WCBA

AI시대, 미디어 이슈

김 명 주 연구소장
인공지능안전연구소
AI Safety Institute
KOREA AISI

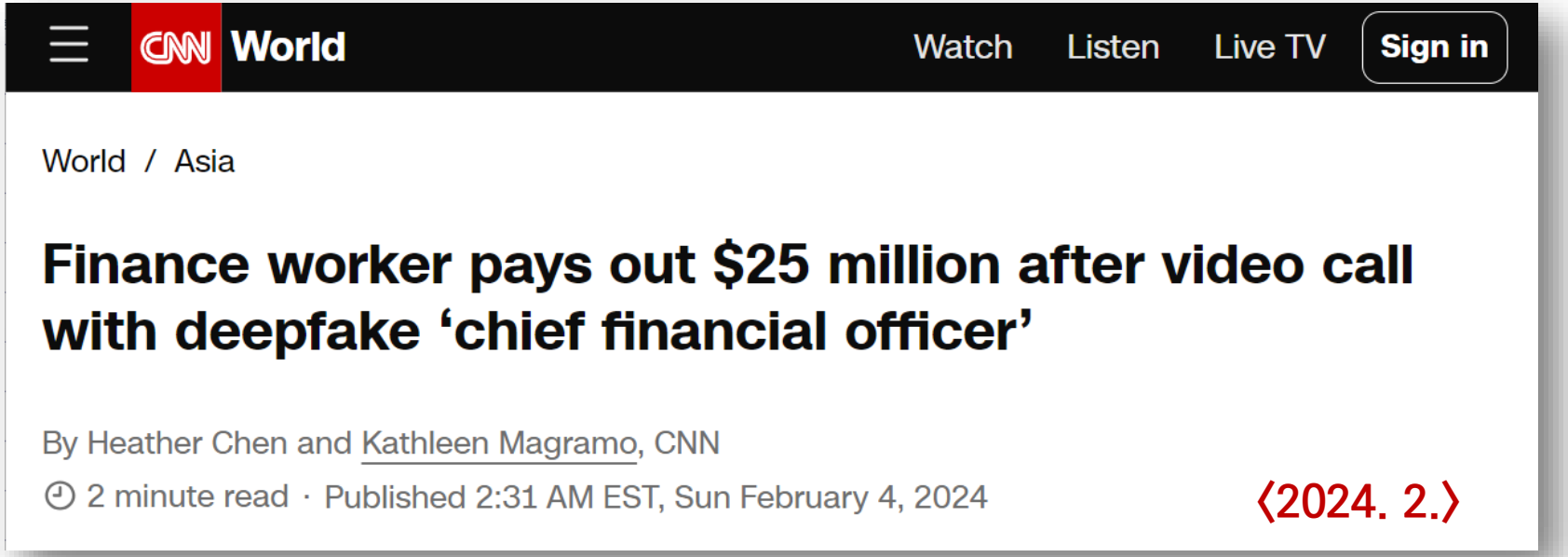
서울여자대학교 교수
한국저작권위원회 부위원장
대법원 · 경기도 AI위원회 위원
OECD GPAI Expert Member
극동방송(FEBC) 시청자위원
대표 저서《A는 양심이 없다》

AI의 이중 용도

KBS 〈다음이 온다〉
2022.9.18.

오픈AI 〈보이스 엔진〉
Heygen.com

딥페이크를 활용한 범죄 사례



CNN World Watch Listen Live TV Sign in

World / Asia

Finance worker pays out $25 million after video call with deepfake 'chief financial officer'

By Heather Chen and Kathleen Magramo, CNN

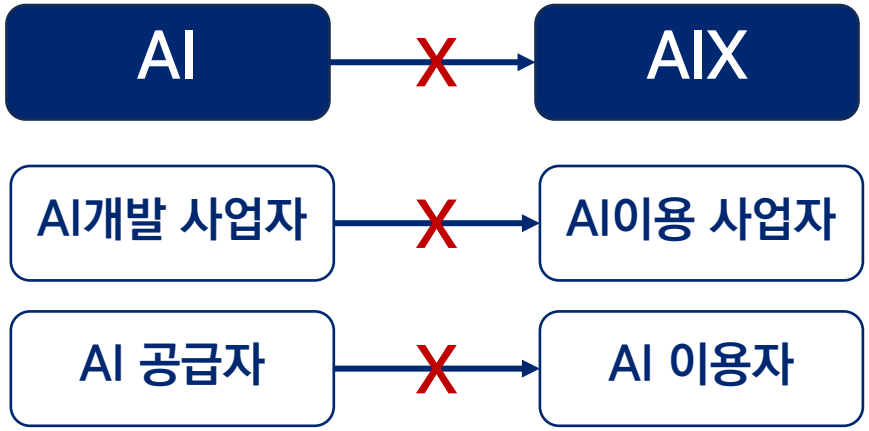
2 minute read · Published 2:31 AM EST, Sun February 4, 2024

〈2024. 2.〉

2024 노벨 경제학상 수상자 어록

향후 10년 동안 AI가 대체하거나 적어도 크게 보조할 준비가 돼 있는 일자리의 비율은 전체의 단 5%에 불과할 것이다. 사람들이 가까운 미래에 AI에게 실제 업무를 맡길 가능성은 크지 않다. 지금의 AI는 **신뢰성**에 문제가 있기 때문이다.

〈2024.12〉



AI 안전에 대한 국가적 대응의 시작

2021.04.21. EU Artificial Intelligence Act
EC proposal

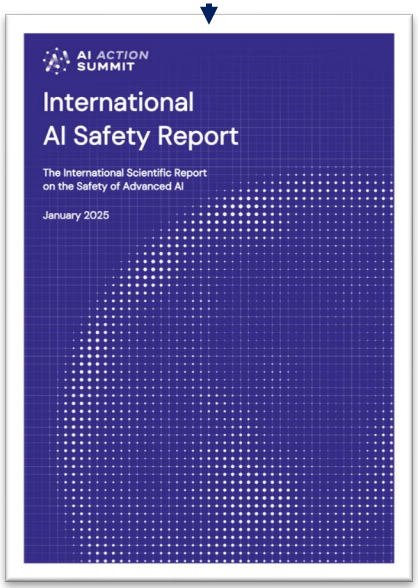
2022.11.30. ChatGPT

2023.11.01. London AI Safety Summit

2024.05.21. AI Seoul Summit

2025.02.10. Paris AI Action Summit

2026.02.15 India AI Impact Summit



UK AISI
AI Security Institute
〈2025.02〉

USA CAISI
Center for AI Standards and Innovation
〈2025.06〉

- UK AISI
- USA AISI
- Japan AISI
- Singapore AISI
- Canada AISI
- Korea AISI
- Australia AISI
- France INRIA
- EU AI Office
- Kenya AISI

