관한 정보를 제공할 수 있다. 또한 추천 시스템의 역할이나 조회수 발생 이전에 삭제된 불법 혐오 발언 콘텐츠의 비율 등 추가 지표가 존재하는 경우 이를 함께 제시할 수 있다.

서명 기관은 불법적 혐오 발언 정책을 위반하는 콘텐츠를 플랫폼 차원에서 어떻게 관리

하는지에 대한 추가 설명을 제공해야 하며, 이에는 콘텐츠 관리자 교육, 사용된 기술, 지원 언어 등에 관한 정보가 포함될 수 있다. 더불어 신뢰 및 안전 관련 조직의 구성과 운영 방식, 해당 기간 중 도입된 신기술에 관한 정보도 해당되는 경우 제시될 수 있다.

3) 투명성, 책임성 및 모니터링

서명 기관은 본 부속서 2와 관련된 모니터링 실행 결과와 그에 따른 대응 사항을 참조하여, 투명성 및 책임성 확보를 위한 조치 이행 현황을 설명한다.

4) 업계 내 및 다중 이해관계자 간 협력

서명 기관은 불법적 혐오 발언 관련 정책과 관행을 수립·운영하는 과정에서 파트너와 협력한 내용에 관한 정보를 제공해야 한다. 이에는 파트너십의 성격, 취해진 조치, 운영 규모 및 대상 국가, 기존 파트너십의 변화 가능성 등이 포함될 수 있다.

서명 기관은 정기 교류 포럼 참여 여부 및 기여 내용에 관한 정보를 제공해야 하며, 여기에는 연례 회의 참여, 업계 내 모범 사례 교류, 체계적 협의 채널 운영, 온라인 지식 허브 참여 여부 등이 포함된다. 아울러 DSA 제36조에 정의된 위기 상황과 관련한 논의에 참여했는지 여부도 해당되는 경우 확인되어야 한다.

5) 인식 제고, 온라인 예의 및 반론 대응 이니셔티브

서명 기관은 미디어 리터러시 및 비판적 사고 능력 향상을 위한 추가 조치에 관한 정보를 제공해야 한다. 이에는 온라인 예절과 안전을 증진하기 위한 인식 제고 캠페인 지원, 또는 플랫폼 내에서 불법적 및 비불법적 혐오 발언 콘텐츠 생성을 억제하기 위한 조치에 대한 설명이 포함될 수 있다.

서명 기관은 모니터링 리포터 및 관련 시민사회단체가 활용할 수 있는 교육, 자원 및 지원에 관한 정보를 제공해야 하며, 여기에는 참가자 수, 교육 평가 결과, 교육 이후의 후속 조치 등에 관한 내용이 포함될 수 있다.

6) 향후 계획 명시

서명 기관은 한 해 동안 자사 플랫폼에서 불법적 혐오 발언 확산과 관련해 나타난 주요 동향을 중심으로 정보를 제공한다. 이 과정에서 인공지능 생성 콘텐츠 등 새로운 기술 환경 변화가 불법적 혐오 발언 확산에 미친 영향 등을 포함한다. 또한 이러한 동향에 대응하

여 자사의 정책과 절차를 어떻게 조정하였는지에 관한 정보를 제시한다. 여기에는 콘텐츠 관리 정책의 개정, 검토 절차의 개선, 탐지 및 대응 방식의 조정 등이 포함될 수 있다.

서명 기관은 행동강령 참여를 통해 얻은 추가적인 교훈 가운데 향후 정책 수립 및 운영에 있어 중요하다고 판단되는 사항을 중심으로 정보를 제공한다. 이러한 내용은 향후 불법적 혐오 발언 대응 역량을 강화하고, 규약 이행의 실효성을 제고하기 위한 참고 자료로 활용된다.

제2절 영 국

1. 개요

가. 제정 배경 및 규제 철학

영국의 온라인 안전법(Online Safety Act 2023)은 디지털 플랫폼 환경에서 확산되는 불법 콘텐츠와 아동 및 취약 이용자에게 유해한 콘텐츠로부터 이용자를 보호하기 위해 제정되었다. 기존의 자율규제 중심 접근이 대형 플랫폼의 영향력 확대, 알고리즘 기반 추천 시스템, 익명성 강화 환경에서 실효성을 확보하지 못한다는 문제의식이 법 제정의 직접적 배경으로 작용하였다. 특히, 아동 성착취·학대 콘텐츠, 테러 선동, 온라인 사기, 자해·자살 유도 콘텐츠 등의 확산은 사회적 피해를 구조적으로 확대시키는 요인으로 지적되었으며, 이에 따라 플랫폼 사업자에게 보다 강한 법적 책임을 부과해야 한다는 공감대가 형성되었다.

영국 정부는 이러한 문제의식을 바탕으로 개별 콘텐츠를 사후 규제하는 방식이 아니라 플랫폼의 설계·운영·위험관리 전반에 대한 구조적 책임을 법률로 규정하는 접근을 채택하였다.[8] 온라인 안전법은 2023년 10월 26일 국왕 승인(Royal Assent)을 받아 공식 발효되었다.

온라인 안전법의 핵심 규제 철학은 '사후 삭제 중심 규제'가 아닌 '사전적 위험관리' (risk-based regulation)에 있다. 즉, 플랫폼이 스스로 서비스 내 위험을 식별·평가하고, 그 결과에 따라 적절한 안전조치를 설계·운영하도록 의무를 부과한다. 법률에 세부 기술 요건을 직접 규정하기보다는 규제기관에 가이드라인 제정 권한을 위임하는 구조를 선택한다. 이를 위해 온라인 안전법은 규제기관인 Ofcom에 광범위한 가이드라인 제정 및 집행 권한을 부여하고, 세부적인 이행 방식은 행동강령(Code of Practice)과 지침(Guidance)을 통해 단계적으로 구체화하도록 설계하였다.

이와 같은 구조는 기술 변화 속도를 고려하여 경직된 규정보다 유연하면서도 강제력이 있는 규제 모델로 평가된다.[9]

---

8) UK Government(2023), Online Safety Act 2023 Explanatory Notes; DCMS.

9) Ofcom(2024), How Ofcom will regulate Online Safety.

나. 법률의 구조 및 규제 대상

온라인 안전법은 12개의 부(Parts) 총 241개조로 구성된 방대한 법안이다. 주요 목차 구성은 다음 표와 같다.

〈표 2－13〉 영국 온라인 안전법 주요 구성

구분	제목	주요내용
Part 1	Introduction	법률의 목적 기술 및 온라인 안전에 관한 기본 원칙 제시
Part 2	Key Definitions	'이용자 간 서비스(U2U)', '검색 서비스', '콘텐츠' 등 법적 적용 대상 정의
Part 3	Duties of Care	'이용자 간 서비스(U2U)', '검색 서비스'에 부과되는 주의의무
Part 4	Other Duties	'이용자 간 서비스(U2U)', '검색 서비스'에 부과되는 기타 의무
Part 5	Duties of Providers of Regulated Services : Certain Pornographic Contents	규제대상 서비스 공급자의 의무 (특정 음란물 콘텐츠)
Part 6	Duties of Providers of Regulated Services : Fees	규제대상 서비스 공급자의 의무 (수수료)
Part 7	Ofcom's Powers and Duties	규제기관인 Ofcom의 책임, 권한 등
Part 8	Appeals and Super-complaints	심판청구 및 최고불만사항
Part 9	Secretary of State's Functions	규제대상서비스 관련 주무장관의 기능
Part 10	Communications Offences	거짓 및 위협적 커뮤니케이션, 자해 조장, 보복성 음란물 공유 등 특정 온라인 행위에 대한 형사 처벌
Part 11	Supplementary and General	보충규정 및 일반규정
Part 12	Interpretation and Final Provisions	해석 및 종결조항

법 적용 대상은 제2조에 따라 크게 세 가지 유형으로 분류된다.

ⅰ) 이용자 간 서비스(User-to-User Services) : 이용자가 생성한 콘텐츠를 다른 이용자가 보거나 공유할 수 있도록 허용하는 서비스

- 예시: 페이스북, 인스타그램, 틱톡과 같은 SNS, 카카오톡·왓츠앱과 같은 메시징 앱, 온라인 게임 커뮤니티, 포럼 등.
- 핵심: 이용자가 올린 텍스트, 이미지, 영상이 타인에게 전달되는 모든 디지털 공간이 포함됨

ii) 검색 서비스(Search Services) : 이용자가 검색어를 입력하면 전 세계 웹사이트의 정보를 색인화하여 결과를 보여주는 서비스

- 예시: 구글(Google), 빙(Bing) 등.
- 특징: 개별 콘텐츠를 직접 호스팅하지 않더라도 검색 결과를 통해 유해 정보에 접근할 수 있으므로 규제 대상에 포함됨

iii) 결합 서비스(Combined Services)

위의 두 가지 기능이 한 플랫폼 내에 공존하는 경우

※ 제외 대상 (Exemptions): 위험도가 낮다고 판단되는 서비스는 규제에서 제외된다(예: 이메일 전용 서비스, 사내 인트라넷, SMS/MMS, 제한적 기능을 가진 리뷰 사이트 등).

법 적용 여부는 단순히 서비스의 본사가 영국에 있는지 여부가 아니라, 영국 이용자와의 실질적 연계성(link to the UK)을 기준으로 판단한다. 영국 이용자를 대상으로 서비스를 제공하거나, 영국 시장을 명시적으로 타게팅하는 경우 법 적용 대상이 된다. 이에 따라 글로벌 플랫폼 사업자 역시 영국 이용자를 보유한 경우 규제 대상에서 제외될 수 없다.[10] 그러나 모든 플랫폼에 과도한 규제 비용이 발생하는 것을 방지하기 위해, 이용자 수와 서비스 기능을 기준으로 카테고리를 분류하여 차등적 의무를 부과한다.

① Category 1: 고위험·대규모 이용자 간 서비스

가장 엄격한 의무가 부여되는 그룹으로, 시장 지배력이 크고 알고리즘에 의한 콘텐츠 추천 기능이 활발한 플랫폼이 해당된다.

10) Ofcom, Regulated Services Guidance, 2024

Category 1
- 기준: 수백만 명 이상의 활성 이용자 수, 추천 알고리즘 보유, 익명성 허용 수준 등(상세 수치는 Ofcom의 2차 입법으로 결정).
- 추가 의무: 사용자 권한 강화(User Empowerment)-성인 이용자가 원치 않는 유해 콘텐츠(법적으로는 허용되나 유해한 경우)를 필터링할 수 있는 도구 제공.
- 투명성 보고: 안전 조치 이행 현황에 대한 정기적인 공시.
- 사기 광고 방지: 유료 광고를 통한 사기 행위 차단 의무.

② Category 2A: 대규모 검색 서비스

이용자 규모가 큰 검색 엔진이 여기에 속하며 높은 월간 활성 이용자 수(MAU)를 보유한 검색 서비스로 검색 결과 내 불법 콘텐츠 노출 위험 평가 및 방지 조치를 강화할 의무가 부과된다.

③ Category 2B: 특정 위험이 있는 이용자 간 서비스

이용자 수는 Category 1보다 적지만, 특정 기능(예: 아동 성착취물 유통 위험이 높은 기능)때문에 집중 관리가 필요한 서비스이다. 따라서 정기적인 투명성 보고서 제출 의무 등이 부과된다.

〈표 2－14〉 카테고리별 의무 사항 요약

의무 사항	모든 서비스	Category 1	Category 2A	Category 2B
불법 콘텐츠 제거 의무	O	O	O	O
아동 보호(유해물 차단)	O*	O	O	O
위험 평가 보고서 작성	O	O	O	O
사용자 필터링 도구 제공	X	O	X	X
사기 광고 방지 의무	X	O	O	X
정기 투명성 보고서 발행	X	O	O	O

* 아동이 접근이 가능한 서비스에 한함

다. 공통적인 의무 사항

온라인 안전법은 서비스 제공자가 자사 플랫폼의 설계 및 운영 전반에서 이용자의 안전을 보장해야 한다는 '주의 의무'를 명문화하였다. 이는 단순히 유해 콘텐츠를 사후에 삭제하는 '통지 후 삭제' (Notice and Takedown) 방식에서 한 걸음 나아가, 서비스 기획 단계

부터 잠재적 위험을 식별하고 이를 원천적으로 방지하거나 최소화할 수 있는 시스템을 구축할 것을 요구한다.

1) 위험 평가 의무

사업자는 자사 서비스의 성격과 기능이 이용자에게 어떠한 위험을 초래할 수 있는지 사전에 평가해야 한다. 서비스 내에서 발생할 수 있는 테러, 아동 성착취, 사기, 마약 거래 등 각종 불법 행위의 발생 가능성과 그 영향력을 평가하며, 평가 항목에는 서비스의 이용자 규모, 추천 알고리즘의 작동 방식, 익명성 보장 수준 등이 포함된다.

아동이 접근할 가능성이 있는 서비스(Likely to be accessed by children)의 경우 아동에게 유해한 콘텐츠(자해 조장, 섭식 장애 권장 등)가 아동에게 노출될 위험을 별도로 평가해야 한다. 이는 아동의 연령대별 발달 단계를 고려한 세분화된 평가를 전제로 한다.

위험 평가는 일회성에 그치지 않고 새로운 기능이 추가되거나 중대한 서비스 변경이 있을 때, 혹은 규제기관인 Ofcom이 위험 수준의 변화를 고지할 때마다 즉시 갱신해야 한다. 또한 평가 결과는 문서화하여 Ofcom의 요청 시 제출할 수 있도록 보관해야 한다.

