서울대학교 인공지능 정책 이니셔티브
http://ai.re.kr/

2019 국제학술대회 보고서

미래를 향한 인공지능 정책: 우리는 AI를 신뢰할 수 있을까?
(AI Policy for the Future: Can We Trust AI?)

공동디렉터 **고학수 (서울대학교) & 임용 (서울대학교)**
보고서 작성 **박도현 (서울대학교 박사과정)**

서울대학교 법과경제연구센터
SNU Center for Law & Economics

일시 2019. **8. 23.**(금), 9:00-17:00

장소 한국프레스센터 국제회의장(20층)

주최 서울대학교 법과경제연구센터
SNU Center for Law & Economics

후원 NAVER Google aws

행사 동영상은 다음 사이트를 통해 볼 수 있습니다.

네이버 채널 https://tv.naver.com/v/10550957/list/524105

유튜브 https://www.youtube.com/playlist?list=PLGEVmkdHcJTKcgVvUhEcCDDiHKDUY6MSn

서울대학교 인공지능 정책 이니셔티브
http://ai.re.kr/

2019 국제학술대회 보고서

미래를 향한 인공지능 정책: 우리는 AI를 신뢰할 수 있을까?
(AI Policy for the Future: Can We Trust AI?)

공동디렉터 **고학수 (서울대학교) & 임용 (서울대학교)**
보고서 작성 **박도현 (서울대학교 박사과정)**

일시 2019. **8. 23.**(금), 9:00-17:00

장소 한국프레스센터 국제회의장(20층)

주최 서울대학교 법과경제연구센터
SNU Center for Law & Economics

후원 NAVER Google aws

행사 동영상은 다음 사이트를 통해 볼 수 있습니다.

네이버 채널 https://tv.naver.com/v/10550957/list/524105

유튜브 https://www.youtube.com/playlist?list=PLGEVmkdHcJTKcgVvUhEcCDDiHKDUY6MSn

속표지

서울대학교 인공지능 정책 이니셔티브
(http://ai.re.kr/)
2019 국제학술대회 보고서

공동디렉터 : 고학수 (서울대학교) & 임 용 (서울대학교)
보고서 작성 : 박도현 (서울대학교 박사과정)

국제학술대회
미래를 향한 인공지능 정책: 우리는 AI를 신뢰할 수 있을까?
(AI Policy for the Future: Can We Trust AI?)

날 짜 : 2019. 8. 23.(금)
장 소 : 한국프레스센터 국제회의장(20층)
주 최 : 서울대학교 법과경제연구센터
후 원 : Naver, Google, Amazon Web Services

행사 동영상은 다음 사이트를 통해 볼 수 있습니다.
네이버 채널 : https://tv.naver.com/aipolicyintiative
유튜브 : https://www.youtube.com/channel/UCKyxSZOtLB1YvkKM2_Mq8gQ

행사일정

전체 진행 : 최경진 (가천대학교)

등 록 9:00–9:30

환영사 9:30–9:35

- 고학수 (서울대학교 인공지능 정책 이니셔티브 공동디렉터)

축 사 9:35–9:45

- 민원기 (과학기술정보통신부 제2차관)

세션 I 9:50–12:40

인공지능의 거버넌스: AI가 대중의 신뢰를 얻기 위해 무엇이 필요한가?

모더레이터 : 임 용 (서울대학교 인공지능 정책 이니셔티브 공동디렉터)

기조연설 I 9:50–10:15

- R. David Edelman (MIT)

기조연설 II 10:15–10:40

- Deirdre Mulligan (UC 버클리 정보대학)

패널토론 11:00–12:40

패널
- R. David Edelman (MIT)
- 김현섭 (서울대학교)
- Jake Lucchi (구글)
- Deirdre Mulligan (UC 버클리 정보대학)
- 송경희 (국장, 과학기술정보통신부)
- 윤정원 (아마존 웹 서비스)

세션 Ⅱ 14:00–16:50

인공지능과 공정: “공정”한 AI는 무엇을 의미하고 어떻게 구현될 수 있는가?
모더레이터 : 고학수 (서울대학교 인공지능 정책 이니셔티브 공동디렉터)

기조연설 Ⅲ 14:00–14:25

- Fredrik Heintz (린코핑대학교, 스웨덴 & EU 인공지능 고위전문가 그룹)

기조연설 Ⅳ 14:25–14:50

- Blaise Ag era y Arcas (구글 AI)

패널토론 15:10–16:50

패널
- Norberto Andrade (페이스북)
- Blaise Agüera y Arcas (구글 AI)
- Gary Chan (싱가포르경영대학교)
- Fredrik Heintz (린코핑대학교, 스웨덴 & EU 인공지능 고위전문가 그룹)
- Malavika Jayaram (디지털 아시아 허브, 홍콩)
- Indrė Žliobaitė (헬싱키대학교, 핀란드)

전체 토론 및 폐회 16:50–17:00

국제학술대회
미래를 향한 인공지능 정책: 우리는 AI를 신뢰할 수 있을까?
(AI Policy for the Future: Can We Trust AI?)

들어가는 말

우리는 인공지능(Artificial Intelligence, AI)을 신뢰할 수 있을까? 자율주행 자동차 시험운행 중 탑승자가 사망하기도 하고, 드론을 통한 공격이 발생하기도 하는 오늘날 시의성을 지닌 중요한 물음이 아닐 수 없다.[1] 공동체와 사회구성원의 '신뢰(trust)'는 인공지능 기술의 수용성과 밀접한 관련이 있다는 점에서, 이러한 물음에 대한 적절한 대처는 앞으로 인공지능 산업의 성패를 결정짓는 요인이 될 것이라고 예측되기도 한다.[2] 그리하여 서울대학교 법과경제연구센터는 지난 2017년부터 빅데이터, 인공지능 기술을 둘러싼 각종 법적, 정책적 이슈를 논의하기 위하여 국내외 관련 분야의 전문가들을 초빙하여 매년 컨퍼런스를 개최해오고 있다.[3] 2018년 제2회 컨퍼런스부터는 지금과 같은 형

1) 최근 미국 연방교통안전위원회(National Transportation Safety Board, NTSB)에서는 지난 해 발생한 테슬라 자율주행 자동차의 소방트럭 추돌 사고에서 자율주행 시스템에 부분적이나마 책임이 있다는 결론을 내렸다고 한다. 연합뉴스 "美교통안전위 "테슬라 자율주행 시스템, 충돌사고 책임있다" "(2019. 9. 5.) 참조. 한편, 최근 사우디아라비아에 위치한 석유시설에 대한 드론을 이용한 테러로 국제유가가 급등하기도 하여 사회적 문제로 비화되었다. 김상욱, "값싼 드론, '막대한 피해주는 테러무기'로 거듭나", 뉴스타운(2019. 9. 16) 참조.

2) High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI", European Commission (2019. 4. 8), pp. 4-5.

3) 제1회 컨퍼런스에서는 '인공지능, 알고리즘, 개인정보보호를 둘러싼 정책적 과제(Policy Issues surrounding AI, Algorithms & Privacy)'라는 제목을 걸고, 오프닝 세션에서 '인공지능, 알고리즘, 개인정보보호를 둘러싼 정책적 과제'를 논의한 뒤 개별 세션에서 '인공지능 빅데이터와 시장경쟁의 문제', '인공지능 의사결정과 법적 사회적 책임(accountability)', '데이터 비식별화'를 다루었다. KDI에서는 제1회 컨퍼런스를 촬영하여 인터넷에 한글자막과 함께 동영상을 업로드 하였다. 그 주소는 아래와 같다.
(http://www.youtube.com/playlist?list=PLOP6ilKzhDLQ_a2hMmD0vxsJn0d-aQco8)
한편, 제2회 컨퍼런스에서는 '인공지능의 시대: 기술 발전에 따른 책임과 규제(Artificial Intelligence Today: Governance and Accountability)'라는 제목을 걸고, 제1세션은 올바른 '데이터 거버넌스(Data Governance)' 구축에 관한 문제를, 제2세션은 '자동화된 의사결정(automated decision-making)'의 책임(accountability)과 윤리에 관한 문제를, 제3세션은 블록체인(blockchain)을 포함하여 다양한 신기술을 올바르게 규율하는 방안을 논의하였다. 서울대학교 자유전공학부 학생인 이지현・박재승・윤건희・조영채는 제2회 컨퍼런스를 촬영하여 인터넷에 한글자막과 함께 동영상을 업로드 하였다. 그 주소는 아래와 같다.
(http://tv.naver.com/aipolicyintiative; http://www.youtube.com/channel/UCKyxSZOtLB1YvkKM2_Mq8gQ)

