인공지능의 위협과 도전

신상규 (이화여대)
이화인문과학원/포스트휴먼 융합인문학 협동과정

AI를 어떻게 이해할 것인가?

- AI, 아직 개 지능에도 못 미쳐…종말론 터무니없다
- SF 소설가 테드 창 "AI가 진짜 지 능이 있다고?…난 그렇게 생각 안 해"
- 테드 창 "AI, 단시간 많이 연습했을 뿐…응용통계에 가까워"(한겨레, 2024.6.13)
  - ✓"챗지피티(ChatGPT)에 '반갑습니다'라고 말하도록 만드는 건 쉽지만, 챗지피티는 여러분을 봐도 반가움을 느끼지 않습니다." .. 테드 창은 인공지능은 '의도'와 '지능'이 없다고 말했다. 강아지와 아기는 말할 수 없거나 서툴러도, 반가운 마음을 사람에게 전달할 수 있지만 챗지피티는 아무것도 느낄 수 없고 아무것도 원하지 않는다. 챗지피티는 감정을 전달하려는 의도가 없기에 언어를 사용한다고 볼 수 없다는 것이 테드 창의 견해다.
  - ✓테드 창은 인공지능이 '예술'에서도 인간을 대체할 수 없다고 주장했다. 테드 창이 바라보는 예술은 '선택의 연속'이다. 테드 창은 "소설을 쓸 때 인공지능에 단어 선택을 맡긴다면, 인공지능은 다른 작가들이 선택한 단어를 평균 내 산출하거나 특정 작가의 형식을 모방할 것"이라며 "특징이 없거나 파생적인 이야기가 될 수밖에 없다. 어떤 경우에도 흥미로운 예술 작품을 만들 수 없을 것"이라고 했다.

낯선지능의 위협

- **지적으로 동등한 존재에 대한 경험의 부재**
- 네안데르탈인: 40만~3만년 전. 마지막 시대에 크로마뇽인이라고 부른 현생 인류와 공존/ 현생 인류의 유전자 중 최소 1~2퍼센트는 네안데르탈인의 것.
- 네안데르탈인의 멸종은 인류가 3만년 전부터는 대등한 지능을 가진 상대를 경험하지 못했음을 의미.
- **지적 능력의 차원에서 인간은 지구상의 다른 존재와 구별되는 특별한 존재라는 생각**
- 인공지능의 등장으로 **그러한 자부심에 위협**

AI 효과

- 인공지능은 과연 “지능”인가?
- AI 효과
  - ✓많은 연구들이 구현 이전에는 인공지능이라 불렸지만, 구현 이후에는 인공지능이라 불리지 않는 현상 (원리를 알면 지능이 아니라고 여김)
  - ✓마빈 민스키(Marvin Minsky) 등의 연구자들의 주장: AI 효과 때문에 인공지능의 공헌이 낮게 평가되고 있다.
- Jack Copeland: **“인공지능이 생각할 수 있는가?”의 질문은 정답의 발견이 아니라 우리가 결정이나 선택으로 답해야 할 문제.**
  - ✓어떤 개념이 근본적으로 새로운 사례에 적용되는 경우에는, 관련 사실을 모두 고려한 다음에도 정답을 발견할 수 없다.