2) 안전 유지 의무

위험 평가 결과를 바탕으로 서비스 제공자는 이용자를 보호하기 위한 구체적이고 비례적인 조치를 시행해야 한다.

서비스 제공자는 플랫폼에서 불법 콘텐츠가 전파되는 것을 방지하기 위해 '비례적이고 합리적인' (Proportionate and Reasonable) 조치를 취해야 하는데, 이는 기술적 필터링, 신고 시스템 강화, 중재 인력 배치 등을 포함하며, 불법 콘텐츠를 발견했을 때 지체 없이 삭제할 수 있는 절차를 갖추는 것을 의미한다.

아동 이용자를 위해 서비스 제공자는 엄격한 연령 확인(Age Verification) 또는 연령 보증(Age Assurance) 기술을 도입해야 한다. 아동에게 유해한 콘텐츠가 알고리즘에 의해 추천되지 않도록 서비스 구조를 설계해야 하며, 유해 콘텐츠로부터 아동을 보호할 수 있는 최고의 안전 설정을 기본값(Default)으로 제공해야 한다.

모든 서비스 제공자는 이용자가 유해 콘텐츠를 쉽게 신고하고, 플랫폼의 결정에 대해 이의를 제기할 수 있는 투명한 불만 처리 절차를 운영해야 한다. 특히 부당하게 콘텐츠가 삭제된 경우 이용자의 표현의 자유를 보호하기 위한 재심사 구조를 갖추어야 한다.

3) 권리 보호와의 균형(Duty to have regard to rights)

안전 조치를 취하는 과정에서 서비스 제공자가 반드시 고려해야 할 '상충하는 권리'를 명시한다. 플랫폼은 안전 의무를 이행하면서도 이용자의 표현의 자유와 개인정보 보호를 침해하지 않도록 주의해야 한다. 이는 과도한 필터링이나 검열이 발생하지 않도록 조치의 수위를 조절해야 함을 의미한다.

라. Ofcom의 집행 권한 및 처벌 규정

1) Ofcom의 조사 및 정보 획득 권한

집행의 첫 단계는 서비스 제공자의 의무 이행 여부를 조사하는 것이다. Ofcom은 법적 의무 준수 여부를 확인하기 위해 다음과 같은 권한을 행사한다.

① 정보 요구권(Information Notices): Ofcom은 서비스 제공자에게 위험 평가 보고서, 알고리즘 설계 방식, 이용자 수 등 규제 준수와 관련된 상세 정보를 제출하도록 요구할 수 있다.

② 조사 및 현장 점검: 필요한 경우 Ofcom은 서비스 제공자의 사업장을 방문하여 관련 서류를 검토하거나 담당자를 면담하는 등의 현장 조사를 수행할 수 있는 권한을 가진다.

③ 숙련된 전문가 임명: 복잡한 기술적 사안의 경우, Ofcom은 외부 전문가를 임명하여 플랫폼의 시스템 및 알고리즘을 감정하고 보고하도록 명령할 수 있다.

2) 단계별 행정 처분 체계

의무 위반이 확인될 경우, Ofcom은 위반의 심각성에 따라 단계별 조치를 취한다.

① 시정 권고 및 경고: 경미한 위반의 경우 공식적인 경고를 통해 자발적인 시정을 유도한다.

② 집행 통지(Enforcement Notices): 위반 사항이 명확할 경우 Ofcom은 특정 기간 내에 위반을 시정하거나 재발 방지 대책을 마련하도록 명시적인 지시를 내린다.

③ 벌금 부과 통지: 집행 통지를 이행하지 않거나 중대한 위반이 발생한 경우 금전적 제재 절차에 착수한다.

3) 강력한 금전적 제재(Financial Penalties)

위반 기업에 대해 최대 1,800만 파운드(약 300억 원) 또는 해당 기업의 전 세계 연간 매출액(Global Annual Turnover)의 10% 중 더 높은 금액을 벌금으로 부과할 수 있다.

벌금 액수를 산정할 때 Ofcom은 위반의 지속 기간, 피해의 심각성, 기업의 시정 노력, 과거 위반 전력 등을 종합적으로 고려한다.

4) 서비스 제한 및 접속 차단(Business Disruption)

금전적 제재만으로 시정이 불가능하거나 극도로 심각한 위험이 지속될 경우, Ofcom은 법원의 명령을 통해 강력한 '비즈니스 중단 조치'를 취할 수 있다.

영국 내 이용자가 해당 서비스의 전부 또는 일부 기능을 이용하지 못하도록 제한하는 서비스 제한 명령(Service Restriction Orders)이나, 인터넷 서비스 제공자(ISP)나 앱스토어 사업자에게 해당 플랫폼으로의 접속을 차단하거나 앱 다운로드를 중지하도록 명령하는 '최후의 수단'으로서 접속 차단 명령(Access Blocking Orders)을 내릴수 있다.

5) 경영진의 형사 책임(Criminal Liability)

이 법은 기업뿐만 아니라 결정권을 가진 개인에 대해서도 책임을 묻는다. 이는 플랫폼 기업들이 규제 당국의 조사를 가볍게 여기지 못하도록 하는 강력한 장치이다.

① 정보 제공 방해: Ofcom의 정보 요구에 고의로 허위 정보를 제공하거나 증거를 인멸한 경우, 해당 업무를 담당하는 고위 경영진(Senior Managers)은 형사 처벌을 받을 수 있다.

② 집행 불응: 아동 성착취물 제거 등 치명적인 안전 의무를 지속적으로 위반하고 Ofcom의 명령을 무시한 경우, 법인은 물론 경영진 개인에게도 징역형이나 벌금형이 부과될 수 있다.

마. 온라인 안전법의 영향 및 의의

온라인 안전법은 플랫폼 기업의 기술·조직·거버넌스 전반에 중대한 변화를 요구한다. 특히, 위험평가와 투명성 보고는 단발성 대응이 아니라 지속적인 내부 관리 체계를 필요로 하기 때문에 이에 따라 플랫폼 내 안전 책임자 지정, 내부 감사 체계 구축, 외부 법률·기술 자문 활용이 일반화되고 있다.[11]

온라인 안전법은 아동 보호와 이용자 안전을 강화하였다는 평가를 받으면서도, 한편으로는 표현의 자유, 프라이버시, 암호화 통신 보호와의 긴장 관계를 지속적으로 야기하고 있다. 그럼에도 불구하고 이 법은 플랫폼 책임을 구조적으로 규정한 선도적 입법 사례로서, 향후 EU, 미국, 한국의 플랫폼 규제 논의에 중요한 참고 사례로 작용하고 있다.[12]

2. 투명성 가이드라인

가. 투명성 가이드라인 제정 배경

온라인 안전법은 온라인 환경에서 확산되는 불법유해 콘텐츠에 대한 플랫폼의 책임을 강화하는 동시에 규제의 실효성과 공공의 감시 가능성을 확보하기 위해 투명성(transparency)을 핵심 규제 수단으로 채택하고 있다. 이 법은 플랫폼 사업자에게 직접적인 안전 의무를 부과하는 데 그치지 않고, 그러한 의무 이행 과정과 결과를 외부에서 검증할 수 있도록 투명성 보고(transparency reporting) 체계를 제도화하였다.

온라인 안전법에서 투명성 의무는 제77조와 제78조를 중심으로 규정되어 있다. 제77조는 규제 대상 서비스 제공자에게 투명성 보고서를 제출하도록 요구하며, 보고 대상에는 불법 콘텐츠 및 유해 콘텐츠에 대한 조치 현황, 위험 평가 결과, 콘텐츠 조정 및 삭제 활동, 내부 정책과 절차 등이 포함될 수 있다. 제78조는 규제기관인 Ofcom에 대해 투명성 보고의 구체적 범위, 형식, 절차를 정하는 투명성 가이드라인을 제정할 법적 권한과 의무를 부여한다. 이 조항에 따라 Ofcom은 플랫폼이 어떤 정보를 어떤 수준으로 공개해야 하는지, 그리고 해당 정보가 규제 및 공공 감시 목적에 어떻게 활용될 수 있는지를 명확히 제시하도록 요구받는다.

이러한 법적 근거에 따라 Ofcom은 온라인 안전법 시행을 지원하기 위한 투명성 보고 가이드라인(Transparency Reporting Guidance)을 마련하였다.[13] 가이드라인은 투명성 보고 의무가 적용되는 서비스 유형을 구분하고, 각 서비스 범주별로 요구되는 데이터 항목과 보고 주기, 보고 형식을 세부적으로 규정하고 있다. 특히, 보고 항목에는 불법 콘텐츠

11) Ofcom(2024), Compliance and Enforcement Overview.
12) OECD(2024), Platform Regulation and Online Safety.
13) Ofcom(2025), Final Transparency Reporting Guidance

신고 및 처리 건수, 자동화 도구와 인적 검토의 활용 현황, 아동 보호 조치, 계정 정지·콘텐츠 삭제 등 집행 결과가 포함되며, 이는 서비스 규모와 위험도에 따라 차등 적용될 수 있도록 설계되었다. 이를 통해 일률적 보고 부담을 피하면서도 규제 목적에 부합하는 정보 수집을 가능하게 한다.

Ofcom은 플랫폼으로부터 제출받은 투명성 보고서를 종합·분석하여, 자체적인 규제 투명성 보고서를 작성하여 공개할 수 있다. 이는 단순한 행정적 보고를 넘어 온라인 안전 규제의 효과성과 한계를 사회적으로 점검하는 수단으로 기능한다. Ofcom은 이러한 투명성 체계가 플랫폼의 책임성을 제고하는 동시에, 향후 규제 개선 및 추가 가이드라인 마련을 위한 증거 기반(evidence-based regulation)을 제공할 것으로 평가하고 있다.[14]

나. 가이드라인의 목적 및 적용 대상

투명성 보고 가이드라인은 온라인 안전법 제77조 및 제78조에 근거하여 마련된 하위 규범으로, 플랫폼 사업자의 안전 의무 이행 현황을 체계적으로 공개·검증하기 위한 구체적 기준을 제시한다. 가이드라인의 핵심 목적은 단순한 정보 공개를 넘어 플랫폼이 수행하는 위험 식별·완화·집행 조치의 실효성을 외부에서 평가 가능하도록 만드는 것에 있다.

가이드라인은 투명성 보고 의무를 모든 규제 대상 서비스에 일률적으로 적용하지 않고, 서비스의 규모, 기능, 위험 수준에 따라 차등화한다. 특히, 대규모 이용자 기반을 보유하고 사회적 영향력이 큰 서비스에 대해서는 보다 상세하고 정기적인 보고를 요구하는 반면, 중소 규모 또는 위험도가 낮은 서비스에는 비례성 원칙(proportionality)에 따라 간소화된 보고 의무를 부과한다.

〈표 2－15〉 서비스 유형별 보고 의무 사항 차등화 원칙

서비스 유형	정의 및 분류 기준	보고의무 차등화 원칙
Category 1	높은 수준의 이용자 수 및 특정 기능을 보유한 이용자 간 서비스(U2U)	최고 위험군: 규모와 영향력이 가장 크므로 가장 광범위한 투명성 보고 의무 부과

14) Ofcom. (2024), How Ofcom will regulate Online Safety.

서비스 유형	정의 및 분류 기준	보고의무 차등화 원칙
Category 2A	일정 기준을 충족하는 검색 서비스(Search Services) 또는 결합 서비스	검색 특화: 검색 엔진 특성에 맞는 검색 결과 관련 투명성 의무에 집중
Category 2B	특정 위험 요소가 있거나 일정 규모 이상의 이용자 간 서비스(U2U)	중점 관리: Cat 1보다는 소규모이나 특정 위험 요소가 있어 차별화된 보고 요구

〈표 2－16〉 서비스 유형별 투명성 보고 항목

구분	Category 1 &2B (이용자 간 서비스)	Category 2A (검색 서비스)
요구 데이터 항목 (주요 내용)	이용자 간 정보 (User-to-user info) • 불법/아동 유해 콘텐츠 발생 및 유포 현황 • 콘텐츠 신고 및 조치(삭제 등) 시스템 운영 • 알고리즘 설계 및 운영 방식(추천/제한 등) • 이용자 신원 확인 조치 이행 현황 • 위험 평가 시스템 및 미디어 리터러시 조치	검색 엔진 정보 (Search engine info) • illegal search content 발생 및 노출 현황 • 검색 결과 관련 정책 및 절차 수립 현황 • 유해 콘텐츠 식별 및 위험 최소화 시스템 • 검색 알고리즘의 설계 및 운영 방식 • 아동 보호를 위한 높은 수준의 안전 조치
보고서 수준	상세 및 심층 보고: 불법 콘텐츠의 전파 속도, 노출도 등 기술적 지표 포함 가능	핵심 지표 위주: 검색 서비스의 특성을 반영한 위험 최소화 조치 보고
보고 주기	연 1회 (Annual): Ofcom이 매년 발부하는 고지에 따라 보고서 작성	동일
보고 형식 규정	독립된 보고서(Standalone Report) 형태 • 타 규제 보고서나 자율 보고서와 분리하여 제출 • Ofcom이 고지(Notice)에서 지정한 형식 및 템플릿 준수 • Ofcom이 지정한 기한 내에 제출 및 일반 대중에게 공개	동일 (단, 보고 내용은 검색 서비스 기능에 맞게 변형가능)

Ofcom은 비례성 원칙을 실현하기 위해, 각 기업에 투명성 고지(Notice)를 보낼 때 다음의 요소들을 개별적으로 검토하여 보고 범위를 조절한다.

