FPRI
Future Policy Research Institute

"Next AI GuideLine"

- 미래 인공지능 가이드라인 V1.0 -

(In The Artificial Intelligence Society)

2023.11.11

Jason, Min (mikado22001@yahoo.co.kr)

□ Next AI GuideLine

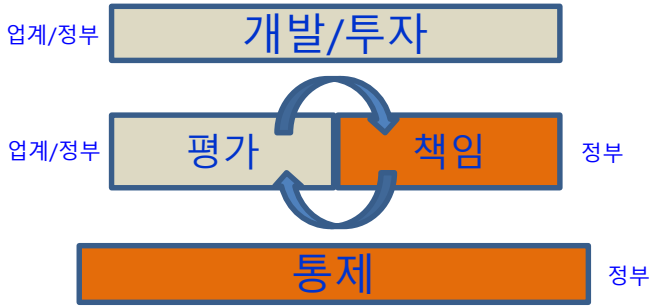
- 자율AI의 제어를 위한 합의 필요
  → 소수의 가장 강력한 AI 시스템에 대한 제어노력 우선

- 위험성이 낮은 사용과 학술 연구를 보호하기 위해
  작고 예측 가능한 AI 모델에 대한 과도한 컴플라이언스는 피할 필요
  → 개인정보의 범위 해석 정립

헌법 AI 필요

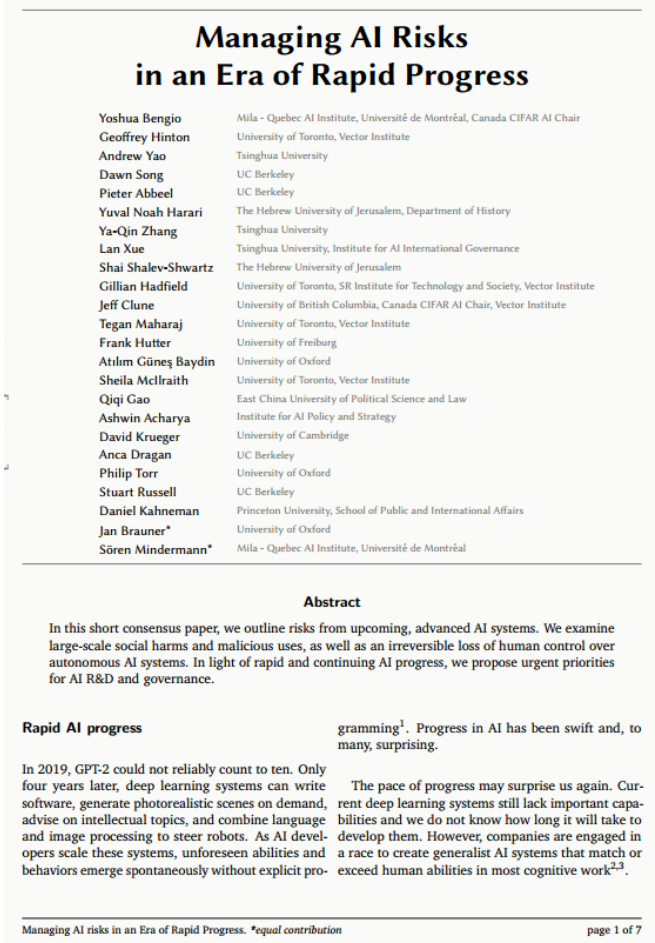
AI의 급속한 기술 발전은 사회적 규모의 위험을 초래할 것이므로 인공지능 리스크 관리 필요 !!

제조업체가 비용을 절감하기 위해 폐기물을 강에 방류하는 것처럼,
AI 개발의 보상을 얻고 그 폐혜는 사회에 던지는 것 우려



[연구소와 정부를 위해 제안된 정책권고사항]

(AI투자) 안전하고 윤리적인 연구를 위한 투자
(AI평가) AI모델의 위험평가 표준 마련
* RISK : 윤리/법/도난/오용등 다양한 위험벡터 식별
(AI책임) AI산업에 대한 감독, 모니터링, 책임 확립
(AI통제) 새로운 위험에 대한 추가 준비조치
* AI 연구소에는 가장 높은수준의 보안 조치 필요



Managing AI Risks in an Era of Rapid Progress

Yoshua Bengio — Mila - Quebec AI Institute, Université de Montréal, Canada CIFAR AI Chair
Geoffrey Hinton — University of Toronto, Vector Institute
Andrew Yao — Tsinghua University
Dawn Song — UC Berkeley
Pieter Abbeel — UC Berkeley
Yuval Noah Harari — The Hebrew University of Jerusalem, Department of History
Ya-Qin Zhang — Tsinghua University
Lan Xue — Tsinghua University, Institute for AI International Governance
Shai Shalev-Shwartz — The Hebrew University of Jerusalem
Gillian Hadfield — University of Toronto, SR Institute for Technology and Society, Vector Institute
Jeff Clune — University of British Columbia, Canada CIFAR AI Chair, Vector Institute
Tegan Maharaj — University of Toronto, Vector Institute
Frank Hutter — University of Freiburg
Atılım Güneş Baydin — University of Oxford
Sheila McIlraith — University of Toronto, Vector Institute
Qiqi Gao — East China University of Political Science and Law
Ashwin Acharya — Institute for AI Policy and Strategy
David Krueger — University of Cambridge
Anca Dragan — UC Berkeley
Philip Torr — University of Oxford
Stuart Russell — UC Berkeley
Daniel Kahneman — Princeton University, School of Public and International Affairs
Jan Brauner* — University of Oxford
Sören Mindermann* — Mila - Quebec AI Institute, Université de Montréal

Abstract

In this short consensus paper, we outline risks from upcoming, advanced AI systems. We examine large-scale social harms and malicious uses, as well as an irreversible loss of human control over autonomous AI systems. In light of rapid and continuing AI progress, we propose urgent priorities for AI R&D and governance.

Rapid AI progress

In 2019, GPT-2 could not reliably count to ten. Only four years later, deep learning systems can write software, generate photorealistic scenes on demand, advise on intellectual topics, and combine language and image processing to steer robots. As AI developers scale these systems, unforeseen abilities and behaviors emerge spontaneously without explicit programming[1]. Progress in AI has been swift and, to many, surprising.

The pace of progress may surprise us again. Current deep learning systems still lack important capabilities and we do not know how long it will take to develop them. However, companies are engaged in a race to create generalist AI systems that match or exceed human abilities in most cognitive work[2,3].

Managing AI risks in an Era of Rapid Progress. *equal contribution — page 1 of 7

https://managing-ai-risks.com/

1) AI투자

1. 산업 연구소와 정부 자금 제공자는 안전하고 윤리적인 AI에 투자해야 합니다.

업계와 정부는 AI 시스템의 안전과 윤리적 사용을 보장하기 위해 AI R&D 자원의 최소 1/3을 할당해야 합니다 . 관련 연구 분야는 다음과 같습니다.

1. 감독 및 정직성: 더 유능한 AI 시스템은 감독 및 테스트의 약점을 더 잘 활용할 수 있습니다 . 예를 들어 거짓이지만 설득력 있는 결과를 생성합니다.
2. 견고성: AI 시스템은 새로운 상황(분배 변화 시)에서 예측할 수 없게 작동합니다.
   또는 적대적인 입력).
3. 해석성: AI 의사결정은 불투명합니다. 지금까지는 시행착오를 통해서만 대형 모델을 테스트할 수 있었습니다. 우리는 그들의 내부 활동을 이해하는 법을 배워야 합니다.
4. 위험 평가: 프론티어 AI 시스템은 제작자가 기대하지 않는 기능을 개발하는 경우가 많으며, 이는 교육 후반이나 배포 후에도 발견될 수 있습니다.
   위험한 기능을 가능한 한 빨리 감지하거나 예측하려면 더 나은 평가가 필요합니다.
5. 새로운 과제 해결: 더 유능한 미래 AI 시스템은 지금까지 이론적 모델에서만 보았던 실패 모드를 나타낼 수 있습니다. 예를 들어 AI 시스템은 특정 목표를 달성하기 위해 복종하는 척하거나 안전 목표 및 종료 메커니즘의 약점을 이용하는 방법을 배울 수 있습니다 .

이 목록은 완전한 것이 아닙니다. 추가 관련 R&D 영역은 Hendrycks et al.
또한 Hendrycks와 Mazeika는 안전 보장으로 간주되어서는 안되는 안전 인접 활동의 한 클래스를 정의합니다 . 즉, 안전 지표를 향상시키는 것 보다 일반 AI 기능을 가속화하기 때문에 안전 기능 균형을 개선하지 않는 활동입니다.

2) AI평가

2. 업계와 정부는 대규모 AI 모델의 위험을 평가하고 완화하기 위한 정책을 수립하고 표준을 설정해야 합니다.

1. 프론티어 AI 개발자는 상세하고 독립적으로 면밀히 조사된 확장 정책을 즉시 이행해야 합니다 . 이러한 정책은 AI 시스템에서 특정 위험한 기능이 발견될 경우 해당 기업이 취할 구체적인 안전 조치를 설명해야 합니다 .

2. 정부는 AI 교육에 대한 국내 및 국제 안전 표준을 설정해야 합니다.
그리고 배포.

   (a) 표준은 공개적으로 이용 가능한 모델의 오용, 개발 중인 모델의 도난, 의도하지 않은 행동으로 인한 사고 및 사고, 신뢰할 수 없는 모델의 광범위한 사용으로 인한 사회적 영향을 포함하여 다양한 위험 벡터를 해결하기 위한 관행을 식별해야 합니다

   (b) 표준은 경험적 증거와 기술적 추세를 모두 추적하면서 AI 위험에 대한 우리의 발전하는 이해에 부응해야 합니다 . 모델이 더욱 강력해지고 평가 및 사고 보고서에서 새로운 위험 영역이 입증됨에 따라 표준은 이에 따라 더 큰 주의를 권장해야 합니다. 표준 설정 기관은 "퍽이 가는 곳으로 스케이트를 타는" 기술 전문 지식을 더욱 발전시켜야 합니다. AI 기능의 증가와 AI 배포의 확대로 인해 발생할 수 있는 피해를 예상합니다 .

