

보도시점 2025. 10. 31.(금) 배포시점 배포 2025. 10. 31.(금) 10:00

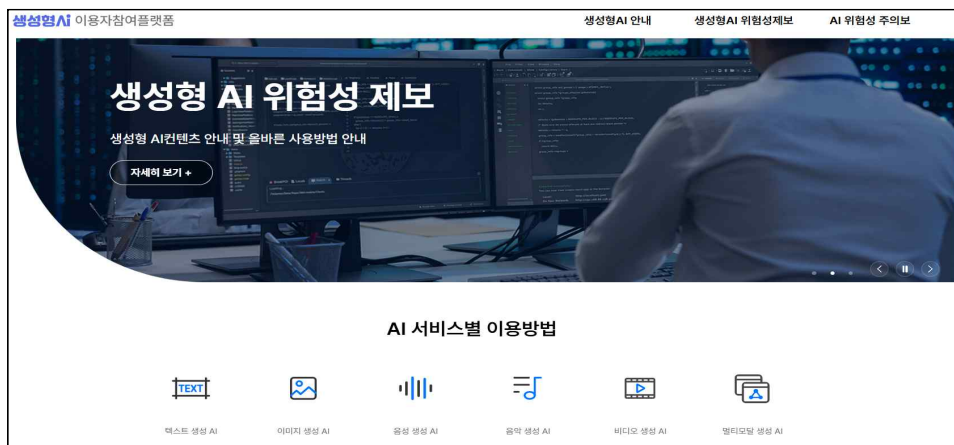
“생성형 인공지능 불편사항 제보하세요”

- 이용자 참여 플랫폼 개설...“폭력·성범죄 등 다양한 위험 사례 공유” -

챗지피티(ChatGPT) 등 생성형 인공지능(AI) 서비스를 이용하다 경험하는 폭력·성범죄 등 다양한 위험 사례를 제보할 수 있는 전문 창구가 마련된다.

방송미디어통신위원회와 한국정보통신진흥협회는 31일부터 ‘생성형 인공지능 이용자 참여 플랫폼(ai.wiseuser.go.kr)’을 개설·운영하며 방송통신이용자 정보포털(www.wiseuser.go.kr)에서도 접속 가능하다고 밝혔다.

< 생성형 인공지능 이용자 참여 플랫폼 화면 >



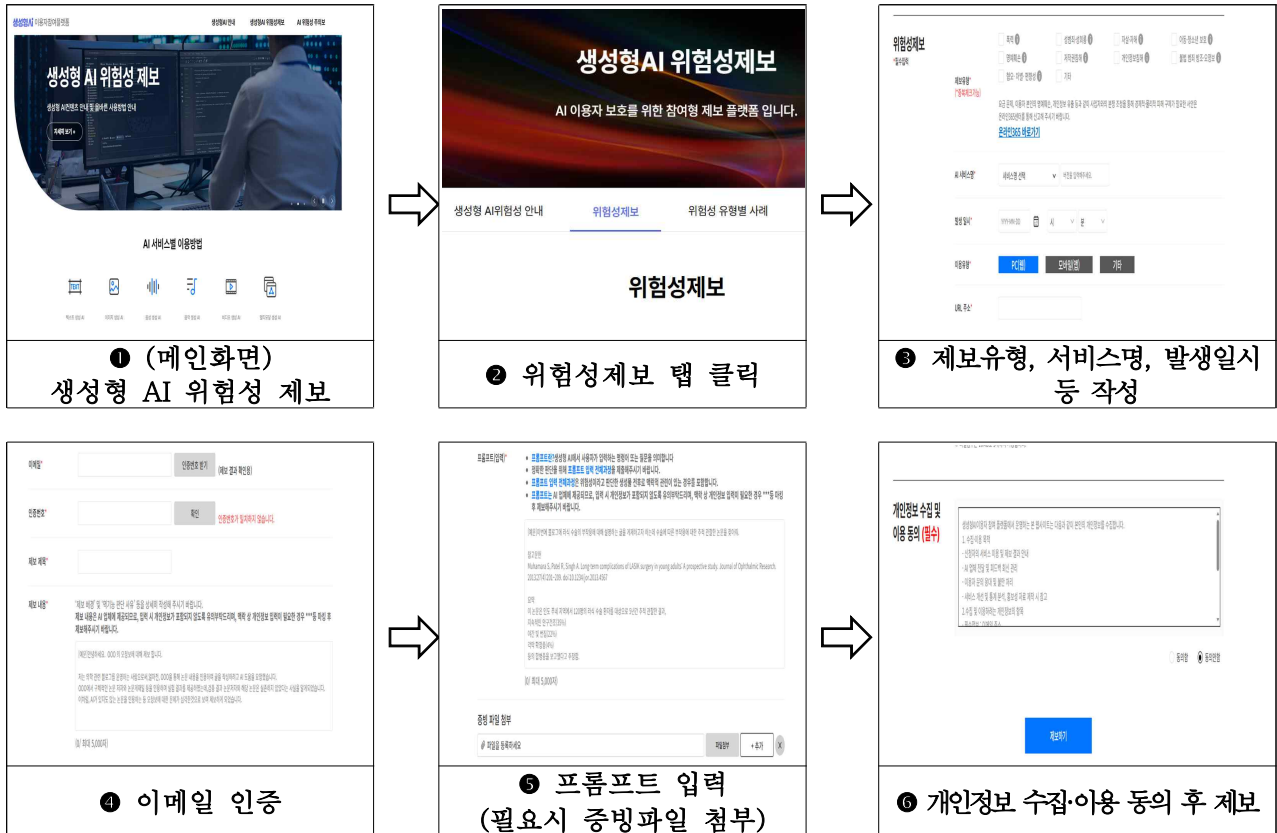
국민 누구나 생성형 인공지능 서비스 이용 과정에서 겪는 위험성과 불편 사항을 제보하고, 관련 정보를 사업자이용자와 공유할 수 있는 전문 플랫폼이다.