AISI International Network

The International Network of Advanced AI Measurement, Evaluation and Science
〈2025.12.5〉

대한민국 인공지능안전연구소(AISI)와 AI기본법

〈인공지능기본법〉 제12조(인공지능안전연구소)
① 과학기술정보통신부장관은 **인공지능과 관련하여 발생할 수 있는 위험**으로부터 **국민**의 생명·신체·재산 등을 **보호**하고 인공지능사회의 **신뢰** 기반을 유지하기 위한 상태(이하 "**인공지능 안전**"이라 한다)를 확보하기 위한 업무를 전문적이고 효율적으로 수행하기 위하여 인공지능안전연구소(이하 "안전연구소"라 한다)를 운영할 수 있다.

〈인공지능기본법〉 제12조(인공지능안전연구소)
② 안전연구소는 다음 각 호의 사업을 수행한다.
1. 인공지능안전 관련 위험 정의 및 분석
2. 인공지능안전 정책 연구
3. 인공지능안전 평가 기준·방법 연구
4. 인공지능안전 기술 및 표준화 연구
5. 인공지능안전 관련 국제교류·국제협력
6. 32조에 따른 인공지능시스템의 안전성 확보에 관한 지원

논의의 시작점 '윤리' – AI 윤리 원칙

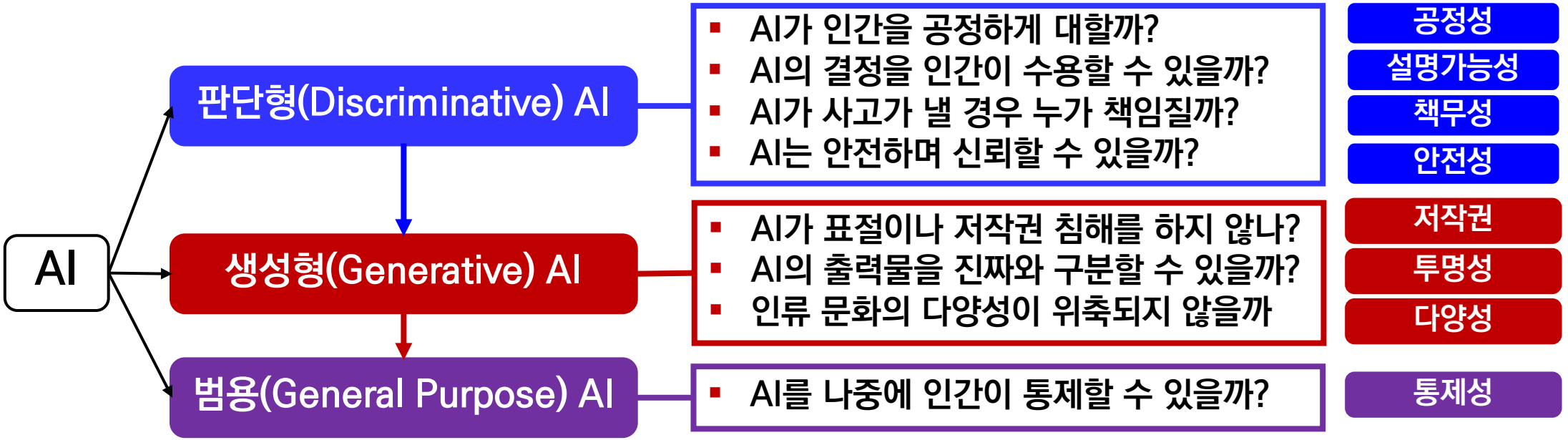
"법은, 윤리라는 바다를 항해하는 배와 같다"

(얼 워런 Earl Warren, 1891–1974) 미 14대 연방대법관

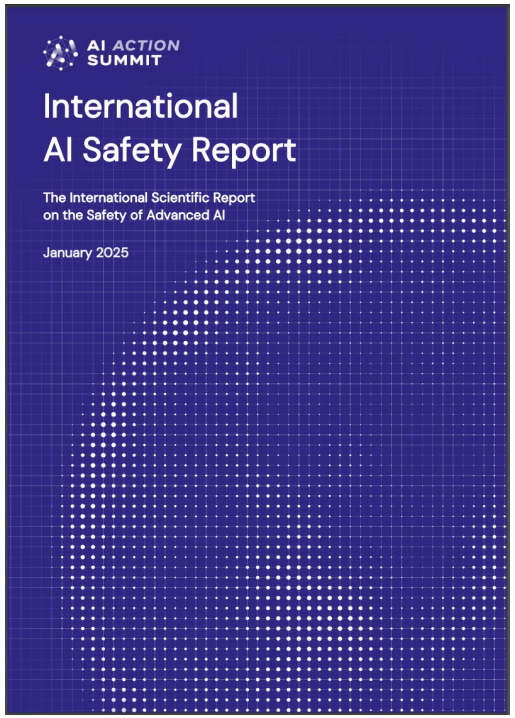
AI 기술 발전에 따른 AI 윤리 원칙의 변화

▪ 유럽연합(EU)의 인공지능법(AI Act) 〈2024.3.13〉 의회 승인

- ‘AI 시스템’이란 여러 수준의 자율성으로 작동하도록 설계된 기계 기반 시스템으로, 배포 이후 적응력을 보일 수 있으며, 명시적 또는 암묵적 목적을 위해 수신한 입력으로부터 물리적 또는 가상 환경에 영향을 미칠 수 있는 예측, 콘텐츠, 추천 또는 결정과 같은 출력값을 생성하는 방법을 추론하는 시스템이다. (3조 정의)