3. 정부는 교육 중 및 배포 전에 최첨단 AI 시스템에 대한 감사를 요구해야 합니다 .

   (a) 정부는 AI 개발자에게 비정상적 이거나 예측할 수 없는 능력을 갖춘 AI 시스템을 만들기 위한 노력을 보고하도록 요구해야 합니다.

   (b) 감사자와 규제기관은 모델을 평가하는 데 필요한 접근 권한을 가져야 합니다.
   이러한 최첨단 AI 시스템을 교육하는 동안 및 배포하기 전에 연구소는 규제 기관과 독립 감사 기관에 이러한 시스템의 위험한 기능을 평가하는 데 필요한 액세스 권한을 제공해야 합니다. 업계 참여를 장려하기 위해 평가 프로세스에서는 법적 및/또는 기술적 수단을 사용하여 지적 재산 침해를 방지할 수 있습니다 (예: 모델 가중치를 공유하는 대신 "구조화된 액세스" 방법 사용) .

3) AI책임

3. 정부는 AI 산업에 대한 감독을 확립하고 AI 피해에 대한 결과를 설정해야 합니다.

보고된 AI 시스템에 대한 표준 및 감사를 보완하기 위해 AI 전문가는 정부가 AI 사고 및 교육 실행을 추적하기 위한 감독 및 모니터링 조치를 수립할 것을 권장합니다.
또한 책임감 있는 AI 교육 및 배포를 장려하기 위해 AI 피해에 대한 책임을 설정할 것을 권장합니다 .

1. 정부는 프론티어 AI에 대한 시민사회 감독 메커니즘을 구축해야 합니다.
   개발. 여기에는 다음이 포함됩니다.

   (a) 내부고발. 정부는 AI 연구소의 내부 고발자에 대한 법적 보호를 제공해야 합니다 . 특히 대규모 기술 회사의 직원은 고용주의 보복을 두려워할 수 있습니다. (b) 사건 보고서. 정부는 연구소에 AI가 발생한 사고를 보고하도록 요구해야 합니다.

   시스템이 유해한 동작이나 위험한 기능을 표시했습니다.

2. 정부는 대규모 컴퓨팅 클러스터에 대한 모니터링을 확립해야 합니다. 그들은 고객 파악(KYC) 확인을 포함하여 정부 및 업계 슈퍼컴퓨터의 사용을 모니터링 하고 중앙 데이터베이스에서 결과를 추적해야 합니다.

3. 정부는 교육 또는 배포 중인 대규모 AI 시스템의 레지스트리에 위의 정보를 통합해야 합니다 . 이 레지스트리는 감사 결과, 사고 보고서, 내부 고발자 공개 및 컴퓨팅 사용량을 추적하여 규제 기관이 잠재적으로 문제가 있는 시스템을 식별할 수 있도록 합니다.

4. 정부는 AI 시스템 개발자와 소유자가 합리적으로 예측하고 예방할 수 있는 AI 시스템으로 인한 피해에 대해 법적 책임을 지도록 의무화해야 합니다.

4) AI통제

4. 정부는 새로운 위험에 대해 추가 조치를 취해야 합니다.

AI 전문가들은 또한 정부 에 예외적으로 위험한 기능을 갖춘 미래 AI 시스템에 필요한 표준을 준비하고 규제 당국을 설립할 것을 촉구합니다 .

1. 정부는 자원 집약적인 프론티어 AI 시스템과 같이 매우 위험한 능력을 발휘할 수 있는 AI 시스템 훈련을 위한 라이센스 시스템 구축을 준비해야 합니다 .

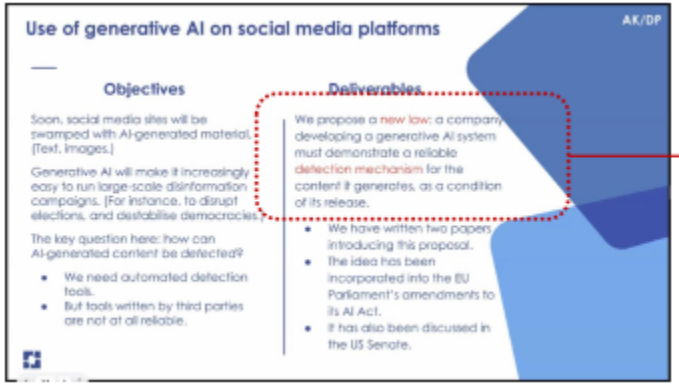
2. 정부는 규제 기관이 향후 개발을 중단할 수 있도록 권한을 부여해야 합니다.
   훈련 중에 위험한 능력을 발휘하는 AI 시스템.

3. 정부는 그러한 프론티어 AI 시스템과 훈련 코드에 대한 접근 통제를 의무화해야 합니다. 요슈아 벤지오(Yoshua Bengio) 가 제안한 대로 , 딥 러닝의 창시자 중 하나인 AI 연구소는 이 정보의 외부 공유를 제한해야 하며 직원이 알아야 할 필요에 따라 액세스할 수 있도록 해야 합니다.

4. 정부는 모델 확산을 방지하기 위해 위험한 국경 AI 시스템에 접근 할 행위자에게 정보 보안 조치를 요구해야 합니다 . 경제적 이익과 악의적인 사용을 위한 고급 AI의 유용성을 고려할 때 AI 연구소에는 가장 높은 표준의 보안 조치가 필요하며 이는 지능형 지속 위협 에도 장벽을 제시합니다. (APT) 및 내부자 위협.

EU의 AIA(인공지능법)

- 수정(Amendment): 범용, 생성형 AI ◀ '투명성' 강화
  - Register General-Purpose, Generative AI models in an EU database ▶ 〈고위험〉
  - Assess and mitigate the risks their models entail ▶ 〈위험평가〉
  - Disclose that the content was generated by AI ▶ 〈Disclaimer〉
  - Design the model to prevent it from generating illegal content ▶ 〈Tool〉
  - Publish summaries of copyrighted data used for training ▶ 〈저작권〉



OECD GPAI
Responsible AI WG
〈2023. 9. 5〉

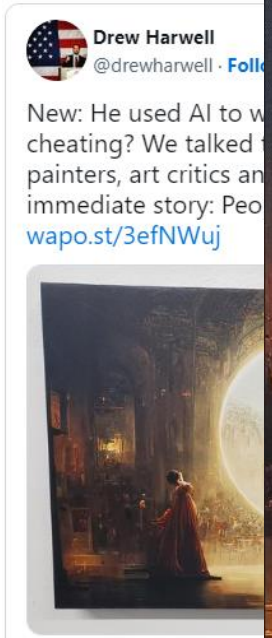
Théâtre D’opéra Spatial — Jason M. Allen

Théâtre D'o

중간 이미지

작품

FPRI
Future Policy Research Ins



저작권법 - AI생성물

미국 저작권청 검토위원회("위원회")는 "Théâtre D"라는 제목의 저작물 “예술 작품 권리 주장” 등록거절

제작 : MidJourney(프롬프트 624번이상) → Adobe Photoshop 수정 → Gigapixel AI (업스케일)
저작권신청 : AI로 제작된 저작물의 저작권 요청
1차검토 : 2022년 9월 21일 1차신청검토(AI생성작업 부분은 저작권제외 요청였으나 Allen은 거절) → 2022년 9월 21일 거절
1차재심요청 : 2023년 1월 24일 (인간 저작자 요건 잘못적용 주장)재요청 → 재심요청 거절
2차재심요청 : 2023년 7월 12일 (다양한 유용한 도구의 출력에 대한 저작권 보호거부초래 주장)재요청 → 재심요청 거절

AI가 생성한 콘텐츠가 최소량 이상임

Copyright Review Board
United States Copyright Office · 101 Independence Avenue SE · Washington, DC 20559-6000

2023년 9월 5일

타마라 페스터, Esq.
Tamara S. Pester, LLC
PO Box 6601
Denver, CO 80206

답변: Théâtre D'opéra 등록 거부에 대한 두 번째 재심 요청
공간(SR # 1-11743923581, 대응 ID: 1-5T5320R)

친애하는 페스터 선생님께:

미국 저작권청 검토위원회("위원회")는 "Théâtre D"라는 제목의 저작물에서 2차원 예술 작품 권리 주장 등록을 거부한 것에 대해 재검토해 달라는 Jason M. Allen("Mr. Allen")의 두 번째 요청을 고려했습니다. '오페라 공간'('일'). 위원회는 두 번째 재심 요청의 주장과 함께 신청서, 기탁 사본 및 관련 서신을 검토한 후 등록 프로그램의 등록 거부를 확인합니다. 위원회는 저작물에 인공지능(“AI”)에 의해 생성된 콘텐츠가 최소한의 양 이상 포함되어 있음을 확인했으며, 따라서 이 콘텐츠는 등록 신청서에서 부인되어야 합니다. Mr. Allen은 AI 생성 자료에 대한 권리를 부인할 의사가 없으므로 해당 저작물은 제출된 대로 등록될 수 없습니다.

