과학 · 기술
발표논문 04

공동선 추구를 위한 기독교 세계관에 입각한 AI모델

김수환 총신대 교수

요 약

2020년대 인공지능 기술은 데이터를 기반으로 한 기계학습(Machine leanring)이 주류를 이루고 있으며, 기계학습의 한 분야인 딥러닝(Deep learning)은 학습, 추론, 예측, 자연어 처리에서 두각을 나타내고 있다[1]. 최근 등장한 챗GPT는 2022년 11월, 일반인이 사용할 수 있도록 서비스를 시작했는데, 5일 만에 사용자 수가 100만명을 돌파하였다. 2023년 10월에는 이미지를 인식하고 답하는 수준까지 발전했으며, 명령에 따라 이미지를 그려주는 DALL · E 3도 추가서비스로 제공되어서 멀티모달(Multimodal)로 발전하는 양상이 가속화되고 있다[2]. 멀티모달은 시각, 청각을 비롯한 여러 가지 형태의 정보를 주고받을 수 있게되는 것을 의미하는데, 이렇게 되면 사람이 구축해 놓은 지식과 문화유산을 학습할 수 있게 된다. 인공지능의 발전단계에서 보면 '인간의 데이터로 학습하는 시기'를 지나서 '보고 읽으며 배우는 시기'가 진행되고 있다. 이 시기가 지나면 '실행하고 책임지는 시기'로 발전하게 된다.[마야책] 지금까지 말과 글로 축척해 놓은 지식 뿐만 아니라 이미지, 영상으로 구축된 정보와 지식도 인공지능의 학습에 이용되어 사람에게 필요한 정보와 지식을 추천, 예측, 판단하는 사회로 발전하게 될 것이다[3].

인공지능의 발전은 양면성을 가지고 있어서 사람의 편리와 행복을 도울 수 있다는 유토피아적인 발상과 사람을 대체하고 배척하는 데까지 이르게 될 것이라는 디스토피아적인 발상이 함께 나타나고 있다[4]. 두가지 주장은 인간 사회의 공정성과 다양성, 신뢰성 등 여러 측면을 고려하는데, 인류의 발전을 도모하는 방향성을 가지고 있다. 인공지능은 데이터 수집, 데이터 전처리, 데이터를 통한 학습, 모델 생성, 모델 검증, 모델 활용의 과정을 거쳐 구현되는데, 각 과정에서 사회에서 통용하는 가치나 규범을 기준으로 삼는다[5] 를 들어 챗GPT를 구현하는 과정에서 인간 피드백을 반영한 강화학습(Reinforcement Learning Human Feedback, 이하 RLHF)을 거쳤는데, 이때 사회의 공정성, 공평성을 훼손하거나 인류에게 해를 가할 수 있는 질문에는 답변을 회피하도록 학습시켰다[6].

인공지능을 활용하면 공동선을 추구하는 데 도움을 받을 수 있다. 이해관계자들의 입장, 관점, 이익을 분석할 수 있고, 해결책에 대한 다양한 아이디어 도출에 활용할 수 있다. 챗GPT의 사례처럼 인류가 만들어 놓은 지식을 함께 사용하여 효율성과 효과성을 높일 수 도 있다. 인공지능을 교육에 적용하면 맞춤형 교육을 지원하여 교육의 격차를 줄이는 데 도움을 주기도 한다[7].

공동선의 목적은 모든 사람에게 공평하고 공정한 사회를 만드는 것으로 귀결된다. 기독교의 원리는 하나님 사랑, 이웃사랑이라는 대강령을 통해 이땅에서 하나님의 나라를 이루고자 노력하는 것이다. 하나님은 공의하시고 공평하시므로 하나님의 통치가 이땅에 이루어지면 공공선이나 공동선도 이루어질 수 있다. 공동선을 추구하는 데 인공지능을 활용하려면 데이터 생성이나 수집 단계에서부터 공동선의 가치가 담긴 데이터를 만들거나 수집해야 한다. 이때 기독교적 관점을 적용하면 일반사회에서 추구하는 공동선의 기준과 다른 데이터가 필요하다. 예를 들어 일반 사회에서의 공동선은 철학을 기반으로 모두에게 이익이 되는 결과에 가치를 두나 기독교를 기반으로 한 공동선은 공의로우신 하나님의 통치와 섭리에 기반을 두게 된다. 즉, 기독교인을 위한 공동선 추구 가능성은 데이터의 생성단계부터 달라야 한다.

