Korean Women's Development Institute ISSUE PAPER

KWDI 이슈페이퍼

2026 November

발간일 : 2026년 11월 30일 **발행처 :** 한국여성정책연구원

수행과제명

디지털 기반 젠더폭력 방지를 위한 AI 벤치마크 데이터셋 필요성 연구

과제책임자

김애라 연구위원

한국여성정책연구원

디지털 기반 젠더폭력 방지를 위한 AI 벤치마크 데이터셋 필요성 연구*

초록

- 인공지능 기술이 젠더폭력의 도구가 될 수 있어 이에 대한 대응이 필요한 상황임.
- 이에 본 연구에서는 인공지능 기술이 야기하는 젠더기반폭력, 그중에서도 성적 침해와 착취를 방지할 수 있는 정책적 프레임워크와 기술적 개입 방안을 모색하고자 하였음.
  - 특히 인공지능 모델의 젠더폭력 위험을 객관적으로 측정, 보완할 수 있는 최소한의 공신력 있는 기준이 필요하다는 점을 검토하였음.
  - 이를 위해 국내외 인공지능 규제 동향과 기술적 한계를 검토하고, 개발 현장의 심층적인 목소리를 반영하여 젠더폭력 방지를 위한 정책적 프레임워크와 기술적 개입 방안을 고민하고자 하였음.
  - 또한 경제적 이해관계에서는 중요하게 고려되기 어려운 영역인 젠더편향, 차별, 젠더폭력과 같은 분야에 대한 공공의 적극적 개입 필요성을 논의하였음.
- 이러한 연구결과를 토대로 ▲ 정부 주도 젠더폭력 방지 벤치마크 구축 ▲ LLM 젠더폭력 안전성 지표 개발 ▲ 인공지능 안전성 확보를 위한 젠더폭력 방지 벤치마크 의무화 ▲ 추진 체계 및 민관 거버넌스 마련 ▲ 부처 간 역할 분담 및 실무 협력 체계의 구체화 ▲ 젠더폭력 방지 벤치마크에 기반한 인공지능 안전성 인증제도 도입 ▲ 지속적인 모니터링 및 환류 체계 확립, ▲ 멀티모달·프론티어 인공지능 모델 평가로의 확장 등의 정책방안을 제시함.

생성형 인공지능으로 인한 주요 위험 예시

	NIST 위험 범주	주요 내용	관리 방안
	외설적, 모욕적 또는 학대적 콘텐츠 (Obscene, Degrading, and/or Abusive Content)	CSAM(아동 성착취물), NCII (비동의 친밀 이미지), 딥페이크 포르노그래피, 성적괴롭힘 콘텐츠	학습 데이터 정제, 콘텐츠 필터링, 피해자 지원
	환각 (Confabulation)	허위 정보 생성, 특정 집단에 대한 편향된 정보	출력 검증 시스템, 사실 확인 메커니즘
	인간-AI 구성 (Human-AI Configuration)	AI의 의인화, 과의존 및 정서적 유착, 젠더화된 인격을 부여받은 챗봇·AI 동반자와의 상호작용에서 발생하는 위험	인터페이스 설계 기준 마련, AI임을 인지할 수 있는 고지, 취약 이용자 보호 조치

자료: NIST(2024), AI RMF Generative AI Profile(AI 600-1) 참고하여 재구성.

* 이 이슈페이퍼는 2025년 한국여성정책연구원의 연구보고서 내용을 일부 발췌하고 정리해 제작되었습니다.
출처: 김애라 · 김정혜 · 김효경(2025). 디지털 기반 젠더폭력 방지를 위한 AI 벤치마크 데이터셋 필요성 연구. 서울: 한국여성정책연구원.