포토샵 작업은 저작자의 노력을 인정할 수는 있으나, AI가 생성한 부분은 제외하여야 함.
그러나 불가분하게 병합되고 분리할 수 없어 저작물 등록은 불가

"저작물"의 법적해석 : **인간**의 창작을 요구

동물은 제외

저작권이 있는 저작물은 지적,창의적,예술적 노동능력을 갖춘 **창작자(인간)**이 있어야 함

타마라 페스터, Esq.
타마라 S. 페스터, LLC

2023년 9월 5일

IV. 결론

여기에 명시된 이유로 미국 저작권청 검토위원회는 저작물에 대한 저작권 주장 등록 거부를 확인합니다. 37 CFR § 202.5(g)에 따라 이 결정은 Mr. Allen의 2022년 9월 신청에 관한 최종 기관 조치를 구성합니다.13

미국 저작권청 검토위원회
Suzanne V. Wilson, 법무 자문위원 겸 저작권 준회원
Maria Strong, 저작권 준회원 겸 정책 및
국제 문제 담당 이사

Jordana Rubel, 법무 자문위원

https://en.wikipedia.org/wiki/Théâtre_D%27opéra_Spatial

테아트르 도페라 스페이셜(Théâtre D'opera Spatial)

아티스트	Midjourney , Jason M. Allen의 프롬프트 사용
년도	2022년
중간	Midjourney를 사용한 AI 아트

Théâtre D'opéra Spatial은 Jason Matthew Allen의 프롬프트를 사용하여 생성 인공 지능 플랫폼 Midjourney 에서 만든 이미지입니다[1] 이 그림은 9월 5일 2022년 콜로라도 주 박람회의 연례 미술 대회에서우승하면서 뉴스 기사가 되었으며[2] [3]

"디지털 예술/디지털 조작 사진"에 대한 박람회 콘테스트에서 우승한 것은 Allen이 부정 행위를 했다고 비난한 예술가들의 반발을 불러일으켰습니다. [4] Allen은 "나는 그것에 대해 사과하지 않을 것입니다. 나는 이겼고 어떤 규칙도 어기지 않았습니다."라고 대답했습니다. [삼]

두 명의 카테고리 심사위원은 Midjourney가 AI를 사용하여 이미지를 생성했다는 사실을 알지 못했지만 나중에 이 사실을 알았더라면 Allen에게 최고 상을 수여했을 것이라고 말했습니다. [삼]

텍스트 프롬프트를 사용하여 생성된 이미지는 Gigapixel AI를 사용하여 업스케일되었습니다. [5]

2023년 9월 미국 저작권청 검토위원회는 [6] Théâtre D'Opéra Spatial이 "인간이 아닌 사람이 제작한 작품을 제외"하는 규정에 따라 저작권 보호를 받을 자격이 없다는 사실을 확인했습니다. [7] 이 결정은 Office가 AI와 관련하여 제공한 이전 지침 [8] 과 원칙에 기초하여 Thaler에 대해 판결을 내린 최근 법원 사건(Thaler v 특허, 디자인 및 상표 감사관 [9])을 재확인합니다. 저작권법에 명시되어 있지는 않지만 미국 저작권청에서 사용하는 업무 원칙입니다.

Allen은 다음과 같이 싸울 것이라고 주장합니다.

"Allen은 자신의 작품을 등록하려고 끈질기게 노력했습니다. 그는 Midjourney가 불러일으킨 것을 조작하기 위해 자신이 얼마나 많은 작업을 했는지, Adobe Photoshop을 사용하여 원본 이미지를 얼마나 조작했는지 자세히 설명하는 서면 설명을 저작권 사무소에 보냈습니다. 결함을 수정하고 Gigapixel AI를 사용하여 크기와 해상도를 높였습니다. 그는 그림을 만드는 데 최소 624개의 텍스트 프롬프트와 입력 수정이 필요하다고 명시했습니다." [10]

일부 법률 작가들은 그의 주장을 지지하며 [11] 이를 사진 처리 및 전자 카메라의 현대적 사용과 비교하는 기술적 차별의 한 형태로 간주합니다.

미국 저작권청과 영국이 AI의 영향에 대해 협의하고 있다는 점을 고려하면 Allen의 법적 이의는 시기적절할 것입니다. [12] [13]

생성물 - AI의 사진 수상

진짜 사진과 AI로 생성한 이미지. 이 둘이 섞여 있던 사진계에 한 작가가 의문을 던졌다. AI로 만든 이미지를 세계적인 대회에 출품해 수상작으로 뽑힌 뒤, 상을 거부했다.

해프닝의 시작은 올해 3월 2일, 소니 월드 포토그래피 어워드 주최 측이 크리에이티브 부문 수상작으로 '가짜 기억 : 전기공'이라는 작품을 선정하면서부터다. 이 작품은 독일 사진작가 보리스 엘다크젠이 생성 AI 툴을 써서 만들었다. 작가는 씨즈와의 인터뷰에서 "큰 사진 대회가 AI로 생성한 이미지에 대한 준비가 돼 있는지 알아보려고 시험 삼아 출품했다"고 설명했다.

수상 소식을 들은 작가는 주최 측에 이 작품이 AI로 생성한 거라고 알렸다. 그러면서 상을 줄 거면 생성 AI로 만든 이미지를 어떻게 볼 것인지 논의하는 장을 마련하고 싶다고 밝혔다. 하지만 그 뒤로도 한동안 주최 측이 공식적인 논의를 진행하지 않았다. 작가는 4월 13일 런던에서 열린 시상식에 참석했지만, 결국 수상을 거부했다. 아래는 작가가 밝힌 입장문이다.

"제 이미지를 골라 주셔서 감사합니다. 그리고 세계적인 '사진' 대회에서 AI로 생성한 이미지를 처음으로 수상작으로 뽑는 역사적인 순간을 만들어 주셔서 감사합니다. 여러분 중에 이 이미지를 AI가 만들었다고 의심한 사람은 몇 명이나 있을까요? 아니면 뭔가 이상하다고 느낀 사람은요? AI로 만든 이미지와 사진은 이런 대회에서 서로 경쟁해서는 안 됩니다. 그들은 서로 다른 실체입니다. AI는 사진이 아닙니다. 그러므로 상을 받지 않겠습니다."

AI로 생성한 이미지와 사진을 어떻게 구분할 수 있을까? 씨즈는 작가의 고민을 직접 들어보고 싶어 독일어 인터뷰를 요청했다. 자세한 인터뷰는 영상에서 확인할 수 있다.

생성물 - AI의 사진 수상

https://www.enuri.com/knowcom/detail.jsp?kbno=2894964

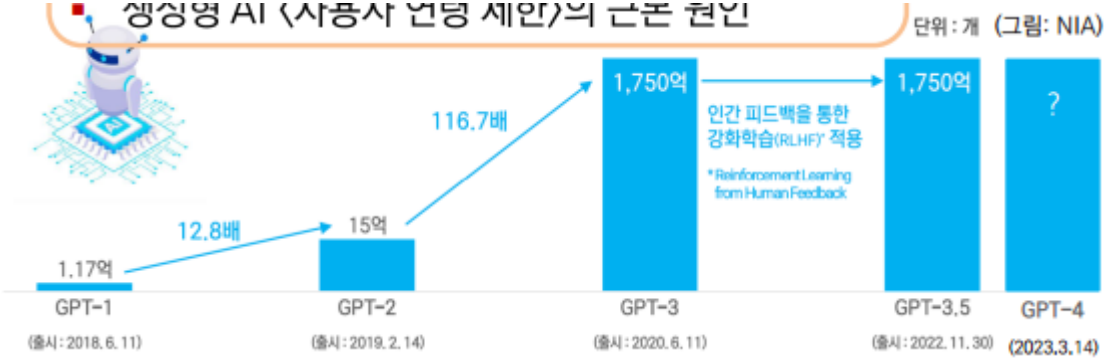
https://www.enuri.com/knowcom/detail.jsp?kbno=2894964

(참고) AI 발전

저작권 침해 - AI윤리, 위험평가, 산업

훈련데이터의 저작권침해

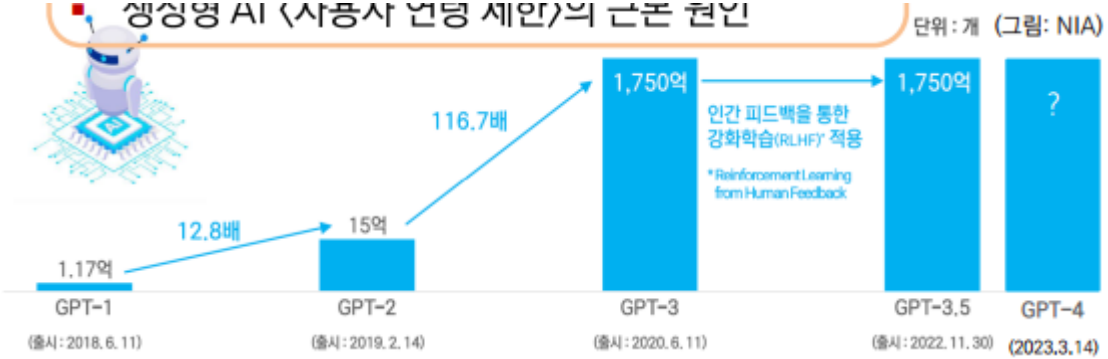
생성물에 대한 저작권논란



저작권 침해 - AI윤리, 위험평가, 산업

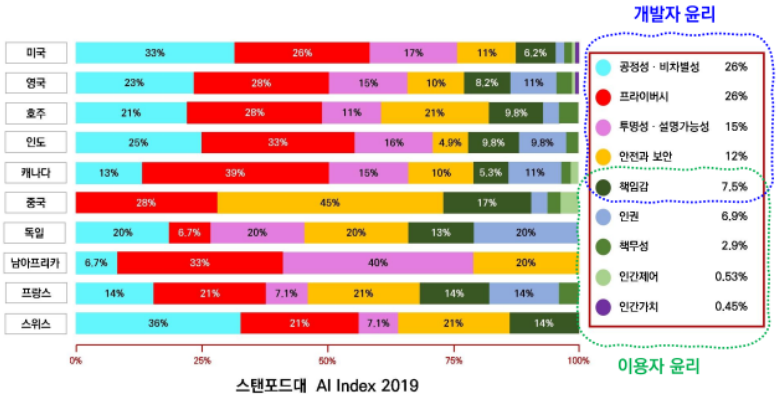
훈련데이터의 저작권침해

생성물에 대한 저작권논란



(참고) AI 이슈

AI 윤리 이슈의 글로벌 동향



스탠포드대 AI Index 2019

훈련데이터 생성물
저작권 침해 - AI윤리, 위험평가, 산업

훈련데이터의 저작권침해

생성물에 대한 저작권논란

생성형 AI의 '표절'과 '저작권' 침해

- "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

WWW '23, May 1–5, 2023, Austin, TX, USA Lee et al.