태의 학술대회 보고서도 제작하여 컨퍼런스의 효과를 극대화하기 위한 다각적 노력을 기울여왔다.[4] 한편, 서울대학교 법과경제연구센터는 2017년부터 인공지능 기술의 지속적 발전과 활용 증가로 나타나고 있는 사회·경제적, 법적, 윤리적 문제들을 고민하고 정책적 함의를 모색하는 연구 프로그램인 서울대학교 인공지능 정책 이니셔티브(SNU AI Policy Initiative)를 운영해오고 있는 중이다.[5]

2019년 제3회 인공지능 컨퍼런스에서는, 점차 많은 관심과 주목을 받고 있는 '인공지능 시대에서의 신뢰성(trust)과 공정성(fairness)의 구현'을 대주제로 삼았다. 인공지능의 신뢰를 얻기 위해서는 인공지능이 가져오는 파급효과가 인류의 공정성이라는 가치체계에 부합할 필요가 있다는 점에서 양자는 밀접한 관련이 있다. 제1세션에서는 주로 전자에 집중하여 인공지능이 대중의 신뢰를 받기 위해 필요한 거버넌스 구조를, 제2세션에서는 주로 후자에 집중하여 공정한 인공지능의 의미와 구현방안을 논의하였다.[6] 올해는 관련 분야 권위자의 통찰력을 바탕으로 토론의 장을 보다 활발하게 하기 위한 차원에서, 세션마다 토론에 들어가기에 앞서 기조연설을 2개씩 배치한 점이 특기할만하다.

이번 제3회 컨퍼런스에서도 명망 있는 국내외 관련 분야 인사들을 초빙하여 논의의 수준과 깊이를 더할 수 있었다. 축사를 담당한 민원기 과학기술정보통신부 제2차관은 인공지능 분야의 전문가이자 경제협력개발기구(OECD)의 경제정책위원회 의장으로서, OECD 인공지능 원칙을 수립하는 과정에 있어서 핵심적인 기여를 한 권위자로 손꼽히고 있다. 그밖에 기조연설자들도 학계와 실무 영역에서 각기 최고 수준의 권위자로 인정받고 있는 인사들을 섭외하였다. 제1세션 기조연설자들은 현재 인공지능 분야의 학계를 선도하는 이들로, MIT 교수 에델만(R. David Edelman)은 수년간 백악관에서 IT 정책을 직접 입안하였을 정도로 실무와 학계를 두루 경험한 권위자로 널리 알려져 있고, 버클리 교수 멀리건(Deirdre Mulligan) 역시 프라이버시와 공정성을 중심으로 많은 연구업

4) 박도현, "서울대학교 인공지능 정책 이니셔티브 2018 국제학술대회 보고서", 『인공지능의 시대: 기술 발전에 따른 책임과 규제』(2018). 보고서는 서울대학교 정책 이니셔티브 홈페이지에서 다운로드 받을 수 있다.

5) 제2회 컨퍼런스 이후에 발간된 보고서로는 "해외 비식별조치 가이드라인 등에 대한 비교·분석", "프로파일링 관련 기술 동향 분석 및 개인정보 정책 방안 연구"가 있고, 단행본으로 『데이터 오너십 : 내 정보는 누구의 것인가?』가 출간되었으며, 그밖에 "인공지능과 미래사회"라는 제목의 이슈페이퍼를 발간하였다. 보다 자세한 내용은 http://ai.re.kr/who-we-are/ 참조.

6) http://ai.re.kr/%ea%b3%b5%ec%a7%80%ec%82%ac%ed%95%ad/?uid=22&mod=document&pageid=1 참조.

적을 쌓은 명망 있는 학자이다. 반면, 제2세션 기조연설자들은 현재 인공지능 분야의 실무를 이끄는 이들로, 하인츠(Fredrik Heintz)는 린코핑(Linköping) 대학교 교수이면서 유럽연합 인공지능 고위전문가 그룹(EU High-Level Expert Group on Artificial Intelligence)에 소속된 전문가이고, 아게라 이 아카스(Blaise Agüera y Arcas)는 구글의 인공지능 실무를 이끄는 수석과학자이자 기술 전문가로 널리 인정받고 있다.

그밖에도 미주, 유럽, 아시아 등지에서 널리 인정받는 중진 및 신진 학자들을 고루 초빙하고, 기술이나 법 분야 전문가 외에도 철학을 전공한 서울대학교 철학과 김현섭 교수를 섭외함으로써 학문의 장벽을 넘나드는 종합적 사고가 가능하도록 노력하였다. 나아가 학계를 넘어 정부나 기업의 입장도 경청할 수 있도록 초청연사의 배경을 다양화할 수 있도록 많은 신경을 기울였다. 예를 들어, 앞서 언급한 민원기 차관과 함께 과학기술정보통신부 송경희 국장처럼 공직에 몸담고 있는 인사들과, 구글, 아마존, 페이스북을 비롯한 글로벌 사기업에 재직 중인 인사들을 고루 초빙하여 컨퍼런스의 논의가 각계각층의 관점을 반영할 수 있도록 노력하였다. 이에 법조계뿐만 아니라 공학, 경제학, 철학 등 다양한 분야의 배경을 가진 400명이 넘는 청중들이 컨퍼런스에 참석하여 자리를 빛내주었다.

I. 제1세션 – 인공지능의 거버넌스: AI가 대중의 신뢰를 얻기 위해 무엇이 필요한가? (Governance of AI: What Needs to be Done to Achieve Public Trust in AI-Impacted Outcomes?)

1. 기조연설 I – 인공지능의 거버넌스: 머신들은 어떻게 우리의 신뢰를 확보할 수 있는가? (Governing Artificial Intelligence: How Machines Can Earn Our Trust?)

기조연설자 에델만은 '파괴적 혁신(disruptive innovation)'이라는 화두를 던지면서 연설을 시작하였다.[7] 실제로 언론에서는 인공지능을 비롯한 제4차 산업혁명의 산물에 대해 파괴적 혁신이라는 용어를 빌려가면서 온갖 장밋빛 미래를 예측하고 있다. 과연 그렇기만 할까? 에델만은 이처럼 언론의 헤드라인을 장식하는 인공지능 기술에 대해 어떤 의심도 없이 수용하는 자세의 위험성을 지적하였다. 인공지능 기술에 대한 맹목적 믿음은 사회질서를 저해하는 요인이 될 수 있을 뿐 아니라, 인류에게 실제로 많은 편익을 가져다줄 수 있는 인공지능 산업의 발전을 저해할 수 있다는 것이다.

에델만은 먼저 유명한 SF 작가인 아서 클라크(Arthur C. Clarke)가 남긴 "고도로 발달한 기술은 마법과 구별되지 않는다(Any sufficiently advanced technology is indistinguishable from magic)"는 문구를 인용했다. 실제 과거 스티브 잡스(Steve Jobs)가 새로운 제품을 출시할 때마다 대중과 언론은 마법과 같다면서 열광적 반응을 보인 바 있다. 지난 2016년 알파고(AlphaGo)가 세계 바둑계를 10년 이상 지배해왔던 이세돌 9단을 무너뜨린 사건에서 우리나라 대중과 언론이 보인 모습도 여기에서 크게 다르지는 않은 듯싶다.[8] 그러나 이러한 생각에는 커다란 맹점(blind spot)이 있다. 예를 들어, '인공지능이

---

7) '파괴적 혁신(disruptive innovation)'이라는 개념을 자세하게 설명한 문헌으로는, Clayton M. Christensen・Michael E. Raynor・Rory McDonald, "What Is Disruptive Innovation?", Harvard Business Review (2015. 12.) 참조.

8) 알파고는 2016년 3월 이세돌을 이긴 알파고 리(Lee), 이듬해인 2017년 3월 당시 세계랭킹 1위 커제(柯洁)를 이긴 알파고 마스터(Master), 그해 10월 완전한 자가학습만으로 알파고 마스터에게 89%의 승률을 거두면서 세계 최강의 바둑 인공지능으로 평가받은 알파고 제로(Zero)라는 세 가지 버전으로 구분된다. 이와는 별개로, 그해 말에는 단일 알고리즘으로 바둑뿐만 아니라 다양한 종류의 게임을 동시에 습득한 자가학습 알고리즘인 알파 제로(Alpha Zero)가 등장하기도 하였다. 자세한 내용은 David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search", Nature 529 (2016); David Silver et al., "Mastering the game of Go without human knowledge", Nature 550 (2017); David Silver et al., "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm", arXiv:1712.01815 (2017) 참조.

