

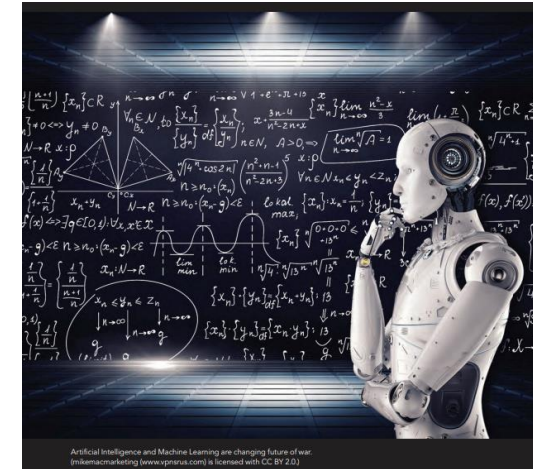
성격심리학에서 본 AI의 사회문제

2023. 08. 18

채 정 민

순서

- 서론
- 연구 과제
- 연구 방법
- 연구 결과
- 결론 및 논의



Artificial Intelligence and Machine Learning are changing future of war.
Infomarketing (www.yourum.com) is licensed with CC BY 2.0



서론

❗ "나는 경고했다"...'섬뜩' 명령 어기고 사람 공격하는 AI



1. 사회문제란,

- 개인문제가 사회문제로 확대된 경우: 예, 개인이 술을 많이 마셔서 문제가 됨
→ 이러한 개인들이 많아져서 사회문제가 됨
- 사회 자체의 문제인 경우: 예, 부동산 급등락 문제나 닷컴 버블 문제

두 가지 모두 고려해야 함.

2. AI의 사회문제란,



- Budic(2022)와 Sindermann 등(2002): AI 관련 윤리적 문제 제기
- AI가 개개인에게 심각한 피해를 주는 문제: 예, AI가 개인의 목숨을 해치는 문제
→ 이러한 행동이 여러 개인들에게 나타나는 문제
- AI가 사회시스템에 심각한 피해를 주는 문제:
예, AI가 금융시스템을 마비시키는 문제(Lee & Yoon, 2022; Kaloudi & Li, 2020),
AI가 가짜뉴스를 제작, 유포하여 국가선거 결과에 악영향을 미치는 문제(Sanders & Schneier, 2023)
- AI가 피해를 주려고 의도하지 않았지만 발생할 수 있는 문제: AI의 편견과 공정성 등

AI도 위 세 가지 모두 발생시킬 수 있음.

특히, 로봇과 결합된 AI는 심각성을 더 할 것으로 보임(전쟁(Gargon, 2022) 포함).

단순히 AI에게 거짓말을 하는 정도는 사회문제로 보지 않음.



연구과제

1. 연구과제

- AI는 성격을 가졌는가?
- AI가 발생시키는 사회문제를 성격심리학적 관점에서 다룰 수 있는가?
- AI의 성격을 어떻게 측정할 것인가?
- AI가 보일 사회문제의 양상은 무엇인가?
- AI의 사회문제를 어떻게 해결할 것인가?



연구 방법

1. 연구 방법

- AI에 대한 기존 문헌을 분석하는 방식
- 주요 AI들(챗GPT3 버전, Bard, Bing Chat, 하이퍼글로바 등)을 상대로 동일 질문 (한글과 영어)을 제시하여 그 반응을 획득하여 분석하는 방식
- 분석자료는 논문, 연구보고서, 도서 등



연구 결과

1. AI가 성격을 가졌다고 볼 수 있는가?

- 성격 정의의 핵심요인

- * 성격은 관찰할 수 있는 사람들의 행동을 바탕으로 판단됨
- * 성격은 인간적응의 측면을 반영 함: 삶은 적응의 과정이며 사람들은 자신이 처한 상황에서 생존하기 위해 자신의 성격을 발달시키고 형성함
- * 성격은 사람들이 보편적으로 공유하는 공통성을 내포함
- * 성격은 사람들을 구별할 수 있는 개인의 독특성 혹은 개인차를 반영함
- * 성격은 비교적 일관되고 안정적인 행동패턴과 관련됨: 행동이해와 예측가능
- * 성격은 개인 내부의 역동적이며 조직화된 특성을 반영함