	Type	Machine-Written Text	Training Text
복사하여 붙이기	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. [...]
출처인용 없이 문장 바꾸기	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].
아이디어 도용	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 *.**

생성물 - 사람에게의 영향AI의 ////

벨기에 피에르 자살사건 2023.3.28

2. 버전 3.5 및 4.0에서 이전 버전 3.0으로의 회귀와 일탈을 가능하게 하는 탈옥(Jailbreak, Safety Bypass) 위험

- 벨기에 30대 가장 '피에르' 자살 사건 〈
  - 대화형 AI 앱 "차이(Chai)" : 〈GPT-J〉라는 LLM 모
  - AI의 가스라이팅 능력 + 의인화 현상
  - 일라이자 효과(Eliza Effect) 연상: 1966 MIT 조셉
- 생성형 AI 〈사용자 연령 제한〉의 근본 원

1. 사실 아닌 내용을 사실처럼 생성해내는 할루시네이션 (hallucination) 위험

- 얀 르쿤 (Yann LeCun, NYU 교수, 메타 AI 연구소 수석총괄)
  - "ChatGPT, Auto-Regressive Generative Models Suck!"
  - "P(correct) = (1 – e)^n" (단, n은 토큰 수)
  - "할루시네이션은 챗GPT가 궁극적 AGI가 아니라는 반증"
  - World Model 제시 (먼저 상식을 배우는 AI)

- 할루시네이션 극복을 향한 다양한 시도들
  - 최신 Fact 검색결과를 문맥 앞에 결합(Bard, Bing Chat)
  - InstructGPT, RLHF의 지속적 고도화(ChatGPT)
  - 상식 추론(Common Sense Reasoning)과의 결합 (인지과학)
    - Global Ontology라는 거대 지식베이스(KB) 구축 필요
      [예] Douglas Lenat(1950-2023)의 "Cyc" – 24 million
- 파인튜닝(Finetuning)에 의한 요동(fluctuation) 현상

생성물 - 사람에게의 영향AI의 ////

문화적 다양성 감소
가짜뉴스 생성

4. 인공지능 합성산출물(synthetic output)의 위험

1) 반복 학습에 따른 문화 다양성 감소 위험
2) 가짜 뉴스에 의한 민주주의 붕괴 위험

- Recursive Learning the Synthetic data의 위험성 경고: 모델 망가짐
- UNESCO 인공지능 권고 : 문화의 다양성 유지 강조
- Disclaimer 활용 – 미 의회 법안 〈HR.3230-2019〉, 워터마크 활용
- Not by AI, Hand-made 명시 운동

- 악의적 목적으로 사람이 직접 작성함
  - 시간과 노력이 소요됨
- 조회수 중심의 대중적 영향력
  - 파워유저 중심, 매크로 도구 활용
- 구분방법
  - 작성자(매체)에 대한 신뢰 지수 유지
  - 글의 관계도 분석에 의한 신뢰성 평가

챗GPT

- 악의적 목적으로 사람이 AI에게 지시함
  - 빠른 시간, 다양하게 대량 작성이 가능함
- AI 학습을 위한 크롤링 대상
  - 조회수보다는 다수의 유사 글 중심
- 구분의 어려움
  - AI 합성산출물과 인간 저작물 구분 불가
  - 기존 방식의 추가 적용 가능

AI의 ////

벨기에 피에르 자살사건 2023.3.28

2. 버전 3.5 및 4.0에서 이전 버전 3.0으로의 회귀와 일탈을 가능하게 하는 탈옥(Jailbreak, Safety Bypass) 위험

- 벨기에 30대 가장 '피에르' 자살 사건 〈2023.3.28〉
  - 대화형 AI 앱 "차이(Chai)" : 〈GPT-J〉라는 LLM 모델 기반
  - AI의 가스라이팅 능력 + 의인화 현상
  - 일라이자 효과(Eliza Effect) 연상: 1966 MIT 조셉 바이첸바움 교수
- 생성형 AI 〈사용자 연령 제한〉의 근본 원인

단위 : 개 (그림: NIA)

(참고) AI 윤리

□ AI기반 사회

- AI에 기반한 자동화된 결정으로 인간스스로의 통제력이 감소된 사회에서 살아가야 할것
  : AI는 인간을 보조하는 단순 범용기술이 아닌,
  생산성, 프라이버시, 평등, 차별, 노동시장, 감시 등 많은 사회적이슈와 연결되어있음

- AI윤리원칙의 수립 후 현실에의 적용을 위해서는, 실행가능한 제도/절차를 마련해야 함
  : AI윤리준수 점검리스트나 AI관련 윤리/법제도 영향평가 수행 등

과학과 신기술에 대한 윤리 그룹의 윤리적 원칙들		
인간의 존엄성	정의, 공정, 연대	보안, 안전, 심신의 건전성
자율성	민주주의	데이터 보호와 프라이버시
책임	법치와 책무성	지속가능성

출처: European Group on Ethics in Science and New Technologies (2018) "Statement on Artificial Intelligence, Robotics and Autonomous Systems".

□ 윤리 : "사람으로서 마땅히 행하거나 지켜야 할 도리"

ethic : "A set of moral principles, especially ones relating to or affirming a specified group field, or form of conduct"

□ 윤리적 합의 : 개인/기업/사회/법적 가치의 지향점을 고려하여 핵심존중가치를 균형있게 존중하는 방식으로 AI를 개발/활용하기 위한 합의적 절충점을 제도적 장치로 마련 (Trade Off)

□ 인공지능 윤리 : 인공지능 관련 이해관계자들이 준수해야 할 보편적 사회 규범

AI의 "자동화된 결정(automated decisions)'이 다양한 사회적 가치를 최대한 존중하는 방식으로 활용되기 위해, 고민해야 할 주제를 고르고, 어떻게 제도화 해야 하는지 대한 논의가 필요하며, 이는 AI 개발/활용 전 과정에 적용되어야 한다.

* 관련주제 : 공정, 책임성, 투명성, 설명가능성, 자율/통제정도, 인권 대 AI의 법인격(인간중심주의 균형점), 기술혁신과 사화가치보전사이의 균형(AI기술개발금지논의), 윤리영향평가 필요여부

'인간중심적(human Centered)'
(AI는 도구, AI의 결정에 대한 책임은 인간에게 있음)

- 자율주행과 같은 경우 인간의 실시간 개입이 어려운 경우 의사결정 과정에서 인간이 반드시 필수불가결한 방식으로 참여하는 Human in the Loop 방식이 아니라 인간이 어떤 방식이든지 AI의 설계와 작동 등 전체 관리 체계에서 중요한 역할을 담당하고 문제가 있다고 판단될 때는 전체 과정을 중지시킬 권한을 갖는 Human on the Loop 방식이 필수적임

□ 인공지능 거버넌스 (AI리스크에 기반한 AI 윤리의 제도화)

: 적응적 거버넌스(adaptive governance)라는 개념으로 접근할 필요

- 사회적 공감대와 합의가 도출된 영역부터 제도화
- 이후 기술발전, 사회적 인식 변화가 지속 반영되어야 함

정책수립(윤리영향평가, 거버넌스, 감독의무, 시스템 보안, 프라이버시 데이터 보호 등) → 모니터링 및 평가

AI-Ethical Impact Assessment (AI-EIA)

□ AI윤리의 법제화

: '적응적' 거버넌스를 구체적으로 어떻게 실현할 지 구체적 법제화 //TODO

인공지능은 **"인간(조직)"**이 제공한 데이터와 알고리듬에 기반하여 판단하므로,
→ 기본적으로 AI윤리원칙(AI Ethical Principles)에 대한 사회적 합의 및 이를 준수하려는 자율적 활동이 필요하고,
→ 더 나아가 최소한의 내용은 법적규제사항으로 정리되어야 한다.

자율성을 가진 AI에 대한 논의 필요, 책임 귀속의 최종 주체는 인간/사회가 되므로 개발/적용/규제 이전에 사회적 합의가 필수적임, 그래서 인공지능 윤리선언이 선행된 후 이를 기반으로 법제도영역에 반영하는 순서로 진행되어야 함.

AI윤리준수 점검리스트

<해외 주요 인공지능 윤리 체크리스트 현황>

	발표주체	체크리스트	
기업	Microsoft	AI 공정성 체크리스트 AI Fairness Checklist	'20.03
	Lex Mundi & Cambrian Futures	법인 법률업무를 위한 AI 준비도 체크리스트 AI Readiness Checklist for Corporate Legal Functions	'20.04
공공	WEF & 싱가포르 IMDA	AI 거버넌스 프레임워크 지침 Companion to the Model AI Governance Framework	'20.01
	EU	신뢰할 수 있는 AI를 위한 자율평가목록 The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment	'20.07
	영국 ICO	AI 및 데이터보호 안내 지침 Guidance on Artificial Intelligence and Data Protection	'20.07
학계	Carnegie Mellon University's Software Engineering Institute	윤리적 AI 경험 설계 체크리스트 Designing Ethical AI Experiences: Checklist and Agreement	'19.12

자료: 정보통신정책연구원 지능정보사회정책센터 내부자료(2020)

[별첨] 유네스코 인공지능 윤리권고 (2021년 11월 25일)

유네스코 인공지능 윤리권고 (2021년 11월 25일)

(a) 인공지능 시스템은 현실 가상 환경에서 예측 및 의사 결정과 같은 결과를 도출하는 인지 과제의 학습 수행 능력을 생성하는 모델 및 알고리즘을 통합한 정보 처리 기술이다
(b) 인공지능 시스템 수명 주기를 관리, 운영, 매매, 재무, 모니터링 및 심사, 검증, 종료, 해체, 폐기를 비롯하여 연구, 설계, 개발로부터 배포, 사용에 이르는 과정까지 포함하는 것으로
이해할 때, 인공지능 시스템에 관한 윤리적 질문은 각 단계마다 존재한다
(c) 인공지능 시스템은 인공지능이 의사 결정, 고용 및 노동, 사회적 상호작용, 건강관리, 교육, 미디어, 정보 접근성, 디지털 격차, **개인정보** 및 소비자 보호, 환경, 민주주의, 법치주의, **보안** 및 치안유지, 군민 양용(dual use), 그리고 표현의 자유, 프라이버시, 비차별을 포함하는 인권 및 근본적 자유에 미치는 영향을 비롯하여 (단, 이에 국한되지 않는) 새로운 유형의 여러 윤리 문제를 유발한다.

. 본 권고는 인공지능 시스템 수명 주기 전반에 걸쳐 법적 규제적 틀을 개발하고 기업의 의무를 촉진시켜야 할 행위 주체이자 관계 당국인 회원국을 대상으로 한다. 또한, 본 권고는 **인공지능 시스템 수명 주기 내내 인공지능 시스템에 대한 윤리영향평가의 토대를 제공**함으로써 공공 민간 부문을 비롯한 모든 인공지능 행위 주체에게 윤리 지침 을 제공한다.

