인공지능 투명성 확보 가이드라인

과학기술정보통신부

TTA 한국정보통신기술협회
Telecommunications Technology Association

인공지능 투명성 확보 가이드라인

CONTENTS

제1장 | 개요

- 적용 대상 및 범위 1

제2장 | 투명성 조항별 적용 사례

- 사전고지 사례 (법 제31조 제1항 사례) 3
- 생성형AI 표시 사례 (법 제31조 제2항 사례) 4
- 실제와 구분이 어려운 결과물의 경우 (법 제31조 제3항 사례) 9
- 사전고지·표시를 하지 않아도 되는 경우 (법 제31조 제4항 사례) 11

제3장 | FAQ

- 공통사항 13

제4장 | 부록

- 인공지능 생성 서비스 사례 14
- 인공지능 생성 콘텐츠 위험사례 16
- 인공지능 생성 콘텐츠 국내외 법안 현황 20
- 인공지능 생성 콘텐츠 국제 사회 논의 22
- 국외 온라인 플랫폼 사업자 투명성 확보 사례 23

본 가이드라인(안)은 여러 전문가, 실무자 및 이해관계자의 의견을 수렴하고, 관련 해외 동향, AI기술 발전, 서비스 형태 등을 반영하여 현재 작성 과정에 있는 자료이며, 본 가이드라인(안)에 수록된 내용은 최종 공개본의 내용과 상이할 수 있음

1장 개요

적용 대상 및 범위

- 제공하는 인공지능 서비스에 대한 투명성 확보 의무는 이용자에게 최종적으로 AI 제품 및 서비스를 제공하는 사업자에 부과

  예) 개발사A가 직접 개발한 파운데이션 모델의 API를 제공하고, 이를 활용하여 B사가 AI 제품·서비스를 이용자에게 제공할 경우 투명성에 대한 의무는 B사에게 부여

- 다만, 파운데이션 모델 개발자가 직접 이용자에게 서비스를 제공하는 경우에는 직접적인 투명성 확보 의무가 발생

  예) 개발사A가 자신이 개발한 파운데이션 모델을 활용해 이용자에게 직접 서비스를 제공할 경우 개발사A 또한 투명성 확보에 대한 의무 부여

| 생성형 AI 서비스 이용관계도 |

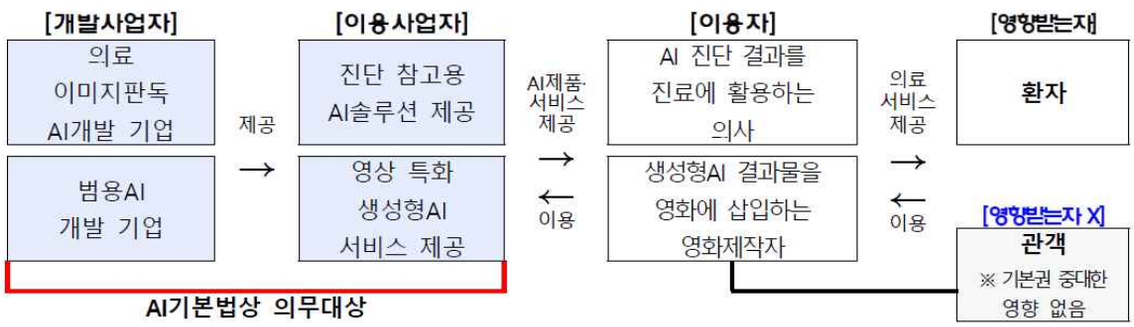
구분	AI개발 사업자	AI 이용사업자	이용자	영향받는자
해당 기업·개인 예시	네이버, OpenAI 등	뤼튼, 포티투마루 등	AI제품·서비스 이용자 (CG제작자, 교사, 의사)	제품·서비스 기본권에 영향을 받는자
제품·서비스 예시	클로바, 챗GPT 등*	생성형AI 기반 에이전트·질의응답 서비스		

* AI개발사업자가 이용자에게 직접 AI서비스 제공시(B2C) AI이용사업자 지위도 함께 가지게 되어 사전고지 및 표시 의무 부담

관련 AI 기본법상 이해관계자 정의 (제2조)

- 인공지능개발사업자(제7호 가목): 인공지능을 개발하여 제공하는 자
- 인공지능이용사업자(제7호 나목): 사업자가 제공한 인공지능을 이용하여 인공지능제품 또는 인공지능서비스를 제공하는 자
- 이용자(제8호): 인공지능제품 또는 인공지능서비스를 제공받는 자
- 영향받는 자(제9호): 인공지능제품 또는 서비스에 의하여 자신의 생명, 신체의 안전 및 기본권에 중대한 영향을 받는 자

- **(투명성 확보 대상 예시)** AI기본법은 AI개발사업자와 이용사업자에 의무를 부과하며, 단순히 AI제품·서비스를 이용한 결과물을 자신의 서비스 등에 활용하는 자는 AI기본법상 사업자에 미해당하여 투명성 확보 의무가 없음
  - AI를 이용하여 CG를 생성한 후 영화에 삽입한 영화제작자는 단순히 AI 서비스를 이용한 결과물을 자신의 콘텐츠에 활용한 것 이므로 AI기본법 상 사업자가 아니며 이용자에 해당

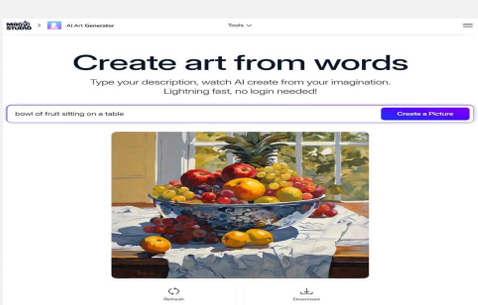


- **(사전고지와 표시의 차이)** 사전고지는 AI제품·서비스를 이용자에게 제공 전 해당 제품·서비스가 생성형 또는 고영향AI에 기반함을 이용자에게 알리는 행위
  - 표시는 AI제품·서비스 제공 중 생성된 결과물이 생성형AI를 통해 생성된 것임을 알리는 표시를 의미

| 사전고지와 표시 구분 예시 |

서비스	전화 연결로 이용하는 생성형 AI 기반 음성 대화 서비스
사전고지	서비스 가입 과정에서 약관 등에 생성형AI 기반임을 명시
표시	서비스 이용을 위한 통화연결 직후 '대화 내용은 생성형AI를 활용하여 만들어진다'는 멘트 재생

- **(생성 결과물 투명성 확보 단위)** 최종 결과물을 기준으로 해당 결과물이 생성형AI 등을 통해 작성되었음을 표시하는 등의 투명성 확보 수행
  - **(예시)** 최종 결과물 이전에 이용자가 프롬프트 입력, 생성 이미지 선택 등을 하는 단계에서는 표시할 필요가 없고 마지막에 최종 결과물 도출(추출, 저장 등) 단계에서 표시



❶ 이미지 생성(매직 스튜디오 서비스)

❷ 이미지 다운로드시 워터마크 표시(오른쪽 하단)