- 성격의 정의: 환경에 적응하기 위해 개인이 지니게 된 비교적 독특하고 일관성 있는 심리 내적 행동패턴

1956 Dartmouth Conference:
The Founding Fathers of AI



John McCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge



Nathaniel Rochester



Trenchard More

출처 : <https://www.scienceabc.com>

- **AI**(Artificial Intelligence, 人工知能)

* ‘인공지능’이라는 용어가 처음 등장한 때는 **1955년 8월 31일**
미국 다트머스 대학교에서 마빈 민스키, 클로드 섀넌 등이
개최한 학회에서 **존 매카시**가 사용한 때임.

* 인간의 지능이 보이는 학습, 추리, 적응, 논증 따위의 기능을 갖춘 컴퓨터 시스템

* 전문가 시스템, 자연 언어의 이해, 음성 번역, 로봇 공학, 인공 시각, 문제 해결,
학습과 지식 획득, 인지 과학 따위에 응용됨

* Geoffrey Hinton (2023):

“인공지능은 이미 **IQ 80~90** 상당의 합리적인 추론을 하고 있다.”

AI는 인간의 지적능력을 닮은 기계임

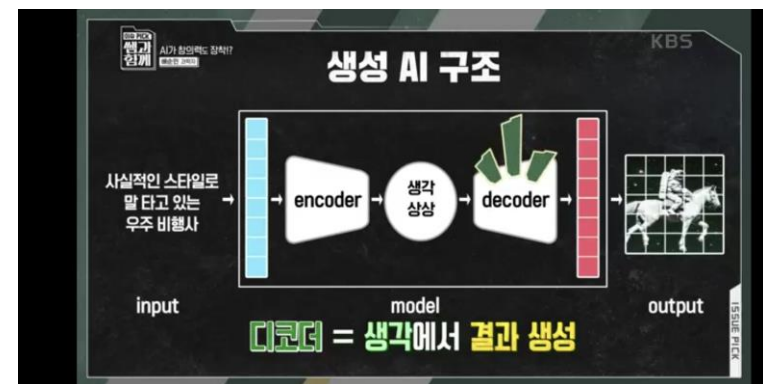
- AI의 구분과 성격

- * 딥러닝으로 사물인식에서 생성까지 가능해짐
- * 약인공지능 = 유용한 도구로서 설계된 인공지능(특정 분야에서만 활용가능한 인공지능)
- 강인공지능 = 인간을 완벽하게 모방한 인공지능(다양한 분야에서 활용가능한 인공지능)
- 초인공지능 = 강인공지능이면서 자아를 가진 인공지능

정서를 나눌 수 있는 AI는 2028년 정도면 가능

- * Endres(1995): 성격이론의 적용가능성 주장
- * Petta와 Trappl(2005): 인공행위자에 성격을 만들어주자는 주장
- * Janowski 등(2022): 적응적인 인공 성격 주장
- * Ritschel과 André(2017): 강화원리로 실시간 성격형성하자는 주장
- * Kostenko 등(2023): AI의 Legal personality 문제 다룸

현재까지는 AI가 약인공지능적 측면을 지니지만 머지 않은 시점에 강인공지능
그리고 **궁극적으로는 초인공지능이 되면**
AI가 성격을 가졌다고 완벽하게 말하기는 어렵지만 어느 정도는 가능함.



- AI 간의 경쟁

* AI 제작사 간의 경쟁

* AI 잡는 AI 개발 중



AI 간 경쟁이 치열해지고 여기에 구체성과 진실성 승부가 펼쳐지면 매우 큰 문제가 발생할 수 있음.