무엇인가?’라는 물음만 생각해보아도 일반인은 물론 전문가조차 어떠한 종류의 합의에 이르지 못한 것이 현실이다.[9)]

그럼에도 불구하고 우리는 부지불식간에 인공지능 기술을 이용하고, 그것의 혜택을 누리면서 살아가고 있다. 에델만은 인공지능 기술이 활용되는 몇 가지 대표적 예시를 들었다. 스팸메일 필터링 기능의 강화, 생체정보를 이용한 보안 기능의 향상, 사진 태깅의 정확도 향상, 행동패턴의 상관관계를 통한 예측 기능의 강화 등은 널리 알려진 인공지능 기술의 응용 사례이다. 인공지능은 과거 인터넷 광고 목적으로 주로 사용되었지만, 이제는 수많은 다양한 영역에 응용되고는 한다. 가령, 우리가 다니는 학교의 모습은 오늘날인 2019년이나 100년 전인 1919년이나 별 차이가 없지만, 인공지능을 활용한 교수법은 과거 오랜 기간에 걸쳐 학습해야 한 분량을 짧은 기간에 습득할 수 있게 하는 맞춤형 학습법 같은 새로운 학습 방법론을 가능케 하고 있다. 병원에서는 인공지능을 활용하여 부작용에 의한 많은 사망자를 내는 질환인 패혈증(sepsis) 발생률을 상당히 감축하기도 하였다.[10)] 아이들이 등교할 때 학교 스쿨버스의 최적의 배차 시간대나 통학경로를 파악하는 데 인공지능이 활용된 사례도 유명하다.[11)]

그렇지만 에델만은 오늘날 인공지능은 어디까지나 터미네이터(Terminator)와 같은 자유의지나 자의식을 가지지 않고, 전지전능하지도 않으며, 기술의 발전 단계는 아직 초창기에 불과하다는 점을 강조하였다. 컴퓨터가 50년 이상의 역사를 가진 것과 비교할 때, 딥러닝(Deep Learning, DL)으로 대표되는 새로운 인공지능 방법론이 정립된 것은 겨우 5년 남짓에 불과하기 때문이다.[12)] 하지만 오늘날 언론을 장식한 헤드라인은 마치 인공

9) 이런 이유로 인해, 세계적으로 널리 읽히는 인공지능 교과서인 러셀(Stuart Russell)과 노빅(Peter Norvig)의 책에서도 인공지능을 적극적으로 정의하는 대신, 인공지능의 접근방식으로 크게 ‘인간적 행위’, ‘인간적 사고’, ‘합리적 사고’, ‘합리적 행위’라는 4가지 접근이 공존하는 상황이라고만 서술할 따름이다. 스튜어트 러셀・피터 노빅(류광 역), 『인공지능: 현대적 접근방식(제3판)』, 제이펍 (2016), 2-6쪽 참조.

10) 자세한 내용에 관하여는, Steven Ashley, “Using Artificial Intelligence to Spot Hospitals’ Silent Killer”, NOVA (2017. 10. 11.) 참조.

11) 자세한 내용에 관하여는, Joi Ito, “What the Boston School Bus Schedule Can Teach Us About AI”, Wired (2018. 11. 5.) 참조.

12) 물론 ‘인공지능’의 역사는 컴퓨터 못지않게 오래된 것이고, 심지어 양자의 기원을 (앨런 튜링 같은 선구자로) 동일하다고 바라볼 여지도 있다. 에델만이 말한 인공지능은 알파고의 등장을 전후로 급부상한 ‘딥러닝’ 기술로 좁게 해석함이 타당할 것이다. 그렇게 본다면, 이미지넷(ImageNet)의 대용량 시각인식 경연대회(Large Scale Visual Recognition Challenge, LSVRC)에서 인공지능의 정확도가 처음으로 인간의 그것을 추월한 2015년을 인공지능 발전사의 이정표로 간주하려는 관점에도 일리가 있다. Yoav Shoham et al., “Artificial Intelligence Index 2018 Annual Report” (2018), p. 47 참조.

지능이 인간과 같은 방식의 자율성(autonomy)을 가지고 행위하는 것처럼 오해를 불러일으킬 수 있다. 그러나 인공지능 시스템에는 설계자를 비롯한 수많은 인간의 의지가 반영되어 있고, 따라서 인공지능의 잘못된 학습에 의하여 오류(error)나 차별(discrimination)과 같은 문제가 발생한 경우, 그에 대한 책임은 인공지능 그 자체가 아닌 인간에게 귀속되어야 한다.[13]

인공지능 기술을 마법처럼 바라볼 때 책임의 문제와 함께 인류에게 발생할 수 있는 또 다른 해악의 위험성이 바로 주체성(agency)의 문제이다. 널리 알려진 것처럼 인공지능은 블랙박스(black-box)에 비견될 정도로 복잡하고 전문가조차도 속속들이 파악하기 어려운 존재이다.[14] 물론 우리가 일상 속에서 접하는 자동차나 휴대폰, 기타 전자기기도 복잡하기는 매한가지라고 생각할 수 있다. 그러나 이러한 기술은 상대적으로 많은 이들에게 친숙하고, 추상적으로나마 작동원리에 대해 이해하고 있는 부분이 많은 편이다. 그러나 (딥러닝) 인공지능은 단지 빅데이터 분석을 통해 그와 같은 결과가 도출되었다는 현상적 사실 이외에는 별로 말해주는 것이 많지 않은 상황이다. 물론 세상 모든 영역에서 작동원리를 이해하고 살 필요도 없고, 그것이 진정한 의미의 주체성이라고 여기지도 않게 마련이기는 하다. 그러나 적어도 인류의 자유와 권리에 중대한 영향을 끼치는 영역에서만큼은 구체적 작동원리에 대한 이해를 포기한 채 단지 마법과 같은 것으로 받아들이는 맹목적 태도는 주체성의 상실과 다름없다.[15]

에델만은 그에 대한 대표적 사례로 인공지능을 이용한 신용평가 의사결정을 언급하였다. 인공지능은 과거 신용이력이 없는 사람에 대하여도 관리비 납부이력, 통신요금 연체이

---

13) 인공지능의 오류에 관하여, 훈련 데이터에 인간의 눈에 보이지 않는 미량의 노이즈(noise)를 첨가하였을 때 인공지능 알고리즘에 막대한 오류를 발생시킬 수 있는 위험성을 지적한 Anh Nguyen・Jason Yosinski・Jeff Clune, “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”, arXiv:1412.1897v4 (2015) 참조. 인공지능의 차별에 관하여, 과거의 훈련 데이터에 내재되어 있는 차별적 행동양식이 인공지능의 의사결정에 고스란히 반영되거나 이를 경로의존적으로 심화시킬 수 있는 위험성을 지적한 고학수・정해빈・박도현, “인공지능과 차별”, 저스티스 통권 제171호 (2019), 242-244쪽 참조.

14) 이러한 문제를 이른바 ‘인공지능의 불투명성(opacity)’이라고 일컫는다. 이에 관한 대표적 선행연구로, Jenna Burrell, “How the machine ‘thinks’: Understanding opacity in machine learning algorithms”, Big Data & Society (2016) 참조.

15) 예를 들어, Michal S. Gal, “Algorithmic Challenges to Autonomous Choice”, Michigan Technology Law Review Vol. 25, Issue 1 (2018), p. 63에서의 선택설계(choice architecture)나 에코챔버(echo chamber)는 일상생활과 정치적 영역에서 개인의 선택이 외부요인에 좌우될 수 있는 대표적 사례일 것이다. 이것이 충분한 정보에 기한 숙고된 동의(informed consent)에 의하였는지 여부에 따라 판단이 달라질 것이다.

력과 같은 비금융정보를 활용하여 신용평가를 가능케 하는 장점을 가진다. 반면, 인과관계(causality)가 아닌 상관관계(correlation)에 기초한 인공지능 의사결정의 특성은 주기적으로 휴대폰을 충전하는지 여부와 같은 인간이 보기에 적절한 관련성이 없는(irrelevant) 요인을 주된 의사결정 변수로 삼을 수도 있다.[16] 이러한 의사결정이 실제로 인간의 의사결정보다 더 정확할 수 있더라도, 그것이 과연 정당한지는 별개의 판단을 거쳐야 할 문제일 것이다.

물론 이러한 문제에 대해 전통적 법제가 아무런 대처를 해오지 않은 것은 아니다. 예를 들어, 미국의 공정신용보고법(The Fair Credit Reporting Act)과 같은 법제는 이미 수십 년 전부터 신용평가가 잘못된 정보에 의한 경우 정정권이나 차별금지 의무를 규정하고 있다.[17] 이러한 법제에서 '인공지능' 기술을 특별히 예정하고 있지는 않지만, '공정성 원칙'과 같은 법원리[18]에 기초한 합리적 법해석과 규율공백이 예상되는 지점에 대한 약간의 법개정을 통해 문제를 해결해나갈 수 있을 것이다. 에델만은 여기서 중요한 것은 인공지능 윤리(ethics)의 차원을 넘어선 거버넌스(governance)임을 강조하면서, 윤리와 거버넌스의 차이는 마치 정치철학(political philosophy)과 규제(regulation)의 차이와 비견될 수 있다고 보았다. 정치철학에서의 논의가 결국 규제로 구체화되는 것처럼, 인공지능 윤리도 어느 시점에서는 거버넌스 형태로 구체화되어야 한다는 것이다. 예를 들어, 2018. 5. 25. 발효된 유럽 일반개인정보보호규정(General Data Protection Regulation, 이하 'GDPR')에서부터 '설명을 요구할 권리(right to explanation)가 도출될 수 있는지, 설명보다 정확성의 효용이 높은 질병예측과 같은 영역에서까지 정확성을 희생하고 설명성을 강화하도록 요구할 것인지와 같은 문제는 사회적 합의와 체계적 규율이 요구되는 사안이다.