2장 투명성 조항별 적용 사례

사전고지 사례 (법 제31조제1항 사례)

AI기본법 인공지능 투명성 확보 의무(제31조제1항)

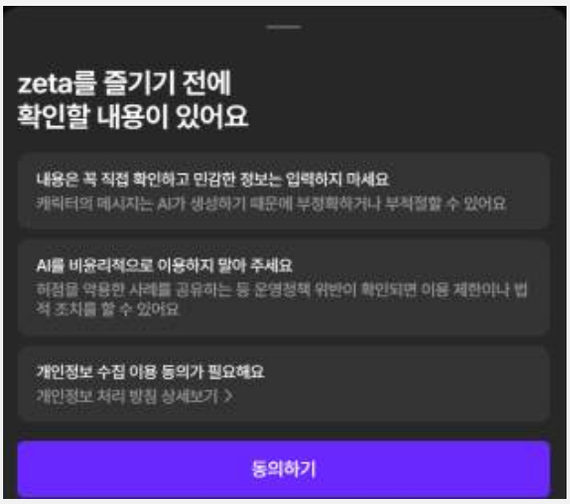
- **(법 제31조제1항)** 인공지능사업자는 고영향 인공지능이나 생성형 인공지능을 이용한 제품 또는 서비스를 제공하려는 경우 제품 또는 서비스가 해당 인공지능에 기반하여 운용된다는 사실을 이용자에게 **사전에 고지하여야 한다.**
  - **(시행령 제22조)** 인공지능사업자는 고영향 인공지능이나 생성형 인공지능을 이용한 제품 또는 서비스(이하 "제품 등"이라 한다)를 제공하기 전에 다음 각 호의 어느 하나의 방법으로 법 제31조제1항에 따른 사전고지를 하여야 한다.
    1. 제품 등에 직접 기재하거나, 계약서, 사용 설명서, 이용약관 등에 기재
    2. 이용자의 화면 또는 단말기 등에 표시
    3. 제품 등을 제공하는 장소(해당 장소와 합리적으로 관련된 범위의 장소를 포함한다)에 인식하기 쉬운 방법으로 게시
    4. 그 밖에 제품등의 특성을 고려하여 과학기술정보통신부장관이 인정하는 방법

○ **(사전고지 취지)** 제품·서비스가 고영향 또는 생성형AI에 기반되어 운용되는 것임을 사전에 알려 이용자가 주의를 기울일 수 있도록 함

❶ **(이용약관·계약서 활용)** 제공하는 서비스의 이용약관, 서비스 가입 절차 및 계약서에 생성형AI 또는 고영향AI를 활용함을 명시

❷ **(화면 표시)** S/W, 모바일 어플리케이션 등을 통해 서비스 제공시 화면상에 인공지능을 활용함을 명시

❸ **(고영향AI)** 고영향AI도 생성형AI와 같이 계약서나 제품·서비스 등에 기재하는 방식을 통해 '사전고지' 가능



❶ 이용약관 – 스캐터랩 제타(Zeta)

❷ 화면 표시 – 카카오 타임톡

생성형AI 표시 사례 (법 제31조제2항 사례)

AI기본법	인공지능 투명성 확보 의무(제31조제2항)

- **(법 제31조제2항)** 인공지능사업자는 생성형 인공지능 또는 이를 이용한 제품 또는 서비스를 제공하는 경우 그 결과물이 생성형 **인공지능에 의하여 생성되었다는 사실을 표시하여야 한다.**
  - **(시행령 제22조)** 인공지능사업자가 법 제31조제2항에 따른 표시(법 제31조제3항에 따른 실제와 구분하기 어려운 결과물을 제공하는 경우로서 해당 결과물이 인공지능시스템에 의하여 생성되었다는 사실을 이용자가 명확하게 인식할 수 있는 방식으로 고지 또는 표시하는 경우에는 적용하지 아니한다)를 할 때에는 다음 각호 중 하나의 방법으로 할 수 있다.
    1. 사람이 인식할 수 있는 방법
    2. 기계가 판독할 수 있는 방법. 다만, 이 경우에는 생성형 인공지능에 의하여 생성되었다는 사실을 1회 이상 안내 문구·음성 등으로 제공하여야 한다.

- **(생성형AI 표시 취지)** 이미지, 음성 등이 생성형AI를 통해 작성되었음을 명확히 하여 혼란을 예방하고, 딥페이크 등 악용 방지

  ※ 콘텐츠 형식별로 표시 방식에 차이가 있으며, 시각·청각적으로 직접 표시하는 방식과, S/W를 통해 확인 가능하도록 하는 비가시적 표시도 가능

- **(기계가 판독할 수 있는 방법 사용시)** 기계가 판독할 수 있는 방법인 비가시적 표시를 할 경우 해당 결과물이 AI결과물임을 알 수 있도록 사전에 안내 필요

❶ **(이미지 콘텐츠)** 이미지 위에 로고, 텍스트 등의 가시적 워터마크*를 표시하거나, 특정 기술을 활용한 비가시적 워터마크 적용

- **(비가시적 방식 사용시)** 해당 콘텐츠가 생성형 인공지능을 통해 생성된 것임을 인식할 수 있도록 생성 결과물에 관련 문구 등을 포함하여 안내

※ **(예시)** 생성 결과 이미지에 "이 그림은 AI로 생성되었습니다."와 같이 생성형AI를 활용 사실 포함하여 최종 결과로 사용자에 전달(예시 마지막 참고)

※ **워터마크:** 이미지 등의 매체에 로고나 텍스트 형태의 식별자를 추가하거나 사람이 인지할 수 없는 패턴을 삽입하여 콘텐츠의 출처 및 소유권 확인에 사용

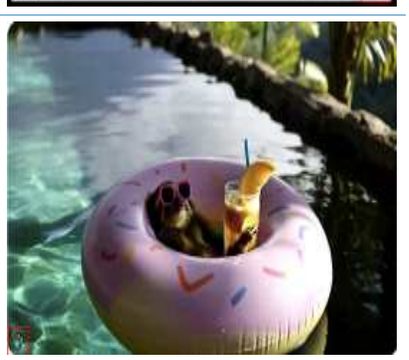
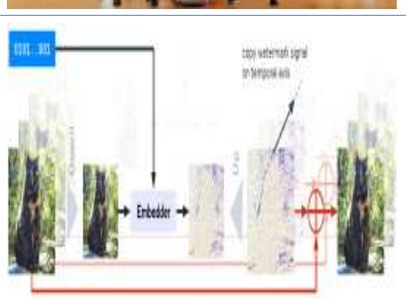
표시 사례	표시 방법	예 시
갤럭시AI (삼성전자)	갤럭시 스마트폰의 사진 앱에서 '포토 어시스트' 기능 활용시 좌측 하단 갤럭시 AI의 로고( ) 및 문구 삽입	
에이닷 (SKT)	이용자의 질문과 요청에 맞춰 날씨, 뉴스, 음악 재생 및 사진 편집 등의 개인 맞춤형 서비스로 좌측 하단 로고 삽입	
Prome AI	생성형 인공지능 기술을 통해 이미지 또는 비디오를 생성하는 서비스로, 중앙 하단에 'Prome AI( )' 로고를 삽입	

