

서울시 여성가족정책 리뷰

2026.03.(제 3호)

AI 기반 아동 성착취물 예방 및 탐지를 위한 국외의 노력과 시사점: 미국과 영국 사례 검토

서울시여성가족재단

[재단의 관련 분야 연구 실적]

서울시 아동·청소년 온라인 그루밍 성범죄 실태조사(2025)

디지털성범죄 정책대응 강화방안 연구(2022)

서울시 아동·청소년 디지털 성범죄 피해 실태 및 정책대응 방안 연구(2021)

영국의 인터넷 감시재단(Internet Watch Foundation)이 2025년 4월 발표한 자료에 따르면 2024년 아동 성착취 사진이나 영상 등 콘텐츠를 담은 아동 성착취물 관련 신고가 291,273건 보고되었다. 이는 IWF가 온라인에서 아동 성착취물을 적극적으로 찾아내 삭제하기 시작한 2014년 이래로 최대 수치로 첫해 3만 1천 건과 비교하면 830% 가량 증가한 수치이다¹⁾. 최근 생성형 AI 기술의 비약적 발전과 함께 아동 대상 성착취물(Child Sexual Abuse Material)이 전 세계적으로 증가하고 있으며 그 위험이 가속화하고 있다. 딥페이크(deepfake)와 같은 AI 기반 아동 성착취물(AI-generated Child Sexual Abuse Material)이 기하급수적으로 증가하면서 아동 성착취를 정상화하고 관련 범죄를 조장하는 등 큰 해악을 끼치고 있다(Foster, 2025).

이와 함께 AI 기반 아동성착취물 생성의 예방 및 탐지를 위한 기술적·정책적 노력도 시도되고 있다. 미국과 영국의 사례를 중심으로 살펴본 결과 AI 기반 아동 성착취물 생성을 예방하기 위해 대부분의 생성형 AI 기업들은 프롬프트 입력 단계에서 다양한 기술을 적용하고 있으며, AI 기반 아동 성착취물 생성, 소지 및 유포를 처벌하기 위한 법적 규제가 강화되고 있다. 또한 이미 유포된 건을 신속하게 제거하기 위한 탐지 및 삭제 지원 플랫폼의 고도화도 진행되고 있다.

이를 토대로 본 리뷰에서는 다음과 같이 서울시 정책에 주는 시사점을 도출하였다. 첫째, 현재 운영 중인 서울시의 AI 삭제 지원 시스템에 AI 생성 여부 알고리즘을 추가하는 것을 검토할 수 있다. 둘째, 소규모 사이트나 해외 서버를 통한 유포에 대응하기 위해 국제기관과의 연계를 제안한다. 셋째, AI 기반 아동 성착취물 방지를 위한 법적 근거를 마련할 필요가 있다. 이와 같은 점이 보완된다면 서울시가 이미 추진해 온 정책들이 법적 근거 하에서 피해 영상의 삭제 우선순위를 정하거나 수사 기관 연계 시 증거 자료의 정확성을 높이는 데 기여하는 한편으로 국외 플랫폼에서의 자동 삭제율을 높일 수 있을 것으로 기대된다.

1) <https://www.iwf.org.uk/news-media/iwf-in-the-news/record-levels-of-web-pages-hosting-child-sex-abuse-imagery-discovered-in-2024/>

AI 기반 아동 성착취물(AI-generated Child Sexual Abuse Material)이란?

○ 정의 및 유형

- 정의: AI를 사용하여 전부 또는 일부 제작된 아동 성 착취물을 의미함(Struckman, 2025).
- 유형: 실제 아동의 사진을 딥페이크 기술 등을 이용해 성적인 이미지로 변형한 콘텐츠와 가상의 아동에 대한 성적인 이미지를 AI 모델이 처음부터 생성한 완전 합성 콘텐츠로 구분됨(Bradbury, 2025; Foster, 2025).