기계지능

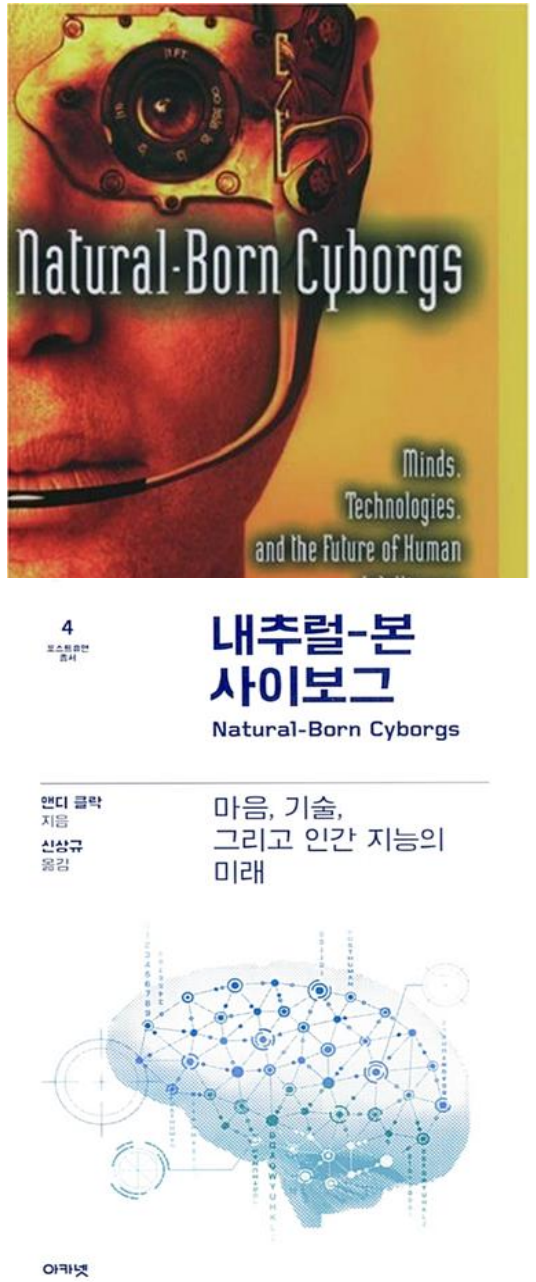
- **‘지능’의 특성을 우리 인간이 갖는 지능의 특성을 토대로 이해하려는 것은 모종의 인간중심적 편견**이 작용하는 것은 아닐까?
- machine intelligence:인공지능은 인간 지능의 특성을 닮아 있지만, 근본적으로는 인간과 매우 다른 낯선 형태의 지능. Jerry Kaplan - 비행기와 인공새(artificial aviation).
- 이미 많은 지능적인 과업에서 AI는 인간을 넘어섬.
  - ✓ 컴퓨터는 엄청난 속도, 정확성, 기억용량에 힘입어 체스나 텍스트 번역 등에서 이미 인간보다 더 효과적으로 작업(일)을 수행하고 있으며, 인공지능이 인간의 영역을 잠식함에 따라, 인간과 기계지능의 구분이 무의미해질 가능성.
  - ✓지금 이루어지는 인공지능 연구는 특정 기능을 수행하는 특수 지능. 일반 지능을 만드는 것보다 특수 지능을 만드는 것이 더 쉬우며 유용함. 인공지능은 매우 뛰어난 “자동기계”.

인간/기계의 경쟁 서사의 극복

- Mark Coeckelberg: 인간과 기계(기술)는 근본적으로 대립적이 아니다. **기술에 대항하여 인간을 지키려는 휴머니즘의 전투가 잘못된 방향을 향하고 있다**.
- 우리는 항상 기술을 사용해왔으며, 기술은 우리의 실존을 위협하는 외부적인 어떤 것이 아니라 우리 실존의 일부다.
- **AI는 인간의 이미지로 만들어질 필요가 없다.**
- AI가 인간을 모방하거나 재건해야 한다는 부담에서 벗어나서, 여러 다른 비인간 종류의 존재, 지능, 창조성 등을 탐색할 수 있다. 우리는 AI가 어떻게 세계에 대한 인간의 관계를 매개하는지 묻고, 올바른 방향의 매개를 적극적으로 형성하려고 노력해야 한다.
- 주어진 과제를 위해 AI와 경쟁하는 대신, 공통 목표를 설정하고 인간과 인공 행위자가 제공할 수 있는 최선의 것들을 동원하고 협력하여 그 목표에 도달하려 해야 한다.

사이보그로서의 인간

- Andy Clark: 기술은 단순히 도구가 아니라, 인간의 정신 및 신체의 확장.
- 유기적 생명과 기술의 병합은 철저하게 인간을 인간답게 만들어주는 인간의 본질적 특성
- 기술은 인간 정체성의 일부.
- 인간과 기술의 경계가 명확하지 않다. 우리는 부분적으로 우리가 사용하는 기술이다.
- 우리가 기술을 형성하고 기술은 우리를 형성.