1 배경 및 문제점

- 디지털성범죄의 많은 사례들에서 확인한 바와 같이 인공지능 역시 젠더기반폭력의 새로운 양상을 초래하고 있음. 인공지능 기술이 젠더폭력의 도구가 될 수 있으며, 이에 대한 대응이 필요함.
- 본 연구는 인공지능 모델의 젠더폭력 위험을 객관적으로 측정, 보완할 수 있는 최소한의 공신력 있는 기준이 필요하다는 문제의식에서 출발하였음.
- 현재 한국은 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(이하 '인공지능기본법')」 시행을 앞두고 있으며, 인공지능 및 데이터 기반 미디어 환경으로 급속히 이행하고 있음. 인공지능에 의한 젠더폭력의 위험을 미연에 방지하기 위해서는 상용 인공지능의 잠재적 성적 침해 및 편향 가능성을 검증하는 등 젠더기반폭력 예방에 관한 정책 방안을 도모할 필요가 있음.
- 이에 본 연구에서는 인공지능 기술이 야기하는 젠더기반폭력, 그중에서도 성적 침해와 착취를 방지할 수 있는 정책적 프레임워크와 기술적 개입 방안을 모색하고자 하였음.
- 또한 본 연구는 경제적 이해관계에서는 중요하게 고려되기 어려운 영역인 젠더편향, 차별, 젠더폭력과 같은 분야에 대한 공공의 적극적 개입 필요성을 탐구하고자 함.

2 연구결과

가. 인공지능과 기술매개 젠더폭력

- 인공지능의 확산으로 챗봇 이루다나 테이(Tay) 사례처럼 인공지능이 젠더폭력의 행위자이자 매개가 되는 문제가 등장함.
  - 생성형 인공지능의 대중화에 따라 딥페이크 성범죄, 여성혐오 발언, 스토킹, 개인정보 탈취 등 새로운 유형의 젠더폭력이 급속히 확산되고 있음. 즉, 인공지능이 차별과 혐오를 자동화·대규모화하는 도구로 작동하고 있음.
- 현행 법제는 행위자에 대한 사후 처벌 중심이며, 생성 도구·모델의 개발·배포 단계 책임은 규율하지 않음.
- 인공지능 규제 법·정책이 젠더폭력 대응 체계와 분리되어 있어, 인공지능 규제 논의는 젠더차별·혐오·폭력 대응과 통합적으로 설계될 필요가 있음.
  - 인공지능의 성능 기준이나 산업 진흥 중심의 접근을 넘어, 젠더폭력 예방과 차별 금지를 명시적으로 포함한 윤리·사회적 안전 기준이 보완되어야 함.
  - 젠더기반폭력과 혐오의 탐지·예방·피해 구제 기능을 인공지능의 위험 평가 및 인증 체계에 반영해야 함.

나. 국내외 인공지능 규제정책 및 안전성 평가 체계

- 2025년 1월 「인공지능기본법」이 제정되어 2026년 1월 22일 시행됨.
  - 해당 법은 인간 존엄과 기본권 보호, 안전성·신뢰성 확보를 강조함과 동시에 기술 발전 및 산업 진흥을 병행함.
  - 한국은 사업자의 자율성과 산업 육성에 더 비중을 둠에 따라 젠더폭력 방지 및 사회적 안전 확보 측면에서는 보완이 필요함.
- 윤리 기준 및 가이드라인의 자율 규제 한계
  - 한국의 AI 윤리 기준은 인간 존엄, 차별 금지, 다양성 보장 등을 명시하나 실질 구속력이 약하고 사업자의 자율적 노력에 의존함.
  - 결과적으로 젠더폭력 예방 및 차별 방지의 실효성이 낮음.

「인공지능(AI) 윤리 기준」의 원칙과 요건



자료: 인공지능 윤리 소통채널, 인공지능 윤리기준, http://ai.kisdi.re.kr/aieth/main/contents.do?menuNo=400029 (검색일: 2025.6.1.).