〈표 2－17〉 보고 의무 차등화 고려요소

분류 요소	구체적인 검토 내용	비례성 적용 방식
서비스 종류 및 기능	서비스의 유형(SNS, 메시징 등)과 개별 기능(라이브 스트리밍 등)	해당 서비스 유형에서 발생하기 쉬운 특정 위험에 한정하여 보고 요구
이용자 규모	영국 내 활성 이용자 수 및 증가 속도	이용자가 많을수록 알고리즘 및 안전 시스템에 대한 상세 보고 요구
사업자 역량	사업자의 재무적 자원 및 가용한 기술적 전문성	중소 규모 사업자에게 지나치게 과도한(Unduly onerous) 보고 부담 방지
아동 이용자 비율	서비스 내 아동 이용자의 비중 및 아동의 이용 행태	아동 비율이 높을수록 아동 보호 조치에 대한 심층적 정보 요구
법적 안전 의무	해당 서비스에 적용되는 온라인 안전법상의 구체적 의무 범위	사업자가 준수해야 할 의무가 없는 항목은 보고 대상에서 제외

다. 핵심 보고항목 체계

Ofcom은 투명성 보고서에 포함되어야 할 핵심 항목을 몇 개의 범주로 구분하여 제시하고 있다. 우선 불법 콘텐츠 및 유해 콘텐츠 대응 현황이 주요 보고 대상이다. 여기에는 이용자 신고 건수, 자동 탐지 및 인적 검토를 통한 조치 내역, 콘텐츠 삭제·차단·노출 제한 등 집행 결과가 포함된다. 둘째, 위험 평가 및 위험 완화 조치와 관련하여, 플랫폼이 식별한 주요 위험 유형과 이에 대응하기 위해 도입한 정책·기술적 조치가 보고 대상이 된다. 셋째, 아동 보호 및 취약 이용자 보호 조치 역시 핵심 항목으로, 연령 확인, 보호 설계(safety by design), 노출 제한 기능 등의 운영 현황이 요구된다.

콘텐츠 조정 과정에서 활용되는 자동화 도구와 알고리즘 시스템에 대해서도 명확한 보고를 요구한다. 플랫폼은 자동화 기술이 어떤 단계에서 활용되는지, 인간 검토와 어떻게 결합되는지, 오류나 편향 위험을 어떻게 관리하고 있는지를 설명해야 한다. 이는 알고리즘 기반 의사결정이 확산되는 환경에서 자동화된 안전 조치의 책임성과 설명 가능성을 확보하기 위한 조치로 볼수 있다.

라. 보고 주기, 형식 및 공개 방식

보고서는 정기적으로 제출되어야 하며, 일정 범위 내에서는 공공에 공개될 수 있다. 다만, 영업 비밀이나 보안상 민감한 정보는 보호될 수 있도록 예외 규정을 두고 있다. 또한 Ofcom은 제출된 보고서를 바탕으로 종합 규제 보고서를 작성·공개할 수 있으며, 이를 통해 전체 온라인 안전 규제의 이행 현황과 문제점을 사회적으로 공유한다.

마. 집행 및 규제 활용 방식

투명성 보고는 단순한 형식적 의무가 아니라, 규제 집행의 핵심 수단으로 활용된다. Ofcom은 보고된 정보를 바탕으로 추가 정보 요구, 가이드라인 보완, 조사 착수 또는 집행 조치 여부를 판단할 수 있다. 특히 반복적 미이행이나 허위·부실 보고가 확인될 경우, 이는 온라인 안전법상 제재 판단의 근거가 될 수 있다. 반대로 충실한 보고와 개선 추세가 확인되는 경우, 이는 플랫폼의 성실한 위험 관리 노력으로 평가될 수 있다.

3. 아동 접근성 평가 가이드라인

가. 개요

온라인 안전법의 핵심 목적 중 하나는 아동을 온라인 유해 환경으로부터 보호하는 것이다. 이를 실행하기 위해 Ofcom은 2024년 5월 초안을 거쳐 확정된 '아동 접근성 평가 지침(Children's Access Assessment Guidance)'을 발표하였다.[15]

아동 접근성 평가는 온라인 안전법의 아동 보호 체계에서 출발점 역할을 수행한다. 온라인 안전법에 따라 아동은 18세 미만의 모든 사람을 의미하며, 법은 특정 서비스가 "아동을 대상으로 한다"고 명시적으로 선언하지 않더라도, 실제 이용 가능성이 존재한다면 아동 보호 의무를 부담하도록 규정한다. 이는 연령 제한을 형식적으로 표기하는 것만으로는 아동 보호 책임을 회피할 수 없도록 하기 위한 장치이다.

플랫폼은 서비스의 기능, 이용자 구성, 콘텐츠 특성, 홍보 방식 등을 종합적으로 고려하여 아동 접근 가능성을 평가해야 한다. 평가 결과는 문서화되어야 하며, 서비스 구조 변경 시 재평가가 요구된다.

---

15) 이 가이드라인은 2025년 4월에 개정되었음.

나. 평가 프로세스: 2단계 제재

Ofcom은 서비스 제공자가 다음의 두 단계를 거쳐 아동 접속 가능성을 평가하도록 규정한다.

① 1단계: 고도의 연령 보증(Highly Effective Age Assurance)

플랫폼이 아동의 접속을 원천적으로 차단하고 있는지를 먼저 확인한다. 단순히 이용자의 생년월일을 입력받는 '자기 선언' (Self-declaration) 방식은 인정되지 않는다.

- 기준: '기술적 정확성, 견고성, 신뢰성, 공정성'이라는 4대 원칙을 충족하는 고도의 연령 보증 기술을 사용해야 한다.
- 인정 기술: 안면 연령 추정(Facial Age Estimation), 신분증 대조, 모바일 네트워크 운영자 확인 등이 포함된다.
- 결과: 이러한 고도의 기술을 통해 아동의 접속을 실질적으로 차단하고 있다면, 2단계 평가로 넘어갈 필요 없이 '아동 접속 가능성 낮음'으로 결론 내릴 수 있다.

② 2단계: 아동 이용자 조건(Child User Condition)

1단계에서 아동 접속을 완벽히 차단하지 못하는 서비스는 다음 두 가지 조건 중 하나라도 충족하는지 검토해야 한다.

- 조건 A (상당한 수의 아동 이용자): 영국 내에서 해당 서비스를 이용하는 아동의 수가 '상당한(Significant)' 수준인지 평가한다. 이는 단순한 수치뿐만 아니라 전체 이용자 대비 비율과 서비스의 영향력을 종합 고려한다.
- 조건 B (아동 유인성): 서비스의 성격상 아동이 이용하고 싶어 할 만한 요소(Appeal)가 있어 상당한 수의 아동을 끌어들일 가능성이 있는지를 평가한다.

다. 아동 유인성 판단의 주요 요소(Indicative Factors)

Ofcom은 서비스가 아동에게 매력적인지(Appealing to children)를 판단하기 위해 다음과 같은 세부 지표를 제시한다.

- 서비스의 내용 및 기능: 아동이 선호하는 게임, 유명인 관련 콘텐츠, 시각적 디자인, 아동용 마케팅 요소가 포함되어 있는지 확인한다.
- 상업적 전략: 기업의 수익 모델이나 광고 타겟팅 전략에 아동이 포함되어 있는지 여부

를 고려한다.

- 시장 데이터 및 사용자 분석: 자체 분석 데이터 외에도 제3의 시장조사 기관 자료를 활용하여 아동 이용자의 실제 이용 행태를 파악해야 한다.

〈표 2－18〉 아동 유인성 판단 세부 지표내용

카테고리	주요 판단 지표 및 세부 항목
1. 서비스 콘텐츠	아동이 선호하는 주제와 인물 • 음악, 비디오, 유머/재미있는 콘텐츠 • 인플루언서, 연예인, 영화, TV, 책, 만화, 애니메이션 등 • 아동이 직접 생성하여 다른 아동에게 매력적인 콘텐츠 • 게임, 스포츠, e스포츠 및 관련 유명인
2. 활동 및 기능	창의적 활동 및 연결 기능 • 예술, 사진, 요리, 메이크업, 패션 등 창의적 활동 지원 • 친구 맺기, 프로필 생성, 다이렉트 메시지(DM) 기능 • 자신의 콘텐츠를 온라인에 게시할 수 있는 기능 • 교육, 숙제 지원 및 새로운 기술 습득 콘텐츠
3. 디자인 및 시각 요소	아동의 시선을 끄는 프레젠테이션 • 밝고 대담한 색상 사용 • 상상력을 자극하는 캐릭터, 만화, 그래픽 요소 • 스토리라인이 있는 인터랙티브 기능 • 아동이 시각적으로 매력을 느낄 수 있는 전반적인 외관
4. 상업적 전략	비즈니스 모델 및 마케팅 방향 • 아동용 서비스로 광고하거나 아동을 타겟으로 하는 마케팅 전략 • 아동을 대상으로 하는 광고, 프로모션, 경합(Competition) 허용 여부 • 수익 모델이 아동 이용자 유입과 직접적으로 연결되는지 여부 • 아동이 주로 이용하는 다른 서비스에 광고를 게재하는 경우
5. 아동에 대한 혜택	아동에게 제공되는 기회와 가치 • 교육적 가치 또는 즐거움/창의성 제공 • 아동 간의 조언, 지원 및 소속감을 느낄 수 있는 환경 조성 • 자기표현 기회 및 친구/관계 형성 촉진

카테고리	주요 판단 지표 및 세부 항목
6. 사용자 데이터 및 분석	실제 이용 행태 증거 • 18세 미만 이용자로부터의 신고 또는 불만 제기 데이터 • 연령 제한 위반으로 삭제된 계정 통계 • 제3자 시장 조사 기관의 아동 미디어 소비 추적 데이터

사업자는 이러한 세부 지표를 판단함에 있어 다음과 같은 유의사항을 준수해야 한다.

- 홀리스틱(Holistic) 관점: 서비스 제공자는 위 지표들을 종합적으로 고려하여 아동 유인성을 판단해야 한다.
- 의도보다 실제성: 서비스의 '의도된 타겟'이 성인이라 하더라도, 실제 콘텐츠나 기능이 아동에게 매력적이라면 '아동 유인성'이 인정될 수 있다.
- 보수적 접근: 법의 취지가 아동 보호에 있으므로, 판단이 모호할 경우 아동 보호 의무를 부과하는 방향으로 보수적으로 평가해야 한다.
- 재평가 의무: 현재 '아동 유인성 없음'으로 판단했더라도, 중대한 기능 변화가 있거나 아동 이용자가 급증했다는 증거가 발견되면 즉시 재평가를 실시해야 한다.

라. 기록 보관 및 사후 관리 의무

평가를 마친 서비스 제공자는 그 결과를 공식적으로 문서화해야 한다.

- 기록 유지: 평가 과정에서 사용된 증거 자료와 판단 근거를 상세히 기록하여 Ofcom이 요청할 시 즉각 제출할 수 있어야 한다.
- 재평가 주기: '아동 접속 가능성 없음'으로 결론 내린 서비스라 하더라도, 최소 12개월마다 주기적으로 재평가를 수행해야 한다. 또한, 서비스에 중대한 기능 변화가 생길 경우 즉시 다시 평가해야 할 의무가 있다.
- 후속 조치: '아동 접속 가능성 있음'으로 판명된 플랫폼은 3개월 이내에 아동 위험 평가(Children's Risk Assessment)를 완료하고, 그에 따른 안전 조치를 시행해야 한다.
- 이행기한: 2025년 1월 16일 이전에 운영 중이던 서비스는 2025년 4월 16일까지 첫 평가를 마쳤어야 하며, 신규 서비스는 운영 시작 후 3개월 이내에 완료해야 한다.
- 제재조치: 적절한 평가를 수행하지 않을 경우 전 세계 매출의 최대 10% 또는 1,800만 파운드 중 더 큰 금액의 벌금이 부과될 수 있다.

아동 접근성 평가 절차를 통해 아동 접근 가능성이 인정될 경우 해당 서비스는 아동 안전 의무 전반을 적용받게 된다. 이는 온라인 서비스에 대해 “아동 전용 서비스”와 “일반 서비스”의 이분법에서 벗어나 실제 이용 행태를 기준으로 보호 의무를 확장하는 효과를 가진다. 따라서 대형 소셜미디어, 동영상 플랫폼, 게임 커뮤니티 등 대부분의 서비스가 아동 보호 규제 범위에 포함되어 아동 보호 의무를 시행하게 하는 결과를 가져왔다는 점에서 큰 의의가 있다.[16)]

4. 아동 위험성 평가 가이드라인(Ofcom, Children's Risk Assessment Guidance, 2024)

가. 개요

온라인 안전법에 따라 자신의 서비스가 ‘아동이 접속할 가능성이 있는 서비스’라고 판단한 사업자는 반드시 아동 위험 평가(Children's Risk Assessment, CRA)를 수행해야 한다. 이를 지원하기 위해 Ofcom은 서비스 제공자가 자사 플랫폼 내의 위험 요소를 식별하고 관리하는 방법을 상세히 담은 가이드라인을 발표하였다.