따라서 기독교적 관점에서 공동선을 위한 인공지능 모델은 다음과 같은 사항을 고려해야 한다. 첫째, 데이터 수집 과정에서 기독교적 세계관에 입각한 데이터를 생성하거나 수집해야 한다. 둘째, 데이터의 전처리 과정에서 정통 교리와 성경에 위배되는 내용은 걸러야 한다. 셋째, 모델 훈련 과정에서 챗GPT의 훈련과정처럼 기독교 세계관에 입각한 강화학습을 적용해야 한다. 넷째, 모델 검증 과정에서도 그 결과가 기독교 세계관에 위배되지 않는지 검토해야 한다[8]. 다섯째, 모델을 적용할 때 기독교 공동체의 선을 이루는 데 사용해야 하며, 영성 영역에서 사용하는 것에는 주의해야 한다.

주제어: 기독교 세계관, 인공지능 모델, 기독교적 공동선

참고문헌

[1] 관계부처합동. (2019). 인공지능 국가전략.
[2] Batker, J. et. al.(2023). Improving Image Generation with Better Captions. OpenAI.
[3] 마야 비알릭, 찰스 페댈, 웨인 홈즈. (2020). 인공지능 시대의 미래교육. (정제영, 이선복 역). 박영스토리.
[4] 박태웅. (2023). AI 강의. 서울: 한빛비즈.
[5] 하정우, 한상기. (2023). AI 전쟁. 서울: 한빛비즈.
[6] Introducing ChatGPT. 2023.10.28. 검색 https://openai.com/blog/chatgpt
[7] 소프트웨어정책연구소. (2023). 생성 AI가 노동시장에 미치는 영향. AI Brief 특집호.
[8] 김수환. (2023). ChatGPT를 대하는 기독교인의 자세. 신앙과 삶. 2023년 5~6월호.

공동선 추구를 위한 기독교 세계관에 입각한 AI 모델

김수환(총신대)

공동선(Common Good)

철학, 경제 및 정치에서 공동선(공동체, 일반 복지 또는 공익)은 특정 커뮤니티의 모든 구성원 또는 대부분의 구성원에게 공유되고 유익하거나 시민권 또는 집단적 하나를 통해 달성되는 것입니다. 정치 및 공공 서비스 분야에 대한 행동 및 적극적인 참여.

공동선의 개념은 철학적 교리에 따라 크게 다릅니다. 공동선의 초기 개념은 아리스토텔레스와 플라톤과 같은 고대 그리스 철학자들에 의해 제시되었습니다.

(학술사전)

unesco

함께 그려보는 우리의 미래

교육을 위한 새로운 사회계약

국제미래교육위원회 보고서

교육의 미래 보고서

- 원제 : 함께 그려보는 우리의 미래 – 교육을 위한 새로운 사회 계약
- 2021년 11월
- 유네스코 국제 미래교육위원회
- 2050년 너머 미래를 조망하는 국제교육담론 및 전세계적 교육 변혁 논의 추진
- **전 생애를 통해 양질의 교육을 받을 권리 보장**
- **공공재이자 공동재로서의 교육 필요**