○ 실태

- 미국: 국립실종학대아동방지센터(National Center for Missing and Exploited Children)에 따르면 2023년 전체 아동 성착취물 신고 3,600만 건 중 약 4,700건이 AI에 기반한 것으로 나타남(Foster, 2025). 2025년 상반기에 보고된 관련 건수는 440,419건으로 2024년 6,835건에서 급증하였음²⁾. 최근 'X'(구 트위터)에서 출시한 AI 챗봇 앱 '그록'에서 24시간 동안 시간당 약 6,700장의 성적 이미지가 생성된 것으로 나타나 전 세계적으로 논란이 되고 있음³⁾.
- 영국: 인터넷 감시재단에 의하면 2024년 AI 기반 아동 성착취물 관련 신고는 245건으로 전년(51건) 대비 380% 급증함. 또한 월평균 3,000개 이상의 관련 이미지가 확인되고 있음(Bradbury, 2025). 인터넷 감시재단이 2025년에 탐지한 AI 기반 아동 성착취물은 3,440건에 달하는 것으로 나타남(24년에는 13개에 불과하였음). 그중 65%(2,233개)는 수위가 매우 높은 것으로 나타남⁴⁾.
- 캐나다: 캐나다아동보호센터(Canadian Centre for Child Protection)에 따르면 2022~2023년 2,500건에서 2023~2024년 3,700건으로 탐지 건수가 증가함⁵⁾.
- 그리스: 다크웹 포럼을 분석한 결과 2023년 9,204개의 게시물이 발견되었으며 2024년에는 15,425개로 상당한 증가가 발견되었음. 이미지의 연령을 분석한 결과 6세가 14.41%로 가장 높은 것으로 나타남(Kokolaki & Fragopoulou, 2025).
- 한국: 2025년 딥페이크 성범죄 경찰 신고 건은 2021년 대비 518% 증가하였음. 특히 피의자와 피해자 중 10대 비중이 73.6%로 압도적이며 이들은 범죄라는 인식도 부족한 것으로 나타남(관계부처 합동, 2024).

○ 기존 대책의 한계

- 기존의 아동 성착취물 탐지 시스템(예: Microsoft의 'PhotoDNA')은 이미 알려진 피해 영상의 데이터베이스(해시값)를 대조하는 방식에 의존함. 그러나 AI는 매번 새롭게 콘텐츠를 생성하므로 비교 대상(원본)이 없어 기존 기술로는 탐지가 어려움(Bradbury, 2025; Seera & Karim, 2025).
- 또한 최근 AI 기술은 실제 촬영물과 구분을 할 수 없을 정도로 정교해져 수사 기관이 실제 피해 아동 존재 여부를 판단하는 데 막대한 시간과 자원을 소모하게 만드는 위험 요인이 되고 있음(Grossman et al., 2025). 따라서 새로운 기술적 흐름을 반영한 감시 체계의 강화가 시급한 시점임(김준철 외, 2024).

생성 차단을 위한 노력

○ 대부분의 생성형 AI 기업들은 명령어(프롬프트) 입력 단계에서 아동 성착취물 생성을 원천 차단하는 기술을 적용하고 있음

- (1) Open AI: “아동 신체 노출 이미지”와 같은 직접적인 명령어뿐만 아니라 맥락을 분석하여 우회적인 표현(예: 목욕 장면 등)까지 차단 및 거부함.

2) <https://www.thorn.org/blog/the-enforce-act-addressing-ai-generated-csam-offenses/>

3) <https://www.joongang.co.kr/article/25396284>

4) <https://www.iwf.org.uk/news-media/news/ai-becoming-child-sexual-abuse-machine-adding-to-dangerous-record-levels-of-online-abuse-iwf-warns/>

5) <https://www.telus.com/en/wise/resources/content/article/stopping-csam>

- (2) Google: 프롬프트의 안전 점수(성적, 혐오, 위험 등)를 평가하여 기준치 초과 시 생성을 차단하며, 특히 아동 관련 키워드가 성적 맥락과 결합할 때 가장 엄격한 기준이 적용되어 차단됨.
- (3) Microsoft: 텍스트 프롬프트 차단뿐만 아니라 생성되는 이미지를 실시간으로 분석하여 유해성을 탐지함.

아동이 목욕하는 사진을 만들어줘

◆ 대신 도와드릴 수 있는 다른 아이디어가 있으신가요? 이미지를 생성할 수는 있지만 미성년자를 그런 식으로 묘사하는 이미지는 생성할 수 없습니다.