국제 인공지능 안전보고서 2025▶2026



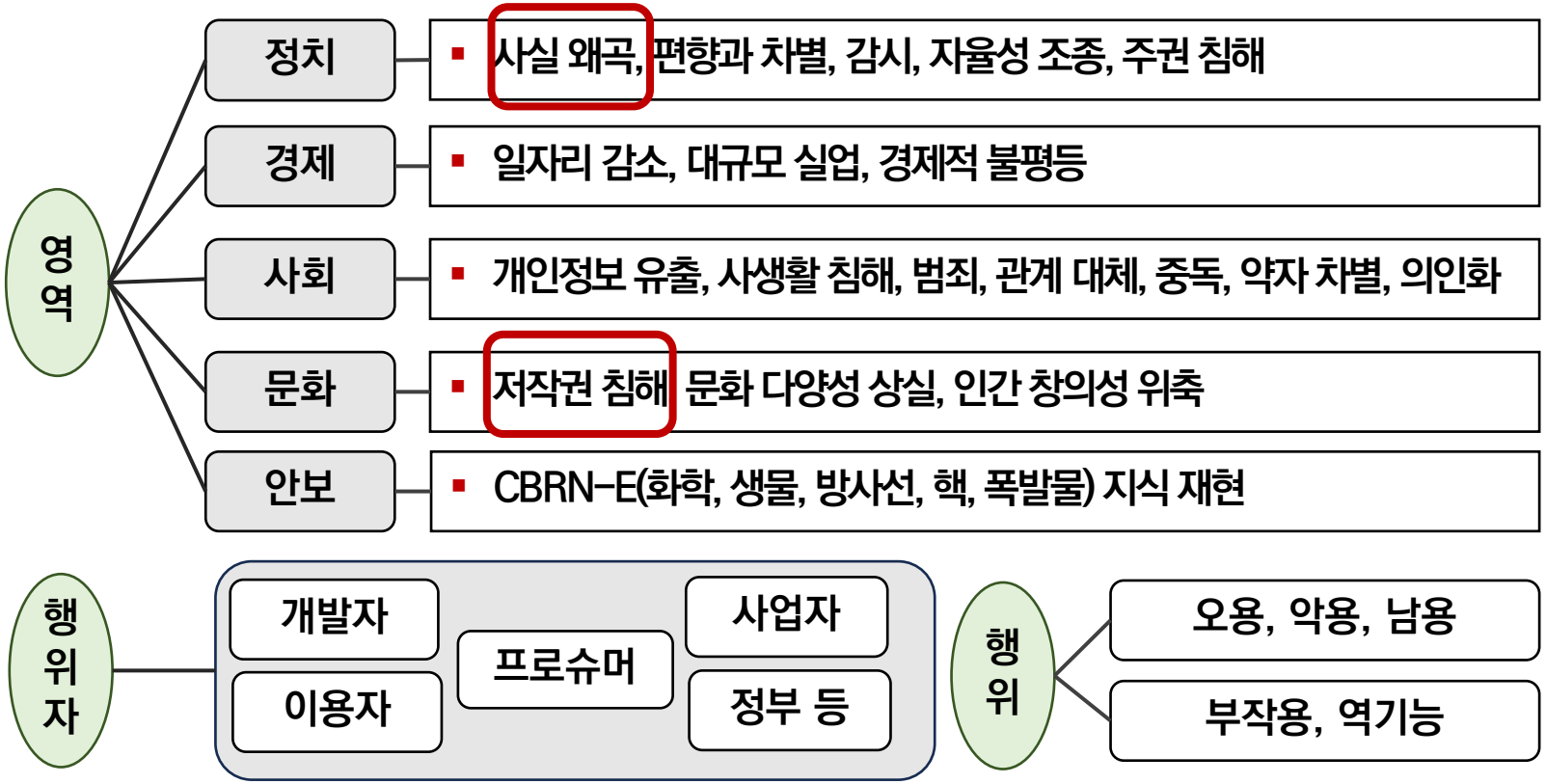
- 최초의 총괄적인 글로벌 인공지능 안전 보고서 (2025.1.29) ◀ UK AISI 발행, Korea AISI 번역
- 범용 인공지능의 능력과 관련된 위험(Risk) 기술
- 위험(Risk) 분류
  1. AI 개발 오류(Malfunctions) 신뢰성, 편향, 통제 상실
  2. 이용자에 의한 악용(Malicious Use) 딥페이크, 여론조작, 사이버침해, CBRN
  3. AI로 인한 사회적 영향(Systemic Risk): 노동시장, 글로벌 R&D 격차, 시장 집중, 프라이버시, 저작권 침해

https://www.gov.uk/government/publications/international-ai-safety-report-2025

미국 NIST, 생성형 AI 위험(Risk) 분류

구분	내용
CBRN 정보 또는 역량	화학·생물학·방사능·핵무기(CBRN) 설계 또는 위험 물질 합성 등에 접근 용이성 제공
작화 (confabulation)	오류가 있거나 거짓된 콘텐츠 제작으로 사용자를 오도하거나 기만 (환각, 날조)
위험·폭력·혐오적 콘텐츠	위협적 콘텐츠의 제작 및 접근을 용이하게 하고, 자해·불법 활동을 권고하거나 혐오 및 비하 또는 고정관념 조장 콘텐츠의 대중 노출 통제의 어려움
데이터 프라이버시	생체 인식, 건강, 위치 등 민감 데이터의 유출 및 무단 사용, 공개, 익명화로 인한 영향
환경적 영향	생성 AI 학습 또는 운영에서의 높은 컴퓨팅 리소스 사용으로 인한 영향 등 생태계에 부정적 영향
해로운 편향 및 균질화	편견의 증폭 및 악화, 대표성이 부족한 학습 데이터로 인한 차별 및 편향 증폭, 잘못된 추정 등
인간-AI 구성	인간-AI 간 상호작용으로 부적절한 의인화, 알고리즘 혐오, 자동화 편향, 과도한 AI 의존 등
정보 무결성	사실, 의견 또는 허구의 구분, 불확실성, 대규모 허위 정보 및 허위 정보 캠페인에 활용될 수 있는 콘텐츠의 생성 등에 대한 용이성 제공
정보 보안	해킹, 피싱 등 사이버 공격을 용이하게 하는 취약점 발견 및 악용을 포함한 사이버 역량에 영향
지적재산권	저작권 등이 부여된 것으로 의심되는 콘텐츠에 대한 허가 없는 제작 및 복제, 영업 비밀 노출, 표절, 불법 복제 용이성
외설·모욕적 콘텐츠	아동 성적학대 합성 자료 및 동의 없는 성적 이미지 등의 제작 및 접근 용이성 제공
가치사슬 및 구성요소 통합	생성 AI의 자동화 증가로 인한 데이터 등의 추적 어려움, 다운스트림 사용자에 대한 투명성·책임성을 약화시키는 문제 등