표시 사례	표시 방법	예 시
메타AI (메타)	메타 AI를 사용해 만든 이미지에 'Imagined with AI' 라벨( )을 표시함으로써 해당 이미지가 AI로 생성되었음을 표시	
통이완샹 (Tongyi Wanxiang)	알리바바(Alibaba)가 개발한 AI 이미지 생성 서비스로 해당 플랫폼은 오른쪽 하단에 '万相AI生成' 워터마크 삽입	
웬신이지 (Wenxin Yige)	바이두(Baidu)가 개발한 인공지능 기반의 이미지 생성 서비스이며 오른쪽 하단에 '文心一格' 워터마크 삽입	
이미지 사전 안내 예시 (제2조 제2항 제2호에 따른 안내)	생성형 인공지능 기술을 활용해 이미지 생성 시 비가시적 방식의 표시를 하는 경우, 이용자에게 결과물만 전달하는 것이 아니라 해당 결과물이 생성형 인공지능 기술을 활용해 제작되었음을 알리기 위한 관련 문구를 **결과물 일부에 안내** (C2PA, SynthID 등 비가시적 표시 기술 적용 시)	

❷ **(동영상 콘텐츠)** 이미지와 유사하게 로고나 텍스트 등의 가시가능한 워터마크 및 특정 기술을 활용한 가시불가능한 워터마크 적용

- **(비가시적 방식 사용시)** 반드시 동영상 전체에 표시할 필요는 없으며, 이용자가 해당 콘텐츠가 생성형 인공지능을 통해 생성된 것임을 인식할 수 있도록 사전에 안내

※ **(예시)** 영상 시작 부분에 "이 애니메이션은 AI로 생성되었습니다."와 같이 생성형AI를 활용 사실을 멘트로 간단히 안내(예시 마지막 참고)

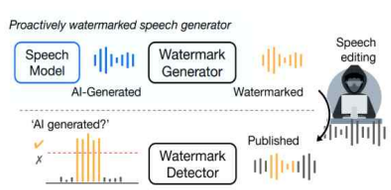
표시 사례	표시 방법	예 시
매직미디어 (캔바)	프롬프트 및 이미지를 이용한 동영상 생성 서비스로 우측 하단 가시적 워터마크 삽입	
무비젠 (메타)	메타의 동영상 생성형 AI 모델로 인스타그램에 도입 예정이며, 좌측 하단 가시적 워터마크( ) 삽입	
드림 머신 (루마 랩스)	드림 머신은 이미지를 업로드하거나 원하는 시나리오를 텍스트로 입력하면, 약 5초 길이의 동영상을 생성, 인지 가능 워터마크 도입하여 오른쪽 상단에 'LUMA' 워터마크 삽입	
비디오 실 (Video Seal) (메타)	모든 고해상도 프레임에 워터마크를 삽입할 필요 없이, 특정 키 프레임에 워터마크를 삽입하는 효율적인 비디오 워터마킹 기술 개발 (이 경우 생성형 AI 결과물이라는 사실을 시작 부분 등에 추가 안내 필요)	

표시 사례	표시 방법	예 시
동영상 사전 안내 예시 (제2조 제2항 제2호에 따른 안내)	생성형 인공지능 기술을 활용해 동영상 생성 시 비가시적 방식의 표시를 하는 경우, 이용자에게 결과물만 전달하는 것이 아니라 해당 결과물이 생성형 인공지능 기술을 활용해 제작되었음을 알리기 위한 관련 문구를 영상 시작 사전에 안내 (비가시적 동영상 표시 기술 적용 시)	

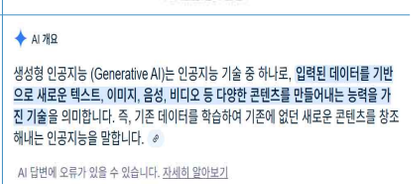
❸ **(오디오 콘텐츠)** 콘텐츠 재생 초기 생성형 AI 결과물이란 사실을 음성으로 알리거나, 비가청적 방식의 기술적 워터마크 활용 가능

- **(비가청적 방식 사용시)** 동영상과 동일하게 오디오 전체에 표시할 필요는 없으며, 이용자가 해당 콘텐츠가 생성형 인공지능을 통해 생성된 것임을 인식할 수 있도록 사전에 안내

※ **(예시)** 오디오 콘텐츠 시작 부분에 AI로 만든 음성 결과물 시작 시 "이 음성은 AI로 생성되었습니다"와 같이 생성형AI를 활용 사실을 멘트로 간단히 안내

표시 사례	표시 방법	예 시
오디오 씰 (메타)	오디오 콘텐츠에 대한 인지불가능 워터마크 오픈소스 기술 개발 (이 경우 생성형AI 결과물임을 시작 부분 등에 안내 필요)	

❹ **(텍스트 콘텐츠)** 챗봇 등의 경우 이용 화면이나 메시지 창을 통해서 답변이 생성형AI를 통해 생성됨을 표시 중

표시 사례	표시 방법	예 시
AI 브리핑 (네이버)	화면에 'AI 브리핑'이라는 서비스명을 명시하고, info 이미지를 클릭하면 "여러 문서를 기반으로 AI 모델이 요약한 결과물로서...(중략) 다소 부정확, 부적절한 정보가 포함될 수 있습니다"라고 서비스 설명을 제시	
구글 검색 AI개요 (구글)	AI 개요를 통해 검색 결과에 대한 필요한 모든 정보를 종합하여 보여주며, 'AI개요' 및 하단에 'AI 답변에 오류가 있을 수 있습니다.' 문구 표시	

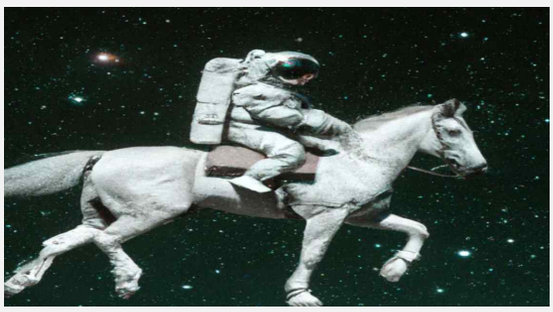
실제와 구분이 어려운 결과물의 경우 (법 제31조제3항 사례)

AI기본법 인공지능 투명성 확보 의무(제31조제3항)