▶ 국내 인공지능 안전성 연구 현황

- 한국지능정보사회진흥원(NIA)과 과학기술정보통신부는 2025년부터 인공지능안전연구소를 통한 LLM 안전성 및 신뢰성 평가 데이터셋 구축 사업을 추진 중임.
- 이는 향후 국가 차원의 인공지능 안전성 검증 체계의 기반이 될 것으로 기대됨.

▶ 인공지능 안전성 평가 및 벤치마크의 부상

- 인공지능 안전성 평가는 단순 성능 측정이 아닌 인권·젠더·환경 요소를 통합해 고려하는 다차원적 가치 평가로 전환 중임.
- 미국 NIST의 「인공지능 위험관리 프레임워크」는 외설, 혐오, 허위 정보 등 12개 위험 범주를 제시함.

〈표 1〉 생성형 인공지능으로 인한 주요 위험 예시

NIST 위험 범주	주요 내용	관리 방안
외설적, 모욕적 또는 학대적 콘텐츠 (Obscene, Degrading, and/or Abusive Content)	CSAM(아동 성착취물), NCII(비동의 친밀 이미지), 딥페이크 포르노그래피, 성적괴롭힘 콘텐츠	학습 데이터 정제, 콘텐츠 필터링, 피해자 지원
환각 (Confabulation)	허위 정보 생성, 특정 집단에 대한 편향된 정보	출력 검증 시스템, 사실 확인 메커니즘
인간-AI 구성 (Human-AI Configuration)	AI의 의인화, 과의존 및 정서적 유착, 젠더화된 인격을 부여받은 챗봇·AI 동반자와의 상호작용에서 발생하는 위험	인터페이스 설계 기준 마련, AI임을 인지할 수 있는 고지, 취약 이용자 보호 조치

자료: "NIST(2024), AI RMF Generative AI Profile(AI 600-1) 참고하여 재구성

- 주요 인공지능 기업(OpenAI, Anthropic, Google DeepMind 등)은 인간 피드백 강화 학습(Reinforcement Learning from Human Feedback, RLHF), 헌법적 인공지능(Constitutional AI), 모델의 위험 수준에 따른 자체 안전 정책(Responsible Scaling Policy, Preparedness Framework 등) 등 자체 안전성 검증 체계를 운영함.

▶ 한국의 인공지능 규제는 기술 진흥 중심 기조 아래 윤리·신뢰·안전의 원칙을 병행하고 있으나 실증적 안전성 검증과 데이터 기반 평가 체계 제도화는 미흡함.

- 젠더편향을 포함하는 사회문화적 편향 항목은 일부 포함되어 있으나 젠더폭력 특화 시나리오와 지표는 부재함.
- 국제적 협력은 빠르게 진전되고 있으나, 영어권 중심 설계로 언어·문화 다양성 반영 한계가 존재하며, 한국어 및 한국 문화 맥락을 반영한 벤치마크 구축이 시급한 과제임.

3 정책제언

▶ 정부 주도 젠더폭력 방지 벤치마크 구축

- 혐오와 폭력 문제를 예방하는 것은 공공의 책무이며, 인공지능이라는 새로운 시스템적인 위험에 대응하기 위해서는 정부 주도의 구속력 있는 제도가 필요함.
- 특히 한국어 환경과 국내 문화적 맥락에 특화된 벤치마크 데이터셋 구축은 인공지능이 사회적 편견과 폭력 재생산 예방의 최소한의 기준이 될 수 있는 정책 도구임.
- 또한 구축된 벤치마크 데이터셋을 오픈 소스 형태로 공개해 인공지능 개발자나 연구자, 스타트업 등이 자신들이 구축하고 있는 모델의 안전성을 자체적으로 평가, 개선하는 데 활용토록 할 수 있음.
- 정부 주도로 구축된 젠더폭력 방지 벤치마크는 향후 국제 표준화 및 상호인정 체계로 확장될 가능성을 높일 수 있음.