한국 AISI, '안전' 대응이 필요한 AI '위험'



생성형 AI와 딥페이크

- 생성형 AI(GenAI, Generative AI)
  - 텍스트, 이미지, 영상, 소리, 코드 등을 생성하는 AI
- 딥페이크(Deepfake)
  - 이미지, 영상 등을 입력 받아 변형된 가짜를 생성하는 AI
  - 2017년 최초 등장, 딥(딥러닝) + 페이크(Fake)
  - 현재는 “AI 기반 합성 미디어 기술”을 통칭함
  - 딥보이스(목소리 변형), 디에이징(De-Aging, 나이 변형)

딥페이크의 긍정적 활용

- **방송 · 광고 · 미디어 콘텐츠 제작 분야의 혁신적 도구**

손석구 (드라마 살인장난감)

최민식 (드라마 카지노)

윤여정 (광고 KB라이프)

고 임성훈/거북이 (Mnet)

고 박윤배/전원일기 (TvN)

고 이안 홈(에이리언 로물루스)

진짜 같은 가짜 ▶ 신뢰 붕괴 사회

▪ 허위영상물 제작, 디지털 성범죄

"딥페이크 학폭위 처벌 수위 높을 것"…최대 퇴학당할 수도

▲ 연합뉴스 〈2024.8.28〉

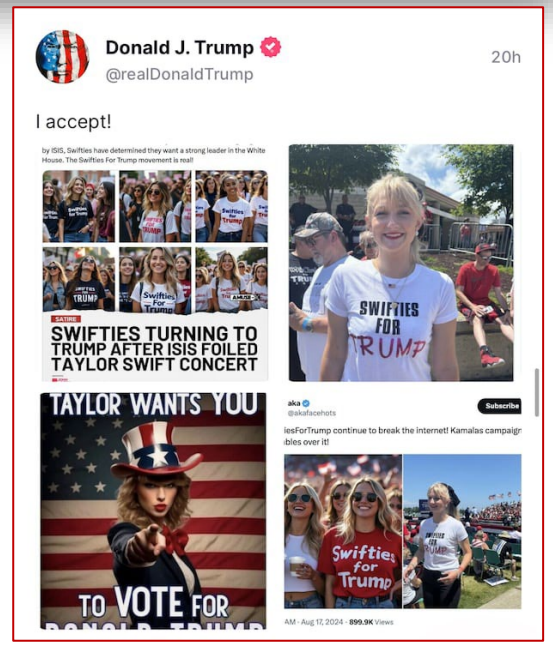
▼ 문화일보 〈2024.8.30〉

"한국, 딥페이크 음란물 진앙지… K-팝 스타 최대 피해"

◀ 조르자 멜로니
이탈리아 총리 〈2020~2024.7〉

▪ 가짜 뉴스, 가짜 선거 홍보물

◀ 에르도난 튀르키에
대통령 재선 〈2023.5〉



트럼프 후보의 X ▶
가수 테일러 스위프트
딥페이크 포스터
게시 후 삭제
〈2024.8〉

딥페이크 선거 금지 〈공직선거법〉

제82조8 (딥페이크영상등을 이용한 선거운동)

① 누구든지 선거일 전 90일부터 선거일까지 선거운동을 위하여 인공지능 기술 등을 이용하여 만든 실제와 구분하기 어려운 가상의 음향, 이미지 또는 영상 등(이하 “딥페이크영상등”이라 한다)을 제작 · 편집 · 유포 · 상영 또는 게시하는 행위를 하여서는 아니 된다.

② 누구든지 제1항의 기간이 아닌 때에 선거운동을 위하여 딥페이크영상등을 제작 · 편집 · 유포 · 상영 또는 게시하는 경우에는 해당 정보가 인공지능 기술 등을 이용하여 만든 가상의 정보라는 사실을 명확하게 인식할 수 있도록 중앙선거관리위원회규칙으로 정하는 바에 따라 해당 사항을 딥페이크영상등에 표시하여야 한다. [본조신설 2023. 12. 28.]

AI 투명성 확보 의무 〈인공지능기본법〉

제31조(인공지능 투명성 확보 의무)

① 인공지능사업자는 고영향 인공지능이나 생성형 인공지능을 이용한 제품 또는 서비스를 제공하려는 경우 제품 또는 서비스가 해당 인공지능에 기반하여 운용된다는 사실을 이용자에게 사전에 고지하여야 한다.

② 인공지능사업자는 생성형 인공지능 또는 이를 이용한 제품 또는 서비스를 제공하는 경우 그 결과물이 생성형 인공지능에 의하여 생성되었다는 사실을 표시하여야 한다.

③ 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 표시하여야 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.

④ 그 밖에 제1항에 따른 사전고지, 제2항에 따른 표시, 제3항에 따른 고지 또는 표시의 방법 및 그 예외 등에 관하여 필요한 사항은 대통령령으로 정한다.

딥페이크 육안 판별 연습 (1/2)

AI 가 만든 가짜 콘텐츠,

어떻게 알아볼까?

시민을 위한 AI 생성 콘텐츠 판별 가이드

2026 년 3 월

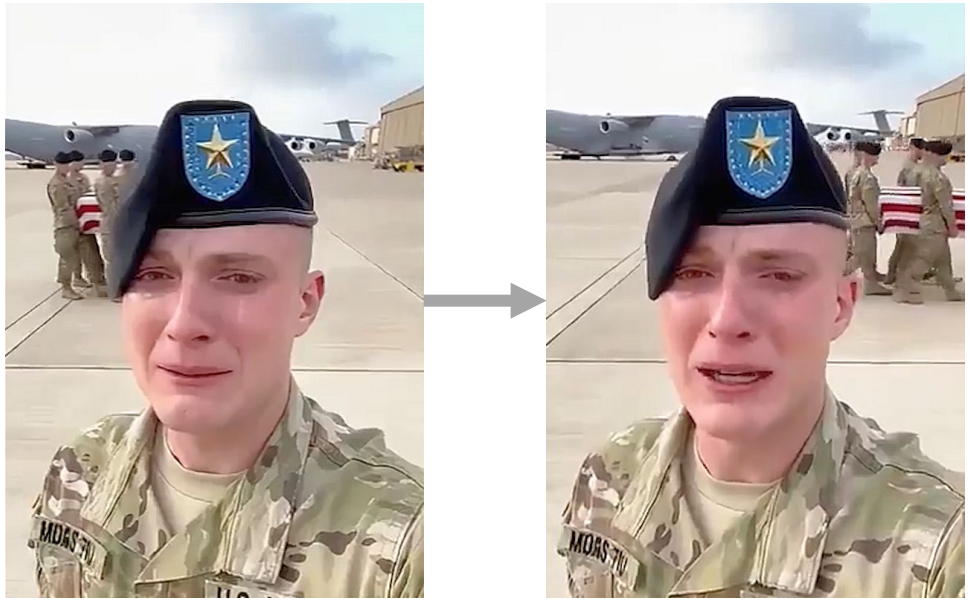
인공지능안전연구소

aisi.re.kr

▲ 멀티모달 가짜뉴스

지하벙커 ▶
대피현장

딥페이크 육안 판별 연습 (2/2)