19세기와 20세기에 확립된 교육을 위한 사회계약

- 첫째, 교육은 학급과 교실의 구조 속에서 교사가 가르치는 수업에 기반한 교육학적 프로젝트로 여겨지며, 비록 학습환경을 공유하지만 **개인의 성취가 우선시된다.**
- 둘째, 교육은 개별 과목들로 구성된 **교육과정을 통해 전달된다.**
- 셋째, 교수활동은 대체로 한 학문 분과 내에서 효과적 학습을 지휘하는 **교사 한 명의 전문 역량에 의존하는 단독 활동으로 인식된다**
- 넷째, 학교는 각각의 상황에 상관 없이 건축, 조직, 진행 절차 면에서 상당한 유사성이 있는 **학교 모델에 따라 구축된다.**
- 다섯째, 교육은 비슷한 연령대의 학생 집단을 가족과 지역사회로부터 상대적으로 떨어진 곳에서 운영되는 특별한 전문기관에서 가르치며, **아동 및 청소년이 성인으로서 미래에 살아갈 준비**가 되었다고 여겨질 때 종료되는 방식으로 조직된다.

새로운 인식

- ❖ 새로운 교육학(pedagogies): 협력과 연대
- ❖ 새로운 교육과정 접근방식: 공유지식, 공동유산 + 새로운 지식
- ❖ 교사들에 대한 새로운 책임: 팀워크, 협력적 직업
- ❖ 학교의 수호와 변혁: 학교의 역할 보장, 시간과 장소 리디자인
- ❖ 서로 다른 시간과 공간에서의 교육: 전생애, 다양한 장소, 미래교육

교육을 *"공동재 (a common good)"*로

인간 능력의 발전을 위한 집단적 노력

*공동재 *(a common good)*: 전사회적 참여와 노력으로 함께 만들어가는 의미

우리가 꿈꾸는 교육

앎의 기쁨
개인의성장

+

사회구성원
진로와직업

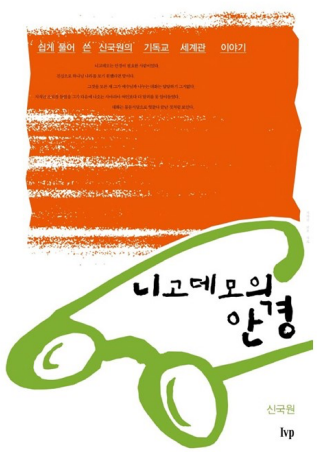
경제적, 문화적, 사회적, 개인적

기독교 세계관

창조

타락

구속



인공지능이란?

앨런 튜링
존 매카시
알파고
챗GPT

AI는 사람의 지적 활동을 컴퓨터를 통해 구현하는 기술

○ AI는 1950년대에 등장한 개념•이나, **혁신적 알고리즘**••, **컴퓨팅파워 혁신, 데이터 폭증** 등으로 최근 급속히 발달

• 수학자 앨런 튜링은 1950년대 「계산기계와 지능」을 통해 AI의 개념적 토대를 놓았으며, AI 용어는 1956년 다트머스 학술회의에서 처음 사용

•• 사람의 뇌를 모방한 **딥러닝 알고리즘** 개발로 학습하는 AI 출현

○ AI의 핵심은 **'학습', '추론', '예측'**으로, 육체노동을 대신하는 이전 기계들과 달리, **지적노동**인지·표현·판단 등이 가능

참고 인공지능 연구와 관련 있는 타 학문 분야의 예

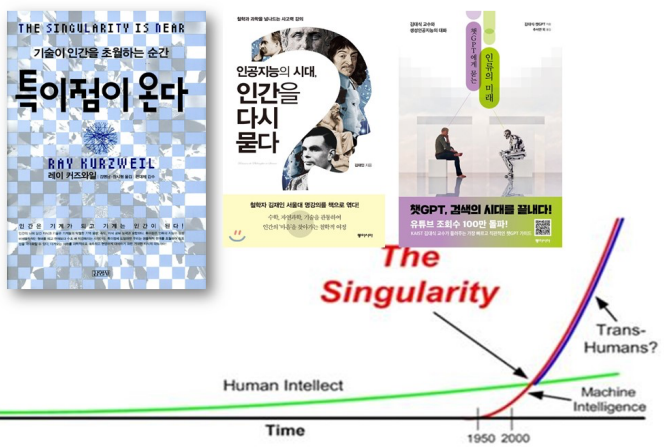
(철학) 도덕적 딜레마와 윤리적 선택 **(통계학)** 데이터 기반 확률적 예측
(뇌인지과학) 기억 등 뇌의 매커니즘 **(심리학)** 인간과 동물의 행동 연구
(언어학) 언어 특성에 따른 사고의 구조
(교육학) 다양한 교수 · 학습 모델 연구 등