에델만은 여기에 더해 협업의 필요성을 강조하였다. 비록 국가나 지역마다 문화나 환경이 다르지만, 인공지능 기술의 특성상 법제도가 수렴할 것으로 예측되기 때문이다. 나

16) 인과관계가 아닌 상관관계에 의존하는 딥러닝 인공지능의 한계를 지적한 Gary Marcus, "Deep Learning: A Critical Appraisal", arXiv:1801.00631v1 (2018), pp. 12-13 참조.

17) 우리나라의 현행 '신용정보의 이용 및 보호에 관한 법률' 제6장도 신용정보주체의 보호에 관한 규정을 두고 있다. 다만, 동법에서 차별금지 의무가 규정되어 있지는 않으므로 신용평가와 관련한 차별행위가 발생한 경우 현재는 '국가인권위원회법' 내지 '헌법'만이 적용될 수 있을 것으로 보인다. 자세한 사항은, 고학수・정해빈・박도현, 앞의 논문, 256-265쪽 참조.

18) 법규범의 두 가지 유형인 법원리(legal principle)와 확정적 법규범(legal rule, 법규칙)의 차이점에 관하여는 김도균・이상영, 『법철학(개정판)』, 한국방송통신대학교 (2011), 39-44쪽 참조.

아가 이러한 법제도적 논의는 본성적으로 제로섬(zero-sum)이 아니라는 점에서 효용가치도 높다. 인공지능은 군비경쟁과 같은 경쟁의 일부라는 세간의 인식과 달리, 각국 정부, NGO, 학계, 기업 등지에서는 이미 수년 전부터 인공지능 거버넌스에 대한 논의를 거듭해오고 있는 것도 이러한 이유 때문이다.[19] 끝으로 이것이 가능하기 위해서는 기술 전문가와 정책 전문가 사이의 의사소통이 필요하고, 가능해야 한다는 점을 강조하면서 발표를 마무리하였다.

2. 기조연설 II – 정책으로서의 공공계약: 머신러닝 기반 행정절차의 관점에서 (Procurement as Policy: Administrative Process for Machine Learning)

두 번째 기조연설자 멀리건의 발표는 정부가 인공지능(멀리건은 머신러닝 방식으로 국한)을 선택하는 상황에서의 거버넌스, 다시 말해, 공공계약과 조달(procurement)에 관한 문제였다.[20] 주지하듯, 정부의 의사결정은 기본적으로 민주주의 원리에서 비롯되므로 사회적 가치를 반영하여야 하고, 또한 그것이 올바른 것이 될 수 있도록 법치주의 원리를 비롯한 견제장치에 의하여 조정될 수 있어야 한다. 한편, 정부의 의사결정은 가시적이어야 하고, 합리성과 전문성을 구비하여야 하며, 나아가 대중의 참여와 감시가 뒷받침되어야 한다.[21] 특히 부당한 국가권력 행사의 대상이 된 국민의 입장에서 항변 가능한(contestable) 것이어야만 비로소 적법절차원리(due process)나 재판청구권이 올바로 구현된 것이고, 이를 통해 사회적 신뢰가 싹틀 수 있을 것이다.

멀리건은 이러한 쟁점이 널리 문제된 루미스(Eric L. Loomis) 사례를 언급하였다. 에릭 루미스는 콤파스(COMPAS)라는 재범예측 프로그램의 평가를 참조한 1심 법원에 의하여 6년형을 선고받게 되었다. 루미스는 이러한 결론이 도출되는 과정이나 판단이 되는 입력 데이터나 가중치와 같은 근거에 대한 정보를 얻을 수 없었다. 정부에서는 그러

19) 대표적 사례로 지난 2017년 1월 인공지능 분야의 각계각층 인사들이 1975년 생명윤리 논의를 시작한 곳인 아실로마(Asilomar)에 모여 23가지의 인공지능 윤리원칙을 발표한 것을 들 수 있다. 자세한 윤리원칙 내용은 Future of Life Institute, “Asilomar AI Principles” (2017) 참조.

20) Deirdre K. Mulligan・Kenneth A. Bamberger, “Procurement as Policy: Administrative Process For Machine Learning”, Berkeley Technology Law Journal Vol. 34 (2019).

21) 우리나라 법제에서 이와 같은 원칙은 헌법원칙뿐만 아니라, ‘행정절차법’ 제4조 신의성실 및 신뢰보호, 제5조 투명성 원칙이나, 객관성・투명성・공정성을 규정한 ‘행정규제기본법’ 제5조 제3항 등에 잘 드러나 있다.

한 정보를 가지고 있지 않았고, 콤파스 제작자는 지식재산권을 이유로 그와 같은 정보를 제공할 의무가 없었기 때문이다. 이에 루미스는 자신에게 내려진 판결이 적법절차 원리 위반을 비롯한 여러 이유로 위법한 것이라고 주장하였으나, 위스콘신주 대법원은 루미스의 주장을 받아들이지 않았다.[22] 멀리건은 결국 이 사안의 핵심은 인공지능을 이용한 의사결정의 불투명성이고, 이러한 불투명성은 인공지능을 이용한 국가의 모든 행정행위로 확장될 수 있다고 지적하였다.

문제는 인공지능 알고리즘이 결코 중립적이고 객관적이기만 하지 않는다는 것이다. 관련 변수나 가중치의 의미를 규정하는 것은 물론이고, 모델을 선택하거나 배치할지 여부를 결정하는 등 다양한 상황에서 정책적 이슈가 나타날 수 있다. 가령, 앞서 에델만이 지적한 신용평가 사안의 경우, '휴대폰 충전여부'라는 변수는 적어도 휴대폰을 구매할 수 있는 경제력을 전제하기 때문에 가치중립적으로 보기 어렵다. 마찬가지로 인간의 자유의지나 여타 사회적 특성이 포함된 데이터를 (주로 자연과학적 예측에 활용되는) 알고리즘에 적용했을 때 나타난 결과를 과연 어디까지 신뢰할 수 있고 어떻게 해석하여야 할지와 같은 물음에 대한 간단한 해답은 적어도 지금까지는 존재하지 않는다.[23] 예를 들어, 안면인식 기술의 역치(threshold)를 결정하는 경우, 일반적 상황에서와 범죄자 판정에 활용하는 상황에서의 민감도는 상당히 다를 수밖에 없을 것이다. 그렇기에 인공지능을 이용한 의사결정을 할 때 알고리즘의 출력값을 그대로 이용자에게 제공할지, 이를 보정할 수 있는 환경설정을 제공할지와 같은 결정에는 모종의 가치평가가 이루어지게 마련이다.

멀리건은 그럼에도 불구하고 인공지능이 조달의 대상이라는 점 때문에 마치 가치중립적인 것처럼 오해를 하여 그것의 정치적, 정책적 측면이 충분히 부각되지 않는다고 지적하였다.[24] 나아가 발표의 주제인 정부조달의 맥락에서 이러한 문제가 더욱 부각된다

---

22) State of Wisconsin v. Eric L. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016). 나아가 연방대법원은 이 사건에 대한 상고허가신청을 불허하였다고 한다. 보다 자세한 설명은 한애라, "'사법시스템과 사법환경에서의 인공지능 이용에 관한 유럽 윤리헌장"의 검토 - 민사사법절차에서의 인공지능 도입 논의와 관련하여 -", 저스티스 통권 제172호 (2019), 65-66쪽 참조.

23) 예측적 치안활동에 주로 활용되는 프레드폴(PredPol)을 비롯한 인공지능 알고리즘의 작동원리는 기본적으로 범죄와 지진의 발생패턴이 가진 유사성에 기초한다. George O. Mohler et al., "Self-exciting point process modeling of crime", Journal of the American Statistical Association Vol. 106, Issue. 493 (2011) 참조.