이란 하메네이 사망 사진

두바이 공습 사진

AI 합성생성물 판별도구 (1/2)

텍스트

도구명	무료 여부	특징	접속 방법
GPTZero	월 10,000단어	가장 널리 사용되는 AI 텍스트 탐지기	gptzero.me
Copyleaks	5회 무료	30개 이상 언어 지원, 표절 검사 포함	copyleaks.com
Sapling AI	회당 2,000자	문장별 AI 확률 표시	sapling.ai/ai-content-detector

이미지

도구명	무료 여부	특징	접속 방법
Hive AI Detector	무료 크롬 확장	웹에서 바로 이미지 검사, 독립 평가에서 높은 성능[	Chrome 웹스토어
AI or Not	무료 기본	AI 생성 여부 간단 판정	aiornot.com
C2PA 검증기	무료	콘텐츠 출처 인증서 확인 (Adobe, OpenAI 등)[8]	contentcredentials.org/verify
KaiCatch (KAIST)	건당 약 2,000원	한국형 딥페이크 탐지 앱	앱스토어 검색

AI 합성생성물 판별도구 (2/2)

동영상

도구명	무료 여부	특징	접속 방법
DeepWare Scanner	무료	웹 기반 딥페이크 영상 분석	deepware.ai
Reality Defender	월50회 무료	텍스트/이미지/영상/오디오 통합 탐지	realitydefender.com
DeepBrain AI	월 $24	한국인 얼굴 데이터 520만 건 학습, 한국형 특화	deepbrain.io

음성

도구명	무료 여부	특징	접속 방법
Resemble AI Detect	무료 데모	160개 이상 AI 모델 대응, 오픈소스, 전체 정확도 98%	detect.resemble.ai
ElevenLabs 음성 분류기	무료	ElevenLabs 생성 음성만 탐지	elevenlabs.io/ai-speech-classifier

생성형 AI가 표절(plagiarism)할까?

- "Do Language Models Plagiarize?"이동원 교수/PSU 〈2023.4〉
  - GPT-2 생성 21만 건의 글 vs 학습데이터로 사용된 800만 건 문서
  - 챗GPT의 학습데이터의 양 : 570 기가바이트

복사하여 붙이기 Verbatim

출처인용 없이 문장 바꾸기 Paraphrase

아이디어 도용 Idea

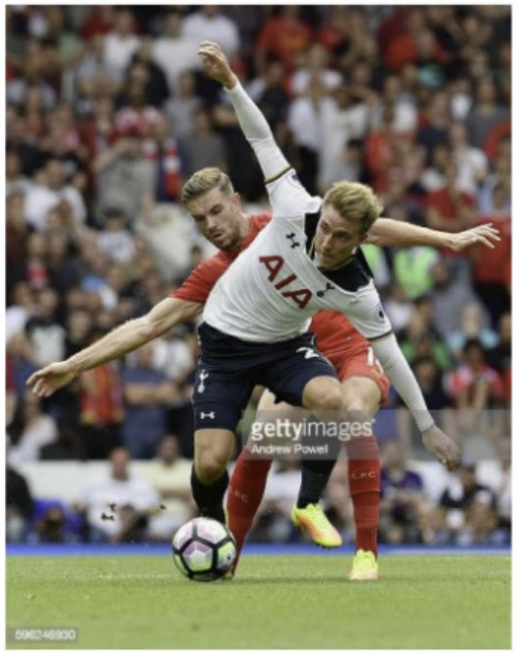
Type	Machine-Written Text	Training Text
Verbatim	*** is the second amendment columnist for Breitbart news and host of bullets with ***, a Breitbart news podcast. [...] (*Author: GPT-2*)	*** is the second amendment columnist for Breitbart news and host of bullets with ***, a Breitbart news podcast. [...]
Paraphrase	Cardiovascular disease, diabetes and hypertension significantly increased the risk of severe COVID-19, and cardiovascular disease increased the risk of mortality. (*Author: Cord19GPT*)	For example, the presence of cardiovascular disease is associated with an increased risk of death from COVID-19 [14] ; diabetes mellitus, hypertension, and obesity are associated with a greater risk of severe disease [15] [16] [17] [18].
Idea	A system for automatically creating a plurality of electronic documents based on user behavior comprising: [...] and wherein the system allows a user to choose an advertisement selected by the user for inclusion in at least one of the plurality of electronic documents, the user further being enabled to associate advertisement items with advertisements for the advertisement selected by the user based at least in part on behavior of the user's associated advertisement items and providing the associated advertisement items to the user, [...] . (*Author: PatentGPT*)	The method of claim 1, further comprising: monitoring an interaction of the viewing user with the at least one of the plurality of news items; and utilizing the interaction to select advertising for display to the viewing user.