아동 위험 평가는 플랫폼의 설계, 기능, 운영 방식이 아동에게 미칠 수 있는 잠재적 해악을 사전에 파악하기 위한 법적 의무이다. 이는 단순히 콘텐츠를 모니터링하는 것을 넘어 서비스의 구조적 결함이 아동을 위험에 노출시키고 있지는 않은지 분석하는 데 목적이 있다. 따라서 플랫폼 사업자가 주관적 판단이 아니라 Ofcom이 제시한 객관적 프레임워크에 따라 위험의 심각성과 발생 가능성을 측정하도록 요구한다.

나. 주요 평가대상 : 아동에게 위험한 콘텐츠

아동 위험성 평가는 아동 접근 가능성이 인정된 서비스에 대해, 아동에게 발생할 수 있는 구체적 위험을 식별·분석하는 절차이다. 위험은 콘텐츠 자체뿐만 아니라 추천 알고리즘, 상호작용 기능, 이용자 간 접촉 구조 등 서비스 전반에서 발생할 수 있다.

위험 평가 가이드라인은 위험 평가 시 반드시 고려해야 할 ‘아동에게 유해한 콘텐츠’를 크게 세 가지 범주로 분류하여 제시한다.

---

16) White & Case LLP. (2025, August 4), UK Online Safety Act: Protection of Children Codes come into force.

- 불법 콘텐츠(Illegal Content): 아동 성착취(CSAE), 테러, 사기, 마약 거래 등 모든 연령에서 금지되는 콘텐츠.
- 최우선순위 콘텐츠(Primary Priority Content): 아동에게 극도로 위험한 콘텐츠로, 자해 조장, 자살 권유, 섭식 장애(거식증 등) 조장, 성적 유해 콘텐츠 등이 포함된다.
- 우선순위 콘텐츠(Priority Content): 사이버 불링(괴롭힘), 혐오 발언, 폭력적 콘텐츠 등 아동의 정서적·신체적 안녕을 해칠 수 있는 정보이다.

다. 위험 평가의 프로세스

플랫폼은 이러한 위험을 단순 나열하는 것이 아니라, 발생 가능성과 심각성을 종합적으로 평가해야 한다. Ofcom은 체계적인 위험 평가를 위해 다음과 같은 4단계 방법론을 준수할 것을 권고한다.

1) Step 1: 평가 대상 유해 콘텐츠의 이해

첫 번째 단계의 핵심은 서비스 제공자가 '아동에게 유해한 콘텐츠(CHC)'의 종류를 명확히 파악하고, 자사 서비스와 관련된 위험 요소를 확인하는 것이다.

- 해악 범주 식별: 사업자는 법에서 규정한 우선순위 유해 콘텐츠(PPC, 예: 자살, 자해, 섭식 장애), 우선순위 유해 콘텐츠(PC, 예: 사이버 불링, 혐오 선동), 그리고 비지정 유해 콘텐츠(NDC)를 구분하여 파악해야 한다.
- 위험 프로필 참조: Ofcom이 제공하는 서비스 유형별(U2U 또는 검색) 위험 프로필을 반드시 참조하여, 자사의 비즈니스 모델이나 기능(알고리즘, 익명성 등)이 어떤 유해 콘텐츠와 강력하게 연결되는지 확인해야 한다.
- 초기 기록: 어떤 위험 프로필을 참조했는지와 서비스에 해당하는 위험 요인을 공식적으로 기록해야 한다.

2) Step 2: 아동 해악 위험 수준 평가

두 번째 단계에서는 식별된 유해 콘텐츠가 실제 아동에게 도달할 발생 가능성(Likelihood)과 그 영향(Impact)을 분석하여 위험 수준을 결정한다.

- 발생 가능성 및 영향 평가: 자사의 추천 알고리즘, 검색 기능, 성인-아동 간 접촉 가능성 등이 유해 콘텐츠 노출 빈도와 심각성에 미치는 영향을 평가한다.

- 연령별·취약계층별 세분화: 아동을 하나의 집단으로 보지 않고, 발달 단계에 따른 5개 연령 그룹(0-5, 6-9, 10-12, 13-15, 16-17세) 및 취약 아동 그룹별로 해악의 영향이 어떻게 다른지 개별적으로 평가해야 한다.
- 증거 기반 분석: 모든 사업자가 검토해야 하는 '핵심 증거' (Core Inputs)와 대형 사업자가 추가로 고려해야 하는 '강화된 증거' (Enhanced Inputs)를 활용하여 객관적으로 분석해야 한다.
- 위험 등급 할당: 분석 결과를 바탕으로 각 콘텐츠 해악별로 High/Medium/Low/Negligible 중 하나의 위험 등급을 할당한다.

3) Step 3: 조치 결정, 이행 및 기록

세 번째 단계에서는 평가된 위험을 완화하기 위한 구체적인 안전 조치를 결정하고 이를 시스템에 반영한다.

- 조치 선택: 기존의 통제 수단이 충분한지 점검하고, 부족할 경우 Ofcom의 '아동 안전 코드' (Children's Safety Codes)를 참조하여 추가 조치를 결정한다.
- 위험 수준에 따른 대응: 평가 결과가 'High'나 'Medium'인 경우 해당 위험을 직접적으로 겨냥한 강화된 조치(예: 추천 시스템 제한 등)를 반드시 시행해야 한다.
- 공식 기록 보관: 어떤 조치를 왜 선택했는지, 해당 조치가 식별된 위험을 어떻게 줄이는지에 대한 상세한 설명을 포함하여 전체 평가 프로세스를 문서화해야 한다.

4) Step 4: 보고, 검토 및 업데이트

마지막 단계는 평가 결과를 내·외부에 보고하고, 변화하는 환경에 맞춰 평가의 유효성을 지속적으로 유지하는 사후 관리 단계이다.

- 내부 거버넌스 및 보고: 위험 평가 결과를 사내 적절한 거버넌스 채널(최상위 경영진 등)에 보고하여 승인을 받아야 한다.
- 외부 보고 및 공개: 카테고리 1 및 2A 서비스는 평가 기록 사본을 Ofcom에 제출해야 하며, 카테고리 1 서비스는 서비스 약관에 요약본을 공개해야 한다.
- 정기 및 수시 업데이트: 최소 12개월마다 정기적으로 재평가를 수행해야 한다. 또한 서비스 설계에 중대한 변경이 있거나 Ofcom의 위험 프로필이 수정되는 등의 트리거 발생 시 즉시 업데이트해야 한다.

Ofcom은 개별 기업의 평가 부담을 줄이고 일관성을 유지하기 위해, 플랫폼 유형별로 공통적인 위험 요소를 정리한 '위험 프로필' (Risk Profiles)을 제공한다. 사업자는 Ofcom의 위험 프로필을 참고하여 자신의 서비스가 속한 카테고리의 일반적인 위험을 먼저 확인한 뒤, 자사 서비스만의 고유한 위험을 추가로 분석하는 방식을 취할 수 있다. 이는 특히 자체적인 연구 역량이 부족한 중소 규모 플랫폼에게 가이드라인 역할을 한다.

라. 기록 보관 및 갱신 의무

위험 평가는 완료 후 즉시 보고서를 작성하여 보관해야 하며, 다음의 경우 반드시 재평가를 실시해야 한다.

- Ofcom이 위험 평가 가이드를 대폭 수정하거나 새로운 위험 요소를 발표했을 때.
- 서비스에 새로운 기능(예: 새로운 소셜 기능, 생성형 AI 결합 등)이 도입되어 위험 프로필이 변했을 때.
- 최소 연 1회 이상의 정기적 검토를 통해 평가의 유효성을 확인

마. 아동 위험평가 지침의 의의

온라인 안전법에 의해 제정된 본 지침은 플랫폼의 비즈니스 모델이나 알고리즘 구조 자체가 아동에게 해를 끼칠 수 있음을 법적으로 인정하고, 이를 사전에 방지하도록 하는 '디자인에 의한 안전' (Safe by Design) 원칙을 실질적으로 구현하였다. '주의 의무'라는 추상적인 개념을 '위험 식별 - 발생 가능성 및 심각성 평가 - 완화 조치'라는 구체적인 프로세스로 정립하여 증거에 기반한 위험관리 체계를 도입하도록 하였다.

이 가이드라인은 특히 아동에게 유해한 콘텐츠의 범위를 자해, 섭식 장애 등으로 구체화하고 이에 대한 알고리즘적 책임을 묻고 있으며 아동을 유해 콘텐츠로 유인하는 '토끼굴' (Rabbit Hole) 현상을 위험 요소로 명시하였다. 이에 따라 기업들은 체류 시간 극대화만을 목적으로 하던 알고리즘 지표에 '안전성'과 '아동 보호' 가중치를 필수적으로 반영하는 기술적 변화를 따르게 되었다.[17]

'아동이 접속할 가능성이 있는 서비스'를 측정하기 위해서는 매우 높은 수준의 연령 확

17) Osborne Clarke. (2024, May 22). UK Online Safety Act: Ofcom publishes guidance on age assurance and children's access assessments.

인 기술을 사용하게 된다. 이는 안면 연령 추정, 제3자 데이터 인증 등 연령 확인 시장의 성장을 촉진하기도 하며, 플랫폼 이용자 경험(UX)에서 연령 확인 절차가 보편화되는 결과를 초래하고 있다.[18] 그밖에 규제 컴플라이언스 비용의 증가, 중소서비스 제공자들에 대한 엄격한 평가 의무 부가에 따른 규제 장벽 등이 타 국가에서 유사 입법을 추진함에 있어 추가적으로 살펴볼 요소라고 할 수 있다.

18) White & Case LLP. (2025), Implementing the UK Online Safety Act: What platforms need to know about Ofcom's latest Codes.

제 3 장 플랫폼 자율규제 개선방안

제 1 절 정보 무결성

1. 개요

디지털·AI 기술 혁신과 디지털 환경의 급진적 확산으로 디지털 정보 유통의 영향력이 지대해짐에 따라 표현의 자유를 확대시킨 이점이 있는 반면, 허위조작정보 등 불법유해정보의 대량 확산이라는 새로운 사회적 위험을 수반하였다. 디지털 정보의 비대칭성과 플랫폼 기반 확산 구조는 피해 회복의 곤란, 신속 대응의 어려움, 사회적 혼란 등의 문제를 야기한다. 생성형 AI 기술의 급속한 발전은 콘텐츠 다양화 등에 기여하지만, 사실과 허위의 경계가 모호해지는 정보 불확실성이 심화될 우려가 있다.

허위조작정보는 사회적 혼란을 야기하고 사회의 건전한 토론을 왜곡하여, 특히 선거를 앞두고 여론을 부적절하게 반영하거나 사회적 분열과 양극화를 심화시킬 가능성이 있다. 또한, 정부와 정책에 대한 신뢰를 훼손하며, 특히 코로나-19와 같은 정부 위기 상황에서 국가 보건정책에 대한 혼란을 야기할 수 있다.

허위조작정보 대응 프레임워크는 세 가지 층위로 나눌 수 있다. 첫째, 정보의 투명성/책임성/다양성을 높이기 위해 대형 플랫폼 기업의 자율규제 강화와 지역 독립 언론을 활성화하여 정보 생태계의 무결성을 촉진하여야 한다. 둘째, 시민들이 복잡한 정보 환경에서 비판적인 접근을 할 수 있도록 다양한 교육을 통해 허위조작정보에 대한 사회적 회복력을 키워야 한다. 셋째, 정보 무결성 확보를 위해 정부가 충분한 기술과 자원을 확보하고 아울러 민간과의 다양한 협력방안을 강구해야 한다.

정보통신망법은 권리침해정보에 한하여 피해자의 삭제 요청 등 구제 수단 제공에 집중하고 있으며, 허위조작정보를 포함하여 다른 유형의 불법유해정보에 대한 대응 및 플랫폼 사업자의 적극적 의무와 책임 규정은 미비하다. 플랫폼 사업자들은 자율적으로 다양한 대응을 하고 있으나, 대응이 제각각이고 전반적으로 불충분하며 국내외 이해관계자들 간 연

계·협력도 미흡한 상황이다. 이에 이 절에서는 허위조작정보 대응을 위한 플랫폼 사업자의 역할과 책임 강화 방안에 대해 살펴본다.

2. 규제 및 사업자 대응 현황

가. 규제체계 현황

허위조작정보로 인해 침해되는 법익의 유형에 따라 규제수준이 상이하다. 개인적 법익 침해와 관련해서는 유포자는 허위사실 적시 명예훼손죄(형법), 출판물 등에 의한 명예훼손죄(형법), 사이버 명예훼손죄(정보통신망법) 등 형사처벌, 재산상 손해, 인격권 침해 기타 정신적 고통에 대한 손해배상 청구, 표시광고법 등에 따른 행정처분, 언론보도의 경우 정정보도/반론보도 청구(언론중재법) 등이 적용된다. 플랫폼 사업자의 경우 제한적 조건 하에 손해배상 청구가 가능한데 정보를 삭제한 사업자에 대해서도 손해배상 청구가 가능하며(감면규정 있음), 정보통신망법 등 법률에 따른 행정기관의 삭제/차단 명령에 응할 의무가 있다.

사회적 법익 침해와 관련해서는 유포자는 허위사실공표죄/후보자비방죄(공직선거법)가 적용되고, 플랫폼 사업자는 관할 선거구관리위원회의 삭제 또는 취급의 거부/정지/제한 요청에 따를 의무가 있다(공직선거법).