24) 인공지능 프로그램과 인공지능이 탑재된 하드웨어는 경제적 생산물인 동시에 과학기술의 산물이다. 그렇기에 인공지능 역시 과학기술의 객관성이나 가치중립성 문제, 과학기술과 사회발전의 관계를 둘러싼

고 보았다. 많은 사람들이 인공지능을 조달하는 상황을 일반 생산물을 조달하는 상황과 유사하게 바라보고는 하지만, 앞서 지적하였듯이 인공지능 알고리즘은 우리가 세상을 이해하고 형성하게 하는 매개체라는 점에서 단순한 물건으로 취급되어서는 곤란하다는 것이다. 그렇기 때문에 일반적 조달과정에서 중요하게 여기는 가격, 성능과 같은 이슈뿐 아니라, 투명성, 공정성, 민주성 같은 다양한 가치가 올바르게 반영되고 있는지 여부까지도 인공지능 조달 과정에서 중요한 이슈로 다루어져야 한다. 그러나 오늘날 정부조달을 담당하는 실무자는 인공지능 기술에 대한 이해는 물론, 인공지능이 가진 정책적, 정치적 함의에 대한 이해가 부족하게 마련이다. 이는 사실상 민간업체의 견해가 인공지능의 정책에 반영되고, 나아가 일반 국민에게 제공되는 행정서비스가 민주적 정당성이 부족한 기술전문가의 사적 판단에 놓여 당사자인 개인의 절차적 권익이 침해될 위험으로 이어진다.[25]

원칙적으로 국민의 권익에 영향을 주는 어떠한 행정처분이 이루어지는 과정에서는, 행정절차법상 투명성 원칙(제5조), 사전통지(제21조), 의견청취(제22조), 이유 제시(제23조), 의견제출 및 청문(제27조 내지 제37조), 공청회(제38조 내지 제39조의2) 같은 다양한 절차적 보호규정이 적용되게 마련이다. 이는 국내뿐만 아니라 세계 각국에서 대체적으로 그러하다. 이러한 절차규정은 행정처분의 적법성과 정당성을 자체로 어느 정도 담보하고 사후적 문제제기를 용이하게 하여 책임성을 강화할 수 있다.[26] 멀리건은 단순 도구로 활용되는 경우를 넘어 인공지능의 의사결정 자체가 개인의 권익에 중대한 영향을 미치는 사안에서는 적어도 이러한 엄격한 기준을 충족할 필요가 있다고 역설하였다. 나아가 멀리건은 행정부가 가진 일정한 정도의 재량(discretion)으로부터 착안하여, 행정기관은 인공지능을 활용할 때 시스템의 다양한 기준을 조정하는 소위 '팅커링(tinkering)'을 적절하게 행할 수 있고, 행해야 한다고 주장하였다.[27]

기술결정론과 사회결정론의 대립을 비롯한 일련의 철학적 물음에 직면한다. 자세한 내용은, Maarten Franssen · Gert-Jan Lokhorst · Ibo van de Poel, "Philosophy of Technology", Stanford Encyclopedia of Philosophy (2018) 참조.

25) 이와 같은 문제의식을 소위 '기술 영역의 적법절차(technological due process)'라는 용어로 풀어낸 선구적 논의로, Danielle Keats Citron, "Technological Due Process", Washington University Law Review Vol. 85 (2008) 참조.

26) 행정절차법을 위반한 처분은 많은 경우 취소사유에 해당되지만, 어떤 경우는 당연무효 사유가 되기도 한다. 가령, 대법원 2019. 7. 11. 선고 2017두38874 판결 참조.

27) 팅커링 개념을 다룬 논의로는, Angèle Christin, "Algorithms in practice: Comparing web journalism and criminal justice", Big data & Society (2017) 참조.

인공지능이 활용된 행정에서의 민주화를 달성하기 위해서는 앞서 언급한 불투명성에 맞서고 문제를 개선할 수 있는 다양한 전문가 집단의 참여가 요구될 것이다. 멀리건은 미국 국세청의 납세자보호관(National Taxpayer Advocate) 제도를 원용하면서 국세청 내의 인공지능에 대한 피드백이 이들을 통해 이루어진다고 언급하였다. 이에 더하여 멀리건은 영향평가, 프로토타입 제시, 시뮬레이션 등을 통해 문제 상황에 처한 이들이 항변할 수 있는(contestable) 인공지능 설계를 탑재하도록 하는 것이 진정한 인공지능 조달과 그에 뒤이은 행정절차의 투명성을 달성할 수 있는 길이라고 제시하면서 발표를 마무리하였다.

3. 패널토론 (Panel Discussion)

제1세션의 패널은 기조연설을 담당한 에델만과 멀리건 이외에 서울대학교 철학과의 김현섭 교수, 구글의 루치(Jake Lucchi), 과학기술정보통신부의 송경희 국장, 아마존 웹서비스의 윤정원 대표로 구성되었다. 먼저 김현섭 교수는 제1세션의 핵심 키워드 '신뢰' 측면에서, 현재 인간의 인지과정에서 '감정(emotion)'이 담당하는 역할이 실마리가 될 수 있다는 견해를 제기했다.[28] 체스 챔피언 카스파로프(Garry Kasparov)를 이긴 딥 블루(Deep Blue)의 승리 비결을 두고 인간과는 달리 감정의 소용돌이에 휘말리지 않는다는 점이 꼽혔는데, 나아가 도덕적 판단에도 감정의 개입이 없는 기계가 인간이 상상하지 못하는 뛰어난 성과를 가져올 수 있지 않을까 하는 물음이 제기되고 있다.[29] 이에 대하여 김현섭 교수는 감정이 때로는 올바른 의사결정을 방해하는 요인이 될 수 있다는 점을 인정하는 한편, 때로는 개념적 평가가 누락하는 정보를 의사결정에 반영할 수 있도록 하는 기제가 될 수 있다고 지적하였다.

철학적으로 감정은 주체가 주목하는 대상에 대한 감각이나 느낌을 넘어선, 평가적 요소를 포함하는 개념이다. 전자의 관점에 따르면 그에 대한 어떠한 옳고 그름을 논의한다는 것이 불가능하지만, 후자의 관점에 따르면 그러한 평가에 대한 옳고 그름을 논할 수 있다는 차이가 발생한다. 가령, 어떤 사물을 보고 두려움을 느꼈는데 그것이 진짜 뱀

28) 감정의 의미와 인공 감정의 실현가능성에 대한 최근 논의로, 천현득, "인공 지능에서 인공 감정으로 - 감정을 가진 기계는 실현가능한가? -", 철학 제131집 (2017), 224-233쪽 참조.

29) 예컨대, Colin Allen · Gary Varner · Jason Zinser, "Prolegomena to any future artificial moral agent", Journal of Experimental and Theoretical Artificial Intelligence Vol. 12, No. 3 (2000), p. 260 참조.

일 때와 장난감 뱀일 때 두려움이라는 평가에 대한 옳고 그름은 달라진다. 그렇다면 감정에 의한 인지과정은 '잘못된 것'이라기보다, 감정에 의하지 않은 인지과정을 보완하는 소중한 정보획득 메커니즘의 일종일 수 있다. 인터넷에서는 안전하다고 설명된 지역을 막상 야간에 거닐었을 때 두려움이 엄습하였다면, 이 두려움은 실체가 없는 오류일 수도 있지만, 생명을 구할 수 있는 사실일 수도 있다. 때문에 직관이나 감정에 근거한 인지체계를 자동화된 의사결정 과정에서 완전히 배제하기보다 이를 적절히 반영할 수 있는 방안을 강구할 필요가 있다는 생각에 이른다.

다만, 김현섭 교수는 이성에 대한 감정의 우위보다는 감정이 성찰과 숙고의 계기를 제공할 수 있다는 점을 강조하였을 뿐임을 분명히 하였다.[30] 그리고 인공지능의 의사결정 시스템과 인간의 감정 시스템이 항상 별개의 것이 아닐 수 있고, 감정 시스템의 역할을 개념화함으로써 인공지능 의사결정에 탑재하는 일도 얼마든지 가능하다고 보았다.[31] 종래 인공지능과 인간의 협업은 인간의 주체성이나 인공지능의 책임성을 강화하기 위한 수단적 측면에서 강조되었지만, 김현섭 교수는 '감정을 통한 인공지능의 자동화된 의사결정 능력의 향상'이라는 보다 실질적 측면이 또 하나의 이유가 될 수 있지 않을까 하는 화두를 제기하면서 발표를 마무리하였다.