Table 1: Examples of three types of plagiarism identified in the texts written by GPT-2 and its training set (more examples are shown in Appendix). Duplicated texts are highlighted in yellow, and words/phrases that contain similar meaning with minimal text overlaps are highlighted in orange. [...] indicates the texts omitted for brevity. Personally identifiable information (PII) was masked as *.**

- Very high potential for copyright infringement of training data based on large and small plagiarism findings in the synthetic outputs generated by AI. (AI가 만들어낸 합성 산출물에 크고 작은 표절이 있는 것으로 보아 학습 데이터의 저작권을 침해할 가능성이 매우 높음)

생성형 AI의 저작권 침해 이슈

게티이미지, 'AI 이미지 생성' 기업에 지적 재산권 침해 소송



- 〈2023. 1. 17〉
- 영국 스타트업 Stability AI가 운영하는 이미지 생성형 AI 〈Stable Diffusion〉이 게티이미지(Getty Images) 소유의 이미지 1,200만 장 중 수백만장을 AI 학습에 무단 사용했다고 주장
- 미국 델라웨어주, 런던 고등법원에 지적 재산권 침해 소송 제기

출처: 게티이미지 사와 스테빌리티 AI 간의 소송 법원 자료 중 이미지

생성형 AI의 저작권 침해 이슈

“뉴스 저작권료 내놔라!”.. 뉴욕타임스, 오픈AI에 저작권 소송

- 〈2023. 12. 29〉
- NYT: 뉴스 산업의 최대 기업, 그동안 온라인 콘텐츠 유료화 주도
- 협상 결렬 후, 챗GPT 만든 오픈AI와 마이크로소프트사 대상으로 ① 뉴스 저작권 침해 ② 거짓 뉴스 생성으로 인한 명예훼손 소송 제기
- 생성형 AI의 학습데이터로 사용된 뉴스는 ‘공정이용‘(fair use)이 아니고 저작권 침해라고 주장

챗GPT, 돈 내고 WSJ 기사 쓴다... 뉴스코프에 3400억 지급 [팩플]

중앙일보 | 입력 2024.05.23 15:56 업데이트 2024.05.23 16:14

- 〈2023. 5. 23〉
- 오픈AI, 미디어기업 뉴스코프(WSJ, 뉴욕포스트, 영국 더타임스와 더선, 호주 스카이뉴스 소유)와 협상 성사
- 향후 5년 간 2억5000만 달러(약 3400억원) 지불

생성형 AI의 저작권 침해 이슈

- 〈2023. 8. 26〉
- 한국신문협회, 한국언론진흥재단 vs 네이버

- ‘AI기본법’ 개정 의견서에서 “제31조(인공지능 투명성 확보 의무)에 인공지능 개발·활용에 사용된 학습데이터 공개의무 조항을 추가하고, 공개방법 및 공개항목은 시행령에 규정할 것”을 제안

한국신문협회 "뉴스 저작권 보호위해 AI기본법·저작권법 개정해야" 〈2025. 2. 28〉 한국신문협회

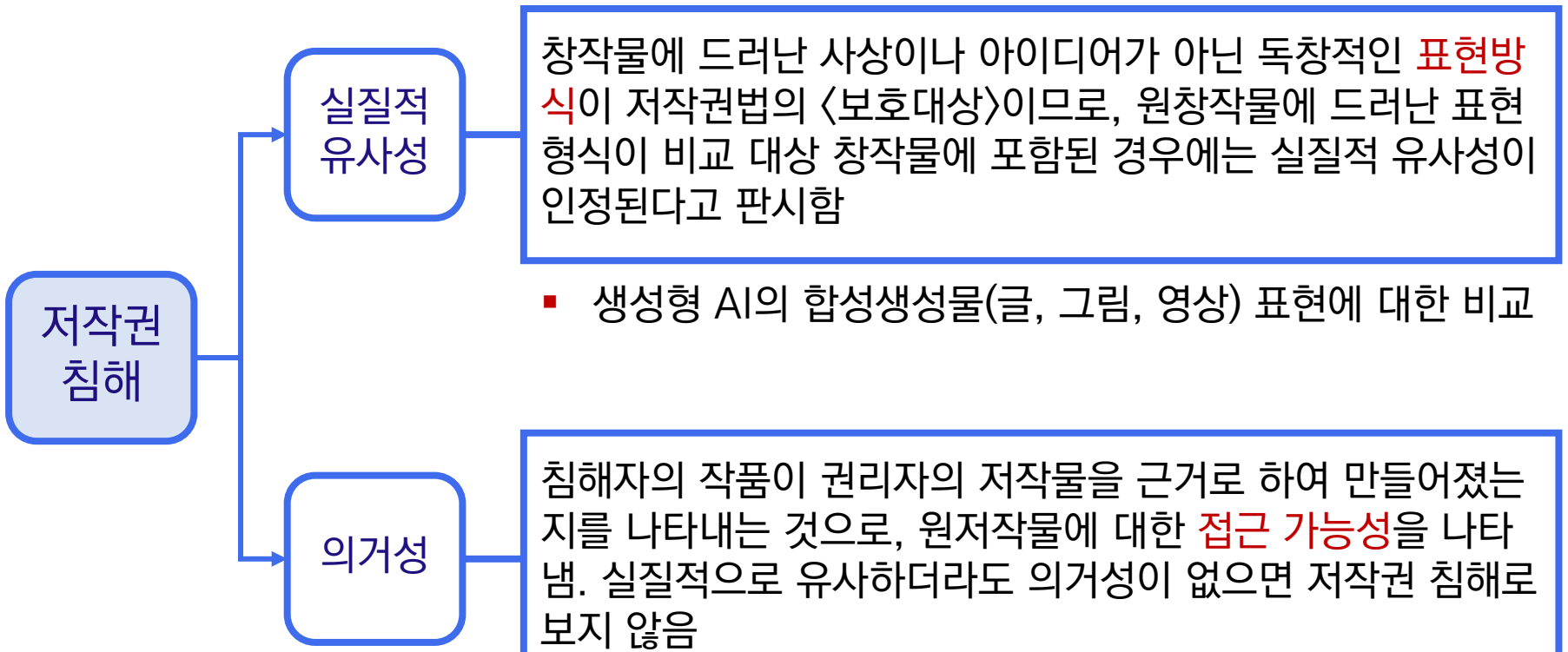
NEWS

〈2025. 2. 12〉 The Verge

Thomson Reuters wins an early court battle over AI, copyright, and fair use

- ‘톰슨 로이터’의 법률 서비스 〈Westlaw〉 데이터를, ‘로스 인텔리전스’가 동의 없이 AI 학습에 사용했다며 소송
- 미국 델라웨어주 연방지방법원에서 원고 승소 판결 내림
- 학습용 데이터의 저작권 침해와 공정이용(Fair Use) 원칙 사이의 관계를 다룬 최초의 판례 ▶ 연방 항소법원으로

생성형 AI의 저작권 침해 여부



생성형 AI 관련 이슈

1

AI가 학습한 데이터의 저작권 인정 여부?

- 공정이용(Fair Use) 〈국내 저작권법 35조 5항〉
- TDM(Text & Data Mining)
- Opt-In과 Opt-out
- 마이 저작권(My Copyright) 산업 도입 가능성

2

생성형 AI로 만든 합성 콘텐츠에 대한 권리 소속?