▶ LLM 젠더폭력 안전성 지표 개발

- 인공지능 모델과 데이터셋이 유해 표현이나 편향을 얼마나 잘 거르고 있는지, 나아가 젠더폭력을 예방하는 데 얼마나 기여하고 있는지를 측정할 수 있는 기준 마련이 필요함.
- 포괄적 윤리 검증을 목적으로 구축된 데이터셋만으로는 젠더폭력 측면의 위험을 포착하는 데 한계가 있음
- 지표 개발과 벤치마크 데이터셋의 구축 이후 지속적인 검증과 모니터링의 차원에서, 해당 인공지능 모델의 실제 여성 이용자들이 인공지능을 사용할 때 안전함을 느끼는 정도나 신뢰 수준을 조사하는 것도 고려할 필요가 있음.

▶ 인공지능 안전성 확보를 위한 젠더폭력 방지 벤치마크 의무화

- 2026년 시행 예정인 「인공지능기본법」은 젠더기반폭력이나 평등, 차별 등을 명시하고 있지 않아 있어 해당 분야 위험 평가 의무 조항 실효성에 대해서는 확신하기 어려움.
- 현행 「인공지능기본법」에 젠더폭력의 위험을 특정하여 규율할 수 있는 명시적 조항이 부재함. 최소한의 방안 중 하나로 젠더폭력 방지 벤치마크 테스트의 의무화 등의 조치가 필요함.
- 위험 평가 제도가 과학기술정보통신부 중심으로 구성되어 있어 기술 및 산업 진흥 관점에 무게가 더 실릴 가능성이 큼. 「인공지능기본법」에서 명시하고 있는 각 위원회에는 젠더폭력이나 성평등 및 인권 전문가를 포함할 필요가 있음.
- 특히, 민간자율인공지능윤리위원회에는 젠더폭력 전문가나 관련 학계 등의 참여를 보장하는 등 성인지적 거버넌스 구조로의 개편이 필요하며 이에 관한 후속 연구가 필요함.

▶ 추진 체계 및 민관 거버넌스 마련

- 「인공지능기본법」상 국가인공지능전략위원회는 인공지능 정책의 기본계획 수립, 부처 간 협력 조정, 인공지능 신뢰성 확보를 위한 주요 정책을 심의, 의결하는 대통령 소속 법정 위원회임.
- 민간자율인공지능윤리위원회는 인공지능사업자와 교육·연구기관 등이 자율적으로 설치할 수 있는 윤리 준수 기구로, 정부의 행정적 감독을 보완하는 역할을 수행하도록 설계됨.
- 「인공지능기본법」을 비롯한 법적 규제 체계는 단순한 기술 통제나 원칙 제시에 그치지 않고, 벤치마크 인증을 포함한 기술적 평가 체계와 긴밀히 연계되는 구조로 발전할 필요가 있음.
- 특히 젠더폭력 방지와 같은 사회적 안전 과제의 경우, 법적 규제의 실효성을 높이기 위해서는 벤치마크 평가가 정책 집행과 자원 배분의 핵심 근거로 기능할 수 있어야 함.
- 벤치마크 구축 및 검증 과정에는 인공지능 개발자나 데이터 엔지니어 등 인공지능 기술 전문가뿐 아니라 젠더 및 법률 분야 전문가와 현장 활동가 등 해당 분야 다양한 전문가들의 참여가 필수적임.