다음으로 루치는 대중의 신뢰를 높이기 위한 구글의 거버넌스 원칙에 대해 설명하였다.[32] 구글은 지난 해 인공지능 기술을 통해 이루고자 하는 사항과, 하면 안 되는 사항에 대한 대원칙을 수립하였고[33], 루치는 그중 편향성을 극복하기 위한 노력에 대하여 특히

30) 이러한 견해는 인간의 인지체계를 별로 노력을 들이지 않은 채 저절로 빠르게 작동하는 단순한 정신활동에 대한 이른바 '시스템 1'과 이와는 대조적으로 많은 노력을 들여야만 하고 느리게 작동하는 복잡한 정신활동에 대한 이른바 '시스템 2'로 구분하면서, 후자의 역할을 강조하는 주장과도 일맥상통한다. 자세한 설명은 대니얼 카너먼(이창신 역), 『생각에 관한 생각(제2판)』, 김영사 (2018), 1부 참조.

31) 이러한 생각은 최근에 대두되는 '인공적 도덕행위자(Artificial Moral Agent, AMA)' 관련 논의에도 시사점을 준다. 인공적 도덕행위자에 관한 대표적 연구로, 웬델 월러치・콜린 알렌(노태복 역), 『왜 로봇의 도덕인가』, 메디치미디어 (2014) 참조.

32) Google, "Perspectives on Issues in AI Governance" (2019. 1. 22.).

33) Sundar Pichai, "AI at Google: our principles" (2018. 6. 7.).
구글이 특히 강조한 사항은 사회적 혜택, 편향성 방지, 안전성, 책임성, 프라이버시, 과학적 탁월성이다.
1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.

강조하였다. 예를 들어, 구글이 사용하는 'What-If'라는 툴은 특정 데이터 포인트를 변화시켰을 때 모델이 어떤 결과를 산출하는지를 직관적으로 나타냄으로써 일종의 반사실적 설명(counterfactual explanation)을 제공한다.[34] 나아가 특정 데이터 세트가 과다대표 혹은 과소대표 되었는지 여부를 검증하거나, 사전적으로 합의된 정책적 제약을 모델에 적용하는 등 다양한 툴이 활용되고 있다. 물론 이처럼 투명성 내지 설명성을 높이기 위해서는 앞서 에델만이 지적한 것처럼 일정 정도의 정확성에 대한 희생이 요구되게 마련이다. 하지만 루치는 책임성 같은 구글의 또 다른 중요한 원리와 균형 잡힌 시각에서 바라본다면 일정 정도의 정확성 감소가 이윤을 추구하는 기업의 입장에서도 정당화될 수 있을 것이라고 보았다.

물론 다양한 원리 사이의 형량이 요구되고, 어느 하나의 가치만을 지나치게 강조한다면 곤란할 것이다. 가령, 널리 알려져 있듯 강화된 투명성은 이것을 노린 악의적 사용자의 조작적 행위(gaming)에 대한 취약성을 높이게 마련이다.[35] 공정성을 제고하기 위하여 한 테스트가 개인의 프라이버시를 침해할 수 있는 위협을 비롯한 다양한 원리 사이의 상충관계(trade-off)에 대한 고려도 요구된다. 이처럼 수많은 고려요소를 올바로 판단하기 위해, 구글은 앞서 말한 다양한 기술적 툴과 함께 이를 뒷받침하는 교육과 전담 인력의 확보, 문제가 생겼을 때 상층부에 신속히 상황을 전달할 수 있는 프로세스, 나아가 제도적 노력을 넘어 이러한 사고를 문화로 반영할 수 있도록 하는 다양한 거버넌스적 노력을 병행하는 중이라고 한다. 끝으로 루치는 이러한 구글의 노력이 현실에 적용되는 과정에서 실효성을 높이기 위해서는 인공지능이 활용되는 구체적 맥락에 대한 고려와, 각

---

7. Be made available for uses that accord with these principles.

반대로 구글이 하면 안 된다고 강조한 사항은 해악과 위험, 무기, 감시, 국제법과 인권 위반이다.

1. Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.
2. Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.
3. Technologies that gather or use information for surveillance violating internationally accepted norms.
4. Technologies whose purpose contravenes widely accepted principles of international law and human rights.

34) 'What-If' 툴에 대한 설명으로, http://pair-code.github.io/what-if-tool/ 참조. 반사실적 설명은 현재 상태를 전제로 원하는 결과가 도출될 수 있도록 하려면 어떤 변수를 얼마나 변화시켜야 할지 알려주는 것을 말한다. 박도현, 앞의 글, 16쪽 참조. 반사실적 설명에 대한 대표적 연구로는, Sandra Wachter · Brent Mittelstadt · Chris Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR", Harvard Journal of Law & Technology (2018) 참조.

35) Jenna Burrell, 앞의 논문, pp. 3-4.

국 정부 및 국제규범과의 조화가 중요하다는 점을 강조하면서 발표를 마무리하였다.[36]

송경희 국장은 인공지능이 신뢰를 형성하는 과정에서 우리나라 정부가 담당하는 역할에 대해 소개하였다. 인공지능의 중요한 특징 중 하나는 불확실성(uncertainty)의 증대이다. 주지하듯, 인공지능에 대하여는 다양한 장밋빛 미래에 대한 전망과 동시에, 실업, 차별, 불평등의 강화를 비롯한 음울한 미래에 대한 전망이 점쳐지고 있다. 앞서 언급한 정부와 민간을 불문한 국제사회의 노력과 윤리원칙은 이와 같은 부작용을 최소화하려는 움직임의 일환으로 이해할 수 있을 것이다. 이것이 단일한 당사자나 영역에서 행한다고 될 일이 아니고, 이해당사자 모두가 모여 힘을 합쳐야 할 인류 전체의 문제라는 것은 말할 나위가 없다. OECD에서 내놓은 윤리 가이드라인이나 G20 정상 선언문은 그러한 노력의 대표적 산물인 셈이다.[37] 앞서 언급한 민원기 차관뿐 아니라 우리나라 정부는 이와 같은 사항이 도출되는 과정에서 중요한 기여를 하였다.

우리나라 정부는 이러한 국제사회의 노력에 발맞추기 위해, 인공지능 기술에 대한 연구개발과 함께 윤리 가이드라인을 마련하였다.[38] 가이드라인은 국제적 논의를 반영하여 공공성(Publicness), 책무성(Accountability), 통제가능성(Controllability), 투명성(Transparency)이라는 일명 'PACT' 원칙으로 대표되고, 각각의 원칙은 개발자, 공급자, 이용자라는 3면의 주체에 적용된다. 여기에 더해 국가 인공지능 전략이 조만간 발표될 예정으로, 현재는 다양한 이해당사자의 의견을 수렴하는 과정에 있다. 그밖에 인공지능 연구개발 생태계를 마련하기 위하여 5년 간 2조 5천억 원에 달하는 투자를 계획하고 있다고 한다. 여기에는 인공지능 대학원 신설, 프로젝트형 교육, 허브 구축 등 다양한 노력이 포함된다. 송경희 국장은 앞으로도 프랑스, 러시아, 영국, 브라질을 비롯한 해외 각국과의 교류를 통해 국제적 협력을 더욱 강화하려는 계획을 모색하고 있다고 언급하면서 발표를 마무리하였다.

---

36) 루치는 싱가포르 사례를 특히 강조했다. 자세한 내용은 Singapore Personal Data Protection Commission, "A Proposed Model Artificial Intelligence Governance Framework" (2019. 1.) 참조.

37) 이에 대하여는 OECD, "Recommendation of the Council on Artificial Intelligence" (2019. 5. 22.); G20, "G20 Ministerial Statement on Trade and Digital Economy" (2019. 6.) 참조.

38) 지난 해 출간된 인공지능 연구개발 전략에 관한 내용으로는 과학기술정보통신부, "I-Korea 4.0 실현을 위한 인공지능(AI) R&D 전략"(2018. 5.) 참조. 지능정보사회 윤리 가이드라인과 지능정보사회 윤리헌장의 내용에 관하여는 정보문화포럼・한국정보화진흥원, "지능정보사회 윤리 가이드라인"(2018) 참조.