딥페이크에 대해서는 선거운동을 위한 경우 선거일 전 90일부터 선거일까지 금지하고(그 외의 기간에는 AI를 이용해 만든 가상의 정보라는 사실을 명확하게 인식할 수 있도록 표시)(공직선거법), AI 시스템에 의해 생성되었다는 사실을 이용자가 명확하게 인식할 수 있는 방식으로 고지/표시해야 한다(AI기본법).

나. 사업자 대응현황

사업자 공동의 자율규제로는 언론사 명의/직책 등을 사칭/도용한 기사 형태의 허위 게시물에 대한 규제가 있다. 한국인터넷자율정책기구(KISO)에서 '가짜뉴스 신고센터'를 운영하며, '코로나19 관련 허위조작정보에 관한 정책'을 한시적으로('21.4.15~'23.3.31) 시행하였다.

네이버는 운영 정책상 인터넷 공간의 신뢰성을 심각하게 위협하고, 사회적 소통의 합리성을 저해할 뿐만 아니라 다양하고 자유로운 소통을 오히려 어렵게 할 우려가 있는 게시물은 게재가 제한될 수 있다고 규정한다. '다른 이용자를 기만할 목적으로 타인을 사칭하

거나 허위 사실을 주장하는 내용의 게시물', '언론사의 명의/직책 등을 사칭 또는 도용하여 기사 형태를 갖춘 허위 게시물' 등이 그러한 예이다. 신고 접수는 신고센터 내 '허위조작정보 신고' 탭에서 이루어진다. 선거 관련허위조작정보는 '중앙선거관리위원회 선거 위반행위 신고'를 안내하고, 딥페이크 기술을 이용한 이미지·영상 등은 성범죄 또는 아동·청소년 보호를 저해하거나 청소년유해 또는 음란성 게시물에 해당하는 내용은 네이버에 신고하거나 '방송통신심의위원회 디지털 성범죄 신고'를 이용하도록 안내한다. 본인의 명예훼손/초상권 등 권리침해 내용은 네이버에 신고하도록 하고, 공직선거법 등 선거 관련 법 위반 내용은 '중앙선거관리위원회 선거 위반행위 신고'를 안내한다. 기사 형태의 허위게시물은 앞서 살펴본 'KISO 가짜뉴스 신고센터'를 안내하고, 그 외에는 '네이버 게시물 신고', '사칭 피해 신고' 등의 탭에서 신고할 수 있다.

네이버는 운영 정책상 게시물 이용제한 현황, 불법정보/권리침해/저작권침해 콘텐츠에 대한 조치 등 이용자 권리보호 현황을 공개한다. 운영정책 위반 유형에는 스팸/홍보, 욕설/차별/혐오, 청소년 유해물, 불법정보 등이 있으며, 허위조작정보 관련 통계는 제공하지 않는다.

카카오는 2024년 3월 「딥페이크 허위조작정보 근절을 위한 카카오의 노력」을 발표하였으며, 주 내용은 선거 관련 허위조작정보 방지에 관한 것으로 이용자 대상 캠페인 전개 및 허위조작정보 신고/조치로 구성되었다. 2024년 7월 「피싱 피해확산 방지 및 근절 위한 카카오의 노력」을 발표하였으며, 기술적 해결책 개발, 정부기관 협력을 통한 대응 강화, 정책적 조치 강화 및 피해 예방을 위한 이용자 교육으로 구성되었다. 2024년 8월 사칭 사기, 피싱 범죄 예방 위해 카카오톡에 '페이크 시그널' 기능을 도입하였다. 이는 '사칭 의심 계정'을 자동 탐지하여 이용자들이 주의를 기울이도록 경고 표시 및 문구 등을 노출한다.

카카오의 운영 정책상 언론사의 명의/직책 등을 사칭 또는 도용하여 기사 형태를 갖춘 게시물 중 그 내용이 허위로 판단되는 게시물을 게시하는 행위, 타인의 개인정보 또는 기기를 도용·탈취하거나 허위의 정보를 입력하여 계정과 아이디를 생성·이용·탈퇴하는 행위, 정상적인 서비스 이용으로 볼 수 없는 다량의 계정 생성 및 반복적인 계정 생성과 탈퇴 행위 등 및 이와 유사한 행위가 금지된다.

카카오가 제공하는 통계에는 피싱예방 도구를 통한 계정 차단, 일반채팅/오픈채팅의 이용제한 조치 계정 건수, 사유별 조치 건수 등이 있다. 일반채팅의 경우 사기·사칭, 도박

등 불법 스팸, 유사투자자문 등, 음란·성적 행위로 구분되고, 오픈채팅은 욕설·혐오·발언·폭력·혐오, 음란·성적 행위, 불법 스팸 및 서비스 운영 방해, 유사투자자문, 불법 상품·서비스, 사기·사칭, 개인정보 무단 수집·유포로 구분된다.

유튜브는 운영 정책상 혼동을 야기하는 콘텐츠 또는 사기성 콘텐츠로 심각한 피해를 입힐 중대한 위험이 있는 특정 유형의 콘텐츠는 허용되지 않으며, 여기에는 실제적인 위험을 초래할 수 있는 특정 유형의 잘못된 정보, 기술적으로 조작된 콘텐츠, 민주적 절차를 방해하는 콘텐츠가 포함된다. 허위 선거 정보 정책 및 허위 의료 정보 정책은 별도 섹션으로 제공한다. 운영 정책 위반 시 위반 횟수와 심각성에 따라 '주의', '경고', '채널 또는 계정 폐쇄' 등의 조치를 취할 수 있다.

유튜브는 외부 평가자와 전문가에게 콘텐츠가 근거 없는 음모론이나 부정확한 정보를 조장하는지에 대한 평가를 요청하여 평가자들의 합의된 의견을 바탕으로 검증된 머신러닝 시스템을 사용하여 추천 모델을 구축한다.

유튜브가 제공하는 통계에는 삭제사유별 건수 및 비중, 감지소스 및 조회수별 비중 등이 있다. 국가별 삭제 건수를 제공하나, 국가별 삭제사유별 건수/비중은 제공하지 않는다. 글로벌 기준 '잘못된 정보'로 인한 채널/동영상 삭제 통계를 포함하고 있으며, 최근('25년 4~6월) 채널 삭제사유는 스팸, 기만행위, 사기(76.0%), 아동 보호(7.5%), 과도한 노출 및 성적인 콘텐츠(6.1%), 잘못된 정보(3.1%), 유해하거나 위험한 콘텐츠(2.3%) 등의 순이고, 동영상 삭제사유는 아동 보호(60.0%), 괴롭힘/사이버 폭력(11.6%), 유해하거나 위험한 콘텐츠(9.5%), 폭력적/노골적 콘텐츠(6.9%), 과도한 노출 및 성적인 콘텐츠(4.9%), 폭력 조장 및 폭력적 극단주의(2.7%), 잘못된 정보(1.1%) 등의 순이다.

메타는 운영 정책상 신체적 상해 또는 폭력, 건강, 유권자/인구조사 방해 관련 허위조작 정보 삭제, 조작된 미디어가 규정위반이 아닌 경우 정보성 레이블 배치, 허위 계정/허위 배포/허위 타겟 구축/외부 허위 행동/허위 참여 금지를 규정하고 있다.

메타는 전 세계 90곳 이상의 인증된 팩트체크 기관과 60개 이상의 언어로 협력하며, 한국의 팩트체크 기관은 JTBC, AFP Korea이다. 팩트체크 기관의 평가에는 '거짓', '조작됨', '일부 거짓', '컨텍스트 누락', '풍자', '사실'이 있다. 게시물에 팩트체크 기사를 첨부하며, 사람들이 이 콘텐츠를 공유하려고 시도하거나 이전에 이미 공유한 경우 해당하는 사람에게 알린다. 메타가 보유한 기술을 활용하여 팩트체크 기관에서 평가한 것과 같거나 거의 같

은 콘텐츠를 감지하고 해당 콘텐츠에도 알림을 추가한다. '거짓', '조작됨' 게시물의 배포를 크게 줄이고, '일부 거짓'게시물은 적게 노출하며, '컨텍스트 누락'게시물은 팩트체크 기관의 정보를 더 많이 표시하는 데 중점을 둔다. '거짓', '조작됨', '일부 거짓' 또는 '컨텍스트 누락'으로 평가한 콘텐츠가 포함된 광고는 거부하며, 해당 콘텐츠를 추천하지 않는다. '거짓' 또는 '조작됨'으로 평가된 콘텐츠를 반복적으로 공유하는 페이지, 그룹, 프로필, 웹사이트, 계정은 지정된 기간 동안 일부 사용을 제한한다. 사람들에게 표시되는 추천 항목에서 해당 계정 제거, 배포 축소, 수익 창출 및 광고 게재 기능 삭제, 뉴스 페이지 등록 기능 삭제 등을 시행한다.

메타가 제공하는 통계에는 운영 정책 위반 콘텐츠 출현율, 조치를 취한 콘텐츠 수, 사전 대응율, 이의제기 및 복원 콘텐츠 수 등이 있다(글로벌 기준). 폭력과 범죄 행위, 안전, 불쾌한 콘텐츠, 정직성과 진정성 영역으로 구분되며, '정직성과 진정성'에서 가짜 계정 및 스팸 통계만 제공한다.

틱톡은 운영 정책상 게시하는 사람이 의도한 바와 상관없이 개인이나 사회에 상당한 피해를 미칠 수 있는 허위 정보를 허용하지 않는다. 그 예로는 거짓 소문, 오해의 소지가 있는 AI 생성콘텐츠, 유해한 음모론 등이 있다. '상당한 피해'란 "심각한 신체적 부상이나 사망, 심각한 심리적 피해(예: 트라우마), 대규모 재산상 피해 또는 사회적 피해(예: 근본적인 사회 시스템 또는 기관의 훼손)"을 의미한다. 신고는 콘텐츠 내 신고 버튼을 누르고 '허위 조작정보/오해유발 콘텐츠' 항목을 선택하여 하고, 사칭계정/댓글/DM 등은 '문제 신고'채널로 신고한다.

틱톡은 독립적인 팩트체크 담당자 및 전문가들과 협력하여 콘텐츠의 정확성을 평가하고, 이를 심사 결정에 반영한다. 허용되지 않는 경우와 추천하지 않는 경우로 구분하며, 경고 라벨을 적용하거나 이용자에게 알림을 전송할 수 있다.

틱톡이 제공하는 통계로는 콘텐츠 삭제 및 복구, 삭제율, 조회수 분포, 신고 대응시간, 정책영역별(개인정보보호 및 보안, 규제대상 물품/상업활동, 민감한 성인 테마, 안전과 예의, 정신/행동 건강, 진실성 및 진정성, 청소년 안전 및 복지) 비중 등이 있으며 국가별로 제공하고 있다. '진실성 및 진정성'항목에 잘못된 정보, 시민 및 선거 공정성, 허위 참여, 편집된 미디어 및 AI 생성콘텐츠, 독창적이지 않은 콘텐츠 등이 있다. 2025년 1분기 글로벌 기준 '진실성 및 진정성'중 '잘못된 정보'가 45.5%로 비중이 가장 컸으며, 다음으로 '시민

및 선거 공정성'(24.5%), '허위 참여'(15.4%), '편집된 미디어 및 AI 생성콘텐츠'(13.8%), '독창적이지 않은 콘텐츠'(0.9%) 순으로 나타났다.

X는 운영 정책상 진위 여부 관련 사항은 플랫폼 조작 및 스팸, 시민 청렴성, 사기성 및 기만적 신원 정보, 합성 및 조작된 미디어(공공의 문제에 대한 광범위한 혼란을 야기하거나 공공 안전에 영향을 미치거나 심각한 피해를 야기할 수 있는 조작된 미디어나 맥락을 벗어난 미디어(오해의 소지가 있는 미디어)를 포함해 허위 미디어의 공유 금지) 등이 있다. 허위 계정, 행동 및 콘텐츠는 전용 인앱 신고 절차를 통해 신고를 받는다.

X는 반기별로 투명성 보고서를 공개하였으나, 2024년 상반기 이후 공개하지 않는다. 기존에는 운영 정책 영역별(학대/괴롭힘, 아동안전, 혐오행위, 불법 또는 규제 상품 및 서비스, 허위 또는 기만적 신원, 비동의 노출, 사적 콘텐츠, 자살 및 자해, 폭력적 및 혐오 단체, 폭력적 콘텐츠)로 신고 건수, 계정 정지 건수, 게시물 삭제/라벨 건수, 게시물 위반율을 글로벌 기준으로 제공하였다. 운영 정책 영역중 '허위 또는 기만적 신원'은 타인을 속이기 위해 신원을 도용하거나 가짜 신원을 사용하는 행위로, 계정 정지 또는 게시물 삭제 조치 통계를 제공하였다.

3. 정책 제언

가. 추진방향

정보의 유형과 확산 구조를 통합적으로 고려한 대응체계가 필요하다. 허위조작정보를 어떻게 정의하느냐에 따라 어느 정보유형(권리침해정보, 법령위반정보, 유해정보)에 해당하는지 달라질 수 있다. 예를 들어, EU 행동강령식 정의는 대체로 유해정보에 해당하지만, 허위성에 초점을 맞추면 모든 정보유형이 가능하다. 사회적 법익 영역(유해정보)에서 허위조작정보에 대한 법적 규제는 표현의 자유와 상충할 수 있다. 권리침해정보는 플랫폼의 신속 삭제 의무 및 피해자 중심 절차 제도화가 중요하다. 법령위반정보는 행정기관의 요청에 따른 우선 대응체계를 갖추어야 한다. 유해정보는 사회적 합의 기반의 자율규제 및 플랫폼 설계 구조를 통한 간접규제가 유용하다.