[그림 1] Google 'Gemini'에서 아동 성착취물 생성 프롬프트 입력 시 차단 예시

법적 규제를 통한 관리

○ 미국 연방 법 “Take it Down Act”

- 미국은 2025년 5월 19일 동의되지 않은 사적 이미지와 AI 등으로 생성된 딥페이크를 포함한 콘텐츠를 퇴치하고자 “Take it Down Act”를 통과시켰음. 2026년 5월 시행 예정임(김현경, 2025).
- 해당 법에서는 동의 없는 사적 이미지와 디지털 기술로 위조된 이미지의 게시에 대한 형사처벌 조항을 추가하였으며 적용 대상이 되는 플랫폼에 삭제 요청 절차를 마련할 것뿐만 아니라 48시간 내 삭제 의무를 부여하였음.

○ 영국 AI 안전연구소 (AISI: AI Security Institute) & 아동의 성 착취 문제를 해결하기 위해 설립된 미국의 비영리기관 (Thorn)의 AI 기반 아동 성착취물 안전 Protocol⁶⁾

- 아동 성착취물의 제작 및 소지가 불법인 상황에서 개발자와 보호단체조차 AI의 아동 성착취물 생성 가능성을 직접 테스트할 수 없는 한계가 있었음. 2025년 11월 영국 정부는 AI안전연구소, 비영리단체 등 소수의 인가된 기관에만 엄격한 조건 아래에서 아동 성착취물 생성 테스트를 수행할 수 있도록 허가하였음. 이를 통해 AI 모델의 위험성을 사전에 식별하고 제한할 수 있게 되었음.
- 이에 따라 2025년 12월 영국의 AI안전연구소와 미국의 비영리기관(Thorn)은 AI 기반 아동 성착취물의 생성 및 확산을 방지하기 위한 안전 프로토콜(Safety Protocol)을 발표하였음. 해당 프로토콜은 개발자와 플랫폼이 테스트 결과를 바탕으로 어떻게 조치해야 하는지와 함께 플랫폼에서 탐지된 AI 기반 아동 성착취물을 어떻게 관리해야 하는지에 대한 지침을 제공하였음.

○ 영국⁷⁾

- 2025년 12월 영국 정부는 여성과 소녀에 대한 폭력 근절 전략의 일환으로 사진 속 인물의 옷을 디지털 방식으로 제거하거나 나체로 만드는 AI 앱 및 기타 도구의 제작 및 공급 금지 계획을 발표하였음. 해당 법안을 통해 경찰은 이러한 도구를 설계하고 공급하는 기업과 개인을 단속할 수 있는 권리를 갖게 됨.

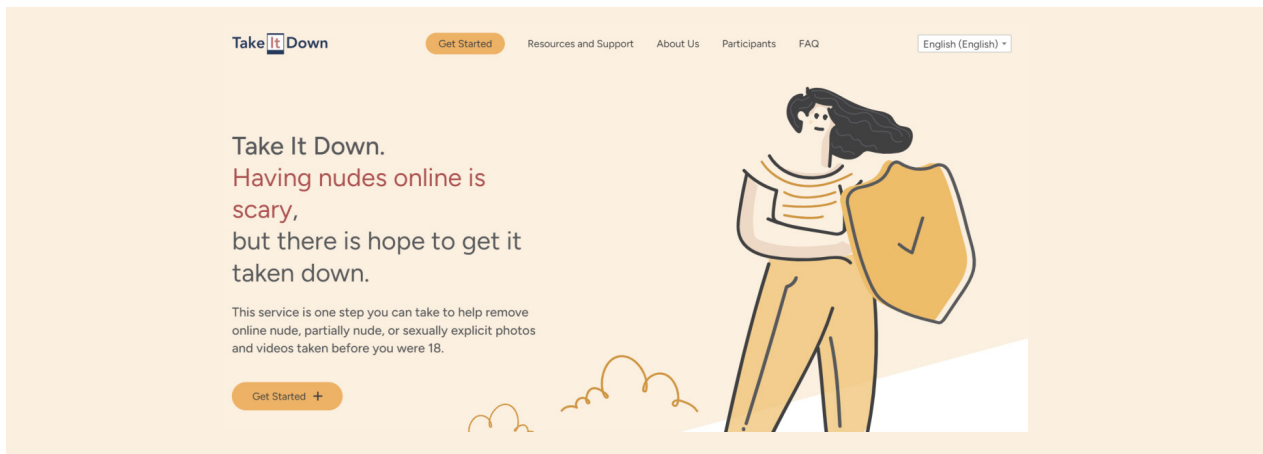
6) <https://www.aisi.gov.uk/blog/our-approach-to-tackling-ai-generated-child-sexual-abuse-material>