교육부(2020). 인공지능시대 교육정책방향과 핵심과제

컴퓨팅 모델

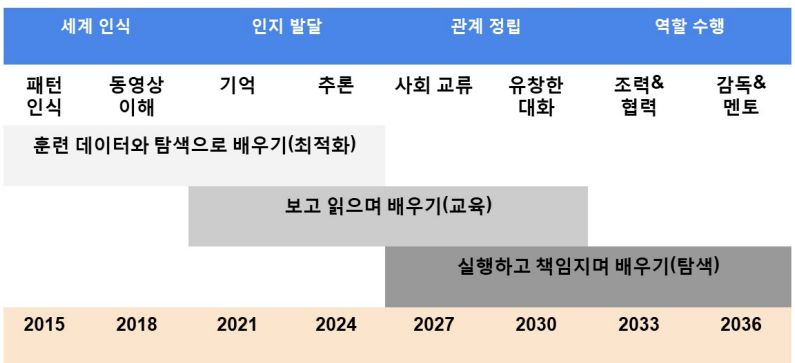
기호주의
연결주의

https://www.junggi.co.kr/mobile/view.html?no=21299



https://www.sramanamitra.com/2017/07/05/man-and-superman-life-near-an-approaching-technology-singularity/

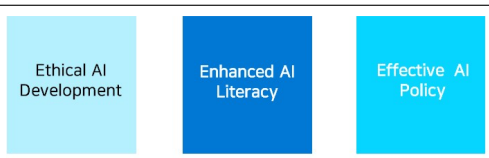
인공지능의 발달

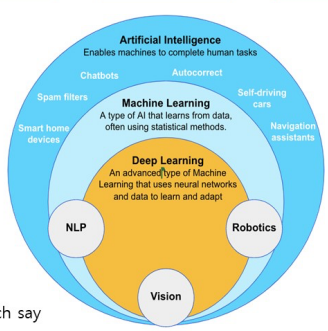


마야 비알릭, 찰스 페댈, 웨인 홈즈(2020). 인공지능 시대의 미래교육

그림 1 인공지능 기술전망

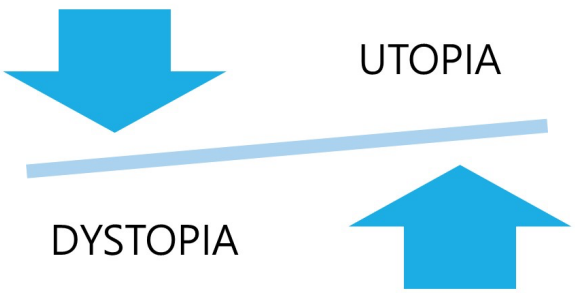
민옥기 외. (2020)/ ATL 1.0: 인공지능 기술 수준 정의. 전자통신동향분석. 35(3), 1-8.