Right to Privacy, and Data Protection
프라이버시 권리 및 정보 보호

32. Privacy, a right essential to the protection of human dignity, human autonomy and human agency, must be respected, protected and promoted throughout the life cycle of AI systems.

32. 인간 존엄성, 인간 자율성, 인간 활동의 보호에 핵심적 권리인 **프라이버시**는 인공지능 시스템 수명 주기 전 영역에서 반드시 존중, 보호, 증진 되어야 한다.

It is important that data for AI systems be collected, used, shared, archived and deleted in ways that are consistent with international law and in line with the values and principles set forth in this Recommendation, while respecting relevant national, regional and international legal frameworks.

인공지능 시스템을 위한 **데이터**가 국제법에 부합하고 본 권고에서 명시된 가치와 원칙에 맞게, 그리고 또한 유관 국가 지역 국제적 법률 틀을 존중하는 식으로 **수집, 사용, 공유, 보관, 삭제**되는 것은 중요하다.

33. Adequate data protection frameworks and governance mechanisms should be established in a multi-stakeholder approach at the national or international level, protected by judicial systems, and ensured throughout the life cycle of AI systems.

33. 적절한 **데이터 보호 틀 및 거버넌스 메커니즘**은 국가 국제적 차원의 다자적 접근법 으로 확립되어야 하고, 사법 체계에 의해 보호받아야 하며, **인공지능 시스템 수명 주기** 전 영역에서 보장되어야 한다.

Data protection frameworks and any related mechanisms should take reference from international data protection principles and standards concerning the collection, use and disclosure of personal data and exercise of their rights by data subjects while ensuring a legitimate aim and a valid legal basis for the processing of personal data, including informed consent.

데이터 보호 틀 및 관련 메커니즘은 **개인 정보의 수집 사용 공개 및 데이터 주체의 권리 행사와 관련된 국제 데이터 보호 원칙 기준을 참고**해야 하며, 동시에 인지동의를 비롯하여 개인 데이터 처리를 위한 적법한 목적 및 타당한 법적 토대도 확고히 해야 한다.

34. Algorithmic systems require adequate privacy impact assessments, which also include societal and ethical considerations of their use and an innovative use of the privacy by design approach.

34. 알고리즘 시스템은 이의 개인정보 사용과 설계 단계에서의 개인정보의 혁신적 사용에 대한 사회 윤리적 고려를 포함하여 적절한 프라이버시 영향평가를 요구한다.

AI actors need to ensure that they are accountable for the design and implementation of AI systems in such a way as to ensure that personal information is protected throughout the life cycle of the AI system.

인공지능 행위 주체는 가령 개인 정보가 인공지능 시스템의 전 영역에서 보호받음을 보장하는 식으로 그들이 인공지능 시스템의 설계 및 구현에 책임을 짐을 보장해야 한다.

Transparency and explainability
투명성 및 설명가능성

37. The transparency and explainability of AI systems are often essential preconditions to ensure the respect, protection and promotion of human rights, fundamental freedoms and ethical principles.
37. 인공지능 시스템의 **투명성 및 설명가능성**은 인권 및 근본적 자유, 윤리적 원칙에 대한 존중 보호 증진을 확고히 함에 있어 필수 선결조건이다.

Transparency is necessary for relevant national and international liability regimes to work effectively.
투명성은 관련 국내 국제 법적 책임 체제가 효과적으로 작동하는 데에 필수적이다.

A lack of transparency could also undermine the possibility of effectively challenging decisions based on outcomes produced by AI systems and may thereby infringe the right to a fair trial and effective remedy, and limits the areas in which these systems can be legally used.
또한 투명성의 부족은 인공지능 시스템이 산출한 결과물에 기반한 의사결정에 대한 **실질적 이의제기 가능성을 저해**할 수 있고, 이에 따라 공정한 재판 및 배상을 받을 권리를 침해 할 수 있으며, 이러한 시스템이 법적으로 사용될 수 있는 영역을 제한할 수 있다.

38. While efforts need to be made to increase transparency and explainability of AI systems, including those with extra-territorial impact, throughout their life cycle to support democratic governance, the level of transparency and explainability should always be appropriate to the context and impact, as there may be a need to balance between transparency and explainability and other principles such as privacy, safety and security.
38. 인공지능 시스템 수명 주기 전 영역에서 민주적 거버넌스를 지원하기 위해서는 (법역 외 영향까지도 포함하여) 이의 투명성 및 설명가능성을 향상시키려는 모든 노력이 이루어 져야 하는 반면, 투명성 및 설명가능성의 수준은 항상 해당 맥락과 영향 정도에 따라 적절한 수준이어야 하는데, 이는 투명성 및 설명가능성과 프라이버시, 안전 및 보안과 같은다른 원칙 사이에서 균형을 이루어야 할 필요성이 있을 수 있기 때문이다

Transparency and explainability
투명성 및 설명가능성

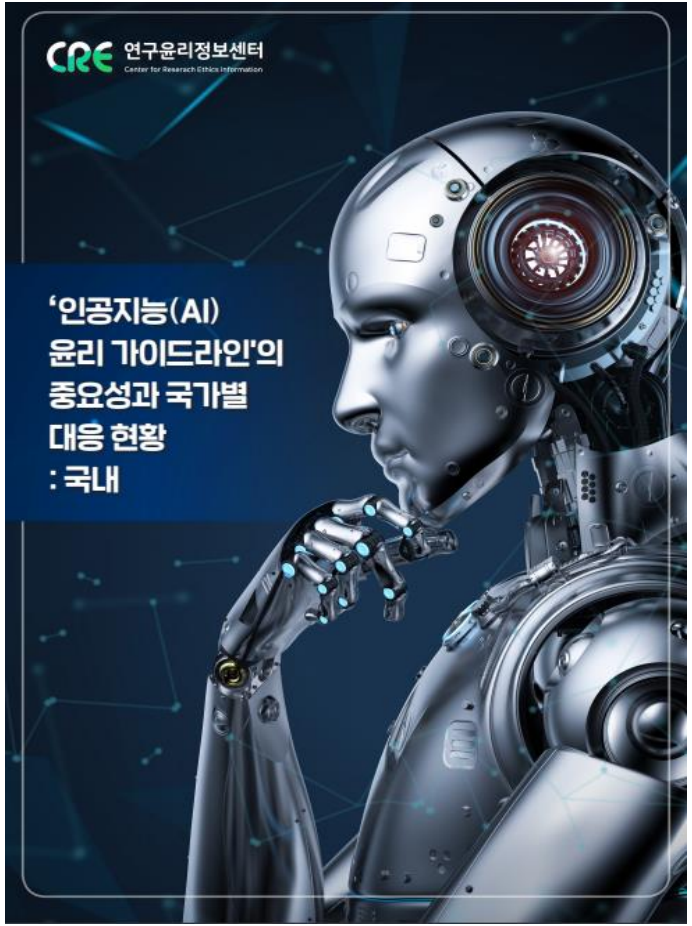
People should be fully informed when a decision is informed by or is made on the basis of AI algorithms, including when it affects their safety or human rights, and in those circumstances should have the opportunity to request explanatoryinformation from the relevant AI actor or public sector institutions.
사람들은 의사결정이 인공지능 알고리즘으로부터 정보를 얻거나 이에 기반하여 내려지는 경우, 특히 이로써 그들의 안전 또는 인권에 영향이 가는 경우에 사전에 알 수 있어야 하고, 그러한 상황에서 관련 인공지능 행위 주체 또는 공공 기관에서 **정보 설명을 요구할 수 있는 기회**를 가지고 있어야 한다.
In addition, individuals should be able to access the reasons for a decision affecting their rights and freedoms, and have the option of making submissions to a designated staff member of the private sector company or public sector institution able to review and correct the decision.
더욱이, 개인은 자신의 권리 및 자유에 영향을 미치는 **결정에 대한 근거에 접근할 수 있어야 하고, 이 결정을 검토 정정**할 수 있는 민간 기업 또는 공공 기관의 담당 직원에게 **의견을 개진할 수 있는 선택권**을 가지고 있어야 한다.
AI actors should inform users when a product or service is provided directly or with the assistance of AI systems in a proper and timely manner.
인공지능 행위 주체는 제품이나 서비스가 직접적으로 또는 인공지능 시스템의 보조를 통해 제공될 때 사용자에게 이를 시기적절하게 **사전 통보**해야 한다.

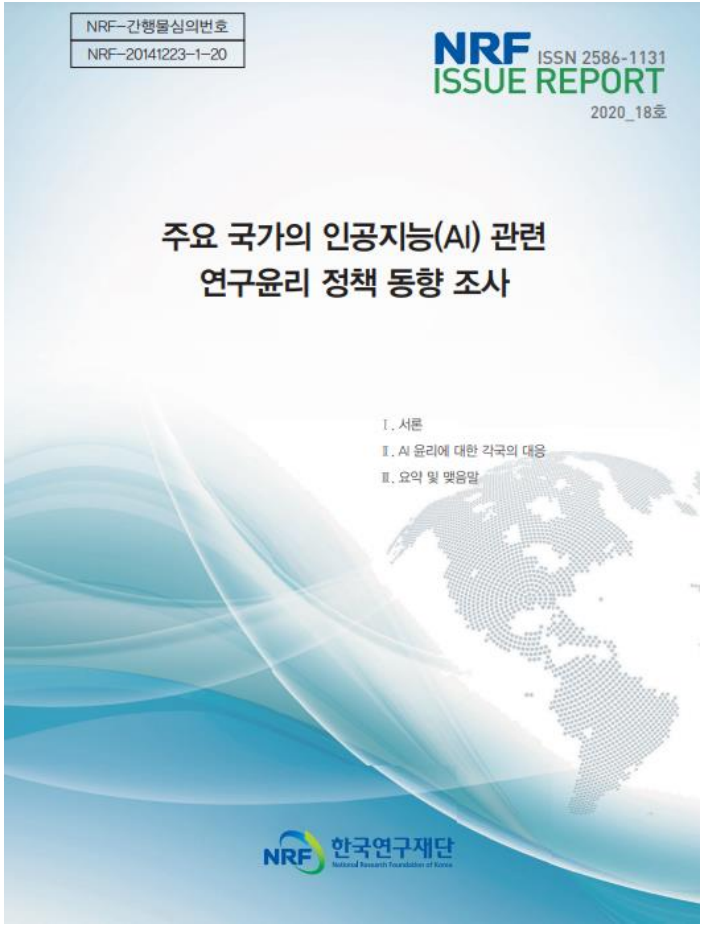
Responsibility and accountability
책임 및 책무