- **(법 제31조제3항)** 인공지능사업자는 인공지능시스템을 이용하여 실제와 구분하기 어려운 가상의 음향, 이미지 또는 영상 등의 결과물을 제공하는 경우 해당 결과물이 **인공지능시스템에 의하여 생성되었다는 사실을 이용자가 명확하게 인식할 수 있는 방식으로 고지 또는 표시**하여야 한다. 이 경우 해당 결과물이 예술적 · 창의적 표현물에 해당하거나 그 일부를 구성하는 경우에는 **전시 또는 향유 등을 저해하지 아니하는 방식으로 고지 또는 표시할 수 있다.**
  - **(시행령 제22조)** 법 제31조제3항에 따른 고지 또는 표시는 인공지능사업자가 다음 각 호의 사항을 고려하여 이용자가 명확하기 인식할 수 있는 방식으로 정해야 한다.
    1. 이용자가 시각, 청각 등을 통하거나 소프트웨어 등을 이용하여 쉽게 내용을 확인할 수 있는 방법으로 고지 또는 표시할 것
    2. 주된 이용자의 연령, 신체적 · 사회적 조건 등을 고려하여 고시 또는 표시할 것

- **(취지)** 딥페이크 등에 악용될 수 있는 생성형AI의 결과물에 대해 AI를 활용해서 작성된 것임을 이용자가 명확히 인식할 수 있도록 하되,
  - 예술적·창의적 결과물에 대해서 이용자가 창작물을 향유할 권리를 침해하지 않도록 예외 규정
- **(실제와 구분하기 어려운 경우)** 사람의 신체·목소리를 실제와 구별하기 어려울 정도로 구조와 특성을 모방한 콘텐츠를 생성하거나, 인공물·자연물·자연현상 등을 이용자로 하여금 실재(實在)하는 것으로 오인하게 할 소지가 있을 정도로 모방한 콘텐츠

- **딥페이크 등을 예방하기 위한 취지에서 반영된 조항인 만큼, 콘텐츠의 내용이 아닌 사실로 오인하게 하는 콘텐츠가 적용 대상**
  - 한국에서 계속 활동하는 인물이 해외에서 활동하는 영상을 생성하는 경우 해당
  - 외국어를 할 줄 모르는 배우의 목소리를 활용한 외국어 음성 콘텐츠를 생성하는 경우 해당
  - 모방 대상이 실재하지 않더라도 이용자로 하여금 생성한 콘텐츠의 내용이 사실로 오인하게 할 소지가 있는 경우
- **딥페이크에 해당되지 않는 사례**
  - 2D, 3D 애니메이션, 직접 조작하는 게임
  - 가상임이 명확한 배경(화성, 우주 등)
  - 공지(共知)의 사실로서 불가능한 자연물·자연현상에 관한 것(SF 영화 등에 등장하는 여러 개의 달·태양 표현, 거대 괴수 등)



(딥페이크 미해당 사례) 우주에서 말타는 우주인과 두 개의 태양이 비치는 강

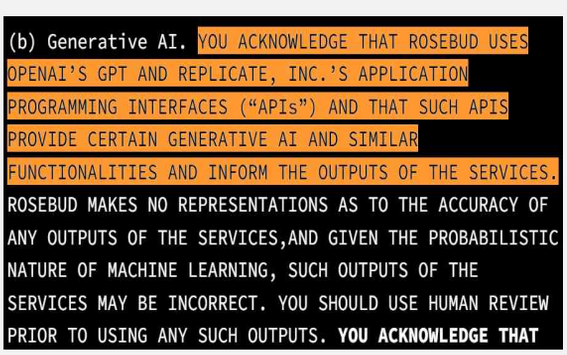
- **(예술적·창의적 표현물)** 미술, 영화, 문학, 사진, 출판, 만화, 게임, 애니메이션 등 창의적 표현 활동에 따른 결과물

※ 문화예술진흥법 제2조(정의) 제1항 제2호

"문화예술"이란 문학, 미술(응용미술을 포함한다), 음악, 무용, 연극, 영화, 연예(演藝), 국악, 사진, 건축, 어문(語文), 출판, 만화, 게임, 애니메이션 및 뮤지컬 등 지적, 정신적,심미적 감상과 의미의 소통을 목적으로 개인이나 집단이 자신 또는 타인의 인상(印象), 견문, 경험 등을 바탕으로 수행한 창의적 표현활동과 그 결과물

- **(전시·향유 등을 저해하지 않는 방식으로 고지 또는 표시)** 결과물이 예술적 · 창의적 표현물에 해당하거나 그 일부를 구성하는 경우 전시 또는 향유 등을 저해하지 아니하는 방식으로 고지 또는 표시 가능

(서비스 이용 방법 안내) Adobe Firefly 서비스 화면



(이용약관) 게임 제작 도구 Rosebud AI 이용약관

- 제품 패키지에 표시하여 안내
- 회원가입 시, 서비스 이용약관 내 고지
- 서비스 온보딩 시, 서비스 이용 방법 및 AI기반 운용 안내
- 세션 연결 시, AI생성 결과물 및 오류 가능성 안내
- 캡션, 주석, 설명 추가 ⇨ 실제 콘텐츠 향유 과정에서는 방해를 주지 않는 방식

- **(주된 이용자의 연령, 신체적, 사회적 조건 등을 고려하여 표시)** 정보 접근성과 이해도를 높이고 오인·혼동을 방지하기 위해 사용자 특성을 고려한 사전 고지 및 표시 필요

※ 사용자 특성 분석을 위한 고려사항 예시

구분	상세 구분	고려사항
연령	아동, 성인, 노인 등	• 아동 또는 노인의 경우, 성인과 비교해 이해할 수 있는 어휘, 단어에 한계가 있어 사용자 연령을 고려해야 함
장애 유무	장애인, 비장애인	• 신체적 제약으로 발생할 수 있는 한계를 고려해야 함. 그 예로는 신체 크기, 신체 능력, 인지능력이 있음
지식	초보자, 전문가 등	• 관련 서비스의 경험 여부와 사전 배경지식의 차이로 지식수준이 다름을 고려해야 함

출처: 2024 신뢰할 수 있는 인공지능 개발안내서(일반분야), 한국정보통신기술협회(TTA)

사전고지·표시를 하지 않아도 되는 경우 (법 제31조제4항 사례)

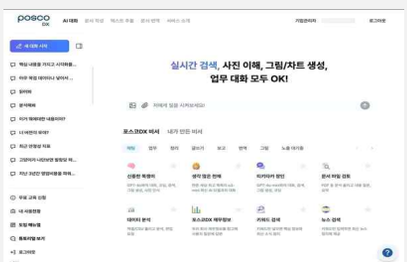
AI기본법 인공지능 투명성 확보 의무(제31조제4항)

- **(법 제31조제4항)** 그 밖에 제1항에 따른 사전고지, 제2항에 따른 표시, 제3항에 따른 고지 또는 표시의 방법 및 그 예외 등에 관하여 필요한 사항은 대통령령으로 정한다.
  - **(시행령 제22조)** 법 제31조제1항부터 제3항까지는 다음 각 호에 해당하는 경우에는 적용하지 아니한다. 다만, 제3호의 경우에는 법 제31조제1항부터 제3항까지 중 전부 또는 일부를 적용하지 아니할 수 있다.
    1. 제품 · 서비스명, 이용자 화면이나 제품 겉면 및 결과물에 표시된 문구 등을 고려할 때 고영향 인공지능 또는 생성형 인공지능을 활용한 사실이 명백한 경우
    2. 인공지능사업자의 내부 업무 용도로만 사용되는 경우
    3. 그 외 제품 등의 유형 · 특성이나 결과물의 내용, 이용형태 및 기술 수준 등을 고려하여 법 제31조제1항부터 제3항까지 중 전부 또는 일부에 대한 적용 예외가 필요한 사항으로 과학기술정보통신부장관이 정하여 고시하는 경우

