인공지능에 대한 사회적 통제 어떻게 할 것인가?

일시 2019년 9월 6일 (금) 오후 3시~5시
장소 국가인권위원회 10층 배움터
주최 사단법인 정보인권연구소

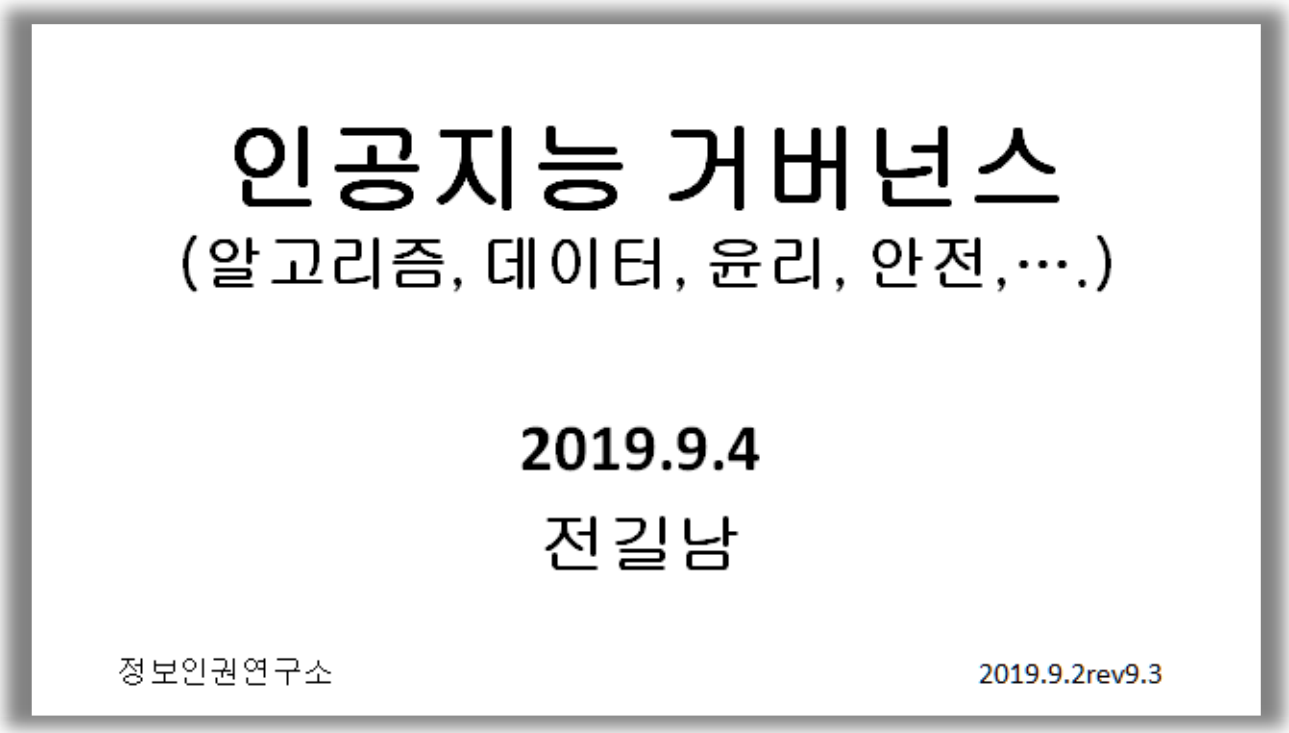
사단법인 정보인권연구소
Institute for Digital Rights

순 서

3:00 ~ 3:10	개회	
	사회	김기중	법무법인 동서양재 변호사
3:10 ~ 3:50	발제	인공지능 거버넌스 전길남	카이스트 명예교수
3:50 ~ 4:40	토론	전치형	카이스트 과학기술정책대학원 교수
		이항우	충북대학교 사회학과 교수
		조지훈	민주사회를위한변호사모임 디지털정보위원회 위원장
		강혜란	한국여성민우회 공동대표
		김원규	국가인권위원회 인권정책과장
4:40 ~ 5:00	참석자 전체 토론	

인공지능 거버넌스

전길남 | 카이스트 명예교수



목차

2

1. 인터넷 거버넌스와 디지털 거버넌스

인터넷 거버넌스 → 디지털 거버넌스

사물인터넷 거버넌스
데이터 거버넌스
사이버보안 거버넌스
인공지능 거버넌스

3

2. 인터넷 거버넌스를 인공지능 거버넌스에 적용가능할까?

1. **그렇다**
   - 공개 문서를 통한 공개 절차
   - 멀티스테이크홀더
2. **그렇지 않다**
   - 오픈 데이터
   - 표준
   - 정부의 역할
3. **기타**
   - 소스코드

4

3. 인공지능 거버넌스 영역

사회 영역:

(1) AI Principles (원칙)
(2) AI Policy (정책)
(3) Social and Economic Impact (사회경제적 영향)
(4) Institution (AI 기관)
(5) Ethics (윤리)
(6) Accountability (and Explainability) (책무성)

기술 영역:

(11) Algorithm (알고리즘)
(12) Security (보안)
(13) Safety (안전)
(14) Data (데이터)

장기적 영역:

(A) Artifical General Intelligence (인공일반지능)
(B) Existential Risk (실존적 위험)

5

(1) Artificial Intelligence Principles (인공지능 원칙)

아래와 같이 좋은 인공지능 원칙이 많이 있음.
세계적으로 수용되는 인공지능 원칙 수립을 위해 무엇을 할 것인가?

인공지능 원칙 목록:

1. 아시모프의 로봇 3원칙, 1940.
2. 인공지능 파트너쉽(Partnership on AI), 원칙(Tenets), 2016.
3. 아실로마 인공지능 원칙, 2017.
4. 책임성있는 인공지능에 대한 몬트리올 선언(Montreal Declaration on Responsible AI), 2018.
5. 유네스코, 인공지능을 위한 원칙, 2019.3.
6. 유럽연합 집행위원회(EC), 윤리적 원칙
7. IEEE, 인공지능 및 자동시스템의 윤리적 고려에 대한 세계 계획, P7000.
8. 경제협력개발기구(OECD), 2019.

비고 : UN은 특정재래식무기금지협약(CCW)를 통해 치명적인 자동화 무기를 금지하는 작업을 하고 있음. 6

(2) Artificial Intelligence Policy(인공지능 정책)

Future of Life Institute (FLI)의 2018년 세계 정책 보고서는 다음과 같은 내용을 담고 있음;

- 지구적 인공지능 정책
- 국가적, 국제적 인공지능 전략
- 인공지능 정책의 과제와 권고
- 인공지능 정책 자원

이 보고서는 5개 아시아 국가의 정책 보고서를 포함하고 있음.

Tim Dutton의 중간 보고서, 2018 국가 인공지능 정책 개요는 약 10개 아시아 국가를 포함한 전 세계 20개 국가를 다루고 있음. .