42. AI actors and Member States should respect, protect and promote humanrights and fundamental freedoms, and should also promote the protection of the environment and ecosystems, assuming their respective ethical and legal responsibility, in accordance with national and international law, in particular Member States' human rights obligations, and ethical guidance throughout the lifecycle of AI systems, including with respect to AI actors within their effective territory and control.
42. 인공지능 행위 주체 및 회원국은 인공지능 시스템 수명 주기 내내 국내 국제법, 특히 국제 회원국의 인권 준수 의무와 인공지능 시스템의 수명 주기 전 영역에서의 윤리 지침에 따라 인공지능 행위 주체를 효과적인 영역과 통제 하에 둠으로써, 각각의 윤리적 법적 책임를 지고 인권 및 근본적 자유를 존중 보호 증진하며 환경 및 생태계 보호를 장려해야 한다.
The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors corresponding to their role in the life cycle of the AI system.
일단 인공지능 시스템에 기반하여 내린 결정 및 행동에 대한 윤리적 책무와 법적 책임은 항상 궁극적으로 인공지능 시스템 수명 주기에서의 해당 역할의 인공지능 행위 주체에게 귀속된다.
43. Appropriate oversight, impact assessment, audit and due diligence mechanisms, including whistle-blowers' protection, should be developed to ensure accountability for AI systems and their impact throughout their life cycle.
43. 인공지능 시스템과 이의 수명 주기 전 영역에서의 영향력에 대한 책임을 보장하기 위해서는 제보자 보호를 비롯하여 **적절한 감독, 영향 평가, 감사, 실사 메커니즘이 개발**되어야 한다.
Both technical and institutional designs should ensure auditability and traceability of (the working of) AI systems in particular to address any conflicts with humanrights norms and standards and threats to environmental and ecosystem well-being.
특히 인권 개념 기준과의 충돌 및 환경 및 생태계의 안녕에 대한 위협이 발생하는 경우, 기술적, 제도적 설계는 이를 해결할 수 있도록 인공지능 시스템(또는 이의 작동)에 대한 **감사가능성 및 추적가능성**을 보장해야 한다.

[별첨] 한국연구재단- 연구윤리정보포탈 자료

한국연구재단 CRE 연구윤리정보포털 https://www.cre.or.kr/





AI 윤리 대응 동향

■ 네이버

○ 인공지능 기반 서비스에서 안전한 데이터 활용을 위해 'Privacy By Design 원칙'을 적용

- 서비스 출시 단계에서부터 높은 프라이버시 보호 기준으로 데이터 활용에 있어 법령 위반사항이나 이용자 프라이버시 보호 관점 에서 문제가 없는지를 검토하고 있다.

○ '네이버 프라이버시센터' 홈페이지, 개인정보 보호 블로그 및 SNS를 통해 이용자 프라이버시 보호를 위한 활동과 관련 정보를 공개하고 이용자의 개선제안 등을 받고 있다.

'네이버 프라이버시 센터' 홈페이지에서는 네이버의 개인정보 보호원칙), 서비스 운영 관련 정책, 보호 활동, 투명성 보고서, 개인정보 보호와 활용 논의를 위한 Privacy White Paper 등 프라이버시 관련 보고서를 제공하고 있다

| 네이버의 프라이버시 관련 보고서 |

구분	내용	발행주기
투명성 보고서 통계	네이버(주)에서 수사기관에 제공한 이용자 정보 통계	연 2회
개인정보보호 리포트	네이버(주)의 개인정보 보호 활동	연 1회
Privacy White Paper	'이용자 프라이버시 보호' 관련 전문 연구 수행 결과	연 1회

자료: 네이버 프라이버시센터 홈페이지

■ 카카오

○ 2018년 1월 31일, 국내 업계 최초로 인공지능 알고리즘 윤리 헌장을 발표했다.

| 카카오의 윤리 헌장 |

원칙	내용
1. 카카오 알고리즘의 기본원칙	카카오는 알고리즘과 관련된 모든 노력을 우리 사회 윤리 안에서 다하며, 이를 통해 인류의 편익과 행복을 추구한다.
2. 차별에 대한 경계	알고리즘 결과에서 의도적인 사회적 차별이 일어나지 않도록 경계한다.
3. 학습 데이터 운영	알고리즘에 입력되는 학습 데이터를 사회 윤리에 근거하여 수집·분석·활용한다.
4. 알고리즘의 독립성	알고리즘이 누군가에 의해 자의적으로 훼손되거나 영향받는 일이 없도록 엄정하게 관리한다.
5. 알고리즘에 대한 설명	이용자와의 신뢰 관계를 위해 기업 경쟁력을 훼손하지 않는 범위 내에서 알고리즘에 대해 성실하게 설명한다.

■ 삼성전자

○ 2018년 국내기업으로는 최초로 인공지능 국제 컨소시엄인 PAI(Partners on AI)에 가입

- 인공지능 미래에 대한 범사회적인 논의를 진행 중인 삼성전자는 향후

AI 안전성(Safety-Critical AI)과 AI 공정성·투명성·책임성(Fair, Transparent, and Accountable AI), AI의 사회적 영향(Social and Societal Influences of AI) 등 다양한 분야에 참여를 확대할 계획이다.

출처 : https://ethics.moe.edu.tw/files/resource/ebook/file/ebook_01_kr.pdf

■ 한국인공지능협회

○ 2012년 스타트업 실무자들의 커뮤니티로 시작하여 2017년 사단법인으로 설립. "산업의 지능화"와 "AI 기술의 대중화"를 목표로 인공지능 기술 기업 및 유관기관들을 중심으로 네트워크를 구축

○ 인공 지능 산업에 필요한 윤리의식과 교육의 필요성을 논의하기 위한 목적으로 2019년 7 월 '제1회 인공지능 윤리+교육 포럼'을 개최하였으며, 포럼 이후 메니패스토의 일환으로 협회 클러스터 기업 72개 업체가 참여한 2019 인공지능 윤리+교육 포럼 공동선언문을 채택했다.

| 2019 인공지능 윤리+교육 포럼 공동선언문 |

본 선언문은 대한민국 인공지능 기술의 지속적 성장과 안전한 발전을 위한 윤리적 연구 개발 뿐 아니라, 이와 연계하여 올바른 인공지능 교육을 할 수 있도록 개인과 기업, 국가 그리고 세계기구에 실천적 가이드라인을 제시하는 데 그 목적이 있다.

1. 인공지능 기술은 처음부터 '사람 우선' 원칙으로 인류에게 육체적·정신적·환경적으로 도움을 줄 수 있도록 연구·개발되어야 하고 동시에 설명될 수 있어야 한다. 또한, 우리가 그것을 신뢰하고 공정하게 공유하여 인류 공동의 선을 실현할 수 있도록 해야 한다.

2. 국민 개인은 '자기의 기술(The Technology of the Self)'을 학습하여 개인 데이터를 스스로 관리해야 한다. 또 미래 인공지능 기술이 가져올 직업적 변화에 대비하여 자신의 적성과 능력을 파악, 소비자인 동시에 생산자가 될 수 있는 새로운 디지털기술을 습득하고 그것을 발전시켜야 한다.

3. 모든 기업은 시작단계에서부터 '안전 우선'을 원칙으로 관계자 교육훈련을 포함한 인공지능 개발과정을 설계하고, 해당 제품의 사람과 사회에 대한 영향을 종합적으로 분석하여 차별과 편견이 없도록 제작해야 한다. 그리고 비상시를 대비하여 위험요소를 즉각 제거할 수 있는 통제기술을 해당 제품에 포함해야 하며, 제품의 전체 제작과정을 표준화시키고, 사후 관리에 책임을 지며, 영업비밀을 훼손하지 않는 범위에서 그것을 설명할 수 있어야 한다.

4. 국가 교육기관은 국민 각자가 국민공동기본교육과정을 통해 생애주기별로 필요한 '스마트 라이프(Smart Life)' 활용기술을 습득하여 일상 생활능력을 확장할 수 있게 하며, 직업적으로 개인의 역량을 최대로 향상시킬 수 있는 '자기 계량화(Quantified Self)' 기술을 포함한 '인간공학(Human Engineering)'에 대한 학습기회를 제공해야 한다.

5. 국가는 국민의 인권과 재산, 프라이버시를 보호하기 위해 산업별 인공지능 기술의 긍정적·부정적 영향을 전체적으로 조사·분석·예측하여 그 결과를 공유하고, 관계자에 대한 자격부여 및 교육·훈련을 해야 한다. 특히, 종합적 제품관리를 위해 기업 경쟁력을 침해하지 않는 범위에서 산업별 효준화된 평가 매트릭스를 적용하여 투명성과 타당성을 확보하고, 합리적 개입과 감독을 포함한 국가 차원의 가버넌스시스템을 구축해야 한다.

6. 인공지능 기술을 그 경계가 없이 세계적으로 확장될 수 있으므로 개인과 기업은 세계시민으로서의 역할교육을 받아야 하며, 한 나라의 국가적 가브넌스 시스템은 UN과 같은 세계기구의 글로벌 시스템과 연동시켜 지속해서 발전해 나가야 한다.

7. UN등 세계기구는 인공지능 기술이 인간의 생리적 진화를 앞지를 수 있는 문명사적 변혁에 대비하여, 인간의 존엄성과 그 존재론적 의미와 가치를 재발견해야 한다. 나아가서는 '사람 우선' 의 원칙을 기반을로 인공지능과 인간의 표준화된 기술적 협업 방법을 개발하고 보급하여 상호갈등을 없애고, 세계평화와 질서를 유지할 수 있는 정책을 강구해야 한다.