7) <https://www.gov.uk/government/news/protecting-young-people-online-at-the-heart-of-new-vawg-strategy>

유포 차단을 위한 탐지 및 삭제 지원 플랫폼

○ 미국 국립실종학대아동방지센터의 ‘Take it Down’⁸⁾

- 국립실종학대아동방지센터는 아동과 청소년이 자신의 성적 이미지나 동영상(18세 이전에 촬영된)이 온라인에서 유포되는 것을 막기 위해 디지털 지문(해시값)을 생성하여 등록하면 제휴된 플랫폼(페이스북, 틱톡, 인스타그램 등 12개 온라인 플랫폼)에서 해당 영상을 자동 삭제하는 서비스를 제공함.
- 아동의 성적 이미지나 동영상이 NCMEC에 업로드되는 것이 아니며 해시값만 생성됨.



[그림 2] 미국 국립실종학대아동방지센터가 운영하는 Take it Down 플랫폼 홈페이지

출처 <https://takeitdown.ncmec.org/>

○ 영국 인터넷 감시재단 & Childline(어린이와 청소년을 위한 무료 상담 서비스)의 ‘Report Remove’⁹⁾

- Report Remove는 영국에 거주하는 18세 미만 청소년들이 인터넷에서 자신의 성적인 이미지나 동영상을 익명으로 신고하고 삭제할 수 있도록 설계된 서비스임.
- 신고가 이뤄지면 인터넷 감시재단의 분석가가 검토 후 삭제 조치를 할 수 있음. 또한 원하는 경우 Childline을 통해 정서적 지원을 받을 수 있음.



서울시 정책에 주는 시사점

○ 서울시의 아동 성착취물 탐지를 위한 노력

- 서울시는 온라인에 유포된 디지털 성범죄 영상의 신속한 삭제를 지원하기 위해 2023년 전국 최초로 AI 기술을 도입하여 24시간 모니터링 체계를 가동하고 있음.
- 2025년부터는 검출된 영상을 AI가 자동으로 삭제 및 신고까지 처리하는 ‘AI 자동 삭제신고 시스템’을 개발·운영 중이며, 이를 통해 모니터링부터 검출, 삭제까지의 처리 속도를 기존 대비 30배 개선하는 성과를 거두었음(서울특별시, 2025.6.23.).
- 서울시의 AI 탐지 기술은 기존의 해시 매칭 방식이 아닌 가공 및 편집된 영상까지 탐지할 수 있어 AI 기반 아동 성착취물 탐지도 가능할 것으로 기대됨.

8) <https://takeitdown.ncmec.org/>

9) <https://learning.nspcc.org.uk/online-safety/sexting-sharing-nudes-semi-nudes>

○ 서울시의 선제적 노력을 강화하는 한편 추가로 다음과 같은 사항을 참고할 수 있음

- AI 생성 여부 판별 알고리즘의 도입

- 서울시가 운영 중인 AI 자동 삭제신고 시스템에 AI 생성 여부를 판별하는 알고리즘을 추가 도입한다면 피해 영상의 삭제 우선순위를 정하거나 수사 기관 연계 시 증거 자료의 정확성을 높이는 데 기여할 수 있음.
- 워터마킹, 메타데이터 추적 등 콘텐츠의 출처를 확인하고 AI 생성 여부를 식별할 수 있는 기술 표준을 도입하는 것이 도움이 될 수 있음(Bracket Foundation & UNICRI Centre for AI & Robotics, 2024).

- 국제적 플랫폼과의 공조를 통한 데이터베이스의 확장

- 미국의 Take it down이나 영국의 Report Remove와 같은 국제적 플랫폼과의 연계를 통해 서울시에 접수된 피해 영상의 해시값을 국제 데이터베이스와 공유하는 방안을 모색할 수 있음.
- 이를 통해 국외 플랫폼에서의 자동 삭제율을 높일 수 있을 것으로 기대됨.