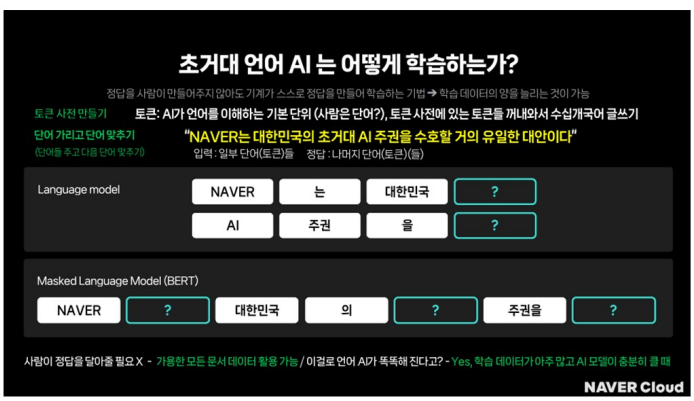


Monica McGill. et al. (2023). What does research say about AI in education?. CSEDCON23.

공동선 + 인공지능 모델



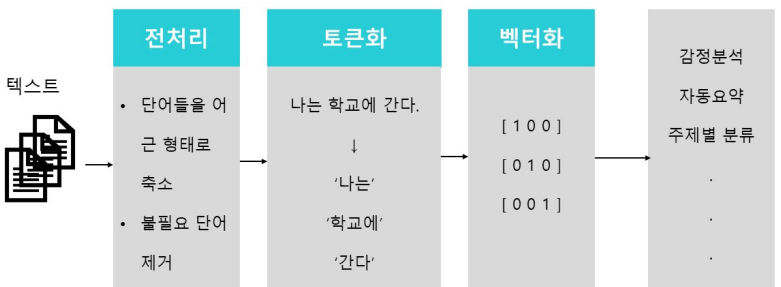
학습 방법



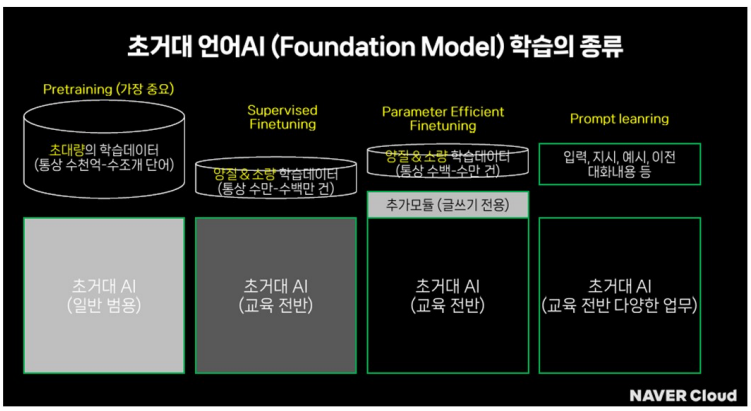
하정우 소장 특강자료(2023). 터치교사단 연수 자료집.

자연어 처리 절차

✓ 맥락을 처리하는 트랜스포머(Transformer)

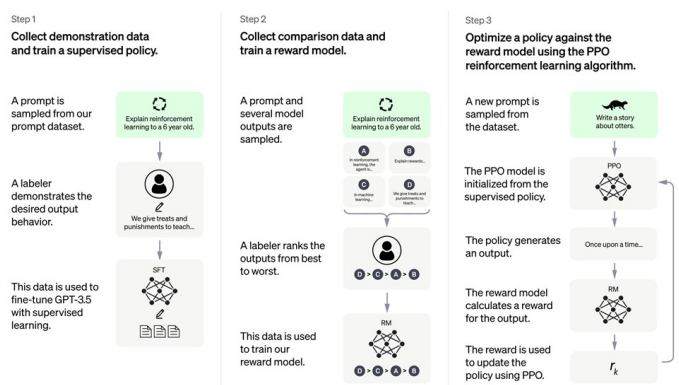


파운데이션 모델



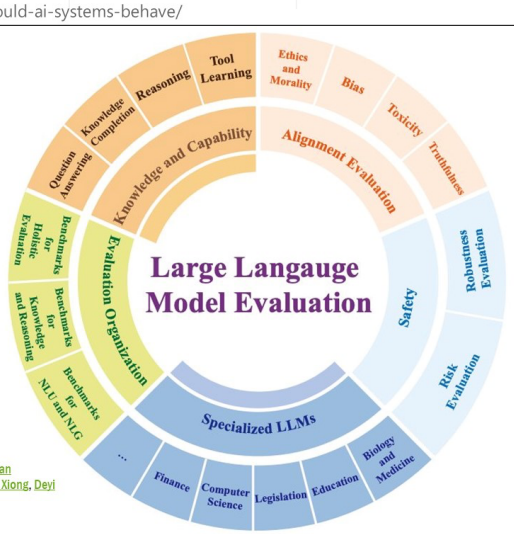
하정우 소장 특강자료(2023). 터치교사단 연수 자료집.