인간의 행위주체성

- 인간의 행위 주체성은 기술의 매개와 얽힘을 본질적인 특징으로 한다.
- 기술(의 사용)은 우리의 삶의 목표나, 기획, 역량이나 잠재성을 확장(유도 affordance) 혹은 제약(constraint)하는 핵심적인 조건이다.
- human agency = hybrid agency (biological agency + technological agency)
- 컴퓨터나 인터넷 정보통신 기술의 발전은 이미 우리가 누구인지, 우리가 무엇을 할 수 있는지를 바꾸어 놓았으며, 우리의 의사결정, 선택, 행위를 조각한다.
  - ✓소셜미디어가 우리의 인식(추천알고리듬)과 행동(커뮤니케이션 방식)에 끼치는 영향, 글 쓰는 과정에서 컴퓨터 워드의 역할
  - ✓ChatGPT와 같은 인공지능의 등장도 비슷한 맥락에서 이해될 수 있다. 업무 수행, 예술 창작, 글쓰기 과정이나 작가로서의 우리 자신을 생각하고 경험하는 방식을 변화시킬 가능성

기술과 인간의 관계

- 인공지능의 기술의 설계와 활용은 결국 인간 행위주체성(행동양식) 및 사회구조, 삶의 형태의 설계이다.
- *Re-Engineering Humanity* by Brett Frischmann & Evan Selinger: 디지털 기술이 우리를 어떻게 re-engineer 하는가?
- 기술은 우리의 능력을 어떻게 확장하는가?
  - ✓하기 싫은 일로부터의 해방
  - ✓인간이 하는 일의 능률 향상
  - ✓전에는 할 수 없었던 새로운 일의 수행 가능
- Design by Ethics (AI 개발 단계에서 윤리/정치적 개입의 필요성.)

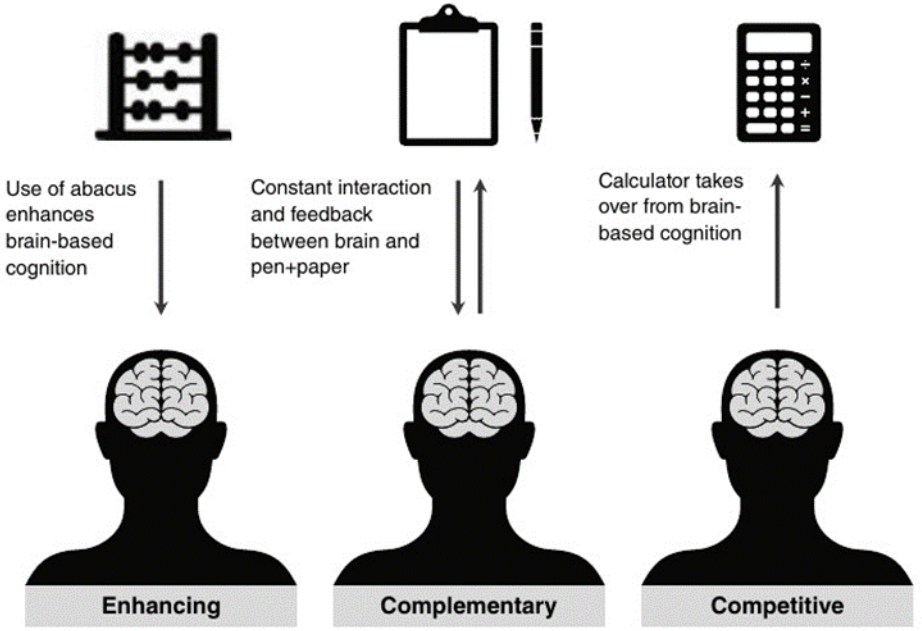
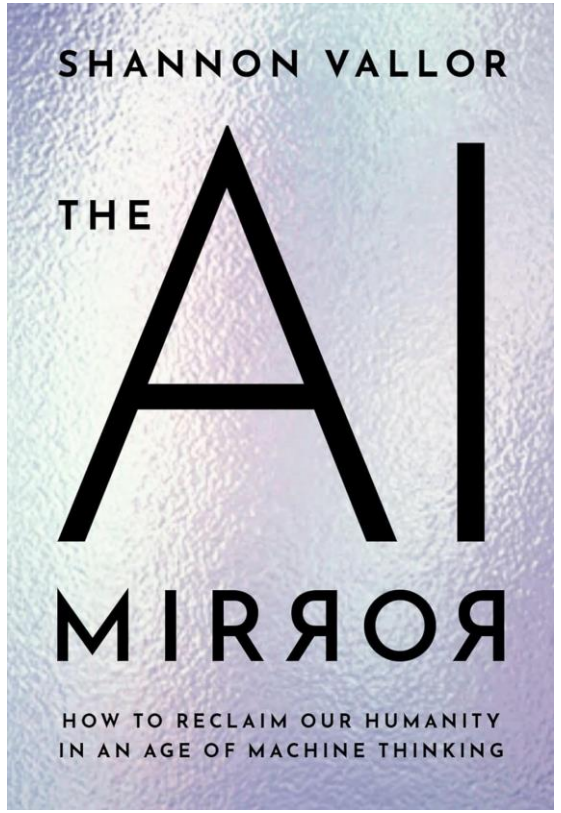