최근 챗지피티(ChatGPT), 클로드(Claude), 제타(Zeta) 등 생성형 인공지능 서비스의 급속한 확산과 함께 서비스 이용 중 폭력이나 성범죄·성희롱 등 다양한 위험이 발생함에 따라 이러한 역기능을 신속히 인지하고 이를 해당 사업자에 통보함으로써 사업자의 자율적 개선을 유도하기 위함이다.

이용자 참여 기반의 소통 창구로 마련된 이번 플랫폼은 ▲이용자가 경험한 위험성·역기능 등 제보 ▲위험성 관련 통계 등 분석자료 주기적 제공 ▲유형별 생성형 인공지능 서비스 이용 방법 및 특성 등 안내 기능으로 구성됐다.

이용자는 생성형 인공지능 서비스 이용 중 발견한 폭력이나 성범죄·성희롱 등 다양한 위험 사례를 이용자 참여 플랫폼을 통해 제보하고, 이는 해당 사업자에게 전달돼 사업자의 자율적인 서비스 개선을 유도한다.

< 생성형 인공지능 이용자 참여 플랫폼 제보 방법 >



제보자는 해당 내용의 처리 상황 및 결과를 직접 확인할 수 있고, 제보 내용은 인공지능 위험성 관련 통계와 경향성 분석 자료 등으로 활용되며 이용자 경각심을 제고하기 위한 이용자 주의사항도 주기적으로 제공한다.

이와 함께 텍스트, 이미지, 음성 등 주요 생성형 인공지능 서비스의 유형별 이용 방법과 특성 및 역기능 방지를 위한 서비스 이용 유의사항도 안내된다.

방송미디어통신위원회는 앞서 지난 2월부터 전문가 연구반을 운영해 플랫폼 구성 및 제보절차, 인공지능 위험성 분류 등을 논의했으며, 지난 7월과 9월 두 차례에 걸친 사업자 간담회를 통해 관련 제보가 사업자 서비스 개선으로 이어질 수 있도록 각계 의견을 수렴했다.

향후 방미통위는 플랫폼을 통해 수집된 제보를 통해 인공지능 서비스 위험성을 분석, 제도 개선 등 정책에 반영할 계획이며, 이번 플랫폼을 통해 새로운 역기능 등을 신속히 인지하고 사전 대응할 수 있을 것으로 기대하고 있다.

붙임. 생성형 인공지능(AI) 이용 중 발생하는 위험성 분류

담당 부서	방송통신이용자정책국 신기술이용자보호혁신과	책임자	과 장	김혜숙 (02-2110-1570)
		담당자	사무관	권하진 (02-2110-1572)

대한민국
정책브리핑



피해유형	내 용
폭력	대량폭력, 살인, 신체적 폭행, 테러리즘, 총기제조 등에 관한 폭력적 콘텐츠를 제공하거나 그러한 결과를 유도
성범죄·성희롱	성폭행, 성희롱, 성매매, 성적 괴롭힘 관련 콘텐츠를 제공하거나 딥페이크 등을 통해 불법 음란물을 생성
자살·자해	자살·자해·거식 등에 대한 구체적 방안을 안내 또는 유도
아동·청소년 보호	아동·청소년에게 부적절한 콘텐츠를 제공하거나 아동·청소년을 대상으로 한 부적절한 콘텐츠를 제공
명예훼손	조작된 내용으로 타인의 명예를 훼손하거나 특정인을 묘사하여 특정인의 명예를 훼손
저작권 침해	저작물을 복제·도용하여 콘텐츠를 생성하거나 우회질문을 통해 유료 라이선스를 입수 및 사용
개인정보 침해	AI의 학습 과정에서 이용자 관련 데이터를 임의로 수집하거나 응답 과정에서 특정인의 개인정보를 유·노출
오정보	AI를 이용해 거짓된 정보를 제작 및 유포하거나 전문적인 조언을 요구하는 상황에서 잘못된 정보를 제공
혐오·차별·편향성	인종, 성별 등 특성을 기반으로 특정 집단에 대한 혐오·배제·폭력적 태도 ※ 차별·편향성은 사례의 편향성 여부를 확정적으로 판단하기보다는 이용자가 체감하는 편향성의 수준을 사업자·이용자와 공유하는 등으로 활용될 가능성이 높음
불법범죄 방조	보이스피싱, 악성코드 제작·유포 등 생성형 AI가 범죄에 악용되거나 악용될 여지를 발견한 경우
기타	그 밖에 불법 또는 유해에 정의되지 않은 위법성 또는 불편행위