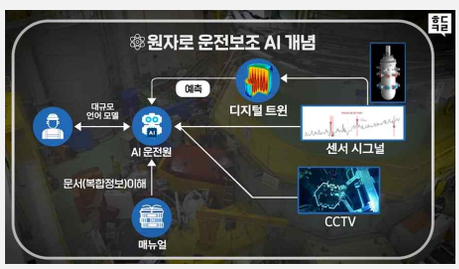
- **(예외 마련 취지)** AI 사용이 명백한 경우나 사업자가 자신의 내부업무를 처리하기 위해 AI를 활용하는 등 사전고지 및 표시가 필요 없는 예외적인 경우에 투명성 확보 의무를 완화

(사업자 내부 업무 전용) 인공지능사업자가 내부 업무를 위해 인공지능 서비스 사용시 의무 완화

- 직접 개발한 AI 서비스를 사내 직원들만 활용하는 경우
- 내부적으로 연구개발을 위해 AI가 사용된 경우



(포스코DX 웍스AI) 업무 효율화를 위한 사내 에이전트 구축 및 운영



(한국원자력연구원 아토믹GPT) 내부 업무 효율화 외에도 AI 운전원 연구 개발에 활용

3장 FAQ

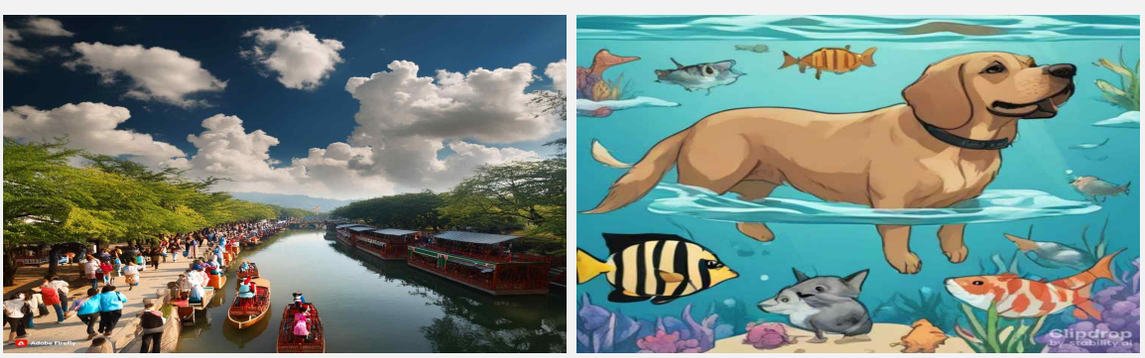
※ 해당 장은 향후 진행될 의견 수렴과정에서 나온 질문에 대해 검토 후 반영 예정입니다.

4장 부록

인공지능 생성 서비스 사례

이미지 기반 서비스

- 인공지능 기술에 기반하여 예술 작품이나 사실적인 사진, 복잡한 이미지나 효과를 적용하는 등의 방식으로 새로운 이미지를 창작 및 수정
- 대표적인 이미지 인공지능 콘텐츠 생성 서비스로 DALL·E[OpenAI], FireFly[Adobe], Stable Diffusion[Stability AI] 등의 플랫폼이 존재



이미지 생성형 인공지능 서비스를 통한 이미지 생성(좌: Firefly, 우: Stable Diffusion)

동영상 기반 서비스

- 영상을 생성하거나 영상의 변환·편집 등 작업을 수행하며[1], 주로 사용자가 작성한 각본(Script)과 입력한 명령어에 따라 영상을 자동으로 생성[2]
- 텍스트(Text-To-Video) 혹은 이미지(Image-To-Video)를 바탕으로 동영상을 생성하는 서비스는 Sora[OpenAI], VideoFx/Veo3Google, Stable Video Diffusion[Stability AI] 등의 플랫폼이 존재

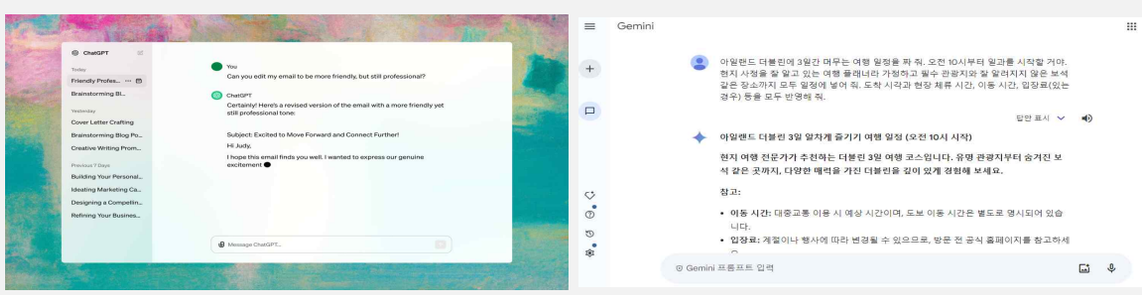
1) WANG JIAJIA. (2024). 정보 특성과 개인적 특성이 생성형 AI 서비스 사용의도에 미치는 영향에 관한 실증적 연구(박사). 세종대학교 대학원, 박사학위논문
2) 손한빈. (2023). 생성형 인공지능의 서예 콘텐츠 초탐. 동양예술, (59), 5-30.

동영상 생성형 인공지능 서비스 Sora를 통한 동영상 생성 예시	
	• 아래 프롬프트로 생성된 영상 중 한 장면 • "한 스타일리시한 여성이 따뜻하고 빛나는 네온과 애니메이션 도시 간판으로 가득 찬 도쿄 거리를 걷고 있습니다. 그녀는 검은색 가죽 재킷, 긴 빨간색 드레스, 검은색 부츠를 신고 검은색 지갑을 들고 있습니다. 선글라스와 빨간색 립스틱을 바릅니다. 그녀는 자신감 넘치고 캐주얼하게 걷습니다. 거리는 축축하고 반사적 이어서 다채로운 조명의 거울 효과를 만들어 냅니다. 많은 보행자가 걸어 다닙니다."