끝으로 윤정원 대표는 인공지능의 활용에 관련된 아마존 웹 서비스(Amazon Web Services, 이하 'AWS')의 입장을 소개하였다. 먼저 소비자 만족도 제고라는 목표를 달성하기 위해서는 개인의 권리를 보호할 필요가 있고, 따라서 기술 혁신은 각국의 실정법 준수를 전제로 이루어져야 한다고 지적하였다. 이를 위해 AWS는 자사제품의 사용을 제한하는 정책(AWS Acceptable Use Policy)을 펼치고 정책을 위반하는 악용우려를 신고할 수 있는 절차도 마련해두고 있다.[39] 그리고 다양한 오픈 커뮤니티에 참여함으로써 이해당사자들과 지속적으로 대화하는 장을 마련하려고 노력하는 중이라고 밝혔다. 특히 기본권 침해우려가 큰 안면인식(face recognition) 기술에 관하여 어떤 상황에서는 사용되어서는 안 되고, 어떤 상황에서는 제한적으로 사용되어야 하는지에 관한 구체적 가이드라인을 제작하기도 하였다.[40]

다른 한편, 윤정원 대표는 인공지능의 진정한 민주화를 달성하기 위해 사용자 집단의 특성을 세분화할 필요가 있다는 점을 지적하였다. 전문가 집단, 데이터 과학자 같은 일정 수준의 이해도를 가진 집단, 일반인 집단에 대해 각기 다른 방식의 툴을 제공하는 것이 진정한 의미의 '설명성'이라는 생각이다. 구체적으로 일련의 노력을 통해 아마존에서 제공하는 기술의 97%가 고객의 수요를 반영한 맞춤형 제품이 될 수 있었다고 제시하였다.[41] 대표적 사례로, 프레드 허치슨 연구소(Fred Hutchinson Cancer Research Center)의 연구사례를 포함하여 미국 암 환자의 10%는 인공지능 기술을 이용한 종양진단 프로그램의 수혜를 받고 있는 상황이다.[42] 이러한 경험은 향후 인공지능 법제화에 있어서도 소비자 수요를 고려하는 일이 얼마나 중요한지에 대한 많은 시사점을 준다.

---

39) AWS 이용정책에 관한 자세한 설명은 http://aws.amazon.com/ko/aup/ 참조. AWS 서비스 악용행위에 대한 신고절차에 대하여는 http://pages.awscloud.com/rekognition-abuse.html 참조.

40) Michael Punke, "Some Thoughts on Facial Recognition Legislation" (2019. 2. 7.).

41) 이러한 수치는 아마존에서 매년 정기적으로 개최하고 있는 컨퍼런스인 re:invent에서 발표된 것이라고 한다. re:invent에 관한 자세한 설명은 http://reinvent.awsevents.com/ 참조.

42) 보다 자세한 내용은, Taha A. Kass-Hout · Matt Wood, "Introducing medical language processing with Amazon Comprehend Medical" (2018. 11. 27.) 참조.

II. 제2세션 – 인공지능과 공정: “공정”한 AI는 무엇을 의미하고 어떻게 구현될 수 있는가? (Fairness in AI: What does it Mean, and How can it be Implemented?)

1. 기조연설 III – 공정성, 설명가능성 그리고 신뢰 가능한 인공지능: 그 현안과 기술적 도전 (Fairness, Explainability and Trustworthy AI: Technical Challenges and State of the Art)

제1세션의 키워드가 ‘신뢰’였다면, 제2세션의 키워드는 ‘공정’이다. 인공지능의 신뢰성을 향상하기 위한 중요한 전제가 되는 공정성을 확보할 수 있는 방안은 무엇일까? 기조연설자인 하인츠는 유럽 집행위원회의 인공지능 고위 전문가 그룹(European Commission High-Level Expert Group on AI)에 참여한 경험과, 그룹에서 작성한 “우리가 작성한 신뢰할 수 있는 인공지능을 위한 윤리 가이드라인(Ethics Guidelines for Trustworthy AI)”에 근거하여 이러한 물음에 대한 자신의 견해를 제시하였다.[43)]

하인츠는 먼저 ‘신뢰할 수 있는 인공지능’이라는 명제에 있어 우리는 현재 중요한 변곡점에 도달한 상황이라고 진단하였다. 일단 대중이 인공지능을 신뢰할 수 없다는 결론을 내릴 경우 재차 신뢰를 회복하기는 어려워 보인다는 것이다. 결국 인공지능의 긍정적 측면을 적절하게 활용하려는 기업이나 정부 모두에 신뢰성의 문제는 중차대한 요인인 셈이다. 특히 유럽연합은 이러한 생각 하에, 각국이 한데 모여 그동안 인류가 개발한 다양한 방법론을 총동원하여 인류에게 편익을 가져다주고, 그로 인해 대중의 신뢰를 얻을 수 있는 인공지능에 도달하기 위한 노력을 하고 있다. 가이드라인은 신뢰할 수 있는 인공지능의 세 가지 구성요소로 합법적이어야 하고(lawful), 윤리원칙을 준수해야 하며 윤리적(ethical), 기술적・사회적 관점에서의 견고성(robustness)을 갖추어야 한다는 점을 제시한다.[44)] 이러한 세 가지 구성요소는 상호보완적인데, 가령, 합법적이고 윤리원칙을 준수한 경우에도 인공지능이 의도치 않은 해악을 초래할 위험이 있기 때문이다.

---

43) 앞에서 언급한 High-Level Expert Group on Artificial Intelligence, “Ethics Guidelines for Trustworthy AI”, European Commission (2019. 4. 8).

44) High-Level Expert Group on Artificial Intelligence, 앞의 논문, p. 5. 다만, 이 보고서에서는 합법성보다 윤리성과 견고성에 주로 초점을 맞추고 있다.

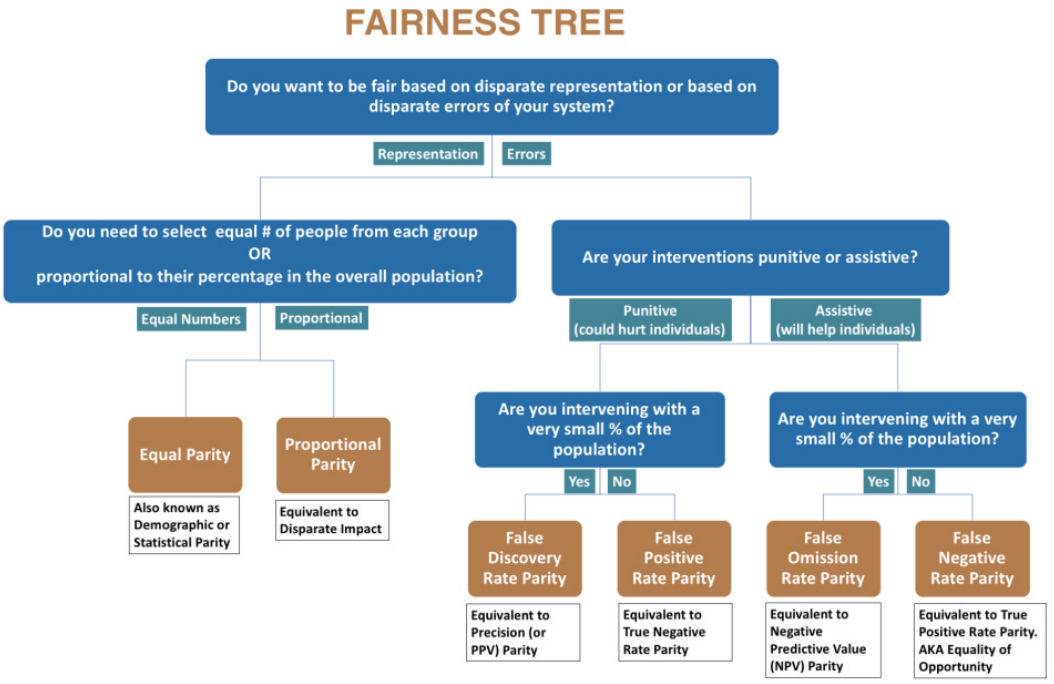
가이드라인은 신뢰할 수 있는 인공지능의 토대가 되는 기본권에 기초하여 자율성의 존중, 해악금지, 공정성, 설명성이라는 4가지 윤리원칙을 제시하고, 이를 현실에서 구현하는 과정에서 필요한 주체성과 감시, 기술적 견고성과 안전, 프라이버시와 데이터 거버넌스, 투명성, 다양성과 차별금지 그리고 공정성, 사회적·환경적 복지, 책임성이라는 7가지 핵심요소를 이끌어냈다. 나아가 현실에서 인공지능 시스템이 이러한 사항을 제대로 준수할 수 있도록 130여 가지의 질의사항으로 구성된 평가 리스트를 제공하였다. 이러한 내용은 잠정적인 것이고, 지속적으로 파일럿 프로젝트와 같은 보완절차를 거칠 예정이라고 한다.[45)]

하인츠는 보고서에 대한 개괄적 설명을 한 뒤 그중 공정성과 설명성에 관한 논의를 심화해 이어갔다. 공정성을 제고하기 위해서는 이를 방해하는 요인인 편향성에 대해 생각해볼 필요가 있다. 앞선 세션에서도 지적되었던 것처럼, 데이터 세트가 모집단을 올바로 대표(represent)하지 못함으로써 비롯된 편향성은 종종 현실을 왜곡하는 요인이 되곤 한다. 다만, 이러한 문제는 대체로 더 많은 표본을 수집하면 해결할 수 있기 때문에 현실적 어려움은 몰라도 근본적 문제라고까지 보기는 어려울 것이다. 반면, 과거의 역사적 차별을 반영한 데이터 같은, 모집단 자체가 가진 편향성이 문제되는 상황은 상대적으로 해결하기가 훨씬 더 어렵다. 물론 그러한 인공지능의 최종적 목적이 현실을 있는 그대로 분석하는 차원에 그친다면 아무런 문제가 되지 않을지도 모르겠지만 그러한 현실을 그대로 용납하기 어려운, 공정성이 중시되는 맥락에서는 그렇지 않을 것이다. 다만, 공정성을 이유로 인공지능의 산출물에 대한 평가나 보정을 한다면 그러한 작업에 대한 권한과 내용의 적정성에 대한 새로운 문제가 제기될 수 있다.[46)]