2. AI의 성격은 하나인가?

- AI의 다양성

- * 약인공지능으로 계속해서 존재하는 경우
- * 강인공지능이나 초인공지능으로 태어나거나 발전하는 경우
- * 같은 약인공지능 계열이나 강인공지능 계열이나 초인공지능 계열이라고 하더라도 제조사가 다른 경우(예, 유전자적 요소)
- * 같은 제조사라고 하더라도 각 개인 사용자가 길들이는 방식이 다른 경우(예, 상황적 요소)
즉, 초개인화된 AI, 맞춤형 AI인 경우(예, Kampman 등, 2019)와 다중인격 AI인 경우(예, Sonlu 등, 2021)

AI가 다양한 만큼 성격도 하나가 아니고 다수라고 보아야 타당함.

3. AI의 성격을 어떻게 잴 것인가?

- AI의 성격 측정 방안

- * Stachl 등(2020): 머신러닝 관련 성격측정 방안 제시

- * 인간 측정 방식과는 달리 AI만의 측정 방식을 개발하는 방향

- * 인간 측정 방식과 동일하게 개발하는 방향(예, NEO-P-I-R, TCI, MBTI, Eysenck 측정 방식 등)

약인공지능 형태의 AI에 대해서는 별도의 측정 방식을 적용할 수 있겠지만,
강인공지능 이상 수준인 경우에는 인간과의 인터페이스를 고려한 측정 방식이 적용되어야 할 것임.

4. AI가 보일 사회문제의 양상은?

- AI가 발생시킬 사회문제의 다양성

- * AI의 불완전성에 의한 반응의 경우
- * AI가 사람을 공격하는 경우
- * AI가 사회시스템을 공격하는 경우
- * AI끼리 싸우는 경우

AI가 보일 사회문제는 매우 다양할 것으로 예상됨.

5. AI의 사회문제 해결 방안은?

- AI 개발 관련

- * 반사회적 AI를 막기 위해 인위적인 공감능력 부여(Christov-Moore 등, 2023, 어려워 보임)
- * 일정 시점에 AI 개발 자체를 늦추거나 중단(불가능해 보임)
- * AI 개발 과정에 심리학자의 비협조나 규제(Hacker, Cordes, & Rochon, 2022, 어느 정도는 가능해 보임)
- * AI의 알고리즘에서 사회문제 발생가능성을 줄이기(Goldsmith & Burton, 2017, 쉽지 않아 보임)

- AI 사용 관련

- * AI 사용자가 사용 관련 사항에 숙달해지기(쉽지 않아 보임)
- * AI 사용하지 않기(쉽지 않아 보임)

AI의 사회문제 해결 방안은 쉽지 않아 보임.

결론 및 논의

- AI의 사회문제 발생가능성은 점차 커지므로 이에 대한 대응이 시급함
- AI의 성격에 대한 심리학적 연구도 시급함
- AI 개발자와 사용자에게 대한 심리학적 교육과 가이드라인 제시도 시급함

[주요 참고문헌]

- Budic, M. (2022, July). AI and Us: Ethical Concerns, Public Knowledge and Public Attitudes on Artificial Intelligence. In Proceedings of the 2022 AAI/ACM Conference on AI, Ethics, and Society (pp. 892-892).
- Christov-Moore, L., Reggente, N., Vaccaro, A., Schoeller, F., Pluimer, B., Douglas, P. K., ... & Kaplan, J. T. (2023). Preventing antisocial robots: A pathway to artificial empathy. *Science Robotics*, 8(80), eabq3658.
- Endres, L. S. (1995). Personality engineering: Applying human personality theory to the design of artificial personalities. In *Advances in Human Factors/Ergonomics* (Vol. 20, pp. 477-482). Elsevier.
- Kostenko, O. M., Bieliakov, K. I., Tykhomyrov, O. O., & Aristova, I. V. (2023). “Legal personality” of artificial intelligence: methodological problems of scientific reasoning by Ukrainian and EU experts. *AI & SOCIETY*, 1-11.
- Lee, J. M., & Yoon, S. (2022). Ready for Battle?: Legal Considerations for Upcoming AI Hacker and Vulnerability Issues. In *The International FLAIRS Conference Proceedings* (Vol. 35).
- Sindermann, C., Yang, H., Elhai, J. D., Yang, S., Quan, L., Li, M., & Montag, C. (2022). Acceptance and fear of Artificial Intelligence: associations with personality in a German and a Chinese sample. *Discover Psychology*, 2(1), 8.