챗GPT 학습 과정



https://openai.com/blog/how-should-ai-systems-behave/

LLM



Evaluating Large Language Models: A Comprehensive Survey
Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong

LLM의 명암

- 대규모 언어 모델(LLM)은 광범위한 작업에서 놀라운 기능을 입증했습니다. 큰 관심을 끌며 수많은 다운스트림 애플리케이션에 배포되었습니다. 하지만 양날의 검과 같이 LLM에는 잠재적인 위험도 존재합니다.
- 개인 데이터가 유출되거나 부적절하고 유해하거나 오해의 소지가 있는 콘텐츠가 생성될 수 있습니다. 또한, LLM의 급속한 발전으로 인해 적절한 보호 장치가 없는 초지능 시스템의 출현 가능성에 대한 우려도 제기되고 있습니다.
- LLM의 역량을 효과적으로 활용하고 안전하고 유익한 개발을 보장하기 위해서는 LLM에 대한 엄격하고 포괄적인 평가를 실시하는 것이 중요합니다.
- 이번 조사 분석에서는 LLM 평가에 대한 종합적인 관점을 제공하기 위해 노력합니다. LLM에 대한 평가를 지식 및 역량 평가, 연계성 평가, 안전성 평가의 세 가지 주요 그룹으로 분류합니다. 이 세 가지 측면에 대한 평가 방법론과 벤치마크에 대한 종합적인 검토와 더불어 전문 영역에서 LLM의 성과와 관련된 평가 개요를 정리하고, 역량, 정렬, 안전성 및 적용 가능성에 대한 LLM 평가를 포괄하는 종합 평가 플랫폼의 구축에 대해 논의합니다.
- 이 포괄적인 개요를 통해 LLM 평가에 대한 더 많은 연구가 촉진되기를 바라며, 궁극적으로는 평가가 LLM의 책임 있는 개발을 유도하는 초석 역할을 하기를 바랍니다. 이를 통해 잠재적 위험을 최소화하면서 사회적 편익을 극대화하는 방향으로 발전할 수 있기를 기대합니다.

에이전트 실험

Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

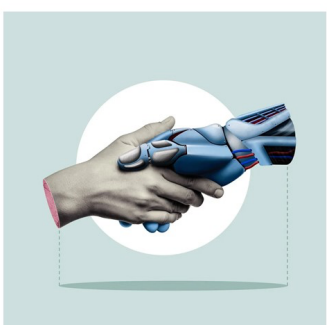
Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 2, 1–22. https://doi.org/10.1145/3586183.3606763

AI and the Next Digital Divide

Brookings Institute



he evolution of the "digital divide":

***The first digital divide*:** The rich have technology, while the poor do not.

***The second digital divide*:** The rich have technology and the skills to use it effectively, while the poor have technology but lack skills to use it effectively.

***The third digital divide?*:** The rich have access to both technology and people to help them use it, while the poor have access to technology only.

Marisa Shuman, Jeremy Bhatia, Kristen DiCerbo. How AI Tools Can Increase Equity and Improve Learning(CSEDcon)

New AI-Powered Sensors Could Tell Teachers What's Really Going on with Students

- The sensors are attached to children's shirts and can also be placed throughout the classroom.
- The data could reveal some important patterns and behaviors that might have been missed otherwise.
- …having clear data will remove any bias from discipline decisions.