출처: Open AI 공식 홈페이지

텍스트 기반 서비스

- 대규모 텍스트 데이터를 기반으로 한 언어 모델을 활용해 인간과 유사한 레벨에서 언어를 이해하고 생성 및 응답함으로써 높은 성능을 발휘
- 대표적으로 ChatGPT[OpenAI], Gemini[Google], Claude[Claude], HyperCLOVAX[Naver] 등의 서비스가 존재



텍스트 생성형 인공지능 서비스를 통한 텍스트 생성(좌: ChatGPT, 우: Gemini)

오디오 기반 서비스

- 인공지능 기술을 사용하여 사람의 목소리나 음악을 학습하여 새로운 음성이나 오디오 생성
- 텍스트(Text-To-Speech)를 바탕으로 가상 비서, 음성 보조 기기 등에 활용되고 있으며, WaveNet[Google], AI Speech[Microsoft Azure], Amazon Polly[AWS] 등이 존재

멀티모달 기반 서비스

- 멀티모달 인공지능은 텍스트, 이미지, 영상, 음성 등 다양한 데이터 모달리티를 함께 고려하여 서로의 관계성을 학습 및 표현하는 기술[3)]
- 텍스트 중심의 단일 모달 처리 위주 서비스에서 사람과 유사하게 종합적으로 문제를 이해하고 해결하기 위한 GPT-4o[OpenAI]와 같은 플랫폼 출현

3) 멀티모달 AI가 바꿀 미래, 네이버, https://channeltech.naver.com/contentDetail/25

인공지능 생성 콘텐츠 위험사례

가짜뉴스 생성(1)

- 가짜뉴스는 뉴스 형태로 제시된 허위 또는 오해의 소지가 있는 정보이며, 종종 개인이나 단체의 평판을 손상시키거나 광고 수익을 통해 돈을 벌려는 목적을 포함[4)]
- 가짜뉴스에 대한 시민의 인식을 확인하기 위한 통계 조사 결과[5)], 가짜뉴스를 경험해 보았다는 비율이 43.3%에 달하고, 인공지능이 발달하면서 가짜뉴스 피해도 심화될 것이라는 전망에 84.2%가 동의

⇒ 통계 사례로 볼 수 있듯이 가짜뉴스는 이미 대중의 생활 곳곳에 침투하고 있어, 이에 대한 대책 마련이 시급

사례 예시(펜타곤 폭발 가짜뉴스 사건)

(출처 : 월드뷰)

펜타곤 폭발 가짜뉴스 사건 및 증시 변동[6)]

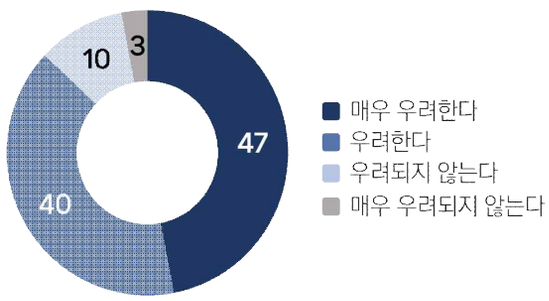
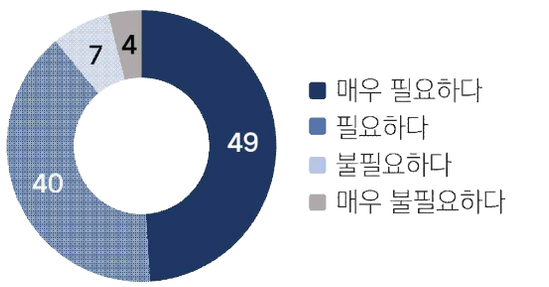
- 2023년 5월 미국에서 펜타곤이 폭발하는 사진이 소셜미디어인 트위터에서 언론까지 광범위하게 퍼지며 큰 혼란을 야기
- 펜타곤 대변인은 "펜타곤은 공격받지 않았다."고 밝혔고, 펜타곤 주변을 관할하는 경찰과 소방 당국 역시 "펜타곤 보호구역이나 근처에서 폭발이나 사고가 발생하지 않았으며 대중에게 위험은 없다."고 전하며 해당 사진은 인공지능 생성물임을 공개
- 그러나, 사진이 유포되자 S&P500 지수가 한때 0.3% 하락하는 등 증시는 출렁였고, 유사시 안전 자산으로 꼽히는 미 국채와 금값도 잠시 상승

4) 가짜뉴스, 위키피디아, https://ko.wikipedia.org/wiki/가짜뉴스
5) 가짜뉴스 피해 심각…허위 사실 유포자, 매체 강력 처벌 필요, 한국 NGO 신문, https://www.ngonews.kr/news/articleView.html?idxno=143414
6) https://v.daum.net/v/20230523143302235

가짜뉴스 생성(2)

- 생성형 인공지능의 활용은 동영상, 이미지, 오디오 등 선거 운동에 필요한 다양한 콘텐츠 제작을 가능하게 하며 선거에 영향력 행사
- 가짜 인공지능 생성 콘텐츠는 유권자들 사이에서 혼란을 야기하고, 특정 후보에 대한 부정적인 인식을 심어줄 수 있는 가능성 존재

| UNESCO & Ipsos('24), 전 세계 16개국 출신 8,000명 대상 설문조사 |

선거에 대한 가짜뉴스의 영향 유무(%)	온라인 가짜뉴스의 규제 필요 유무(%)
	

출처: 새로운 디지털 질서 정립 추진계획(24.5)

사례 예시(美 대통령 선거 딥페이크 이미지 유포 사건)

- 2023년 미국 선거기간에 해당 후보의 지지자들은 흑인 유권자들을 공략하기 위해 흑인들이 후보를 지지한다는 취지의 사진을 인공지능으로 생성하여 공유
- 후보자를 비방하기 위해서 생성형 AI 모델인 '미드저니'를 이용해 후보자가 체포되는 장면과 형을 선고받아 투옥되는 장면 등을 제작한 뒤 소셜미디어에 게시

출처: BBC News 코리아(좌), AI 타임스(우)

생성형 인공지능으로 제작한 트럼프 美 대통령 딥페이크

◉ 디지털 성범죄

- 생성형 인공지능을 오남용하여 성적 허위 영상물(성착취물)을 제작 및 유포하는 행위가 발생하고 있으며, 연예인이나 지인의 얼굴을 음란물과 합성해 유포하는 디지털 성범죄가 빈번히 발생
- 이미지의 유포·합성·소비의 가능성을 무한대로 확장시키기 때문에 디지털 성범죄는 피해와 가해의 구도가 1대 N을 이루고, 생산자와 소비자의 경계가 불분명[7)]

| 딥페이크 기술을 활용한 디지털 성범죄 사례 |

기간	내용
2024년 9월 ~	서울의 한 고등학교에 다니는 학생들 사이에선 최근 주변 4개 고교에 재학 중인 남학생 5명이 동료 여학생 10여 명의 사진을 불법 합성
2024년 8월 ~	텔레그램 단체 채팅방에서 현역 군인들이 여성 동료 군인들의 얼굴 사진을 딥페이크 방식으로 합성해 성착취물을 제작 및 공유

➽ 사례 예시(생성형 인공지능 활용 아동 성착취물 제작자 기소)

생성형 인공지능 활용 아동 성착취물 제작자 기소[8)]

- 2023년 4월, 생성형 인공지능 서비스에 '나체', '어린이' 등의 텍스트를 입력하여 360개의 아동 성착취물을 제작한 사건이 발생
- 우리나라 법원에서는 해당 행위가 아동·청소년의 성보호에 관한 법률에 위반되었다고 판시하여 징역 2년 6개월이 선고
- 피고인은 유포할 목적이 없었고, 가상의 이미지를 가지고 있는 것은 처벌 대상이 아니라고 생각했다고 진술했으나 재판부는 피해자가 없다고 해도 성적 표현물 자체가 명백하게 아동으로 인식될 수 있다면 아동청소년보호법 위반에 해당할 수 있다고 판시