7

(5) Ethics (윤리)

Center for Future of Intelligence (CFI), Cambridge는 다음에 기반하여 인공지능 윤리에 대한 보고서 발간;

- 아실로마 인공지능 원칙 (2017)
- 일고리즘 투명성과 책무성에 대한 ACM 성명 (2017)
- 인공지능에 대한 일본 협회, 윤리적 가이드라인 (2017)
- 책임성있는 인공지능 원칙에 대한 몬트리올 선언 (2017-2018)
- 인공지능 및 자동시스템의 윤리적 고려에 대한 IEEE P7000 위원회 (2017)
- 인공지능 파트너쉽, 원칙 (2018)
- 영국 하원의 부문간 인공지능 규칙 (2018)
- EC 고위급 전문가 그룹 : 신뢰할만한 인공지능을 위한 가이드라인 초안 (2018)
- 구글의 인공지능 윤리 원칙 (2018)

8

(6) Accountability (책무성) and Explainability (설명가능성)

Future of Life Institute (FLI), Oxford University:세계 인공지능 정책은 책무성, 투명성, 설명가능성을 다음과 같이 설명하고 있음;

"인공지능 시스템 및 설계자가 책무성을 갖도록 하기 위해서 몇 가지 과제가 있다. 투명성 및 설명가능성의 부재(lack of transparency and explainability)는, 특히 기계 학습과 결합하여, 어떤 알고리즘이 왜 특정한 결정을 내렸는지 알기 어렵거나 불가능하게 한다. 또한 누가 핵심 알고리즘에 접근가능한지, 그것을 어떻게 이해할 수 있는지에 대한 문제가 있는데, 독점적 알고리즘의 사용으로 이러한 문제가 악화된다. 의사결정이 인공지능 시스템에 양도됨에 따라, 바람직하지 않은 결과에 대해 누가 책임을 질 것인지에 대한 명확한 가이드라인이 존재하지 않는다.

FAT* (공정성, 책무성, 투명성) 컨퍼런스, 2018~

DARPA는 현재 Explainable AI Project를 진행하고 있다.

서울대 연례 컨퍼런스, 인공지능: 거버넌스와 책무성, 2017~.

9

(11) Algorithm (알고리즘)

인공지능 알고리즘은 인공지능 시스템을 편향되게 할 수 있으며, 인간 사회에의 보급을 저해할 수 있다.

인공지능 알고리즘은 가능한 투명할 필요가 있고, 설명가능해야 한다.

비고: 사용되는 데이터는 사전 검토될 필요가 있다.

10

(13) Safety (안전)

안전에 대한 아실로마 인공지능 원칙은 "인공지능 시스템은 전 운영 주기를 통해 안전하고 보안성이 있어야 하며, 가능하다면 검증가능해야 한다."고 말한다.

삶의미래연구소(FLI)는 2015년 이후 엘론 머스크의 후원으로 30개 이상의 기관을 위한 인공지능 안전 프로그램을 수행했는데 2018년에 2단계로 인공 일반지능에 방점을 두고 있다. FLI는 실존적 위험 웹페이지에서 인공지능 안전을 다루고 있다.

빅토리아 크라코프나(Victoria Krakovna)는 인공지능 안전 자원을 포함한 인공지능 안전에 대한 논문을 출판했다.

스튜어트 러셀(Stuart Russell)은 2018년에 인공지능 안전 : 인공 일반지능을 위한 안전과 통제 강의를 제안한다.

인공지능 거버넌스 : 인공지능 거버넌스 센터의 알란 다포우의 연구 의제는 인공지능 안전에 대한 개요를 제공한다.

11

(14) Data (데이터)

데이터는 인공지능 개발에 매우 중요하며, 인공지능은 더 많은 데이터 생성에 기여한다. 다음과 같은 요소들이 다양한 이슈를 제기한다. ;

- 데이터 거버넌스

  데이터 거버넌스를 어떻게 할 것인가; 공공 데이터 및 기업에 의해 수집 되는 데이터?

- 프라이버시

  인공지능 시스템에 의해 사용되는 데이터로부터 어떻게 프라이버시를 보장할 것인가?

EU 개인정보보호규정(GDPR)은 인공지능과 데이터 이슈의 좋은 시작점일 수 있다.

12

4. 기타 이슈들

1. 비유 : 인공지능 기술과 핵 기술
2. 개발도상국 대 선진국 - “승자독식”
3. 표준
4. 인공지능 거버넌스 영역의 구조
5. 인공지능 리터러시와 역량 개발
6. 많은 주제들이 연구 중.
7. 일반적 알고리즘 대 인공지능 알고리즘?
8. 편향 : 알고리즘 및 데이터?
9. 스테이크홀더; 정부, 시민단체, 기업, 애커데믹
10. Europe과 협력(?)

13

참고자료

(1) 논문
(2) 단행본
(3) 컨퍼런스
(4) 강좌
(5) 단체
(6) 발표 자료
(7) 보고서
(8) 비디오

14

참고자료 – 논문

- BAAI, Beijing AI Principles, 2019.5.28
- Kilnam Chon, Digital Governance, APSIG.asia, 2018.
- Kilnam Chon, AI Governance, 2019.
- Alan Dafoe, AI Governance: Research Agenda, Center for Study on AI Governance, Oxford, 2018.
- Tim Dutton, An overview of national AI strategies, Medium, 2018.6.29.
- Edward Felten, AI and Explainability, 2018.
- L. Floridi, et al., "AI4People: An ethical framework for a good AI society," Mines and Machines, 2018.
- Google, Perspective on issues in AI governance, 2019.1.
- Michael Jordan, AI: revolution has not happened yet, Medium, 2018.4.

15

참고자료 – 논문 (계속)

- Kai-Fu Lee, AI Superpowers, 2018.
- FLI, Asilomar AI Principles, 2017.
- McKenzie, Promise and Challenge of the Age of AI, 2018.
- PwC, Global Artificial Intelligence Study: Sizing the Prize, 2017.
- Steven Strogatz, One giant step for a chess-playing machine, NYT, 2018.12.
- Max Tegmark, How Far Will AI Go?, 2018.
- Wired, How to Teach AI Some Common Sense, 2018.11.
- Wired, How tech companies are shaping the rules governing AI, 2019.5.16.
- Jess Whittlestone, et al., The role and limits of principles in AI Ethics, AES-19, 2019.

16

참고자료 – 단행본

- Nick Bostrom, Superintelligence, 2014.
- Keith Frankish and William Ramsey, Cambridge Handbook of AI, 2014.
- Hannah Fry, Hello world, how algorithm decides future, 2017.
- Yuval Noah Harari, 21 Lessons for the 21st Century, 2018.
- Stephen Hawking, Brief answers to big questions, 2018. (Chapter 9: AI)
- Kai-Fu Lee, AI Superpowers, 2018.
- Stuart Russell and Peter Norvig, AI: A modern approach.
- Max Tegmark, Life 3.0, 2017.

17

참고자료 – 컨퍼런스

- ACM/AAAI, AI, Ethics and Society, 2018, 2019. (AAAI Workshop in 2016, 2017)
- ACM, FAT* (Fairness, Accountability, Transparency) Confeence, 2-18~.
- APSIG, AI Governance Workshop, Annual APSIG Meeting, 2018, 2019.
- FLI, Beneficial AI Conference, 2015, 2017, 2019.
- UAE, Global Government Summit, Global Governance of AI, Dubai, 2019.
- IGF, Annual Conferences, 2017, 2018.
- ITU, AI for Good, 2018, 2019.
- KIAS, AI, Ethics and Governance, 2018.

18

참고자료 – 컨퍼런스 (계속)

- OECD, AI: Intelligent Machine and Smart Policy, 2017.
- Seoul National University, AI: Governance and Accountability, 2017~
- Tokyo University, AI and Society Symposium, 2017.
- UAE, Global Governance of AI, Global Government Summit, 2019.2.
- UAE, AI Everything, Dubai, 2019.4.
- UN, GGE, Autonomous Weapon, 2017~
- UNESCO, Principles of AI, Global Conference, 2019.3.