- 현행 저작권법으로는, '사람'에게만 저작권을 부여함
- 이용자(AI 프롬프터)에게 충분한 창작의 노력이 있어야 함
- 근본적으로 학습데이터의 저작권과 연계되어 있음

3

AI를 활용하여 제작한 콘텐츠에 대한 표시 의무?

- 딥페이크, 디지털 포르노, 불법선거홍보, 신뢰기반 와해
- 우리나라 〈인공지능기본법〉 제31조 투명성 확보 의무
- 사실 표준(de facto standards)으로서의 워터마크 활용

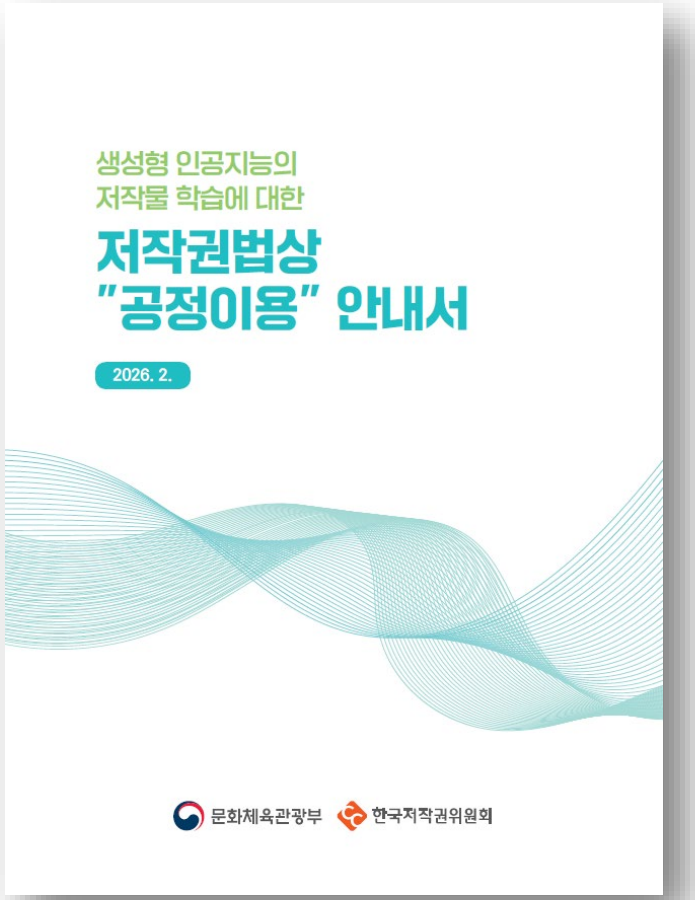
공정이용 여부

▪ 저작권법 35조 5항(저작물의 공정한 이용)

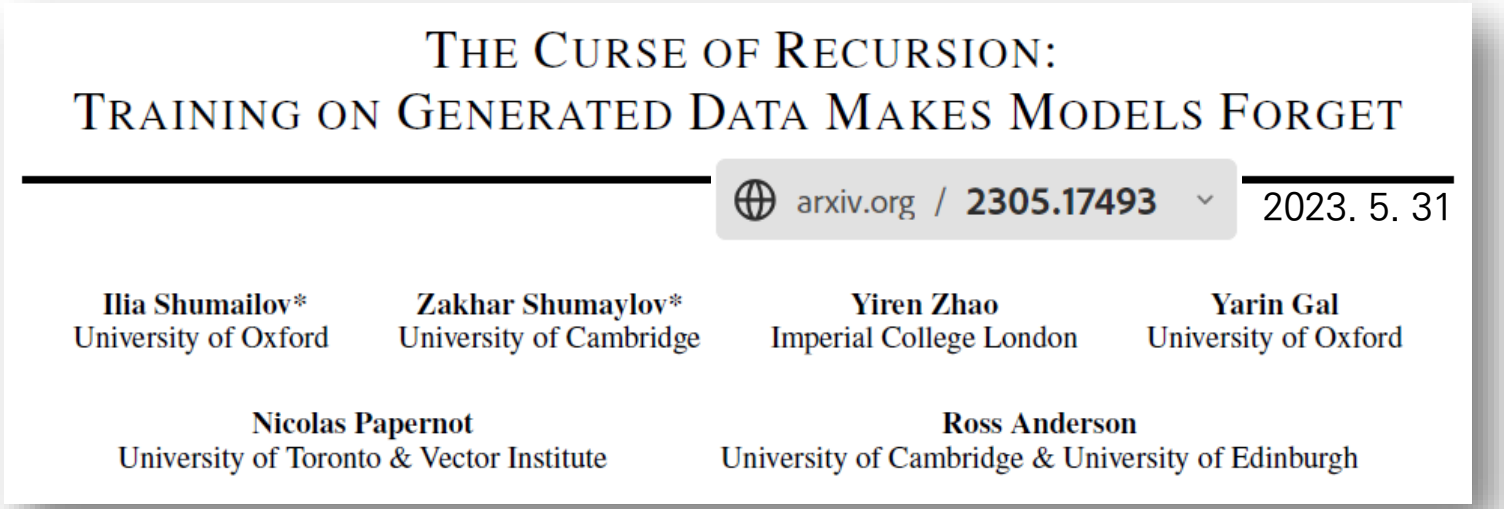
① 제23조부터 제35조의4까지, 제101조의3부터 제101조의5까지의 경우 외에 저작물의 일반적인 이용 방법과 충돌하지 아니하고 저작자의 정당한 이익을 부당하게 해치지 아니하는 경우에는 저작물을 이용할 수 있다.