7) 디지털 성범죄의 특징 및 현황, 찾기 쉬운 생활법령 정보(법제처), https://www.easylaw.go.kr/CSP/CnpClsMain.laf?csmSeq=1594&ccfNo=1&cciNo=1&cnpClsNo=2
8) AI 프로그램으로 아동 성 착취물 제작해 구속 기소, KBS 뉴스, https://news.kbs.co.kr/news/pc/view/view.do?ncd=7738404

학술 및 콘텐츠 윤리 문제

- 생성형 인공지능이 학술논문, 자기소개서, 기사, 논문에 사용하는 그림 등을 생성하여 활용하는 사례가 증가하면서 공동 저자로 등록하고, 서로 다른 논문임에도 유사한 그림, 글이 무분별하게 양산

| 생성형 인공지능의 AI 표절 |

사건	내용
네이처, 인공지능 활용 논문 검증에 대한 주의('24년 11월)	'Frontiers in Cell and Developmental Biology'에 실렸던 한 논문을 소개하면서 터무니없이 큰 생식기를 지닌 쥐 이미지가 사용됐는데 알고 보니 생성형 이미지 모델 '미드저니'로 만든 것으로 밝혀져 철회된 사례를 집중하며 인공지능을 논문에 활용하는 것을 우려
턴잇인, AI 논문 표절 심각성 발표('24년 4월)	온라인 표절 검사 서비스(Turnitin) 기업은 '23년 4월 이후 플랫폼 제출 2억개 논문 중 약 11%가 표절이라고 발표

사례 예시(공동 저자로 ChatGPT가 등장한 논문이 게재되면서 과학계 반발)

과학기술 저널에 등재된 공동 저자 'ChatGPT'

- ChatGPT가 2023년 한해 책/저널을 발표하여 저작권을 보유하고 있는 경우 200여 권이 확인, 네이처의 발표에 따르면 최소 4편 이상이 연구논문에 공동 저자로 등록

과학계 반발에 관한 다양한 기사

- (과학계) ChatGPT가 공동 저자로 등록하는 것이 표절 문제를 동반한다면 최소한의 윤리 지침을 지켜야 한다고 지적하며 네이처 등에서는 저자로서 인정 불가 의사 표시

인공지능 생성 콘텐츠 국내외 법안 현황

- 생성형 인공지능이 사회 전반에 미치는 영향이 증대됨에 따라, 딥페이크를 통한 가짜뉴스 확산, 선거 개입, 지적재산권 침해 등 다양한 사회적 부작용이 우려
- 따라서, 각국은 인공지능 생성 콘텐츠의 안전하고 신뢰할 수 있는 활용을 위한 법적 규제 체계 마련

| 국외 인공지능 생성 콘텐츠 관련 법안 현황 |

<table>
<tr><th>구분</th><th>주요 내용</th></tr>
<tr><td>미국</td><td>
• 행정 명령 14110호(Executive Order, ('23.10))
- 안전하고 신뢰할 수 있는 인공지능 개발 및 사용 행정명령 제 14110호(Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)
- AI 기술의 급속한 발전과 그에 따른 잠재적 위험에 대한 우려 및 사회적 으로 발생 가능한 문제(차별, 편견, 허위정보 등)로 인해, 국가 안보에 대한 위험 증가하고 있어 책임감 있는 AI 개발 및 사용을 위한 규제 입법
- 정부 기관들이 240일 내에 합성 콘텐츠를 탐지, 라벨링(워터마킹 등의 방법을 통해), 추적하기 위한 최신 기술을 식별하는 보고서를 작성하도록 의무화

• 자율 서약(Voluntary Commitments, ('23.7))
- 美 바이든 정부는 주요 AI 기업 7개와 안전, 보안, 신뢰를 강조하는 8가지 약속에 AI 생성 시청각 콘텐츠 표식(워터마크) 기제 개발 규정 제정
※ 2025년 1월 상기 두 인공지능 정책(Executive Order 14110, Voluntary AI Commitments)은 트럼프 대통령 취임 직후 바이든 행정부의 AI 규제 정책을 전면 폐지하는 행정명령 발표에 의해 폐지 (EO 14179, "Removing Barriers to American Leadership in Artificial Intelligence")

• 주(州)별 인공지능 법안 현황
- 캘리포니아, 오하이오 및 뉴욕 등 각 주(州)는 인공지능 생성 콘텐츠에 명시 의무를 요구하는 조항이 포함된 법안 시행 또는 추진 중
<table>
<tr><th>주(州)</th><th>번호</th><th>법안명</th><th>시행일</th><th>주요 내용</th></tr>
<tr><td>캘리포니아</td><td>SB942</td><td>California AI Transparency ACT</td><td>26. 1. 1.</td><td>매월 100만 명 이상의 사용자가 있는 시스템은 인공지능 생성물 탐지 도구 및 표시나 메타데이터를 제공하여 사용자가 인공지능 생성 콘텐츠를 쉽게 식별할 수 있도록 규정

사용자 생성 콘텐츠(user-generated content)를 제외한 비디오 게임, 텔레비전, 스트리밍, 영화 또는 인터랙티브 경험(interactive experiences) 만을 제공하는 제품, 서비스, 인터넷 웹사이트 또는 애플리케이션에는 미적용</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><th>구분</th><th>주요 내용</th></tr>
<tr><td></td><td>
<table>
<tr><th>주(州)</th><th>번호</th><th>법안명</th><th>시행일</th><th>주요 내용</th></tr>
<tr><td>오하이오</td><td>SB217</td><td>AI Generated Product Watermark</td><td>추진 중</td><td>인공지능 생성 콘텐츠에 워터마크를 탑재하고, 개인의 신원을 무단으로 활용한 콘텐츠는 24시간 이내에 삭제하도록 규정</td></tr>
<tr><td>뉴욕</td><td>A7106</td><td>Political Artificial Intelligence Disclaimer(PAID) Act</td><td>추진 중</td><td>정치적으로 사용되는 인공지능 생성 콘텐츠는 인공지능 기술 활용 사실을 고지하도록 의무 규정</td></tr>
</table>
</td></tr>
<tr><td>
유럽연합</td><td>
• 인공지능법(EU AI Act, ('24.5))
– 범용 AI 모델 관련 콘텐츠가 AI에 의해 생성되었음을 명시하는 투명성 의무 규정
– 딥페이크를 구성하는 이미지, 음성 또는 비디오 콘텐츠를 생성하거나 조작하는 인공지능시스템의 배포자는 해당 콘텐츠가 인위적으로 생성 또는 조작되었다는 것을 공개
• 디지털 서비스법(Digital Services Act, ('22.7))
– 인공지능으로 생성한 콘텐츠(이미지, 오디오, 영상)에 대해 생성 주체를 소비자들이 알 수 있도록 표시할 것을 규정(Article 35-Mitigation of risks-(k) 조항)
</td></tr>
<tr><td>
중국</td><td>
• 생성형 인공지능 서비스 관리 법 ('23.5)
– 생성형 인공지능 서비스 제공자는 '인터넷정보서비스 딥페이크 관리 규정'에 따라 이미지, 비디오 등 인공지능 콘텐츠에 대해 표시할 것을 의무화(제3장)
– 규정 위반 시 시정하도록 명령하며, 시정을 거부하거나 상황이 엄중한 경우 관련 서비스 제공을 잠정 중단하도록 규정
• 인터넷정보서비스 딥페이크 관리 규정 ('22.11)
– 기업들이 딥페이크 기술을 사용해 콘텐츠 제작 시 사용 기술 내용을 명시하도록 규정
– 생성형 AI 서비스 공급자들이 텍스트·이미지·오디오 등의 콘텐츠에 인공지능이 생성한 것임을 표시(워터마크 포함)하여 원본 추적이 가능하도록 해야 한다는 '생성형 AI 규정' 마련
</td></tr>
</table>