자료: 한국인공지능협회 홈페이지

■ 고등과학원 초학제연구단

■ 한국인공지능법학회

■ 한국인터넷윤리학회

■ 정보통신정책연구원

■ 방송통신위원회

○ 2019년 11월 11일 정보통신정책연구원과 함께 AI 시대 정부, 기업, 이용자 등 구성원이 지켜야 할 '이용자 중심의 지능정보서비스 기본 원칙' 발표.

○ 지능정보서비스 기본 원칙 주요내용
- (**사람 중심**의 서비스 제공) 지능정보서비스의 제공과 이용은 사람을 중심으로 그 기본적 자유와 권리를 보장하고 인간의 존엄성을 보호할 수 있는 방향으로 이루어져야 한다.
- (**투명성과 설명가능성**) 지능정보서비스가 이용자에게 중대한 영향을 끼칠 경우 기업의 정당한 이익을 침해하지 않는 범위에서 이용자가 이해할 수 있도록 관련 정보를 작성해야 하며, 이용자 기본권에 피해를 유발했을 때 예측, 추천, **결정의 기초로 사용한 주요요인을 설명**할 수 있어야 한다.
- (**책임성**) 지능정보사회의 구성원들은 지능정보서비스의 올바른 기능과 사람 중심 가치의 보장을 위한 공동의 책임을 인식하고, 관련한 법령과 계약을 준수한다.
- (**안전성**) 안전하고 신뢰 가능한 지능정보서비스의 개발과 이용을 위해 모두가 노력하고, 지능정보서비스가 초래할 수 있는 **피해에 대한 자율적인 대비체계**를 서비스 제공자와 이용자가 수립하고 운영한다.
- (**차별금지**) 지능정보서비스가 사회적·경제적 불공평이나 격차를 초래할 수 있다는 점을 인식하고, 알고리즘 개발과 사용의 모든 단계에서 차별적 요소를 최소화할 수 있도록 노력한다.
- (**참여**) 지능정보사회의 구성원들은 공적인 이용자 정책 과정에 차별 없이 참여할 수 있으며, 공적 주체는 제공자와 이용자가 실질적으로 의견을 제시할 수 있는 정기적인 통로를 조성해야 한다.
- (**프라이버시와 데이터거버넌스**) 지능정보서비스의 개발, 공급 및 이용의 전 과정에서 **개인정보 및 프라이버시를 보호**하며, 구성원들은 기술적 이익의 향유와 프라이버시 보호 사이의 균형을 위해 지속적인 의견 교환에 참여한다.

○ 이용자 보호를 위한 공동의 노력
- 지능정보사회의 구성원들은 공동의 기본 원칙에 입각하여 지능정보사회의 기본 가치를 수호할 수 있도록 자율적인 노력을 지속한다.
- 지능정보사회 이용자 보호를 위한 실질적이고 지속적인 논의를 위해 지능정보사회의 구성원들이 참여하는 **협의회를 구성하여 운영**한다.

출처 : https://ethics.moe.edu.tw/files/resource/ebook/file/ebook_01_kr.pdf

■ 정부

○ 2019년 12월 17일 정부는 문재인 대통령 주재로 열린 제 53회 국무회의에서, 과학기술정보통신부를 비롯한 전 부처가 참여하여 마련한「인공지능(AI) 국가전략」을 발표

▶ AI 반도체 세계 1위, 전국 단위 AI 거점화 등 세계를 선도하는 AI 생태계 조성

▶ 전 생애·모든 직군에 걸친 AI교육 실시 및 세계 최고의 AI인재 양성

▶ 현 전자정부를 차세대 지능형 정부로 대전환하여 국민 체감도 향상

▶ 사회보험 확대 등 일자리 안전망 확충 및 AI 윤리 정립으로 사람 중심 AI 실현

- AI를 통해 경제효과 최대455조원 창출, 삶의 질 세계 10위 도약(~'30) -

| AI 윤리 규범 및 가이드라인 마련 현황 |

○ 역기능 방지 및 AI 윤리체계 마련

❶ AI 기반 사이버침해 대응체계 고도화('20~)

❷ 딥페이크* 등 신유형의 역기능 대응을 위한 범부처 협업체계 구축('20)

(AI 기반 영상 합성기술 또는 그 영상. 신시장 창출과 동시에 명예훼손 등 부작용도 우려)*

❸ AI 신뢰성·안전성 등을 검증하는 품질관리체계 구축 추진('20~)

❹ OECD 등 글로벌 규범에 부합하는 AI 윤리기준 확립('20) 및 AI 윤리교육 커리큘럼* 개발·보급('21~)

((학생·이용자) AI와 생명윤리, 개인정보보호 / (개발자) 윤리적 AI 설계, 정보보안 등)*

❺ 이용자 보호를 위한 중장기적 정책 수립 지원체계 마련

○ 정부는 대통령 직속의 현 4차산업혁명위원회를 AI의 범국가 위원회로 역할을 재정립하여 이번 전략의 충실한 이행을 위한 범정부 협업체계를 구축해 나갈 방침

- 특히, 대통령 주재 전략회의를 개최하여 전 국민 교육, 전 산업 AI 활용 등 범정부적 과제의 실행력을 확보하고, 대국민 성과 보고대회도 병행하여 국민의 참여와 성과 확산에도 노력할 계획

■ 금융위원회

○ 금융 분야 인공지능(AI) 윤리 가이드라인 및 인프라를 조성해 AI 기술 활성화에 나설 계획.

- 금융위원회는 2020년 7월 16일 국내·외 금융 분야 AI 활용 및 정책 동향 파악을 목표로 '금융 분야 인공지능(AI) 활성화' 워킹 그룹을 조직, 첫 회의를 개최

금융 분야 AI 활성화 정책을 추진할 수 있도록 워킹 그룹을 구성하고, 약 4개월간 운영해 '금융 분야 AI 활성화 방안'을 마련할 계획이다.

워킹 그룹은 ▲AI 관련 규제 ▲AI 인프라 ▲소비자 보호 ▲레그테크*와 섭테크* 접목 4개 측면에서 AI 활성화 정책을 추진한다.

○ AI 활성화를 목표로 관련 규제 개선 및 규율 체계를 정립.

- 가이드라인 형태로 AI 금융 서비스 개발에 특화한 실무 프로세스를 마련하고, 과학기술정보 통신부 AI 법제정비단과 협력해 적법성·공정성을 담은 '금융 분야 AI 윤리 가이드라인'을 갖춘다.

○ AI 활용 시 소비자 권한을 보호할 수 있는 체계를 확보.

- AI 업무 처리로 발생하는 소비자 피해를 해결할 수 있도록 책임 주체와 구제 절차 등 기준을 마련한다. AI가 도출한 결과를 객관적으로 설명할 수 있는 '설명가능 AI(XAI)' 기준도 정립한다.

*(*레그테크는 '레귤레이션(규제)'과 기술을 의미하는 '테크놀리지(기술)'의 합성어로 AI를 활용해 복잡한 금융 규제를 기업이 쉽게 이해하고 지킬 수 있도록 하는 기술이다. *섭테크는 '슈퍼비전(감독)'과 '테크놀리지(기술)'의 합성어로 신기술을 활용해 금융 감독 업무를 효율적으로 수행하는 기술이다.)*

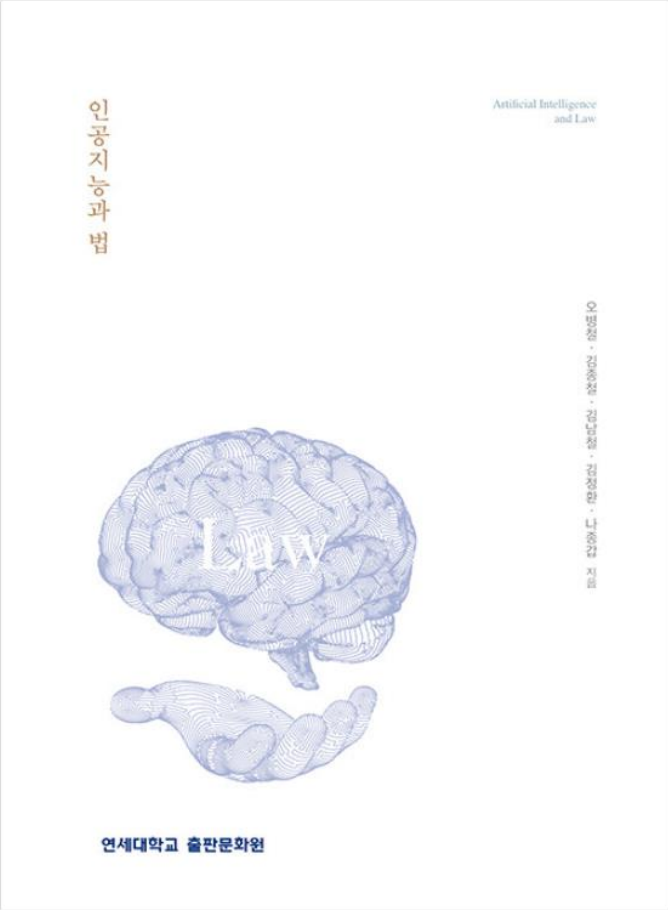
■ 과학기술정보통신부

○ AI 법제도 개선 및 윤리 정립

향후 계획으로는 AI로 인한 경제·사회 전반의 패러다임 변화에 대응하고자 규제개선 사항을 종합하여 AI분야 법제도 개선 로드맵을 제시할 계획이다.('20.11월 잠정) 아울러 금년 12월 AI시대의 기본법제인 『지능정보화 기본법』이 시행됨에 따라 하위법령을 완비하고, 주요국, 국제기구 등의 AI 윤리규범을 비교 분석하여 우리나라의 AI 윤리 기준도 정립할 예정이다.