단순한 삭제 중심의 대응을 넘어 플랫폼 시스템 구조 자체에 대한 재설계가 불법유해정보 대응 전략의 핵심 축으로 부상함에 따라 플랫폼 설계 구조적 대응이 요청된다. 이는 알

고리즘 설계, 콘텐츠 수익화 구조, 계정 생성방식 등 플랫폼 기능 자체에 대한 규범적 설계 요구를 의미한다.

사전 차단과 사후 차단의 양대 축에 의한 대응이 모두 요청된다. 사전 차단은 허위조작정보가 유포되기 전의 대응(예: 리터러시 향상)을 의미하고, 사후 차단은 허위조작정보가 이미 유포된 상황에서의 사후 대응(예: 팩트체크)을 의미한다.

다음으로 '규제화된 자율규제'(공동규제) 도입을 검토해야 한다. 강력한 수준의 공동규제를 택하고 있는 EU의 경우 법률(디지털서비스법)에서 플랫폼 사업자에게 허위조작정보로 인한 시스템 위험 평가 및 경감 조치, 이용자 통제권 보장, 광고 투명성 등 규정하고 있다. 서비스 및 해당 서비스에 통합된 알고리즘을 포함한 아키텍처가 '어텐션 이코노미'하에서 허위조작정보의 유통 및 이용자의 의사결정 자율성에 미치는 영향/위험을 적절한 파악하고 경감 조치를 시행하도록 한다.

공동규제에서 행동강령 준수는 법률 준수의 안전지대 제공, 모범사례의 활용, 평판 위험으로부터의 보호 등 이점을 제공한다. 허위조작정보 개념의 광범위성으로 인해 행동강령이 다른 법률 또는 해당법의 다른 규정에 따른 의무와 중복되는 경우 다른 법률 등이 우선 적용되며, 행동강령은 이를 '보완'하는 규정으로 해석될 수 있다.

플랫폼 사업자 자체 이행평가보고서를 공개하고 제3자(감사)의 검토를 거쳐 정부와 이해관계자들이 함께 평가하고 개선책을 제안하는 순환체계를 구축한다. 정책효과를 명확히 평가하기 위해 측정가능한 목표 및 표준화된 핵심성과지표를 개발한다. 행동강령 이행여부를 자체 평가할 때 사용하는 준수 기준 및 관련 근거자료를 감사기관에 제공한다. 이 경우 서비스별 특성의 차이를 고려하여 사업자 간 벤치마크 기준이 상이할 수 있다.

연구자에게 데이터베이스 및 알고리즘 접근권한을 부여하는 것도 중요하다. 허위조작정보를 연구하는데 필요한 실시간/비실시간 데이터 및 데이터를 효과적으로 분석할 수 있는 API 도구를 제공하도록 한다.

나. 주요 개선 과제

먼저 허위조작정보 유포자들의 주요 수익원인 광고 배치를 차단하는 것이 실질적 대응의 핵심이다. 불법적 허위조작정보를 반복적으로 유포하는 이용자에 대해서는 정보 삭제 및 계정 정지·폐쇄를 확실히 시행하고, 유해한 허위조작정보는 정보의 가시성에 직접적인

영향을 미치지 않는 수익화 차단을 중심으로 한 대응을 통해 신속성 및 표현의 자유와의 균형을 추구한다.

콘텐츠 수익화 및 광고 수익 공유 프로그램에 대한 자격요건과 콘텐츠 검토 절차를 강화하도록 한다. 유튜브의 경우 커뮤니티 가이드라인 위반 시 광고 수익을 제한하고, 반복적으로 문제를 일으킨 계정에 대해 광고 연동 중단조치를 취하고 있다.

광고주/대행사에게 광고가 게재될 콘텐츠 유형에 대한 정보를 제공하고 선택권을 보장한다. 허위조작정보의 출처 등에 관한 정보를 제공하는 출처 평가 기관, 신뢰성 지표, 팩트체커, 연구자 또는 기타 이해관계자들의 정보/분석을 통합할 수 있는 옵션을 제공한다.

다음으로 딥페이크 대응방안을 살펴본다. 딥페이크의 유형은 모욕(비동의 사적 이미지), 사기(사기성 광고, 로맨스 스캠), 허위조작정보(정치적/지정학적/보건관련 등)으로 나눌 수 있다. 허위조작정보를 담은 딥페이크는 주로 선거, 보건, 문화적 이슈와 같은 정치·사회문제에 대한 여론을 형성하는 것을 목표로 한다. 선거를 앞두고 특정 후보의 신뢰성을 떨어뜨리거나 특정 개인·사건과 관련 없는 논쟁적 이슈에 대한 의견 불일치를 조장하거나 극단주의 유포를 위한 경우가 많다. 허위조작정보를 담은 딥페이크의 영향력은 선거와 같은 특정 사건 이후에 약해질 수 있지만, 음모론을 증폭시키거나 끊임없이 혐오를 불러일으키고 오인의 소지가 있는 주장을 하는 딥페이크는 지속적인 영향을 미칠 수 있다.

딥페이크 위험을 완화하기 위해 AI 모델 개발자부터 콘텐츠를 확산하는 공간 역할을 하는 플랫폼에 이르기까지 기술 공급망의 모든 부분에서 대응이 필요하다. 먼저 예방 측면에서 AI 모델 개발자는 프롬프트 필터를 사용하여 특정 유형의 콘텐츠 생성 방지, 학습 데이터셋에서 유해 콘텐츠 제거, 유해 콘텐츠 생성을 자동으로 차단하는 출력 필터 사용 등을 시행할 수 있다. 임베딩 측면에서는 AI 모델 개발자와 플랫폼은 콘텐츠에 워터마크를 임베드하여 딥러닝 알고리즘을 사용하여 감지할 수 있도록 한다. 마지막으로 탐지 측면에서 플랫폼은 맥락 데이터가 첨부되지 않은 경우에도 콘텐츠 검토를 통해 허위조작정보를 식별할 수 있다.

이용자에 대한 정보 제공 및 역량 강화도 필요하다, 첫째, 추천시스템의 투명성과 통제권을 보장해야 한다. 콘텐츠 추천 기준 및 작동 원리를 외부에 설명할 수 있는 메커니즘을 마련해야 하며, 공공기관 또는 제3자의 검증이 필요하다. 통제 기능을 사용하면 허위조작정보에 접하는 것을 줄일 수 있다. 예를 들어, 폭력, 자극적 언어 등 특정 카테고리에 대한

필터링 옵션을 적용할 수 있다. Ofcom 조사('23.10월)에 따르면, 「온라인 안전법」 시행으로 통제권이 부여된 후 이를 사용해 본 적 있는 이용자가 26%인데, 그 이유는 관심사와 선호도에 맞춰 콘텐츠 정렬(36%), 유해하거나 불쾌한 콘텐츠를 보지 않도록 자신을 보호(35%), 유해하거나 불쾌한 콘텐츠를 본 적이 있음(21%) 등의 순이었다. 이후 경험이 향상되었다고 답한 사람은 38%, 경험이 변하지 않았다(44%), 악화되었다(2%), 모른다(16%) 순이었다. 22%는 통제 기능을 알지 못했다고 답하였고, 47%는 알고 있지만 사용해 본 적이 없다고 하였다. 알고 있지만 사용하지 않은 이유는 플랫폼에서 보는 콘텐츠에 만족함(68%), 필요하다고 생각하지 않음(66%), 플랫폼이 콘텐츠를 분류하는 방식을 신뢰하지 않음(26%), 시간이 없음(26%), 흥미로운 콘텐츠를 놓치고 싶지 않음(23%) 등의 순으로 나타났다.

둘째, 이용자에게 허위조작정보를 식별할 수 있는 도구를 제공해야 한다. 이용자가 콘텐츠의 출처, 편집 이력, 진위 여부 또는 정확성을 평가할 수 있는 도구를 제공한다. 예를 들어, C2PA 표준을 활용하여 사진의 출처 및 진위를 확인할 수 있는 기능을 제공하고, 출처가 확인되지 않은 콘텐츠는 경고 표시를 할 수 있다. 또한, 게시자가 자발적으로 AI 생성물임을 표시하도록 유도한다. 유튜브의 경우 변경되었거나 합성된 콘텐츠의 사용을 공개하도록 하고, 지속적으로 위반하면 콘텐츠 삭제, 파트너 프로그램 정지 등 제재를 가한다. 팩트체커가 제작한 콘텐츠의 통합 및 가시성을 향상시킴으로써 팩트체크의 영향력을 증대하는 것도 좋은 방안이다.

셋째, 신뢰할 수 있는 정보와 맥락에 대한 접근성을 향상시킨다. 권위 있는 정보의 가시성을 높이고 허위조작정보의 가시성을 낮추도록 추천시스템을 설계한다. 예를 들어, 추천시스템에 출처의 신뢰성 지표를 포함하거나 이용자에게 선택권을 부여할 수 있다. 특정한 사회적 이슈 또는 위기상황에서 이용자를 권위 있는 출처로 안내하는 기능을 제공한다.

4. 소결

디지털 공간에서의 정보유통과 관련된 위험 문제에 대응하기 위해서는 기술, 서비스, 교육 등 미래의 사회변화를 선제적으로 예측하면서 유연하게 대응하고 다각적인 노력을 지속해야 한다. 과거부터 미래에 이르는 전통적인 과제에 대한 장기전으로서 손쉽게 해결될 수 없다는 인식 하에 구체적 대응책을 마련하고 개선해가는 사회적 체계를 구축해야 하고,

그 과정에서 실태 파악 및 효과 검증이 중요하다.

이에 플랫폼 사업자의 자율적인 노력만으로는 부족하고 표현의 자유와 공공 안전 간 균형을 고려하여 플랫폼 사업자가 적절히 대응할 수 있도록 구체적인 법제도를 정비할 필요가 있다. 이는 우리나라 특유의 과제가 아닌 글로벌 과제로서 다른 국가에서는 이미 다양한 이해관계자가 협력하여 효과적인 대책을 검토하고 시행해 온 성과가 쌓여 있는 상황이다. 디지털 공간에 국경이 없다는 점을 감안하여 국내외 민산학관을 포함한 다양한 이해관계자가 상호 연계·협력하면서 디지털 공간에서의 정보 유통에 관한 거버넌스의 방향에 대해 국제적/안정적/지속적 체계를 갖추어야 한다.

허위조작정보 유통 억제와 신뢰할 수 있는 정보 유통 촉진을 병행하는 대응도 필요하다. 재해 발생 시 등 정확한 정보의 적시 공유가 요구되는 상황에서 국민에게 필요한 정보를 확실하고 편향 없이 전송해야 한다. 이를 위해 정보 발신 측면에서 팩트체크 추진, 저널리즘이나 취재를 통한 검증 보도 및 정보 발신 등 신뢰성 있는 정보 발신을 위한 교육/지원, 이용자 리터러시 향상 등이 필요하다.

제 2 절 아동·청소년 보호

1. 개요

청소년의 SNS 사용은 거의 보편적이며(82.4%) 지속적으로 증가하는 추세이다(전년 대비 11.0% 증가).[19] 숏폼 이용률(전체 평균 81.2%)은 과의존위험군(86.8%)이 일반사용자군(79.6%)보다 높으며, 중학생(91.1%), 고등학생(92.1%)의 이용율이 대학생(93.4%) 다음으로 높다.

SNS는 청소년에게 긍정적 영향과 부정적 영향을 동시에 미친다. SNS는 소외되는 청소년들에게 연결의 장을 제공하고, 자기표현의 공간을 창출하며, 수용받는 느낌을 갖도록 돕는 역할을 한다. 하지만 젊은 세대의 정신건강 위기 상황이 나타나고 있으며, SNS를 주요 유발 요인으로 보는 연구결과도 다수 제시되고 있다. 미국 공중보건서비스단(PHSCC) 의무총

19) 과기정통부, 2024년 스마트폰 과의존실태조사.

감은 SNS가 청소년의 정신건강에 심각한 위해를 미칠 수 있어 강력한 대책이 필요하다는 내용의 권고문을 발표하고(2023년), "의무총감 명의의 경고 표시를 SNS에 노출하도록 요구할 때가 됐다"라고 신문에 기고한 바 있다(2024년, 뉴욕타임즈).

숏폼도 마찬가지로 긍정적 평가와 부정적 평가가 모두 존재한다. 이용자 설문조사 결과, '재미있다'(89.2%), '중독성이 있다'(87.1%), '여가시간을 보내는데 도움이 된다'(80.1%), '자극적/선정적이다'(79.9%)에 대한 동의율이 매우 높고, 다음으로 '생활/시사 정보를 얻는데 도움이 된다'(61.2%), '폭력적이다'(44.3%) 순으로 나타났다.[20]

이 절에서는 SNS/숏폼에 대한 청소년 과의존 현황과 문제점을 살펴보고, 이를 해소하기 위한 사업자 조치현황과 해외 정책동향을 비교 분석한 후 정책 개선방안을 제시하고자 한다.

2. 청소년의 SNS/숏폼 과의존 현황 및 문제점

최근 스마트폰 과의존위험군 조사에서 성인, 60대는 감소하였지만 유아동 및 청소년은 증가하였다.[21]

[그림 4-1] 연령별 스마트폰 과의존 현황

주: 성인은 20~59세를 의미

출처: 과기정통부, 2024년 스마트폰 과의존실태조사

20) 한국언론진흥재단, 2024년 소셜미디어 이용자 조사.