Figure 4.1 Three Kinds of Cognitive Artifact

기술과 인간의 불화

- 기술이 언제나 인간과 조화를 이루지는 않는다.
- Shannon Vallor: *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking (2024)*
  - ✓AI 기술은 새로운 미래를 열기는 커녕 과거를 재현.
  - ✓방대한 데이터를 바탕으로 강력하지만 결함이 있는 거울을 만들어낸 이 기술은 우리가 벗어나고자 노력하는 오류, 편견, 지혜의 실패를 그대로 반영.
  - ✓이 거울은 우리가 이미 가본 곳만 보여줄 뿐, 우리가 처음으로 함께 모험을 떠날 수 있는 곳은 보여주지 않는다.
- 불평등의 악화와 가난한 사람들에 대한 처벌
- 인종, 성별, 계급에 대한 고정관념과 편견의 고착화
- **AI 윤리는 이러한 불화를 피하고 인간과 AI의 행복한 공존을 위한 지침을 제공하려는 노력**



AI의 관계성/비가시성/물질성

- AI는 종종 비가시적. 그러나 AI는 오늘날 이미 편재해 있으며, 일상적인 도구와 복잡한 기술 시스템의 일부로 종종 보이지 않게 은폐되어 있다.
- AI는 이미 소셜 미디어 플랫폼, 검색 엔진 등 이미 일상의 일부가 된 여타 미디어와 기술을 지배하며, 더 넓은 사회적, 경제적, 과학적, 기술적 실천과 절차 속에 embedded 되어 있다.
- AI는 때로 특정한 물질적 인공물이나 인프라 구조와는 관계없는, 추상적이고 형식적인 본성을 갖는 것으로 보임. 그러나 모든 형식화, 추상화 및 부호 조작은 물질적인 도구와 물질적인 인프라 구조에 의존.
- 오늘날 AI는 네트워크와 전자 기기를 통한 대량의 데이터 생성에 의존. 이러한 네트워크와 장치들은 단순한 "가상(virtual)"이 아니며, 물질적으로 생성되고 유지되어야 함.
  - 숨겨진 인간 노동(아마존 메카니컬 터크), 희귀금속 채굴
  - GPT-3훈련: $1,287MWh 전력$소비, 총18만5000갤런(70만 리터)의 냉각수물 사용, 552톤의 이산화탄소 배출(미자동차 110대가 일년에 배출하는 탄소 배출량과 동일)

Automation and Utopia

- <생각을 기계가 하면, 인간은 무엇을 하나? - 일이 없는 미래 세계, 행복한 인간을 위한 철학>, 김동환 역, 뜻있는도서출판, 2023-07
- John Danaher, Sven Nyholm, “Automation, meaningful work, and the achievement gap”, *AI and Ethics* (2021) 1:227–237
- 기계학습과 로보틱스의 발전, 자동화에 따른 구조적 실업의 가능성
- 인간의 Mind Work를 필요로 하는 경제, 투자, 법률, 의료, 복지, 정책, 군사 등의 업무의 대체와 성취 공백

알고리듬 기반 사회의 도래

- 인공 지능의 주요 목표는 기계가 마음 업무를 대신 하도록 만드는 것. 장기적으로는 인간 자율성/행위주체성(agency)의 침식으로 연결될 가능성
- 일상적 생활환경은 컴퓨터 가상 환경의 소프트웨어 행위자나 사물인터넷의 다양한 지능적 장치들이 결합하여 이루어진 인공 행위자들의 네트워크에 의해 재조직화: 아마존 , 넷플릭스, 구글포토, Roon, 자율자동차
- **Algocracy 혹은 '알고리듬 보스' 사회**: 일상적 판단이나 의사결정의 방식을 포함하여 사회의 전체적인 구조나 행동 양식까지도 인공지능 알고리듬의 작동 방식에 맞추어 재편
- → 인류의 집단적 운명에 대한 인간 통제력에 대한 위협. 효율성의 가치가 압도하는 비인간적인 미래의 도래?