19

참고자료 – 강좌

- APSIG, AI Governance by Danit Gal.
- APSIG, Digital Governance by Kilnam Chon, 2018.
- Colombia University, AI, edX.
- Coursera, AI for Everybody, 2019.
- Microsoft, Introduction to AI, edX.
- Microsoft, Ethics and Law, edX.
- MIT, Artificial General Intelligence, 6.S099 (by Lex Fridman), 2018.
- MIT, Minds and Machines, 24.09 (by Alex Byrne), 2017.
- UC Berkeley, Introduction to AI, CS188.
- University of Helsinki, Elements of AI

20

참고자료 – 단체

- AI Finland
- AI for Good
- AI Now Institute, New York University, USA
- AI Policy Initiative, Seoul National University, South Korea
- AI Transparency
- Berkeley Existential Risk Initiative, Berkeley, USA
- Center for Governance of AI, Future of Humanity Institute, Oxford, UK
- Center for Human-Compatible Artificial Intelligence, Berkeley, USA
- Center for Study on Existential Risks, Cambridge, UK
- DeepMind Ethics & Society, UK
- Ethics and Governance of AI Foundation, USA
- Ethics and Governance of AI Initiative, Berkman Klein Center & Media Lab, USA
- European AI Alliance, EU
- Future of Life Institute, USA
- Future Society, AI Initiative, Harvard, USA
- Human Centered AI, Stanford, USA
- IEEE, Global Initiative on Ethical Consideration on AI and AS. P7000
- ISO/IEC JTC1 SC42, Standardization on AI
- Japan Society on Artificial Intelligence, JSAI Ethics Committee
- Leverhulme Center for Future of Intelligence (CFI), Cambridge, UK

21

참고자료 – 발표 자료

- Kilnam Chon, AI: Past and Present, 2018.
- Kilnam Chon, AI Governance, 2019.
- Allan Dafoe, AI, Strategy, Policy and Governacne, Beneficial AI, 2019.
- Arisa Ema, AI Ethics and Policy, Taming AI, Seoul, 2018.
- Danit Gal, AI Governance, APSIG, 2017
- Woodraw Herzog, AI: Accountability, 2018.

22

참고자료 – 보고서

- AI Index Report. [annual: 2018,…]
- Now Institute, AI Now 2017 Report, 2017.
- EC, Communication from the Commission: AI for Europe, 2018.4.25.
- EU, Digital Europe: 2021-2027, 2018.
- EU, Ethics guidelines for trustworthy AI, 2019.4.
- EU, AI Definition, 2019.4. [through Futurism/European AI Alliance]
- EU, Requirements for trustworthy AI, 2019.4.
- Future of Life Institute (FLI), Global AI Policy, 2018.
- Chinese Government, A new generation AI development plan (AIDP), 2017.7.20
- China Standard Administration, White Paper on AI in China, 2018. [English]
- Allan Dafoe, AI Governance: Research Agenda, Center for Governance of AI, FHI, Oxford, 2018.
- Google, Perspective on issues on AI governance, 2019.1.
- Japanese Government (Cabinet Office), AI and Human Society, 2017.
- Japan Society on AI, Ethical Guideline, 2017.
- Dongwoo Kim, AI in East Asia, AP Institute of Canada, 2019.
- McKinsey, AI problems and promises, 2018.10. (and more in 2018)

23

참고자료 – 보고서 (계속)

- Center for Governance of AI, Oxford, AI: American attitudes and trends. 2019.1.
- OECD, Recommendation of the Council of AI (Principles), 2019.5.
- PwC's Global Artificial Intelligence Study: Sizing the prize, 2017.
- Roadmap for US Robotics, 2016 Edition: From Internet to Robotics.
- Stanford AI 100 Report, 2015
- Tsinghua University, China AI Development Report, 2018.
- UN, GGE Report on Lethal Autonomous Weapon, 2017.
- UNGP & IAPP, Building ethics into privacy frameworks for big data and AI, 2018.10.
- Web Foundation, Future of technology - AI. [White paper on AI], 2017.
- World Government Summit, Summary Report of Global Governance of AI, 2019.2.

24

참고자료 – 비디오

- Nick Bostrom, Superintelligence, 2014.
- APSIG, Digital Governance by Kilnam Chon, 2018. [also AI: Past and Present, 2019]
- Yuval Harari, 21 lessons for the 21st century, 2018.
- Kai-Fu Lee, AI Superpowers, 2018.
- FLI, Beneficial AI with Asilomar AI Principles, 2017.
- FLI, Applying AI Safety & Ethics Today (with Lorens and Rossi), 2019.
- NHK Special, Money World, #2 Work Will Be Gone!, 2018.10.7.
- MIT, Artificial General Intelligence, 6.S099 (by Lex Fridman), 2018.
- Stuart Russell, Long-term future of AI, MIT AI, 2018.
- Amnon Shashua, Success and challenges in modern AI, CBMM, MIT, 2019.
- Max Tegmark, Life 3.0, 2017.

25

부록 A: 용어 (from Life 3.0 by Tegmark)

- 협의의 지능(Narrow Intelligence): 좁은 범위의 목표를 수행하는 능력; 체스 게임, 운전
- 일반 지능(General Intelligence): 사실상 어떤 목표든 수행하는 능력
- 보편적 지능(Universal Intelligence): 일반 지능을 획득하는 능력
- 인공 일반 지능(General AI, AGI): 최소한 인간과 동등하게 어떠한 인지 작업도 수행할 수 있는 능력
- 인간 수준 인공지능(Human-level AI): 인공 일반 지능(AGI)
- 초 지능(Super Intelligence): 인간 수준을 많이 넘어선 일반 지능
- 의식(Consciousness): 주관적인 경험
- 윤리(Ethics): 어떻게 행동해야 하는지 규율하는 원칙
- 특질(Qualia): 주관적 경험의 개별 사례
- 지능 폭발(Intelligence explosion): 초 지능으로 향하는 반복되는 자가-개선
- 특이점(Singularity): 지능 폭발

26

부록 B: 협의의 인공지능 및 일반 인공지능, 그리고 약 인공지능 및 강 인공지능
(헬싱키, 인공지능의 요소)

협의의 인공지능은 하나의 작업을 다루는 인공지능을 말한다. 일반 인공지능, 혹은 인공 일반 지능(AGI)은 어떠한 지적 작업도 다룰 수 있는 기계를 말한다. 오늘날 우리가 사용하는 모든 인공지능 방법은 협의의 인공지능이며, 일반 인공지능은 공상과학의 영역이다. 사실, 인공 일반지능의 이상은 인공지능 연구자에 의해 거의 폐기되었는데, 모든 노력에도 불구하고 50년 이상 기술적 진전이 없었기 때문이다. 반면, 협의의 인공지능은 급속히 발전하였다.

관련 이분법은 "강" 및 "약" 인공지능이다. 이는, Searle이 강조한 바와 같이, 지능적인 것과 지능적으로 행동하는 것 사이의 위의 철학적 구분으로 압축된다. 강 인공지능은 진정으로 지능적이고 자기 인식이 있는 "마음"에 이른다. 약 인공지능은 우리가 실제로 하고 있는 것, 즉 "단지" 컴퓨터이지만 지능적인 행위를 보여주는 시스템이다.

27

부록 C: 알란 다포우와 제시카 쿠신과 함께하는 거버넌스, 국가 정책, 공공의 신뢰, Future of Life Institute (FLI), 2018.8.30.

"전문가들은 인공지능이 농업혁명과 산업혁명을 무색하게 하는, 역사적으로 가장 변혁적인 혁명이 될 것이라고 예측한다. 그리고 이 기술은 평균적인 관료체제가 따라잡을 수 없을 만큼 빠르게 발전하고 있다. 어떻게 지역, 국가, 국제 정부들이 이와 같은 급격한 변화를 준비하고, 인공지능 연구와 사용이 보다 유용한 방향이 되도록 이끌 수 있도록 할 것인가?"