21) 과기정통부, 2024년 스마트폰 과의존실태조사.

최근 1년간 이용량이 증가한 콘텐츠는 게임(24.3%), 영화/TV/동영상(17.4%), 메신저(10.3%), SNS(8.8%) 등의 순이고, 부작용 우려 콘텐츠는 '게임'이 27.9%로 가장 높고, '해당 없음'(19.8%), 'SNS'(8.3%), '영화/TV/동영상'(7.5%) 등의 순으로 나타났다.

숏폼 이용자 중 31.9%는 숏폼 시청 조절의 어려움을 겪고 있다. 이 중 청소년의 비율이 42.2%로 가장 높았으며, 다음으로 유아동(35.1%), 성인(32.3%), 60대(19.7%) 순으로 나타났다. 숏폼 이용자 중 40.7%는 알고리즘으로 인해 같은 유형의 숏폼 반복 시청의 경험이 있다. 이 중 청소년의 비율이 47.8%로 가장 높았으며, 다음으로 유아동(41.8%), 성인(41.7%), 60대(28.8%) 순으로 나타났다. 본인의 과의존에 대한 인식 측면에서 이용자 중 37.0%가 본인이 스마트폰에 의존적인 편이라고 응답하였다. 이 중 청소년의 비율이 61.5%로 가장 높았으며, 다음으로 유아동(31.9%), 성인(38.2%), 60대(18.2%) 순이었다.

과의존 문제해결의 주체로는 '개인 및 가정'이라는 응답이 43.3%로 가장 높았으며, 다음으로 기업(21.3%), 교육시설(17.8%), 정부(17.6%) 순으로 나타났다. 개인 및 가정의 해소방안으로 '대체 여가활동의 활용'(59.9%)이 가장 높았으며, 다음으로 '가족 및 친구의 조언과 도움'(28.1%), '사용조절 앱 등 기술적 지원'(11.9%) 순으로 나타났다. 기업의 해소방안으로 '이용시간에 대한 기술적 조치'(38.1%)가 가장 높았으며, 다음으로 '과다 사용에 대한 안내/경고문 제시'(36.1%), '인식제고 캠페인'(25.8%) 순으로 나타났다. 정부의 해소방안으로 '인식제고를 위한 캠페인'(46.1%)이 가장 높았으며, 다음으로 '전문기관을 통한 상담 및 치료'(30.9%), '법, 규제'(23.0%) 순으로 나타났다.

메타는 2021년 내부고발에 따르면, 페이스북과 인스타그램이 청소년의 정신/신체 건강에 미치는 악영향을 파악하고도 아무런 조치를 취하지 않고 오히려 중독성을 강화하였다. 메타가 파악한 'Social Comparison Sweet Spot'은 사회적 비교를 통한 참여자 극대화의 싸이클(Highlight Reel Norm→Pressure to Look Perfect→Feeding the Spiral)이다. 메타는 이러한 설계가 섭식 장애, 신체이형 장애, 우울증 등의 위험을 초래할 수 있음을 인지하였으며, 특히 여성 청소년 1/3에게 외모에 대해 왜곡된 기준을 제공하는 방식으로 악영향을 끼치고 있음을 인지하였다. 2023년 美 41개州는 페이스북과 인스타그램이 의도적으로 중독성이 강한 시스템을 설계하여 청소년의 정신건강에 해악을 끼쳤다는 이유로 소송을 제기하였다. 이들에 따르면, 페이스북과 인스타그램이 '좋아요'와 댓글이 많을수록 상단에 드러내 이용자 간 경쟁과 더 많은 활동을 부추기고, 새로운 댓글, '좋아요'등을 알림으로 알려

줘 확인하고 싶게 만들었으며, 무한 스크롤 도입, 일정 시간 후 사라지는 '스토리', '라이브' 등을 통해 빨리 확인하지 않으면 못 본다는 불안감을 유도하였다는 것이다.

틱톡에 대해서도 2024년 美 14개州가 틱톡이 청소년에게 미치는 피해를 인지하면서도 조치를 취하지 않았다며 소송을 제기하였다. 이 소송 과정에서 공개된 틱톡의 내부 문건에 따르면, 틱톡은 이용자가 습관을 형성하는데 필요한 정확한 시청량을 측정하였다(260개의 동영상, 약 35분). 틱톡의 자체 조사 결과, 아이들이 앱의 끝없이 흐르는 동영상 피드에 가장 쉽게 끌려드는 것으로 나타났다("예상대로 대부분의 참여지표에서 이용자가 어릴수록 성과가 더 좋았습니다"). 또한, 틱톡의 자체 조사 결과, 강박적인 사용은 분석 능력, 기억 형성, 맥락적 사고, 대화의 깊이, 공감 능력의 저하, 불안 증가 등 여러 가지 부정적인 정신건강 영향과 상관관계가 있으며, 충분한 수면, 직장/학교 업무, 사랑하는 사람과의 소통 등을 방해하였다. 틱톡의 자체 조사 결과, 시간 관리 도구가 사용시간을 미미하게 줄이므로(1.5분 감소) 해당 기능을 출시하기로 결정하였다.

아동기 뇌인지 발달을 마치고 청소년기에 진입하는 시기에 과도한 SNS 사용이 인지발달에 미치는 영향에 대한 연구결과 9~10세 SNS 사용이 증가할수록(복용량 효과) 2년 후 인지능력 저하되었다(美 UC 샌프란시스코, 2025). 청소년기는 생후 1년 다음으로 뇌가 가장 급격하게 재편되는 시기이므로, SNS를 자주 사용할수록 즉각적인 피드백과 보상에 민감하게 반응하도록 뇌가 재구성될 가능성이 있다고 한다. SNS상의 '좋아요', 댓글, 반응 등에 과도하게 의존하는 뇌 구조가 형성되면 학습이나 집중력 유지에 필요한 다른 인지 과정이 상대적으로 약화될 수 있다는 것이다.

기업의 투명성 및 데이터 접근성 미흡으로 SNS가 청소년 건강에 미치는 영향분석에는 한계가 있다. SNS의 단기/중장기 영향을 잘 이해하기 위해서는 더 많은 연구가 필요한 한편, 이러한 지식의 공백이 무위의 변명이 되어서는 안 된다. 비벡 머시 美공중보건서비스단 의무총감은 "청소년들에게 충분히 안전하다는 결론을 내리는데 필요한 정도의 증거는 존재하지 않으며, 이들은 자신도 모르는 채 수십 년의 실험에 참가하게 된 것일 수 있다"고 한다.

3. 사업자 조치 현황

틱톡은 기본 설정으로 계정 비공개, 60분 스크린 타임 제한, 알림 음소거(14~15세 계정은 저녁 9시 이후 푸시 알림 비수신, 16~17세 계정은 저녁 10시 이후 푸시 알림 비활성화) 등을 적용하고, 부모 통제 도구로 맞춤형 일별 스크린 타임 제한(요일별 시간 제한, 맞춤 설정), 스크린 타임 대시 보드(사용시간 및 실행 빈도) 등을 제공한다.

인스타그램은 기본 설정으로 계정 비공개, 60분 경과시 앱을 종료할 것을 권하는 알림 제공, 알림 음소거(22시~7시)등을 적용하고, 부모 통제 도구로 사용시간 제한 설정 및 차단, 특정 요일과 시간 사용 제한 설정, 최근 7일간 평균 사용시간 보기 등을 제공한다.

유튜브는 60분마다 휴식 알림 설정, 취침시간 알림, 자동 재생 중지 등 사용관리 기능을 제공하고, 부모 통제 도구로 사용시간 제한, 기록 삭제, 자동 재생 중지 등을 제공한다. 다만, 13세 이후에는 부모 통제 도구 사용에 대해 자녀 동의가 필요하다.

4. 해외 정책동향

유럽연합은 2022년 제정한 「디지털서비스법」에 따라 미성년자에게 안전한 온라인 환경 보장을 추진하고 있다. 이 법에 따라 미성년자가 접근할 수 있는 플랫폼은 높은 수준의 미성년자 프라이버시·안전·보안을 보장하는 적절하고 비례적인 조치를 취할 의무가 있다. 2025년 7월 제정된 '미성년자 보호 가이드라인'에서는 그루밍, 유해콘텐츠, 문제적 및 중독성 행동, 사이버 괴롭힘, 유해한 상업관행과 같은 온라인 위험으로부터 미성년자를 보호하기 위한 비례적이고 적절한 조치의 예시들을 제시한다. 대형 플랫폼은 미성년자의 신체적·정신적 건강에 대한 부정적 영향과 관련하여 알고리즘 시스템, 기능, 서비스 이용을 포함하여 설계에서 비롯된 모든 시스템 위험을 연 1회 이상 성실하게 식별·분석·평가하고, 식별된 구체적인 위험에 맞춰 합리적·비례적·효과적인 완화 조치를 취할 의무가 있다.

틱톡은 틱톡라이트의 리워드 프로그램(동영상 시청, 좋아요, 팔로우, 친구 초대 등 특정 작업을 수행하면 포인트 획득)에 대한 EC의 조사 개시 후 해당 프로그램을 중단한 바 있으며 이후 영구 폐지하기로 동의의결하였다(2024년). 또한, EC는 페이스북과 인스타그램의 인터페이스가 미성년자의 취약성과 경험 부족을 이용하여 중독적 행동을 발생시키고 소위 토끼굴 효과(rabbit holeeffect)를 강화시킬 위험을 평가하고 완화할 의무 위반에 대해 조

사 중이다. EC는 디지털서비스 제공자의 불공정·기만적 영업관행을 근절하고 소비자 권익을 실질적으로 보장하기 위한 통합 규제체계를 만들기 위해 2025년 7월 「디지털 공정성법」(Digital Fairness Act) 제정을 위한 공개 의견수렴 절차 개시하였다. 이 법은 다크패턴, 중독성 있는 디자인, 불공정한 개인화 관행, 인플루언서 마케팅, 구독 서비스 자동 연장 및 해지 등을 규제한다.

영국은 개인정보보호법에 따른 '연령적합 설계 지침'에서 필요 이상으로 오래 머물게 유도하는 설계를 금지하고, 보상, 사회적 압박 높이기 두려움 등 아동의 발달 취약성을 악용하는 설계를 금지하며, 프로파일링은 기본적으로 비활성화하도록 하였다.

프랑스는 2023년 「디지털 성인 확립 및 온라인 혐오 퇴치법」에서 부모 동의 없이 15세 미만 청소년이 SNS에 가입하는 것을 금지하였다. 부모의 동의를 받아 가입하는 경우에는 사용시간 모니터링 장치를 활성화하고 정기적으로 알림을 제공해야 한다.

호주는 2024년 「온라인 안전법」을 개정하여 부모의 동의 여부와 관계 없이 16세 미만 청소년의 SNS 가입을 금지하였다.

미국은 연방 차원에서 미국 상원의 공화당 블랙번 의원과 민주당 블루멘솔 의원 등이 아동 온라인 보호를 위한 초당적 법안인 「아동 온라인 안전법」을 2025년 5월 공동발의하였다. 이 법은 2024년 상원에서 압도적 찬성으로 통과되었으나, 하원에서 부결된 바 있다. 이 법에 대해 마이크로소프트, 스냅챗, X, 애플 등은 지지하는 입장이고, 구글과 메타는 반대하는 입장이다.

이 법은 미성년자가 사용할 가능성이 큰 온라인 플랫폼에 대해 '주의의무'(duty of care)를 부과한다. 미성년자가 해당 플랫폼에서 사용하는 빈도, 시간 또는 활동을 증가시키거나 장려하는 디자인 기능을 제한해야 하며, 여기에는 무한스크롤, 자동 재생, 플랫폼에서 보낸 시간에 대한 보상, 알림 및 미성년자가 해당 플랫폼을 강박적으로 사용하게 만드는 기타 디자인 기능이 포함된다. 또한, 플랫폼에서 소비하는 시간을 제한하기 위한 쉽게 접근 가능하고 사용하기 쉬운 옵션을 제공해야 한다. 부모 도구로 해당 플랫폼에서 소요된 총 시간에 대한 측정 항목을 보고하고, 미성년자가 해당 플랫폼에서 소요한 시간을 제한할 수 있는 기능을 제공해야 한다. 정신 건강 장애(불안, 우울증, 섭식 장애, 자살 등), 중독을 조장하는 사용패턴, 알고리즘 기반 콘텐츠 추천 기능에 대한 보호자 통제권 및 비활성화 기본설정을 의무화한다.

뉴욕주는 2024년 「아동을 위한 중독성 피드 착취 방지법」을 제정하여 부모의 동의 없이 18세 미만의 아동에게 중독성 있는(맞춤형) 피드를 제공하는 것을 금지하고, 부모의 동의 없는 0시~6시 알림을 금지한다. 캘리포니아주는 2024년 「건강 및 안전법」에 소셜미디어 중독으로부터 아동을 보호하는 규정을 추가하였다. 이에 따라 부모의 동의 없이 18세 미만의 아동에게 중독성 있는 피드를 제공하는 것이 금지되고, 부모의 동의 없는 0시~6시 알림, 9월~5월 월~금 8~15시 알림이 금지된다. 기본 설정으로 계정 비공개, 1시간 사용시간(변경시 부모 동의 필요), '좋아요'수 등 피드백을 볼 수 있는 기능을 제한해야 한다. 부모 도구로 사용시간 및 알림 제한, '좋아요'수 등 피드백을 볼 수 있는 기능 제한 등을 제공해야 한다.