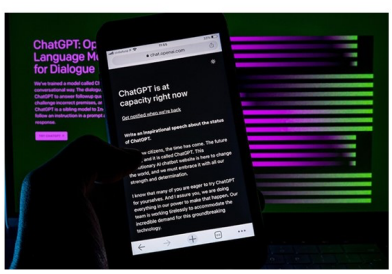
https://www.edweek.org/teaching-learning/new-ai-powered-sensors-could-tell-teachers-whats-really-going-on-with-students/2023/10

Marisa Shuman, Jeremy Bhatia, Kristen DiCerbo. How AI Tools Can Increase Equity and Improve Learning(CSEDcon)

chatGPT : 연구를 위한 조언

Conversational AI is a game-changer for science. Here's how to respond.

* 긴급하고 광범위한 토론 조직 만들기
  - 교육자 : 사용에 대한 윤리 정립
* LLM의 개발 및 사용에 관한 국제적인 포럼
  - 예) 인간 유전자 편집과 같은 기술 사례
* 연구의 다양성과 불평등 고려

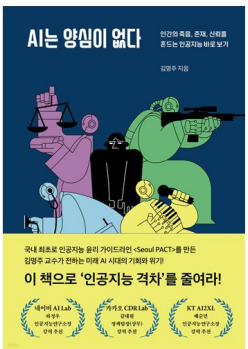


Questions for debate

- Issues for discussion at a forum about conversational AIs.
- Which research tasks should or should not be outsourced to large language models (LLMs)?
- Which academic skills and characteristics remain essential to researchers?
- What steps in an AI-assisted research process require human verification?
- How should research integrity and other policies be changed to address LLMs?
- How should LLMs be incorporated into the education and training of researchers?
- How can researchers and funders aid the development of independent open-source LLMs and ensure the models represent scientific knowledge accurately?
- What quality standards should be expected of LLMs (for example, transparency, accuracy, bias and source crediting) and which stakeholders are responsible for the standards as well as the LLMs?
- How can researchers ensure that LLMs promote equity in research, and avoid risks of widening inequities?
- How should LLMs be used to enhance principles of open science?
- What legal implications do LLMs have for scientific practice (for example, laws and regulations related to patents, copyright and ownership)?

https://www.nature.com/articles/d41586-023-00288-7?fbclid=IwAR2Uz3x1eASIIPUp-2U6O2qX27jCubuimic_xkFWY_pVFe6md5_cF14JrT8

사고의 주체



➢ Embedded Ethics

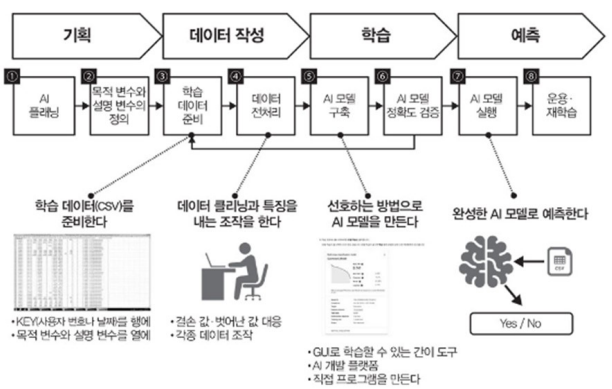
Stanford University

HAI Stanford University Human-Centered Artificial Intelligence

Our Mission

Advancing AI research, education, policy, and practice to improve the human condition.

AI 문제해결 과정



노구치 류치(2020). AI시대, 문과생은 이렇게 일합니다. 전종훈 역, 시그마북스

정보, 지식, 지혜, 진리

정보: 단편적, 비맥락, 비인격적(이데아)

지식: 체계적, 비인격적 (에피스테메)

지혜: 인격적, 전인적 (소피아/호크마)

진리: 영적, 존재론적 (알레테이아/에메스)

신국원(2022). 총신대 세미나 자료.