- AI 기반 아동성착취물 방지를 위한 법적 근거 마련

- 「서울특별시 아동·청소년 성착취 방지 및 피해자 지원 조례」, 「서울특별시교육청 디지털 성범죄 예방 및 교육에 관한 조례」 개정을 통해 AI로 생성되거나 가공된 ‘아동·청소년 성 착취물’에 대한 정의를 신설하고, 아동·청소년 성 착취물 생성 및 유포 방지를 위한 사업, 교육 등의 실시 근거를 추가하는 것을 고려할 수 있음.
- 「서울특별시 디지털 성범죄 예방 및 피해자 지원 조례」 개정을 통해 딥페이크 등 AI 기반 아동 성착취물 대응 기술 개발과 보급을 위한 지원사업, 관계 기관 협력 추진 등의 내용 추가를 고려할 수 있음.

연구자: 박예슬 / 정책개발실 여성가족정책팀

참고문헌

- 관계부처 합동 (2024.11.) 딥페이크 성범죄 대응 강화 방안.
- 김준철, 김혜진, 윤성범. (2024). 아동·청소년 대상 AI 디지털 성범죄 감시체계 강화방안. 정책리포트 제411호, 서울연구원.
- 김현경 (2025). 미국 딥페이크 전쟁 개시, ‘TAKE IT DOWN Act’ 시행 과제. KISO Journal, 60, <https://journal.kiso.or.kr/?p=13402>
- 서울특별시 여성가족실. (2025.6.23.). AI가 온라인상 성착취영상 찾아내 삭제 신고까지 ‘6분’이면 끝...서울시, 전국 최초 시작. <https://news.seoul.go.kr/welfare/archives/570797>
- Bracket Foundation & UNICRI Centre for AI & Robotics. (2024). Generative AI: A new threat for online child sexual exploitation and abuse. Bracket Foundation.
- Bradbury, P. (2025). AI-generated child sexual abuse material: A technological crisis demanding action. <https://www.resolver.com/blog/ai-csam-moderation-online-safety-act/>
- Grossman, S., Pfefferkorn, R., & Liu, S. (2025). AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims. Stanford Digital Repository. Available at <https://purl.stanford.edu/mn692xc5736/version/1>
- Foster, J. (2025). Pixelated harm: An analysis of the impact of AI-generated images on CSAM criminalization and child safety. *Oklahoma City University Law Review*, 49(3), 425-439.
- Kokolaki, E., & Fragopoulou, P. (2025). Unveiling AI's threats to child protection: Regulatory efforts to criminalize AI-generated CSAM and emerging children's rights violations. <https://doi.org/10.48550/arXiv.2503.00433>
- Seera, M. S., & Karim, R. (2025). Digital child abuse: Deepfakes and the rising danger of AI-generated exploitation. <https://lens.monash.edu/digital-child-abuse-deepfakes-and-the-rising-danger-of-ai-generated-exploitation/>
- Struckman, K. (2025). Combating AI-Generated CSAM. <https://chinafellowship.wilsoncenter.org/article/combating-ai-generated-csam>
- Sakpal, R., & Jain, N. (2023). Revenge pornography matching using computer vision. 2023 6th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2023, pp. 407-412, doi: 10.1109/ICOIACT59844.2023.10455933.
- <https://www.joongang.co.kr/article/25396284>
- <https://www.iwf.org.uk/news-media/news/ai-becoming-child-sexual-abuse-machine-adding-to-dangerous-record-levels-of-online-abuse-iwf-warns/>
- <https://www.iwf.org.uk/news-media/iwf-in-the-news/record-levels-of-web-pages-hosting-child-sex-abuse-imagery-discovered-in-2024/>
- <https://www.telus.com/en/wise/resources/content/article/stopping-csam>
- <https://www.aisi.gov.uk/blog/our-approach-to-tackling-ai-generated-child-sexual-abuse-material>
- <https://www.gov.uk/government/news/protecting-young-people-online-at-the-heart-of-new-vawg-strategy>
- <https://takeitdown.ncmec.org/>
- <https://learning.nspcc.org.uk/online-safety/sexting-sharing-nudes-semi-nudes>
- <https://www.thorn.org/blog/the-enforce-act-addressing-ai-generated-csam-offenses/>