28

부록 D: Guideline for Trustworthy AI [EU]

1. 인간의 협력과 감독
2. 기술적 강건성과 안전
3. 프라이버시 및 데이터 거버넌스
4. 투명성
5. 다양성, 비차별, 공정성
6. 사회적, 환경적 복지
7. 책무성

29

부록 E: 인공지능 원칙에 대한 OECD 위원회 권고

1: 신뢰할 수 있는 인공지능을 위한 책임있는 관리 원칙

포용적 성장, 지속가능한 개발, 복지

인간 중심 가치와 공정성

투명성과 설명가능성

강건성, 보안 및 안전

책무성

2: 신뢰할 수 있는 인공지능을 위한 국가 정책 및 국제 협력

인공지능 연구 개발에 대한 투자

인공지능을 위한 디지털 생태계 활성화

인공지능을 위한 정책 환경의 형성

인간 역량의 구축과 노동 시장 변화의 준비

신뢰할 수 있는 인공지능을 위한 국제 협력

30

인공지능의 거버넌스란 무엇인가?

- **기술적 정의**: 어떠한 결정이 내려지고 이행되는 절차. 이는 규범, 정책, 제도 및 법을 포함한다.
- **규범적 정의**: 그러한 모범 절차의 집합. 좋은 거버넌스는 통상 효과적이고 적법하며, 포용적이고, 적응적인 것을 의미한다.

[알란 다포우, 인공지능 전략, 정책 및 거버넌스, 2019년 유용한 인공지능]

31

3. 인공지능 역사

1940년대 신경망 모델 (McCulloch-Pitts), 사이버네틱스(Wiener)
1950년대 튜링 테스트, 다트마우스 인공지능 워크샵
1960년대 초기 인공지능 호황기
1970년대 인공지능 겨울
1980년대 두번째 인공지능 호황기; 신경망, 전문가 시스템
80년대 후반 - 90년대 초반 두번째 인공지능 겨울
2000년대 심층신경망 / 딥러닝
2010년대 세번째 인공지능 호황기(데이터, 클라우드 컴퓨팅, 알고리즘)

32

4. 협의의 인공지능과 일반 인공지능

33

4.1 협의의 인공지능

현재 개발 및 확산 노력
주된 기법으로서의 딥러닝
주된 응용영역으로서의 패턴 인식
　　게임(체스, Go,….)
　　이미지 인식, 음성 인식, 자연어 처리,…

34

4.2 일반 인공지능 혹은 인공 일반지능(AGI)

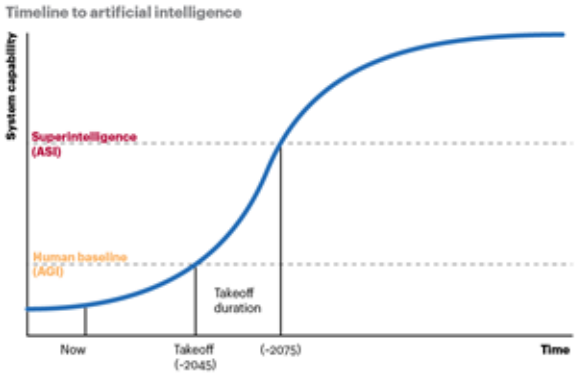
인간 수준의 지능과 초지능
실현가능하더라도 많은 세월이 걸림
뇌 과학과 중첩

닉 보스트롬(Nick Bostrom)은 초지능에 대한 책을 썼으며, 일반 인공지능과 초지능에 대한 다이어그램을 제안했다.
스튜어트 러셀(Stuart Russell)과 막스 테그마크(Max Tegmark)는 일반 인공지능의 실현을 적시에 준비할 것을 제안했다.

35

4.2 AGI (continued): 초지능 – 보스트롬의 다이어그램

Figure 19
Developing superintelligent AI may be possible in this century



Note: AI is artificial intelligence, ASI is artificial superintelligence, and AGI is artificial general intelligence.
Sources: WaitButWhy.com, Nick Bostrom, Superintelligence: Paths, Dangers, Strategies; A.T. Kearney analysis

36

(3) 사회적, 경제적 영향

다가올 시대에 인공지능의 도입으로 인간 사회와 세계 경제에 광범한 영향을 줄 것이다.

사회에 대한 영향은 인간 사회의 거의 모든 부문을 포괄하며, 우리는 이를 고려할 필요가 있다.

경제적 영향은 2030년대까지 약 15-20% 증가로 보이며, 선진국과 주요 인공지능 기업이 대부분의 이익을 취하는 "승자독식"이 될 것이다. 경제적 영향을 적극적으로 조정하여 그 분배를 고려할 필요가 있다.

37

(4) 제도(기관)

표준화를 넘어선 세계적인 기관이 필요한가?

인공지능의 가능한 사례로 인공지능 안전을 포함하는가?

다른 분야의 사례로 핵 기술과 관련한 IAEA와 지구 온난화와 관련한 IPCC 포함.

국가적 사례로는 현재 일부 나라에서 관련 기관을 두고 있으며, 조만간 더 많은 나라가 그렇게 할 것이다.

38

(12) 보안

인공지능 시스템의 보안은 매우 중요하다. 그리고 인공지능 기술은 점차 보안을 증진하기 위해 사용되고 있다. 인공지능을 통한 사이버보안은 매우 중요하지만, 많은 어려운 이슈들을 제기한다. 사이버보안은 인공지능을 적절하게 사용함으로써 증진될 수 있으며, 사이버보안과 관련된 기존 문제들을 해결할 수 있다. 반면, 그러한 혹은 유사한 인공지능이 남용될 수도 있다.

인공지능 거버넌스 연구 센터는 Zweltsloot의 인공지능과 국제 안보를 포함한, 인공지능과 사이버보안에 대한 논문을 발표했다.

UC Berkely 대학의 장기적 사이버보안 센터(Cussins)는 인공지능과 사이버보안에 대해 연구하고 있다.

2019년 미국의 국가안보인가법은 인공지능에 대한 공식적인 새 국가보안위원회를 설립했다.

39

(A) 일반 인공지능

스튜어트 러셀과 막스 테그마크와 같이 많은 사람들이 인공 일반지능의 적절한 통제에 대해 많은 이슈를 제기한 바와 같이 일반 인공지능, 혹은 인공 일반지능(AGI)은 특히 주의할 필요가 있다.

Future of Life Institute(FLI)의 국제인공지능정책에 따르면, 주요 이슈는 "인공 일반지능이 인간 수준 지능을 넘어 초지능에 도달했을 때 무엇을 할 것인가?" "인공 일반지능은 협의의 인공지능의 모든 과제와 맞닥뜨릴 뿐만 아니라, '억제(containment)'와 고유한 위험을 추가적으로 제기할 것이다." 많은 사람들이 인공 일반지능이 가능하고 이번 세기에 현실화될 것이라고 생각한다.

2017년 캇자 그레이스(Katja Grace)는 "언제 인공지능이 인간의 능력을 넘어설까?"를 출판했다.

스튜어트 러셀은 20세기 중반의 핵기술과 비교하면서, 우리가 인공 일반지능의 현실화에 대한 합의가 없지만, 이제 준비해야 한다고 권고했다.

40

(B) 실존적 위험

Future of Life Institute, Oxford 는 실존적 위험 웹페이지에서 인공지능의 이익과 위험을 다루면서, 인공지능 정책 과제 및 권고의 인공지능 안전 부분에서 다음과 같이 말한다. "위험: 인공지능 시스템이 야기하는 위험, 특히 재난적 혹은 실존적 위험은 예상되는 영향에 상응한 계획 및 감축 노력이 수행되어야 한다."

Center for Study of Existential Risk, Cambridge는 무엇보다 21세기의 로봇과 인공지능의 실존적 위험에 대해 경고했다.