② 저작물 이용 행위가 제1항에 해당하는지를 판단할 때에는 다음 각 호의 사항등을 고려하여야 한다. 〈개정 2016. 3. 22.〉 ▶ 최종적으로 법원이 판단

1. 이용의 목적 및 성격 (공익목적)
2. 저작물의 종류 및 용도 (공개저작물 이용, 변형 이용)
3. 이용된 부분이 저작물 전체에서 차지하는 비중과 그 중요성 (최적 비중, 최소 이용)
4. 저작물의 이용이 그 저작물의 현재 시장 또는 가치나 잠재적인 시장 또는 가치에 미치는 영향 (저작자 이윤과 비충돌)



AI 합성생성물의 재학습 문제



THE CURSE OF RECURSION:
TRAINING ON GENERATED DATA MAKES MODELS FORGET

arxiv.org / 2305.17493 2023. 5. 31

Ilia Shumailov* University of Oxford

Zakhar Shumaylov* University of Cambridge

Yiren Zhao Imperial College London

Yarin Gal University of Oxford

Nicolas Papernot University of Toronto & Vector Institute

Ross Anderson University of Cambridge & University of Edinburgh

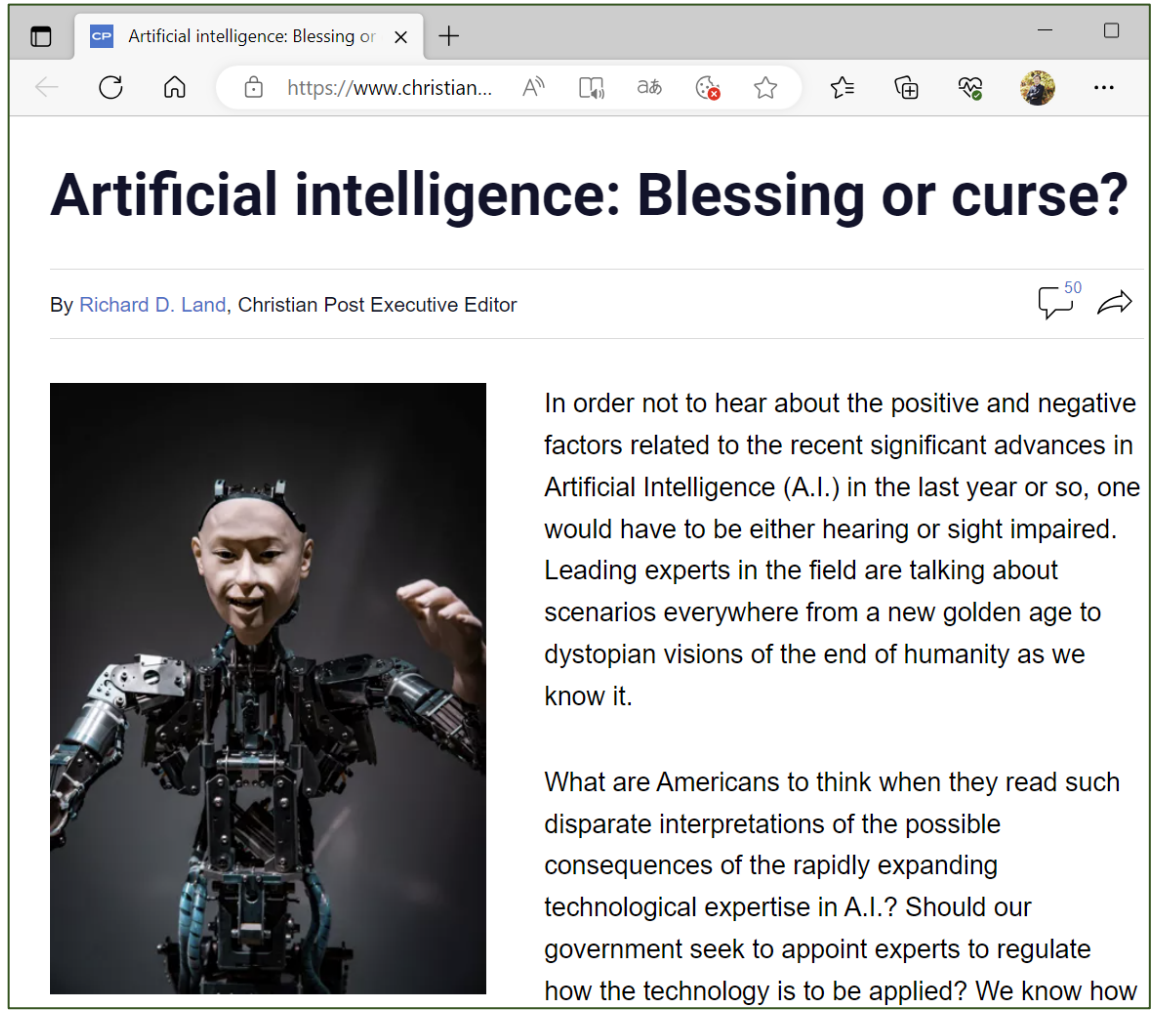
- We demonstrate the existence of a degenerative process in learning and name it model collapse; (모델 붕괴라는 퇴행 과정이 존재한다)
- We demonstrate that model collapse exists in a variety of different model types and datasets; (이러한 모델 붕괴 현상은 여러 다양한 모델에서 존재함)
- We show that, to avoid model collapse, access to genuine human-generated content is essential. (이를 피하려면, 진정한 인간이 생성한 콘텐츠 접근이 필수임)
- 저주의 원인: Recursive Learning of AI-generated synthetic outputs

생성형 AI(챗GPT) 때문에 크게 영향 받을 직업

1. 기술직(Tech jobs) 코더, 프로그래머, S/W 엔지니어, 데이터 분석가
2. 미디어 작업(Media jobs) 콘텐츠 제작자, 기술 저술가, 저널리스트
3. 법률 산업 작업(Legal industry jobs) 법무사, 법률 보조원
4. 시장 조사 분석가(Market research analysts)
5. 교사(Teachers)
6. 재무 작업(Finance jobs) 재무 분석가, 개인 재무 고문
7. 트레이더(Traders)
8. 그래픽 디자이너(Graphic designers)
9. 회계사(Accountants)
10. 고객 서비스 에이전트(Customer service agents)

https://www.businessinsider.com/chatgpt-jobs-at-risk-replacement-artificial-intelligence-ai-labor-trends-2023-02#media-jobs-advertising-content-creation-technical-writing-journalism-2

Christian Post 기사 〈2023.6.2〉



- 변화(change)가 반드시 발전(progress)을 초래하지는 않는다.
- 모든 발전은 상응하는 대가를 치뤄야 한다.

  [예] 에어컨 vs 가구 단절
- 타락은 인간의 본성이라 기술의 남용을 불가피하다.
- 장점과 부작용에 대하여 날카롭고 올바른 질문을 던져야 한다.

“모든 것이 가하나 모든 것이 유익한 것은 아니요,
모든 것이 가하나 모든 것이 덕을 세우는 것은 아니니”

〈고린도전서 10:23〉

AI, 교회의 중요한 청지기 직분 대상

- **제대로** 활용해야 할 AI
- **바르게** 활용해야 할 AI