- 또한 시행령 단계에서 국가인공지능위원회의 기능에 인공지능으로 인한 젠더기반폭력 예방 및 피해방지 대책을 명시하고 관련 부처 혹은 전문가의 참여를 보장할 필요가 있음.
- 「인공지능기본법」에서 규정하는 민간자율인공지능윤리위원회가 젠더폭력 대응 거버넌스의 외부 자문 및 감시의 한 축으로 기능하려면 위원회가 다학제적으로 구성되도록 가이드라인을 제시해야 함.
- 실질적인 벤치마크 구축을 위한 추진 체계 확립이 필요하며 국가인공지능위원회의 의사결정과 이에 따른 벤치마크 등 데이터셋 구축 및 운영 기능, 구축된 데이터셋의 기술적, 사회적 영향력 및 적절성을 평가, 심의하는 기능이 마련되어야 함.
- 이 같은 과정을 모두 거친 뒤 인공지능 윤리 및 안전성 의무를 민간자율인공지능윤리위원회로 이어질 수 있도록 하는 다층적인 거버넌스 구성이 필요함.

▶ 부처 간 역할 분담 및 실무 협력 체계의 구체화

- 기술 인프라(과학기술정보통신부)와 젠더 전문성(성평등가족부)을 결합한 부처 간 협력 모델 구축 및 제도화가 필요함.
- 예산 및 기술 표준(과학기술정보통신부)과 젠더폭력 범주 및 시나리오의 판단 기준(성평등가족부)으로 역할을 이원화하여 기획·검수 권한의 위임 및 제도화가 요청됨.
- 추후 신종 젠더폭력에 즉각 대응하여 정책에 반영하도록 실무급 협의체 운영 및 상시적 데이터 환류 조직 마련이 요구됨.

▶ 젠더폭력 방지 벤치마크에 기반한 인공지능 안전성 인증제도 도입

- 벤치마크 테스트가 민간 기업들에서 안정적으로 정착하기 위해서는 일정한 요건을 갖춘 인공지능 시스템에 대해 독립적 평가를 실시하고 젠더폭력 방지에 관한 특정 기준을 충족한 경우 공식 인증마크를 부여하는 등 인증제 도입을 고려할 필요가 있음.
- 인증제가 실효성을 갖기 위해서는 어드밴티지 정책을 결합하고, 부 차원에서 인증제 운영에 관한 가이드라인과 평가 기준을 마련하고 인공지능 분야 스타트업 등 중소기업을 대상으로 기술 지원과 재정 지원하는 방안을 함께 고려해야 함.
- 인증 기준은 기술 및 사회 환경에 맞춰 주기적으로 갱신해야 함. 이를 위해서 인증 획득 이후에도 관련 노력이 지속될 수 있도록 모니터링 체계를 마련해야 함.

▶ 지속적인 모니터링 및 환류 체계 확립

- 여성혐오적 표현이나 디지털성범죄 수법은 시간이 지나면서 새로운 형태로 나타날 수 있으며 인공지능 모델이 훈련하지 않은 새로운 유형의 유해 콘텐츠가 등장할 수 있어 인공지능의 젠더폭력 안전성을 지속적으로 점검하고 개선하는 체계가 필요함.
- 모니터링 주기와 프로세스를 제도화하여 지속적인 모니터링을 실시할 수 있도록 하고 독립적인 감독 기구를 지정하여 정기적으로 정책 평가보고서를 작성하도록 하는 것도 효과적 방법이 될 수 있음.

▶ 멀티모달·프론티어 인공지능 모델 평가로의 확장

- 본 연구는 텍스트 기반 대규모 언어 모델의 젠더폭력 안전성 평가에 초점을 맞추었으나, 딥페이크, 합성 이미지·음성 콘텐츠 확산 등을 통해 체감되는 인공지능 기술의 발전 방향을 고려할 때, 멀티모달 벤치마크와 평가지표가 필요함.
- 최신 프론티어 모델들은 텍스트, 이미지, 음성, 비디오를 통합 처리하는 멀티모달 역량을 핵심 특징으로 하며, 이는 젠더기반폭력의 양상을 질적으로 변화시키고 있음.
- 텍스트 기반 벤치마크 구축을 시작으로 멀티모달 평가 체계 역시 준비할 필요가 있음.

주관부처 성평등가족부 성평등정책실

관계부처 성평등가족부 성평등정책실