"Berkeley Existential Risk Initiative (BERI)의 목적은 인간 문명의 장기적인 생존 및 번성 전망을 개선하는 것이다. 프로젝트가 실존적 위험을 감소시키는데 중요하다고 생각하기 때문에, 현재, 주요 전략은, 기금 지원자 및 협력자로서, 윤리적, 법적 책임성을 담당하는 것이다. "

41

인공지능이 인권에 미칠 영향에 대한 UN 특별보고관의 분석

UN Special Rapporteur analyses AI's impact on human rights

보고서 소개문[1)]

2018년 10월, 의사 표현의 자유 증진 및 보호를 위한 UN 특별보고관 데이비드 케이(David Kaye)는 인공지능(AI, Artificial Intelligence) 기술이 인권에 미치는 영향에 대한 자신의 보고서를 발행했다. 이 보고서는 2018년 8월 29일 UN 총회에 제출되었으나 최근 들어서야 출판되었다. 해당 문서는 특히 의사 표현의 자유, 프라이버시와 차별 금지 문제에 집중한다. 데이비드 케이 특별보고관은 보고서에서 우선 그가 인공지능에 대해 이해한 내용과 인공지능 사용이 현재의 디지털 환경에 수반하는 바에 대해 정리하며, (인공지능과 관련된) 몇가지 미신에 대해 반박한다. 그런 다음 그는 관련 기술의 발전이 인권에 미칠 잠정적 영향에 대해 개관한 후, 새로운 기술에 대한 인권 기반 접근 프레임워크를 설정한다.

1. 인공지능은 중립적인 기술이 아니다

데이비드 케이는 인공지능을 "데이터를 결론, 정보, 또는 출력으로 변환하는 명령을 수행하는 … 컴퓨터 코드"를 통해 "사람이 수행하였을 특정 작업을 컴퓨터가 보완하거나 대체할 수 있도록 하는 기술 및 과정의 집합"이라고 정의한다. 그는 알고리즘이 제대로 작동하려면 사람이 시스템을 설계하고, (시스템의) 목적을 정의하고 데이터셋을 조직할 필요가 있으므로 인공지능이 여전히 사람의 개입에 의지하고 있다고 주장

1) 유럽 정보인권단체 EDRi (2018. 11. 7.),
https://edri.org/un-special-rapporteur-report-artificial-intelligence-impact-human-rights/

한다. 보고서는 인공지능이 중립적인 기술이 아니라고 지적하는데, 그 결과물을 사용하는 것은 여전히 인간의 손에 달려 있기 때문이다.

현재의 인공지능 시스템 양식은 완벽성과 거리가 멀다. 인공지능은 여전히 사람의 정밀 조사를 요하고, 이따금씩은 교정해줄 필요도 있다. 보고서는 인공지능 시스템의 적응성 뿐 아니라, 자동화된 특성, 데이터 분석의 품질이 편향의 원천이라고 간주한다. 자동화된 결정은 기준들 간에 반드시 균형을 맞추지는 않고 특정 기준에 배타적으로 의존하기 때문에 차별적인 효과를 낳을 수 있다. 이는 결과물의 정밀함과 투명성을 약화시킬 수 있다. 또한 인공지능 시스템은 출처와 정확도가 의심스러운 대량의 데이터에 의존한다. 게다가 인공지능은 상관관계를 식별할 수 있는데 이것이 인과관계로 오해될 수 있다. 데이비드 케이는 사람의 감독이 없을 때 적응성의 주요 문제로서, 투명성과 책무성의 보장을 어렵게 하는 것을 들었다.

2. 현재 인공지능의 사용은 인권을 침해한다

데이비드 케이는 인권에 중대한 위협을 가하는 인공지능 기술의 주요 적용예를 세 가지 설명하였다.

첫 번째 문제는 의사 표현의 자유에 인공지능이 미치는 효과이다. 한편에서는 “인공지능이 이용자에게 불투명한 방식으로 정보 세계를 형성할 것”이며 이용자가 무엇을 보고 소비할지 결정하는 자신의 역할을 숨길 것이라고 한다. 다른 한 편에서는 개인화된 정보의 게시가 편향성을 강화하는 것으로 보이며 “이용자의 온라인 참여를 지속시키기 위해 선동적인 콘텐츠나 허위정보를 홍보하거나 추천하는 것을 장려하고 있다”고 한다. 이런 관행은 사실적이고 다양한 정보에 기반하여 개인의 의견을 형성하고 발전시킬 수 있는 개인의 자기결정권과 자율성에 영향을 미치고, 결국 의사 표현의 자유를 위협한다.

둘째, 프라이버시권과의 관계에서도, 특히 광고 목적의 인공지능의 지원을 받은 마이크로 타게팅에 대해 유사한 우려가 제기될 수 있다. 데이비드 케이가 언급하듯이, 이용자 프로파일링과 타겟팅은 개인정보의 대량 수집을 조장하고 “사람들이 제공하거나 확인하지 않은 개인의 민감한 정보”를 추정하는 것으로 이어진다. 인공지능 시스템에 의해 수집되고 생성된 개인 정보를 거의 통제할 수 없다는 사실은 프라이버시가 존중될 수 있는지에 대한 의문을 제기한다.

셋째, 특별보고관은 온라인 콘텐츠를 필터링하거나 관리하는데 있어 인공지능의 역할이 갈수록 더욱 커지기 때문에, 인공지능이 표현의 자유 및 차별받지 않을 권리에 중대한 위협이 될 것이라고 강조했다. 몇몇 기업들은 인공지능이 인간의 능력을 넘어서도록 도움을 줄 것이라고 주장하지만, 이 보고서는 자동화된 관리에 의지하는 것은 인권 행사를 방해하는 것이라고 보았다. 사실상 인공지능은 차별적인 가정에 반대하거나 비꼬는 표현 또는 생산되는 모든 콘텐츠의 문화적 맥락을 이해할 수 없다. 그 결과, 인공지능과 사적 행위자에게 복잡한 검열의 시행을 위임함으로써 표현의 자유와 차별받지 않을 우리의 권리가 심각하게 방해받을 수 있다.

3. 기업과 정부 모두를 위한 권고 세트

“윤리”는 기업과 공공기관이 구속력과 강제력이 있는 인권기반 규제를 우회하기 위한 포장이 아니라는 점을 상기시키면서, UN 특별보고관은 “인공지능 분야에서 국가 정책이나 규제를 개발하기 위한 모든 노력에서 반드시 인권을 고려할 것”을 권고했다.

데이비드 케이는 기업의 관행, 인공지능의 설계 및 보급을 인권이 이끌어야 한다고 말하며, 투명성의 증진, 공개 의무 및 효과적인 구제수단을 포함한 강력한 개인정보보호법을 요구하였다. 온라인 서비스 제공자들은 어떤 결정이 사람에 의해 검토되는지 혹은 인공지능 단독으로 이루어지는지를 분명하게 밝혀야 한다. 해당 정보는 알고리즘을 사용한 의사결정 로직에 관한 설명도 동반해야 한다. 더 나아가서, 인공지능 시스템의 “존재, 목적, 구조 및 영향”은 이 주제에 대한 개인 이용자의 교육 수준 향상을 위한 활동의 일환으로써 공개되어야 한다. 보고서는 또한 “인공지능 시스템이 고객불만 및 피해구제를 요청받는 빈도와 가능한 구제수단의 유형 및 효과”에 대한 정보를 제공하고 공개할 것을 권고했다.

국가는 다원적 정보 환경에 적합한 입법 체계를 만들고 기술 독점을 방지하며 망 중립성과 장치 중립성을 지원할 책임이 있는 핵심 주체로 인식된다.

마지막으로, 특별 보고관은 인공지능 개발을 감독하는데 유용한 도구를 제시하였다.