거룩한 문화의 기초

그러므로 형제들아 내가 하나님의 모든 자비하심으로
너희를 권하노니 너희 몸을 하나님이 기뻐하시는
거룩한 산 제사로 드리라.
이는 너희의 드릴 영적 예배니라.
너희는 이 세대를 본받지 말고 오직 마음을 새롭게
함으로 변화를 받아 하나님의 선하시고 기뻐하시고
온전하신 뜻이 무엇인지 분별하도록 하라.
(롬 12:1~2)

영적 분별력을 갖추라 (고전2:12)

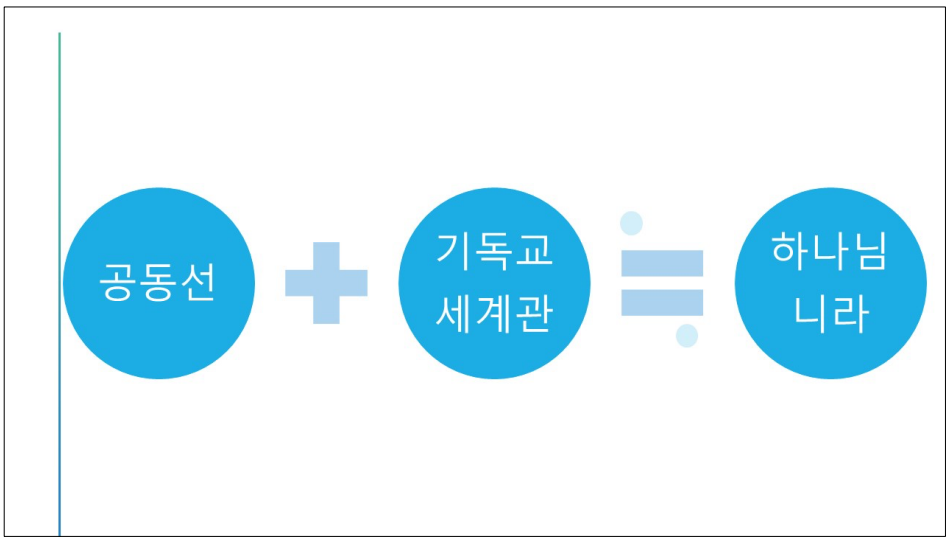
문화는 자연과 달리 인간의 책임이 따른다.
문화의 일꾼으로서 책임감 있는 자세
하나님의 뜻에 대한 영적 분별력
(*holy discernment*)

하나님의 나라와
그의 의를 추구하는 삶

신국원(2022). 총신대 세미나 자료.

기독교 세계관이 입각해서...

- 기독교의 원리는 하나님 사랑, 이웃사랑이라는 대강령을 통해 이땅에서 하나님의 나라를 이루고자 노력하는 것이다.
- 하나님은 공의하시고 공평하시므로 하나님의 통치가 이땅에 이루어지면 공공선이나 공동선도 이루어질 수 있다.
- 공동선을 추구하는 데 인공지능을 활용하려면 데이터 생성이나 수집 단계에서부터 공동선의 가치가 담긴 데이터를 만들거나 수집해야 한다.
- 기독교인을 위한 공동선 추구 가능성은 데이터의 생성단계부터 달라야 한다.



우리의 시도

- 첫째, 데이터 수집 과정에서 기독교적 세계관에 입각한 데이터를 생성하거나 수집해야 한다.
- 둘째, 데이터의 전처리 과정에서 정통 교리와 성경에 위배되는 내용은 걸러야 한다.
- 셋째, 모델 훈련 과정에서 챗GPT의 훈련과정처럼 기독교 세계관에 입각한 강화학습을 적용해야 한다.
- 넷째, 모델 검증 과정에서도 그 결과가 기독교 세계관에 위배되지 않는지 검토해야 한다[8].
- 다섯째, 모델을 적용할 때 기독교 공동체의 선을 이루는 데 사용해야 하며, 영성 영역에서 사용하는 것에는 주의해야 한다.