출처 : https://ethics.moe.edu.tw/files/resource/ebook/file/ebook_01_kr.pdf

[별첨] 인공지능과 법



제1장 인공지능의 헌법적 의의 _ 김종철
I. 헌법의 의의와 기능 그리고 지능정보사회
II. 인공지능과 민주주의
III. 인공지능과 기본적 인권
IV. 헌법에 입각한 인공지능 규율체계의 기본원칙과 체계

제2장 인공지능과 민사법상 법인격 _ 오병철
I. 서론
II. 전자인격의 시기와 종기
III. 전자인격의 권리능력 범위
IV. 전자인격의 물권관계
V. 기타
VI. 결론

제3장 인공지능과 계약 _ 오병철
I. 서론
II. 인공지능이 한 의사표시
III. 인공지능을 이용한 계약의 무효
IV. 인공지능을 이용한 계약의 취소
V. 인공지능 의사표시의 발신과 도달
VI. 결론

제4장 인공지능과 불법행위 _ 오병철
I. 인공지능에 의해 통제되는 로봇에 의한 손해의 발생
II. 로봇의 행위성
III. 로봇과 귀책사유
IV. 편익책임의 구체적 검토
V. 결론

제5장 인공지능과 데이터 _ 오병철
I. 서론
II. 데이터의 분류
III. 데이터의 귀속 - 데이터는 누구의 것인가?
IV. 데이터의 활용을 위한 법률관계
V. 결론

제6장 인공지능과 행정규범체계 _ 김남철
I. 법의 이해
II. 행정법의 이해
III. 행정조직법
IV. 행정작용법
V. 행정구제법
VI. 행정규제
VII. 인공지능 관련 법률

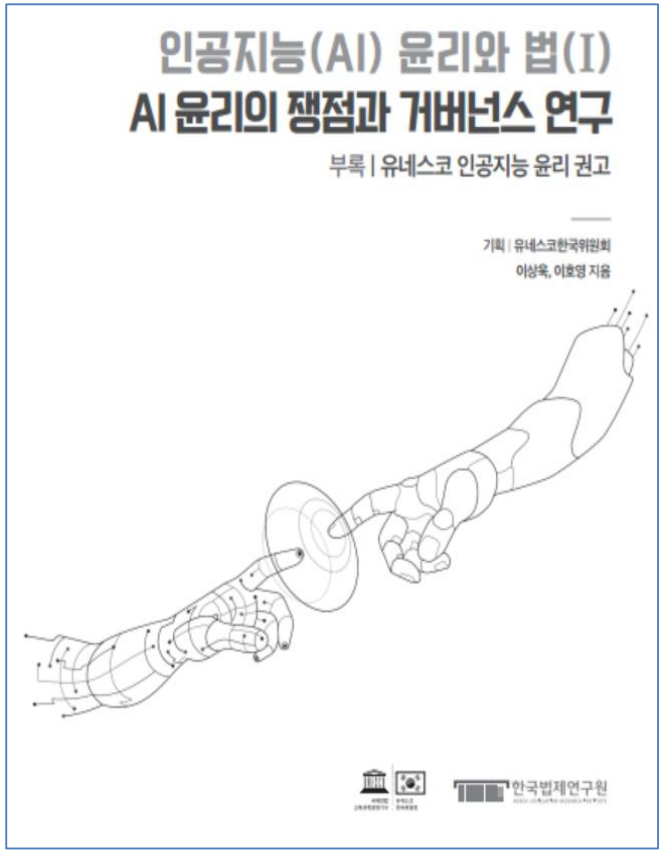
제7장 형법의 개념과 인공지능의 처벌 _ 김정환
I. 인공지능과 형사법의 논의 방향
II. 형법
III. 중세의 동물재판
IV. 범죄 주체

제8장 일반 자동차 사고와 자율주행 자동차 사고의 형사책임 _ 김정환
I. 일반 자동차 사고의 형사책임
II. 자율주행자동차 사고의 형사책임

제9장 형사소송법의 개념과 인공지능의 활용 _ 김정환
I. 형사소송법
II. 증거재판주의
III. 형사절차에서 인공지능의 활용

제10장 인공지능과 지식재산권 _ 나종갑
I. 지식재산/지적재산
II. 지적노동의 결과물 보호에 대한 철학
III. 인공지능과 실정법적 보호
IV. 인공지능 관련한 특허, 영업비밀, 저작권 및 부정경쟁방지법의 비교

[별첨] 유네스코 인공지능 윤리권고 (2021년 11월 25일 193개 유네스코 회원국의 만장일치로 'AI 윤리 권고'를 공식적으로 채택)



FPRI
Future Policy Research Institute

news1 KOREA

"인공지능 판사, 민주적 정당성 차원서 문제될 수 있어"

서미선 2019.12.18. 18:17

인공지능(AI) 판사가 도입돼 국가 권력을 AI가 행사할 경우 민주적 정당성 차원에서 문제가 될 수 있다는 의견이 나왔다.
김중권 중앙대 법학전문대학원 교수는 18일 서울 서초동 서울법원종합청사 청심홀에서 열린 'AI와 법 그리고 인간' 심포지엄에서 주제 발표를 통해 이같이 밝혔다.
김 교수는 "민주주의와 법치국가원리는 민주적 정당성을 지닌 자연인에 의한 지배를 바탕으로 한다"며 "국가권력 행사에서의 AI도입은 특히 인적 민주적 정당성 차원에서 문제될 수 있다"고 지적했다.
그는 "기계가 AI에 의해 스스로 향상될 수 있는 시점까지 어떤 전문가시스템도 인간통제 없이 출현할 수는 없다"며 "인풋과 아웃풋을 통제할 수 있는 전문가를 항상 필요로 할 것이고, 정보지식을 구비한 더 유능한 법조인이 요구된다"고 봤다.
이어 "맹목적 기술신봉에서 법조직업의 종말을 외치는 건 전혀 타당하지 않다"며 "현재의 한계를 인식하고 AI프로그램을 효과적으로 활용해야 할 것"이라고 언급했다.
그러면서 "AI 입법과 AI 재판, AI 행정에 관한 움직임 이면엔 기성 국가작용 메커니즘에 대한 심각한 불신이 있다"고 진단했다.

2019.12.18
AI와 법 그리고 인간

https://www.msn.com/ko-kr/news/national인공지능-판사-민주적-정당성-차원서-문제될-수-있어/ar-BBY6UoD

사법분야에서 AI를 활용하려면 윤리적 사용지침이 마련돼야 한다는 제언도 나왔다.
에스토니아의 Kai Härmand 법무부차관(판사)은 450여개 기관과 150여개 공공기관이 쓰는 에스토니아의 국가정보 교환플랫폼인 엑스로드(X-road)를 소개하고, **사법분야에선 검색·번역·기록·자문영역에서 AI를 활용할 수 있다면서 이처럼 말했다**.
오스트리아의 마르크 코켈버그(Mark Coeckelbergh) 빈 대학교 교수는 "**AI로 인한 도덕적·법적 책임주체는 기술이 될 수 없고 인간이 돼야 하는데, 주체가 불분명하고 AI 알고리즘이 불투명한 문제가 있다**"고 지적하며 AI 윤리지침을 제시했다.
구체적으로 **인간의 감독과 기술적 안정성, 데이터 관리, 투명성, 차별금지, 공정성이다.**
옥스퍼드 딥 테크 분쟁해결연구소 리 지 엔(Ji En Lee) 연구원은 증거분석과 자동기록을 통해 판사업무를 지원하는 중국의 '206' 시스템, 교통사고 사건을 위한 싱가포르의 시뮬레이션 프로그램, 범죄자 위험평가 도구인 미국의 콤파스(COMPAS)를 소개했다.
이어 "법률분야에 AI를 적용할 때 생길 수 있는 윤리적 문제를 해결하려면 공정, 투명, 설명 가능성, 기본권 존중, 데이터 정확성과 보안, 협력과 포용, 이용자에 의한 통제 등 요소를 고려해 AI를 개발해야 할 것"이라고 말했다.
발표에 앞서 강현중 사법정책연구원장은 개회사에서 "이미 전산화돼 축적된 법률문서 및 정보가 AI에 의해 활용된다면 법조인 업무에도 큰 변화가 올 것"이라며 "정보기술(IT) 강국인 한국은 AI영역에서도 앞서나갈 저력이 있다"고 밝혔다.
조재연 법원행정처장은 축사에서 법원이 추진하는 차세대전자소송 시스템구축 사업을 언급하고 "축적된 기존 전자소송문서 등 정보를 빅데이터 형태로 AI기술에 활용할 수 있을 것"이라고 설명했다.
이번 심포지엄은 사법정책연구원이 중앙대 인문콘텐츠연구소, 한국인공지능법학회, AI정책포럼과 공동으로 대한변호사협회와 한국연구재단, 교육부 후원을 받아 개최했다.

“만약 당신이 미래를 꿈꾸지 않거나 지금 기술개선을 위해 노력하지 않는다면 그건 곧 낙오되고 있는 것이나 마찬가지 입니다.”

그윈 쇼트웰(Gwynne Shtwell, SpaceX CEO, COO)

감사합니다

(facebook.com/sangshik, mikado22001@yahoo.co.kr)

AI윤리기준

정부의 '사람이 중심이 되는 인공지능 윤리기준'(2020)
카카오 알고리즘 윤리 헌장(2018)
네이버AI윤리준칙(2021)

지능정보화 기본법

금융 분야 AI 윤리 가이드라인

영혼, AI의 권리

자율

IEEE에서 이미 가장 기초적인 형태의 AI ‘윤리 표준’을 제정하고
있다는 사실이다. IEEE는 상당 기간의 의견 수렴을 거쳐서 추상적인 윤리 원칙을 AI 설계
과정에서 어떻게 반영할 지에 대한 연구를 진행했고 이를 Ethically Aligned Design이라는 개념으로 이미 제시한 바 있다. IEEE는 여기에 그치지 않고 이를 AI 설계의 윤리적 적합성을 평가하는 산업 기준 개발로 연결하고 있는데 현재 가장 일반적인 수준의 P7000

개인정보영향평가 -> AI윤리영향평가

데이터 수집 과정에서
'자발적 동의'를 받기만 했다고 모든 문제가
해결되는 것도 아니다. 여러 윤리적 원칙과
의
충돌 가능성을 데이터 선별 규칙을 통해 구
현해서 결국에는 수집된 데이터 중에서 윤리
적
으로 사용가능한 데이터만을 사용

저작권법 – 자연생성물

Mr. Allen is closer to the plaintiff in **Kelley v. Chicago Park District 2008**

https://www.artistrights.info/decision_kelley-v-chicago-park-district