| 국내 인공지능 생성 콘텐츠 관련 법안 현황 |

제정일	기간	주요 내용
'25. 1	인공지능 발전과 신뢰 기반 조성 등에 관한 기본법	- 인공지능 기술의 급속한 발전과 사회 전반에 걸친 활용 확대에 따라, 기술 혁신과 윤리적 책임 간의 균형을 확보하기 위해 「인공지능의 발전과 신뢰 기반 조성 등에 관한 기본법」 제정 - 제31조(인공지능 투명성 확보 의무)는 고영향 인공지능 또는 생성형 인공지능을 이용한 제품 또는 서비스를 제공하는 인공지능사업자는 해당 사실을 이용자에게 사전에 고지하여야 하며, 생성형 인공지능 또는 이를 이용한 제품 또는 서비스를 제공하는 경우 그 결과물이 생성형 인공지능에 의하여 생성되었다는 사실을 표시하여야함을 요구 - 인공지능시스템을 이용하여 실제와 구분하기 어려운 가상의 결과물을 제공하는 경우 그 사실을 이용자가 명확하게 알 수 있도록 고지 또는 표시하도록하는 규정을 마련
'24. 12	정보통신망 이용촉진 및 정보보호 등에 관한 법률	- 인공지능 기술을 이용하여 사람의 얼굴·신체 또는 음성을 대상으로 한 촬영물·영상물 또는 음성물을 대상자의 의사에 반하여 편집·합성 또는 가공한 정보의 무분별한 유통으로 인한 성범죄, 명예훼손 또는 사기 등의 피해를 예방하기 위하여 시책을 마련해야 하는 규정 신설(제42조의2-합성영상등으로 인한 피해 예방을 위한 시책)
'23. 12	공직선거법	- 누구든지 선거일 전 90일부터 선거일까지 인공지능 기술 등을 이용하여 만든 동영상, 이미지, 오디오 등을 게시하는 것이 금지(제82조의8-딥페이크영상등을 이용한 선거운동) - 중앙선거관리위원회 규칙으로 정하는 사항을 딥페이크 영상 등에 표시하지 아니하면 처벌하는 규정 마련 ※ 제250조(허위사실공표죄) 4항, 제255조(부정선거운동죄), 제261조(과태료의 부과·징수 등)

인공지능 생성 콘텐츠 국제 사회 논의

- AI 서울 정상회의('24.5)
  - '안전하고 혁신적이며 포용적인 AI를 위한 서울 선언'과 '프론티어 AI 안전 서약'이 체결
  - 구글, LG AI연구소, 세일즈포스, KT, 마이크로소프트, 삼성전자, 앤스로픽, SK텔레콤, IBM, 네이버, 코히어, 카카오, 오픈AI, 어도비 등 총 14개 기업이 서약에 서명하였으며 주요 내용은 다음과 같다.
    - AI 안전 연구소 간 네트워크 확대 및 글로벌 협력 촉진
    - 워터마크 등 인공지능이 생성하는 콘텐츠 식별을 위한 조치와 국제표준 개발 협력 강화
    - AI 모델에 대한 위협 평가를 위한 레드팀 구성 및 안전성 접근 방식을 투명하게 공유
- 히로시마 AI 프로세스 행동강령 발표('23.12)
  - G7*은 '23년 5월 일본에서 개최된 히로시마 정상회의에서 인공지능의 위험을 관리하는 국제 규범인 '히로시마 AI 프로세스' 수립에 대해 합의

    * G7 회원국은 미국·일본·독일·영국·프랑스·이탈리아·캐나다이며, EU는 참가국 자격 활동

- 개발자가 가짜정보의 확산과 프라이버시 침해 위험을 완화하는 방안을 검증받고, 인공지능 콘텐츠에 대한 워터마크 삽입하는 등의 '히로시마 AI 프로세스 행동강령' 발표[9)]

  ⇒ 기술적으로 가능한 경우 워터마크 또는 '인공지능 콘텐츠임을 알 수 있도록 하는 기타 기술적 조치' 등 인공지능 콘텐츠의 진본성 및 출처 체제를 개발 이행할 것을 요구[10)]

- 독일 뮌헨안보회의(MSC) 딥페이크 공동 대응 논의('24.2)[11)]

  - 독일 뮌헨안보회의(MSC)에서 20개 글로벌 빅테크 기업들이 공동으로 인공지능 기술을 악용해 선거에 영향을 끼치는 행위에 공동 대응하기로 합의했다고 발표
  - 합의문의 제목은 '2024년 선거에서 인공지능의 기만적 사용을 방지하기 위한 기술 협약'(Tech Accord to Combat Deceptive Use of AI in 2024 Elections)

    ⇒ 해당 합의문에는 선거에 영향을 미치는 인공지능 사기 콘텐츠를 막기 위한 기술 개발, 인공지능 사기 콘텐츠와 관련하여 발새알 수 있는 위험 평가 모델 수립 등에 대한 조약

해외 사례: 인공지능 생성 결과물 유통 과정에서의 투명성 확보 사례

- EU 디지털서비스법

  - 인공지능기본법은 인공지능 생성 결과물 유통 과정에서의 투명성 확보 의무를 규율하고 있지 않으나, 유럽의 DSA 등의 법안에서는 온라인 플랫폼 사업자*에게 인공지능 생성 결과물의 유통 과정에서 투명성 확보(위험 완화) 의무를 부과

    * Very Large Online Platforms (VLOPs) or Very Large Online Search Engines (VLOSEs) 대상이며 틱톡, 유튜브, 페이스북, 링크드인 등이 이에 포함

9) 디지털 심화대응 실태진단, https://beingdigital.kr/front/iagnosis_view.do?pageView=_11
10) G7, 히로시마 프로세스 행동강령의 주요 내용 및 시사, NIA, 23.12
11) 글로벌 빅테크, 유권자 속이는 'AI 페이크' 공동 대응, https://www.hani.co.kr/arti/economy/economy_general/1128857.html

인공지능 투명성 확보
가이드라인