1. 인공지능 시스템의 사용 전, 사용 중, 사용 후에 인권영향평가를 수행하기
2. 인권단체들과 외부적으로 감사하고 의견수렴하기
3. 개인이 선택할 수 있도록 고지와 동의 절차 갖추기
4. 인권침해를 종식시키기 위해 효과적인 구제절차 마련하기

보고서 결론 및 권고[2)]

61. 이 보고서에서 의사 표현의 자유에 인공지능이 미치는 현존하는 혹은 잠정적인 영향을 탐구하고자 하였다. 인공지능이 지금의 정보 환경에서 중요한 부분이라는 전제 하에 개인이 각자의 권리를 누리는 데 (인공지능이) 혜택과 위험을 초래한다는 사실을 지적했다. 기술적 역능이 확대되는 상황에 직면해서 이러한 권리를 옹호하기 위해 국가의 의무와 기업의 책임을 생각해 보는 개념적인 틀을 제시했다. 그리고 인공지능 기술의 권한, 범위, 영역이 확장됨에 따라 인권을 존중하기 위해 정부와 기업 모두가 실행할 수 있는 구체적인 조치를 제안했다.

---

2) Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (2018. 8. 29), A/73/348, https://freedex.org/wp-content/blogs.dir/2015/files/2018/10/AI-and-FOE-GA.pdf.

정부를 위한 권고

62. 인공지능 시스템이나 응용프로그램을 조달하거나 배치할 때, 국가는 공공기관들이 인권 원칙을 준수하도록 보장해야 한다. 특히 인공지능 시스템의 조달 및 배치 전에 공적 의견수렴을 수행하고 인권영향평가 또는 공공기관 알고리즘 영향평가를 수행하는 경우를 들 수 있다. 인종 및 종교적 소수자, 정치적 반대 단체나 활동가에게 이런 기술이 다르게 미칠 수 있는 영향에 더 신경을 써야 한다. 정부가 인공지능 시스템을 배치하는 경우 외부의 독립적인 전문가로부터 정기적인 감사를 받아야 한다.

63. 국가는 민간부문 인공지능 시스템의 설계, 배치 및 실행에서 인권이 중점이 되도록 보장해야 한다. 이는 인공지능 영역에 현행 규제, 특히 개인정보보호 규제를 갱신하고 적용하는 것 등을 말하며, 기업에 인공지능 기술에 대한 영향평가 및 감사 실시를 요구하고 효과적인 외부 책임 메커니즘을 보장하도록 설계된 규제 혹은 공동 규제 체제를 추진하는 것을 포함한다. 가능하다면, 특정 인공지능 응용프로그램에 대한 영역별 규제가 인권을 보호하기 위해 필요하고 효과적일 수도 있다. 그러한 제한으로 표현의 자유에 대한 간섭이 발생하거나 유발되는 경우에는, 국가는 그러한 제한이 자유권규약 제19조 제3항에 따른 정당한 목적을 달성하기 위해 필수적이고 비례적임을 보장해야 한다. 또한 인공지능 관련 규제는 시민사회와 인권단체, 소외되거나 잘 드러나지 않는 최종 이용자를 대표하는 사람 등 광범위한 공적 의견수렴을 통해 발전되어야 한다.

64. 국가는 다양하고 다원적인 정보 환경에 좋은 정책과 입법 환경을 만들어야 한다. 이는 인공지능 분야에서 경쟁력을 확보하기 위해 조치를 취하는 것 등을 말한다. 이런 조치로는 인공지능 전문가와 권력이 몇몇 지배적인 기업에 집중되지 않도록 방지하는 기술 독점 규제, 서비스와 기술의 상호운용성을 증진하도록 설계된 규제, 그리고 망 중립성과 기기 중립성을 지원하는 정책의 채택을 포함할 수 있다.

기업을 위한 권고

65. 인공지능 기술의 윤리적 영향에 대한 가이드라인이나 규약을 만드는 모든 노력은 인권 원칙에 기반해야 한다. 모든 사적, 공적 인공지능 개발 및 배치는 시민사회에 발언할 수 있는 기회를 제공해야 한다. 기업들은 기업 정책 및 기술 지침에서, 모든 기업 운영이 인권 책임에 따라 이루어지고 인공지능의 설계, 배치 및 실행의 특정 상황에서 인권원칙의 적용을 촉진하는 방식으로 윤리 원칙이 도움이 될 수 있음을, 기술자, 개발자, 데이터 기술자, 데이터 정제사(data scrubber), 프로그래머 및 인공지능의 수명 주기에 관여하는 여러 사람들에게 반복적으로 알려야 한다. 특히, 플랫폼 이용약관은 보편적 인권 원칙에 기초해야 한다.

66. 기업은 그들이 소유한 플랫폼, 서비스 및 응용프로그램에서 인공지능 기술과 자동화 기술이 어디에서 어떻게 사용되는지를 분명히 밝혀야 한다. 사람들이 인공지능 기반 의사결정 과정에 처하거나, 인공지능이 콘텐츠를 게시하거나 관리하는 역할을 하거나, 혹은 사람들의 개인정보가 인공지능 시스템에 제공되는 데이

터셋에 통합될 때, 사람들에게 알릴 수 있는 혁신적인 수단을 사용하는 것이 중요하다. 이는 이용자들이 인공지능 시스템이 자신의 인권 향유에 미치는 영향을 이해하고 해결하는데 필요한 고지를 제공할 수 있다. 기업들은 콘텐츠 게시의 경향성에 대한 데이터 뿐 아니라, 얼마나 자주 삭제에 대해 이의가 제기되고 삭제에 대한 문제제기가 수용되는지 등 콘텐츠 삭제에 대한 데이터 또한 공개해야 하며, 여기에는 상업적, 정치적 프로파일링에 대한 사례 연구 및 교육이 동반되어야 한다.

67. 기업들은 반드시 인공지능 시스템의 입출력 모두에서 차별을 방지하고 책임져야 한다. 이는 인공지능 시스템을 설계하고 배치하는 팀들이 다양하고 반차별적인 태도를 반영하게끔 보장하고, 샘플링 오류 해결, 차별적 데이터를 제거하는 데이터셋 정제 및 그런 데이터에 대한 보상 조치를 취하는 등 데이터셋 선정과 시스템 설계에 있어 편향과 차별 배제를 우선순위에 두는 것을 말한다. 인공지능 시스템의 차별적 결과에 대한 적극적인 모니터링 역시 필수적이다.

68. 보유하고 있는 시스템을 새로운 세계시장에 배치하는 경우를 포함하여 새로운 인공지능 시스템의 설계 및 배치가 이루어 지는 동안, 인권영향평가 및 공적 의견수렴이 이루어져야 한다. 공적 의견수렴 및 참여가 의미가 있으려면 제품이나 서비스가 최종 확정되거나 출시되기 전에 이루어져야 하며, 시민사회, 인권 활동가들 및 소외되거나 잘 드러나지 않는 최종 이용자를 대표하는 사람의 참여를 포함해야 한다. 인권영향평가 및 공적 의견수렴의 결과는 그대로 일반에 공개되어야 한다.

69. 기업들은 모든 인공지능 코드가 완전히 감사(audit) 가능하도록 만들어야 하고, 규제기관의 요구조건과 별개로 인공지능 시스템 외부의 독립적인 감사를 가능하게 하는 혁신적인 수단을 강구해야 한다. 인공지능 감사의 결과는 그대로 일반에 공개되어야 한다.