AI의 정치적 위협

- Coeckelbergh, <Why AI Undermines Democracy and What To Do About It>(2024).
- 정치의 부재와 민주주의의 침식, 전체주의의 위협
- **생각의 외주화와 사유 무능력:** 새로운 아이히만 들의 출현
- **공통 세계와 공통 감각의 부재**: 필터버블, 에코챔버. 너무 빠르게 일어나는 변화와 변화에 적응하는 개인별 차이와 세대간 갈등.
- 진실의 문제: 공론장의 붕괴와 post-truth, 가짜뉴스, 개소리(bullshit)의 만연, 분파화와 혐오의 확산
- 신뢰의 붕괴: 사적 이익이 공공선을 압도, 공공선을 대변하지 않는 정치인과 정부, 민주주의의 실패

권력(힘)의 문제

- 누구를 위한 기술인가? 기술의 사용과 발전은 언제나 사회적, 정치적, 문화적 맥락 안에서 발생. 기술은 언제나 정치적이다(Feenberg).
- AI 기술은 개인의 삶 뿐만 아니라 사회-경제 구조, 그리고 정치의 작동 방식도 변화시킨다.
- 누가 AI의 미래를 설계하고 지배하는가? 소수의 거대 기업에 의한 독점
  - ✓이러한 기술을 통해 누가 이익을 얻고, 누가 손해를 입는가?
- 우리가 어떤 가치에 입각하여 어떤 방식으로 AI 기술을 사회-문화-정치적 기술 환경에 embed 시킬 것인지를 고민해야 한다. 이에 대한 선택을 빅-테크 기업의 소수 권력자에게 양도하거나 위임해서는 안된다.
- 어떻게 민주주의의 증진을 위한 AI를 만들 것인가? 공동선의 증진, 소통과 신뢰의 회복, (더 많은) 공통 세계의 구축에 도움이 되는 AI는?

윤리적 원칙과 헌장?

- AI의 위험에 대한 지배적인 대응은 공공 및 민간 부문에서 발표된 수많은 AI 가이드라인과 윤리 강령
- Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. : 수많은 AI원칙들은 beneficence, non-maleficence, autonomy, justice, and explicability의 다섯 가지 핵심 원칙으로 압축
- Luke Munn, “The uselessness of AI ethics”
  - ✓무의미한 원칙: 구체적인 권장 사항을 거의 제공하지 않고 핵심 개념에 내재된 근본적인 규범적, 정치적 긴장을 해결하지 못하는 모호하고 고차원적인 원칙과 가치 진술. “인간”이라는 보편적인 개념과 “인류”를 위해 AI를 설계해야 한다는 공허한 진리를 보다 비판적으로 성찰할 필요.
  - ✓고립된 원칙: 소프트웨어 공학, 컴퓨터 과학 및 기타 AI 개발자 교육은 기술적 과제와 그 해결책에 집중되며, 윤리교육은 거의 부재. AI 개발은 윤리적 고려가 부재한 환경에서 이루어진다.
  - ✓이빨 빠진 원칙: AI 윤리 프레임워크는 이러한 가치와 원칙을 준수하도록 강제하는 메커니즘을 결여. 윤리 원칙에 초점을 맞추는 것은 논란이 되는 기술에 대한 법적 강제력을 피하려는 실리콘밸리의 노력과 전략적으로 일치, 윤리 세탁의 혐의

어떻게 할 것인가?

- AI윤리 원칙의 실패는 높은 수준의 이상과 현장의 기술 개발 사이의 괴리, 원칙과 실천 사이의 간극을 반영
  - ✓개발자에게 윤리 강령을 고려하도록 명시적으로 지시하는 것이 아무런 효과가 없다는 연구
  - ✓개발자의 태도를 구체적으로 조사한 연구에서도 AI 개발 과정에서 윤리적 고려가 완전히 무시된다는 결과
- AI 정의(justice)에 대한 보다 폭넓은 접근: 알고리즘의 도덕적 속성은 모델 자체에 내재되어 있는 것이 아니라 오히려 알고리즘이 배치된 사회 시스템의 산물이라는 사실에 주목. 사회정치적 역학과 억압 시스템의 고려. 다양한 이해 당사자들의 관심을 반영할 필요.
- 구체적인 문제에 초점을 맞추어 보다 좁게 생각할 필요. ' 윤리 ' 라는 모호한 개념을 측정 가능한 지표와 개별적인 목표로 세분화. 투명성의 확보와 감시. 그에 따른 구체적인 개선책, 책임 소재, 거버넌스의 수립과 입법화.