5. 정책 제언

청소년은 도파민을 지속적으로 제공하는 중독성 피드에 특히 취약하여 충동을 통제하는 데 어려움을 겪으며, 발달을 저해하고 장기적인 피해를 초래할 수 있다. 또한, 아직 정체성이 확립되지 않은 시기에 개인화된 콘텐츠를 제공하는 것이 다양한 관점을 접할 기회를 제한한다는 비판도 존재한다.

플랫폼의 특성, 규모, 목적 및 이용자 기반에 따라 청소년에게 다양한 유형의 위험을 초래할 수 있음을 인지하는 위험기반 접근법을 택하되, 플랫폼의 조치가 아동의 권리를 부당하게 제한하지 않도록 해야 한다.

먼저 기본설정으로 과도한 사용을 야기하는 기능을 기본적으로 비활성화해야 한다. 동영상 자동 재생 및 라이브 스트리밍 호스팅 기능을 비활성화하고, 푸시 알림은 기본적으로 비활성화하며, 연령에 맞춰 설정된 핵심 수면시간 동안에는 항상 비활성화한다. 과도한 사용을 유발할 수 있는 기능(예: '좋아요' 수, 읽음 확인)도 비활성화한다.

기본설정을 청소년이 변경하는 경우에는 기존 기본설정으로의 쉬운 복귀를 가능하게 해야 한다. 설정을 변경하는 시점에 경고 신호를 표시하여 변경의 잠재적 결과를 명확히 설명하고, 변경의 잠재적 결과에 대해 주기적으로 알림을 제공하며, 기본설정으로 복귀할 기회를 제공해야 한다.

인터페이스 설계/운영 측면에서 주로 참여 유도 목적으로 설계된 요소에 청소년이 노출

되는 것을 방지해야 한다. 예를 들어, 무한 스크롤, 업데이트된 정보를 받기 위해 특정 행동을 수행해야 하는 불필요한 요구사항, 동영상 자동 재생, 인위적 알림, 희소성/긴급성을 전달하는 신호, (반복적인) 행동을 수행할 때 가상 보상 제공 등이 그러하다. 정기적으로(연 1회 이상), 인터페이스 설계에 중대한 변경을 가할 때 기타 인터페이스 설계/운영에 영향을 미치는 상황을 인지할 때마다 위험 평가 및 완화 조치를 시행해야 한다. 제3자의 실효적 감사를 위해 모든 관련 데이터와 기술사항에 접근을 허용하고 질의에 답하는 등 협력과 지원을 제공한다.

보호자 통제 기능은 기본설정 및 설계 조치와 보완적 관계로 사용한다. 보호자 통제 기능에는 기본설정 관리, 시간제한 설정, 소통하는 계정 확인, 계정설정 관리, 지출한도 설정 등을 포함해야 한다. 다만, 청소년의 프라이버시를 과도하게 제한하지 않도록 모니터링 기능이 활성화될 때마다 청소년에게 실시간으로 명확한 표시를 제공하는 등 적절한 조치를 취한다.

제4장 결 론

본 연구는 온라인 플랫폼을 주축으로 한 ICT 환경변화 속에서 새롭게 대두되는 이용자 보호 이슈에 대한 자율규제 방안 마련 및 이행을 지원함으로써 협력적 생태계 구축에 기여하는 것을 목표로 하였다. 이를 위해 시장 환경 변화와 해외 정책동향 등을 참고하여 플랫폼 자율규제 개선방안을 제시하였다.

먼저 허위조작정보 대응은 정보의 유형(권리침해정보·법령위반정보·유해정보)과 확산 구조를 통합적으로 고려하는 체계가 필요하다. 허위조작정보의 정의 방식에 따라 규율 대상이 달라지며, 특히 사회적 법익 영역(유해정보)에서의 직접 규제는 표현의 자유와 충돌할 소지가 크다. 권리침해정보는 플랫폼의 신속 삭제 의무와 피해자 중심 절차의 제도화가 핵심이고, 법령위반정보는 행정기관 요청에 따른 우선 대응체계를 갖추며, 유해정보는 사회적 합의에 기반한 자율규제와 플랫폼 설계 구조를 활용한 간접규제가 유용하다.

허위조작정보를 포함하여 불법유해정보에 대한 대응은 단순 삭제 중심을 넘어 플랫폼의 시스템 구조 자체를 재설계하는 접근이 중요해지고 있다. 이는 알고리즘 설계, 콘텐츠 수익화 구조, 계정 생성 방식 등 기능 전반에 대해 규범적 설계 요구가 필요하다는 의미다. 이를 위해 '규제화된 자율규제'(공동규제) 도입을 검토해야 한다. 공동규제의 행동강령 준수는 법 준수의 안전지대, 모범사례 활용, 평판 리스크 완화 등 인센티브가 있다. 이행 체계로는 플랫폼이 자체 이행평가보고서를 공개하고, 제3자 감사를 거쳐 정부·이해관계자가 함께 평가·개선하는 순환형 거버넌스를 구축한다. 정책효과 측정을 위해 측정가능한 목표와 표준화된 KPI 개발이 필요하며, 준수 기준과 근거자료를 감사기관에 제공해 실효성을 담보해야 한다.

주요 세부과제로는 첫째, 허위조작정보 유포의 실질적 동기를 약화시키기 위해 수익화 인센티브를 차단하는 것이 핵심이다. 불법적 허위조작정보를 반복 유포하는 이용자에 대해 정보 삭제와 계정 정지·폐쇄를 엄정히 적용하고, 유해한 허위조작정보는 표현의 자유와의 균형을 위해 가시성에 직접 영향을 미치지 않는 수익화 차단 중심으로 대응한다. 콘

텐츠 수익화 및 광고 수익 공유 프로그램의 자격요건과 콘텐츠 검토 절차를 강화하고, 광고주와 대행사에게 광고가 게재될 콘텐츠 유형에 대한 정보를 제공해 선택권을 보장한다.

둘째, 딥페이크 위험 완화는 AI 개발자부터 플랫폼까지 공급망 전 단계 대응이 필요하다. 예방 측면에서 AI 모델 개발자는 프롬프트 필터를 사용하여 특정 유형의 콘텐츠 생성 방지, 학습 데이터셋에서 유해 콘텐츠 제거, 유해 콘텐츠 생성을 자동으로 차단하는 출력 필터 사용 등을 시행할 수 있다. 임베딩 측면에서는 AI 모델 개발자와 플랫폼은 콘텐츠에 워터마크를 임베드하여 딥러닝 알고리즘을 사용하여 감지할 수 있도록 한다. 마지막으로 탐지 측면에서 플랫폼은 맥락 데이터가 첨부되지 않은 경우에도 콘텐츠 검토를 통해 허위조작정보를 식별할 수 있다.

셋째, 이용자 측면에서는 정보 제공과 역량 강화가 중요하다. 추천시스템의 투명성 및 통제권을 보장하면 허위조작정보에 접하는 것을 줄일 수 있다. 이용자가 콘텐츠의 출처, 편집이력, 진위 여부 등을 평가할 수 있도록 도구를 제공하고, 게시자의 AI 생성물 표시 반복 위반 시 제재 등을 통해 자발적으로 AI 생성물임을 표시하도록 유도한다. 팩트체커 콘텐츠의 통합과 가시성 제고로 팩트체크의 영향력을 강화한다. 권위 있는 정보의 가시성을 높이고 허위조작정보의 가시성을 낮추는 방향으로 추천시스템을 설계하고, 위기 상황에서는 이용자를 신뢰할 수 있는 공식 출처로 안내하는 기능을 제공한다.

다음으로 청소년의 SNS/숏폼 과의존과 관련하여 청소년은 도파민을 지속적으로 제공하는 중독성 피드에 특히 취약해 충동 통제가 어렵고, 이는 발달 저해와 장기적 피해로 이어질 수 있다. 또한, 정체성이 형성되는 시기에 개인화된 콘텐츠가 과도하게 제공되면 다양한 관점을 접할 기회가 제한된다는 비판도 제기된다. 이에 플랫폼은 각자의 특성·규모·목적·이용자 기반에 따라 위험 수준이 다르다는 점을 고려한 위험기반 접근을 취하되, 조치가 아동의 권리를 부당하게 제한하지 않도록 균형을 유지해야 한다.

구체적으로는 과도한 사용을 유발하는 기능을 기본설정에서 비활성화해야 한다. 동영상 자동 재생과 라이브 스트리밍 호스팅을 끄고, 푸시 알림은 기본적으로 비활성화하되 연령에 맞춘 핵심 수면시간에는 항상 차단한다. 좋아요' 수 표시나 읽음 확인 등 참여 경쟁과 압박을 유도하는 기능도 기본적으로 비활성화해야 한다. 청소년이 이러한 기본설정을 변경하려는 경우에는 변경 시점에 경고를 제공해 잠재적 결과를 명확히 알리고, 주기적으로 안내하며 언제든 쉽게 기본설정으로 복귀할 수 있도록 해야 한다.

인터페이스 설계·운영 측면에서 무한 스크롤, 불필요한 반복 행동 요구, 인위적 알림, 희소성·긴급성을 강조하는 신호, 가상 보상 등 참여 유도 요소에 청소년이 과도하게 노출되지 않도록 해야 한다. 플랫폼은 정기적으로 그리고 인터페이스에 중대한 변경이 있을 때마다 위험 평가와 완화 조치를 시행하고, 제3자의 실효적 감사를 위해 관련 데이터와 기술 정보에 대한 접근과 협력을 제공해야 한다.

보호자 통제 기능은 기본설정 및 설계 조치를 보완하는 수단으로 활용하되, 기본설정 관리, 이용 시간 제한, 소통 계정 확인, 계정·지출 관리 등을 포함해야 한다. 이 과정에서 청소년의 프라이버시를 과도하게 침해하지 않도록, 모니터링 기능이 작동할 경우 청소년에게 실시간으로 명확히 알리는 등 적절한 보호 조치를 병행해야 한다.

참 고 문 헌

〔국내 문헌〕

과학기술정보통신부, 한국지능정보사회진흥원(2025), 2024년 스마트폰 과의존 실태조사 보고서.

김현수(2024), “국내외 통신규제 제도 비교 연구” 정책연구 24-39.

한국언론진흥재단(2025), 2024 소셜미디어 이용자 조사결과 발표.

〔해외 문헌〕

European Commission(2016), Code of conduct on countering illegal hate speech online.

European Commission(2018), Code of conduct on disinformation.

European Commission(2020), European democracy action plan.

European Commission(2022), Strengthened code of practice on disinformation.

European Commission(2023), Strengthening the code of conduct on countering illegal hate speech online (“Code of Conduct+”)

European Parliament Research Service(2023), The EU approach to tackling disinformation.

OECD(2024), Platform regulation and online safety.

OECD(2024a), Transparency and accountability in platform regulation. OECD Publishing.

OECD(2024b), Platform regulation and online safety: International trends. OECD Publishing.

Ofcom(2024), Codes of practice on illegal content.

Ofcom(2024), Compliance and enforcement overview.

Ofcom(2024), How Ofcom will regulate online safety.

Ofcom(2024), Illegal content risk assessment guidance.

Ofcom(2024), Regulated services: Guidance on scope and application.

Ofcom(2024), Children's access assessment guidance.

Ofcom(2024f), Children's risk assessment guidance.

Ofcom(2024g), Transparency reporting guidance.

Ofcom(2024i), Categorisation and additional duties roadmap.

Ofcom(2024j), Enforcement powers under the Online Safety Act.

Ofcom(2025), Final Transparency Reporting Guidance.

Osborne Clarke(2024), UK Online Safety Act: Ofcom publishes guidance on age assurance and children's access assessments.

UK Government(2024), Statement on the implementation timeline of the Online Safety Act.

UK Government, Department for Culture, Media and Sport(2019), Online harms white paper.

UK Government, Department for Culture, Media and Sport(2022), Online Safety Bill: Policy statement.

UK Government, Department for Culture, Media and Sport(2024), Child safety measures under the Online Safety Act.

UK Government, Department for Science, Innovation and Technology(2023), Online Safety Act 2023: Explanatory notes.

UK Home Office(2023), Online harms policy framework.

UK Parliament, House of Commons Library(2024), Online Safety Act 2023: Briefing paper.

White & Case LLP(2025), Implementing the UK Online Safety Act: What platforms need to know about Ofcom's latest Codes.

White & Case LLP(2025), UK Online Safety Act: Protection of Children Codes come into force.

〔웹페이지〕

법제처 세계법제정보센터, https://world.moleg.go.kr/web/main/index.do.

유럽집행위원회 보도자료, https://ec.europa.eu/commission/presscorner/home/en.
영국법령정보, legislation.gov.uk.
영국정부, https://www.gov.uk/government/news/.

● 저 자 소 개 ●

김 현 수

- 고려대 법학 박사
- 현 정보통신정책연구원 선임연구위원 디지털플랫폼경제연구실 실장

이 경 은

- 서울대 경영학 석사
- 현 정보통신정책연구원 부연구위원

정책연구 25-23

지능정보사회 플랫폼 자율규제 활성화

2025년 12월 일 인쇄
2025년 12월 일 발행

발행인 방송미디어통신위원회 위원장
발행처 방송미디어통신위원회
경기도 과천시 관문로 47
Homepage: www.kmcc.go.kr
인 쇄 경성문화사


방송미디어통신위원회
Korea Media and Communications Commission