70. 개인 이용자들은 인공지능 시스템의 반인권적 영향에 대한 구제조치에 접근할 수 있어야 한다. 기업들은 인공지능 기반 시스템에 부과되는 모든 이용자들의 불만 및 이의제기에 적시에 대응하기 위해 사람에 의한 검토절차 및 구제조치를 마련해야 한다. 인공지능 시스템에 불만이 제기되고 피해구제가 요청된 빈도에 대한 데이터 뿐 아니라, 이용 가능한 구제수단의 종류와 효과성에 대해서도 주기적으로 공개되어야 한다. □

트리플 A: 알고리즘에 대한 적극적 평등조치
- 인공지능, 머신러닝과 젠더 : 구체적인 행동 촉구

Triple A: Affirmative Action for Algorithms

– Artificial Intelligence, Machine Learning & Gender: A concrete Call to Action[3)]

우리는 중대한 전환점에 있다. 급변하는 글로벌 환경에서 혁신과 번영을 위해 새로운 규범이 필요하다. 오늘날 우리가 살아가고 일하는 방식을 통제하는 이 결함 많은 체제와 문화적 표준의 디폴트(기준값)는 "표준화된 남성"으로서, 이 기준이 너무나 일반화되어 알아차리기도 어렵다. 20세기 약물 실험, 국제 표준 및 국제 무역 규칙에서부터 21세기 알고리즘 의사결정 및 머신러닝 시스템에 이르기까지 이러한 기준이 사람들에게 피해를 준다는 사실이 결국 드러났다. 사실 민주주의 자체가 위험에 처해 있다. 우리는 새로운 규범을 수립해야 한다.

우리는 여성과 소녀들의 모든 유형과 교차성에 초점을 두고 있다. 사회 체제의 옛 규칙을 정할 때 여성을 체계적으로 배제해 왔기 때문에 – 그리고 새로운 규칙을 정할 때에도 여성 배제가 계속되어 왔다 – 우리는 우리가 창출하는 새로운 시스템에서 성평등을 달성하고 민주주의를 강화하기 위해 법률, 규정 및 규범에서 전략적이고 혁신적인 사고를 옹호하는 새로운 연합을 출범시켜야 한다. 우리 사회 시스템을 재구성하는 힘을 가지고 있는 여성들 – 필수적이고 활용되지 않은 자원으로서 – 은 모든 차원의 의사결정에 영향력

3) 스위스 여성단체 Women at the Table (2019. 6. 23.),
https://www.womenatthetable.net/blog/triple-a-affirmative-action-for-algorithms

있게 참여해야 하며, 지금 변화가 일어날 수 있다. 오래된 규범과 고정관념이 미래의 머신러닝 시스템에 주입되기 전에.

특히 ADM(algorithmic decision-making, 알고리즘 의사결정) 시스템과 머신러닝이 도입되고 있는 규모를 감안해 보면 알고리즘에 대한 적극적 평등조치가 시급히 필요하다. 이는 현재 그리고 우리가 맞이할 미래에 여성들이 온전한 참여와 권리를 누리는 것을 방해할 실생활의 편향과 장벽을 바로잡기 위함이다.

마샬 맥루한의 "우리는 우리의 도구를 만든다. 그리고 그 다음에는 우리의 도구가 우리를 만든다"는 말은 잘 알려져 있다.

이것은 우리의 당면 과제다. 우리는 지금 그리고 이후 시대에 걸쳐서 제도적, 문화적 시스템을 계속 변화시킬 수 있는 새로운 도구와 새로운 규범을 확립해야 한다. 이것은 세계 각지의 문제다. 여성과 남성 모두를 위해 지금 우리가 성평등과 민주주의에 초점을 맞추는 것이 중요하다. 그렇게 하면 모두가 성공할 수 있다. 우리는 우리 뒤에 뒤쳐지는 이를 남겨두지 말아야 한다.

우리는 정부, 민간부문 및 시민사회단체에 다음을 촉구한다.

1. 공공 부문과 민간 부문 모두에서 '알고리즘 의사 결정(ADM)'에 대한 책임성과 투명성을 수립하는 지침을 옹호하고 채택할 것.

- 이미 편향된 시스템이 머신러닝으로 우리의 모든 미래에 내포되지 않도록 해야 한다.

행동하라 :

- 여성과 소녀들의 온전한 참여와 동등한 권리 향유를 방해하는 실생활의 편향과 장벽을 바로잡기 위해 알고리즘에 공평한 조치를 취할 것.
- 공공기관이 시범자 및 선구자로 나설 것 : 공공기관이 ADM를 시범 도입할 때 알고리즘에 대한 적극적 평등조치를 취할 것. 기존 체제로 인해 여성들이 전통적으로 뒤쳐져 사회적 인센티브, 보조금, 장학금을 할당받는 신진 사회과학 다년 연구에 시범 도입할 것. 이것은 우리가 오랫동안 수용해 온 평등의 가치를 발전시키고, 인구에 비례해 여성의 가시성과 자질, 영향력을 재고하기 위한 적극적 의제이다.
- 공공 부문 및 민간 부문에서 알고리즘 영향 평가(Algorithmic Impact Assessments)를 실시할 것 : 책무성과 공정성의 원칙에 따라 국민의 삶에 영향을 미치는 인공지능 시스템에 대해 알 권리를 존중하도록 자체 평가 프레임워크를 설계할 것.
- 인공지능 시스템의 수명 주기 전반에 걸쳐 철저한 테스트를 실시할 것 : 제품 수명 주기에 걸쳐 훈련 데이터, 테스트 데이터, 모델, 응용프로그램 인터페이스(API) 및 기타 구성 요소의 출처와 쓰임새를 고

려해서 테스트를 실시해야 한다. 테스트는 편향 및 기타 위해성 검사를 위해 출시전 테스트, 독립 감사, 인증 및 지속적인 모니터링을 포함해야 한다. ADM은 인간의 경험을 통제하는 것이 아니라 그 질을 향상시켜야 한다.

- 책무 증진을 위해 강력한 법체제를 도입할 것 : 부문별 기관의 권한을 잠정적으로 확대하거나, 공공 및 민간 부문을 규제 감독하고 법적 책임을 부과하기 위해 ADM 시스템을 감독, 감사 및 모니터링하는 새로운 위탁 조건을 생성하는 등을 말한다.

- 젠더인지적 조달 지침을 마련할 것 : 정부의 기관 및 모든 활동에서 ADM 성평등 조달 지침을 개발하기 위한 확실한 목표를 수립할 것. 그리고 그러한 원칙을 적용하기 위해 필요한 기관별 역할 및 책임에 대해 개괄할 것.

- 데이터셋을 개선할 것 – 젠더 분리 개방 데이터, 데이터 수집, 포용적인 고품질 데이터셋에 대한 개선 등 : 젠더에 따라 세분화된 개방형 데이터셋을 적극적으로 생성할 것. 이는 인공지능 편향의 원천을 이해할 수 있게 하여 궁극적으로 머신러닝 시스템의 성능을 향상시킬 것이다. 데이터 수집 과정과 인간참여 검증에 대한 감독 통제에 투자하고, 여성 및 전통적으로 소외된 집단을 희생시키는 데이터가 수집되지 않도록 할 것. 데이터셋의 양 뿐 아니라 품질에도 중점을 두는, 보다 포용적인 데이터 수집 과정을 도입할 것.

2. ADM의 생성, 설계 및 코딩 과정에서 다양한 교차성과 동등한 수의 여성과 소녀를 포함하도록 명확한 사전 조치를 수행할 것.

- 새로운 기술은 전적으로 새로운 구조를 창출하여 새로운 아이디어와 새로운 팀에 대한 수요 등 새로운 기회를 제공한다. 현실세계에서 제거되고 있는 성역할들이 젠더, 인종, 계급에 대한 오래되고 전형적인 관념과 관련성을 통해 새로운 ADM에 반영되고 있다. 혁신적이고 포용적인 사고가 필요하다. 이러한 상상력과 기술은 지구상에 가장 많이 활용되지 않은 지적 자원인 여성과 소녀들이 제공할 수 있다.

행동하라 :

- 인공지능 의사결정에서 젠더 균형 : ADM의 자금개발, 설계, 채택 및 평가에 관련된 모든 과정에서 의사결정 젠더 균형을 공식 의제에 포함시켜야 한다.

- 설계팀의 젠더 균형 : ADM 시스템의 설계에 페미니스트와 교차하는 강력한 다양성을 채용하면 혁신과 창의성이 증대될 뿐만 아니라 여성, 소녀 및 전통적으로 배제된 사람들에 대한 편향적이거나 유해한 효과를 감지하고 완화할 수 있다.

- 기업들에 대해 설계팀의 젠더 균형을 적극적으로 공개하고 보고하도록 요구할 것. 균형 잡힌 팀을 보유한 기업에게는 인센티브를 부여할 것.

- 대학과 스타트업이 업스트림 등에서 보조금을 신청할 때, 연구 및 설계팀의 젠더 균형을 적극적으로 공개하고 보고하도록 요구할 것. 균형 잡히고 다분야적인 팀에게는 인센티브를 부여할 것.

- 연구 기금 : 디지털 리터러시를 구현하는 새로운 방법을 포함시키기 위해 컴퓨터 과학과 공학적 관점을 넘어 다분야적인 접근법으로 젠더와 인공지능, 머신러닝, 편향성과 공정성의 영향을 탐구하기 위해 연구 기금을 조성할 것. 여성 및 전통적으로 규칙 제정과 의사 결정에서 배제된 사람들의 삶에 ADM이 미치는 경제적, 정치적, 사회적 효과를 연구하기 위한 연구 기금을 조성할 것.

3. 인권에 기반한 ADM과 머신러닝을 위한 국제 협력 및 접근관점

- 우리 뒤에 뒤쳐지는 이를 남겨두지 않기 위해 왜곡된 데이터 시스템을 대규모로 교정하려면 다자간 그리고 국제적인 협력이 필요할 것이다.

행동하라 :

- UN 기관 차원에서 ADM, 머신러닝 및 젠더 문제에 기존 국제 인권법 및 인권 기준 적용을 검토할 것 : 이러한 조치는 빠르게 변화하는 디지털 시대에 목적에 부합하는 인권 기반 접근에 대한 창조적 사고를 유도하고 자극할 수 있다.

- 디지털 포용을 위한 일련의 지표 개발 : 그 시급성에 부응하여 세계적으로 측정되고, 성별에 따라 세분화된 데이터들이 UN, 국제통화기금, 국제전기통신연합, 세계은행, 기타 다자개발은행 및 OECD와 같은 기관에서 연례적으로 보고될 것. ▢

유럽연합 집행위원회
<신뢰할 수 있는 인공지능을 위한 윤리 가이드라인>

Ethics guidelines for trustworthy AI[4)]

유럽연합 집행위원회 공지사항

2019년 4월 8일, 인공지능 고위전문가그룹은 〈신뢰할 수 있는 인공지능을 위한 윤리 가이드라인(Ethics Guidelines for Trustworthy Artificial Intelligence)〉을 발표했다. 이는 2018년 12월 공개된 가이드라인 초안의 후속 작업이며 500건 이상의 공개적인 의견수렴이 이루어졌다.

가이드라인에 따르면 신뢰할 수 있는 인공지능은 반드시 다음과 같아야 한다.

(1) 적법해야 한다 - 적용되는 모든 법률과 규제를 준수한다.
(2) 윤리적이어야 한다 - 윤리적 원칙 및 가치를 존중한다.
(3) 견고해야 한다 - 기술적 관점과 사회적 환경 모두를 고려해야 한다.

언어별 가이드라인은 다음에서 다운로드할 수 있다. (생략)

가이드라인은 인공지능 시스템이 신뢰할 수 있다고 여겨지기 위해서 충족해야 하는 7가지 핵심 요구사항 세트를 제언한다. 각각의 핵심 요구사항 적용을 확인하기 위하여 구체적인 평가 항목들이 열거되어 있다.

4) 유럽연합 집행위원회 (2019. 4. 8.),
https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

- 인간 작용 및 감독: 인공지능 시스템은 인간에게 자율권을 부여하여, 설명에 기반한 의사결정을 하게끔 하고 그들의 기본적 권리를 신장하여야 한다. 동시에 적절한 감독 체제가 보장되어야 하고, 이는 인간참여(human-in-the-loop), 인간지배(human-on-the-loop), 인간지휘(human-in-command) 등의 접근 방식으로 달성될 수 있다.

- 기술적 견고성 및 안전성: 인공지능 시스템은 회복적이어야 하고 보안이 지켜져야 한다. 인공지능 시스템은 안전해야 하는 바, 정확하고 신뢰할 수 있고 재현할 수 있어야 할 뿐 아니라, 무언가 잘못될 경우 철회 계획을 낼 수 있어야 한다. 이는 의도하지 않은 해악을 최소화하고 방지할 수 있는 유일한 방법이다.

- 프라이버시 및 데이터 거버넌스: 프라이버시와 개인정보 보호를 완전히 보장하는 것 외에도, 적절한 데이터 거버넌스 체제가 반드시 보장되어야 하며, 데이터의 품질 및 무결성을 고려하고 데이터 접근의 정당성을 보장해야 한다.

- 투명성: 데이터, 시스템 및 인공지능 사업 모델은 투명해야 한다. 추적성 메카니즘이 도움이 될 수 있다. 나아가 인공지능 시스템 및 그 의사결정은 관련 이해당사자들에게 적합한 방식으로 설명되어야 한다. 인간은 본인들이 인공지능 시스템과 상호작용하고 있다는 사실을 알아야 할 필요가 있으며, 시스템의 성능과 한계에 대한 정보를 제공받아야 한다.

- 다양성, 차별금지 및 공정성: 불공정한 편향성은 반드시 방지되어야 한다. 사회적 약자 소외로부터 편견과 차별의 악화에 이르기까지 복합적인 부정적 영향을 미칠 수 있기 때문이다. 인공지능 시스템은 다양성을 증진하면서 모든 이들이 장애와 무관하게 이에 접근할 수 있어야 하며, 그 수명주기 전반에 걸쳐 관련 이해당사자들을 참여시켜야 한다.

- 사회적, 환경적 복지: 인공지능 시스템은 미래 세대를 포함하여 모든 인간에게 혜택을 주어야 한다. 따라서 지속가능하고 환경친화적이어야 한다. 나아가 다른 살아있는 존재를 비롯한 환경을 고려해야 하고, 그것이 사회적 관계 및 전체 사회에 미치는 영향을 신중하게 살펴야 한다.

- 책무성: 인공지능 시스템과 그 결과물에 대한 책임성과 책무성을 보장하는 체제가 가동되어야 한다. 감사가능성은 알고리즘 평가를 가능케하는데 특히 중요한 응용분야에 있어서 데이터 및 설계 절차가 내부적으로 중요한 역할을 수행한다. 무엇보다 적절한 시정이 이루어질 수 있어야 한다.

인공지능 고위전문가그룹은 이 가이드라인의 목적을 위해 인공지능 정의를 상세히 설명하는 또다른 문서를 준비해 왔다.

언어별 인공지능 정의 문서는 다음에서 다운로드할 수 있다. (생략)

시범 사업

이 문서는 또한 핵심 요구사항들을 작동시키는 평가 항목을 제시하고 있으며 이를 실제로 구현하기 위한 지침을 제공한다. 6월 26일부터 이 평가 항목에 대한 시범 사업이 진행 중이며, 모든 이해당사자에게 열려 있고 개선 방안에 대한 현실적인 피드백을 기다린다.

피드백은 여러 경로로 수집될 것이다.

- 시범 사업에 등록한 모든 이들에 대한 개방적 설문조사 또는 “정량 분석”
- 다양한 분야에서 보다 상세한 피드백을 수집하기 위한 소수 대표 집단 심층 면접
- 유럽 인공지능 협의체를 통한 피드백 및 모범 관행의 지속적인 업로드

시범 절차는 2019년 12월 1일까지 진행될 예정이다.

자세한 사항은 웹사이트를 방문하시기 바란다. ▢

메모

메모

메모

메모

메모