나아가 '공정성' 자체에 대한 이해방식이 제각기 다르고, 설혹 이를 합치할 수 있다고 하더라도 인간의 직관적 이해를 어떻게 인공지능이라는 공학적 산물에 반영할지가 문제된다. 예를 들어, 시카고대학교의 데이터 사이언스 및 공공정책 센터(Center for Data Science and Public Policy, University of Chicago)에서는 편향성 감사 툴인 'Aequitas'라는 오픈소스 프로그램을 제공하고 있는데, 여기서 활용하는 공정성 나무(fairness tree)는

45) 자세한 내용은 High-Level Expert Group on Artificial Intelligence, 앞의 논문, p. 9 이하 참조.
46) 표본의 대표성 문제에 의한 차별과 모집단의 역사성 문제에 의한 차별 발생의 경로를 구분하고, 유형별로 발생하는 문제의 특성을 비교분석한 연구로, 고학수·정해빈·박도현, 앞의 논문, 265-266쪽 참조.

6가지 유형의 공정성 개념을 규정하고 있다.[47] 이들 유형 중 어느 하나가 항상 옳거나 그르다고 말할 수 없기 때문에, 구체적 맥락에 따른 가치평가가 필연적으로 요구되는 셈이다. 공정성에 대한 정의만 어떻게든 이루어지면, 의사결정에 대한 보정절차는 다음과 같이 행해질 수 있다. 단순하게 말하면 '입력과 출력'이라는 프로세스의 어느 한 가지 혹은 그 이상에 대한 보정을 하는 것이다. 데이터의 존재나 내용에 일부 변형을 가하는 사전처리를 하는 방안, 알고리즘에 사전·사후적 개입을 가하는 방안 등이 그것이다.



[그림 1] 공정성 나무(출처 : 시카고대학교 데이터과학 및 공공정책 센터)

다음으로 하인츠는 설명성에 대한 논의를 이어갔다. 설명성은 자체로 흥미로운 문제이지만, 공정성을 제고하는 한 가지 방편이기도 하다. 하인츠는 앞선 세션에서 지적한 것처럼 설명성과 정확성 사이에 상충관계가 존재하는 것은 사실이지만, 그렇다고 하여 지나치게 설명성이 뒤떨어지고 정확성만 뒷받침되는 인공지능이 만연하게 될 것 같지는 않다고 전망하였다. 정확성은 인공지능이 활용되는 구체적 맥락과 결부될 여지가 있고, 그렇다면 일부 상황에서는 극도로 정확하고 여타 상황에서는 그렇지 않은 인공지능의

47) Aequitas 소개 글, http://www.datasciencepublicpolicy.org/projects/aequitas/ 참조.

설명성이 뒤떨어진 경우, 그와 같은 인공지능에 대한 신뢰가 형성되기 어렵기 때문이다. 다만, 이와 같은 생각이 인공지능이 활용되는 '모든 상황'에서 상당한 설명성이 뒷받침되어야 한다는 결론으로 이어지지 않는다는 것은 물론이다. 하인츠는 인간의 의사결정에 요구되는 설명성조차 그렇지는 않다고 지적하였다.

앞서 멀리건이 언급한 것처럼, 설명성의 가장 중요한 기능은 불만족스러운 결정에 대한 항변을 용이하게 하는 것이고, 그렇다면 여기서 말하는 '설명'이란 인간 입장에서 이해할 수 있는 것이어야 할 것이다. 따라서 인공지능의 구체적 작동을 이해하기 힘든 일반인에게 소스코드 전체를 공개한다고 '투명성'은 몰라도 '설명성'이 높아졌다고 보기는 어렵다. 미국 방위고등연구계획국(Defense Advanced Research Projects Agency, DARPA)에서 진행하는 '설명가능 인공지능(explainable AI, XAI)' 프로젝트에서 인공신경망을 보다 설명성이 높은 모델과 결합하거나, 애초에 모델을 생성할 때 설명성이 높은 방식을 활용하거나, 시각화된 유저 인터페이스를 강화하여 직관적 이해를 도모하려는 이유도 그로부터 설명할 수 있을 것이다.[48] 이와 같은 작업의 궁극적 목적은 정확성을 최대한 보존하면서 설명성을 도모하는 것이라고 말할 수 있다.[49] 그리고 설명성이 높아진 인공지능은 개발자가 시스템을 개선하는 데도 도움이 된다. 끝으로 하인츠는 향후 상관관계로부터 인과관계에 기반한 접근법으로의 전환이 필요하고, 이러한 연구가 공정성이 무엇인지에 대한 인간의 이해 자체를 도모하는 효과도 있을 것이라는 점을 강조하면서 발표를 마무리하였다.

48) DARPA의 설명가능 인공지능 프로젝트는 제2회 컨퍼런스에서 이미 다루어진 바 있다. 이에 관하여는 제2회 컨퍼런스 보고서인 박도현, 앞의 글, 13-14면; David Gunning, "Explainable artificial intelligence (XAI)", Defense Advanced Research Projects Agency (DARPA) (2017) 참조.

49) 하인츠가 언급한 대표적 모델은 '라임(Local Interpretable Model-agnostic Explanations, LIME)'이다. 라임은 기본적으로 입력값에 약간의 교란(perturbation)을 주고 출력값에 나타난 국지적(local) 변화를 통해 독립변수와 종속변수 사이의 관계를 파악하는 방법이다. 보다 자세한 설명으로는, Marco Tulio Ribeiro · Sameer Singh · Carlos Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", arXiv:1602.04938v3 (2016) 참조.

2. 기조연설 IV – 책임있는 인공지능의 개발: 프라이버시에 관한 Federated Learning의 사례를 통한 접근 (A Responsible Development of AI: With an Example of Federated Learning on Privacy)

마지막 기조연설자인 아게라 이 아카스는 주로 기술적 측면에서 인공지능의 공정성과 프라이버시 문제를 해소하는 방안에 대해 논의하였다. 공정성 문제가 대두되는 이유를 설명하기 위해, 아게라 이 아카스는 2차대전 이후의 컴퓨터 기술과 최근 논의되는 딥러닝 기반 인공지능이 작동하는 방식의 차이를 간단히 소개하였다. 컴퓨터 프로그램 언어로 작성된 (소스)코드는 기본적으로 엄격하게 정해진 규칙에 의해 순차적으로 작동되는 방법론을 취한다.[50] 그 과정에서 여러 가지 이유로 오류가 발생할 수 있고, 코드의 양이 많아질수록 가독성이 떨어지고 디버깅(debugging)이 어려워진다. 컴퓨터는 반복적 연산을 빠른 속도로 수행함으로써 수리적 계산은 물론 수많은 영역에 응용되어 오늘날 인류의 삶을 풍요롭게 해주고 있다.

컴퓨터가 발전하는 과정에서 우리는 컴퓨터가 인간보다 비교우위를 가지는 영역과 반대로 인간이 컴퓨터보다 비교우위를 가지는 영역이 무엇인지를 알게 되었다. 가령, 딥블루나 알파고가 보여주었듯 수리적 연산과 그것으로 쉽게 환원될 수 있는 구조를 가진 문제는 컴퓨터가 인간보다 우월한 능력을 발휘하게 마련이다. 졸음이나 배고픔, 감정적 동요와 같은 육체적 한계에 보다 덜 취약한 것은 물론이다. 반면, 걷고, 달리는 것과 같은 매우 기본적 운동능력에 있어서는 인간이 보다 우월한 능력을 발휘하고 있다.[51] 나아가 인간은 컴퓨터와는 달리 유추(analogy)와 같은 유연한 사고에 능하여 기존에 보지 못한 대상이나 개념에 대한 이해와 창조적 사고가 가능하다. 전통적 방식의 하드 코딩(hard coding)만으로는 "생각하는 기계"라는 튜링의 이상을 구현하기 어려운 이유가 바로 이것 때문이었다.[52]

---

50) 이는 이른바 '튜링머신(Turing machine)'의 고유한 특성이다. 튜링머신을 최초로 고안한 앨런 튜링의 고전적 문헌으로, A. M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem", Proceedings of the London mathematical society Series 2 Vol. 42 (1936) 참조.

51) 이 문제를 지적한 로봇공학자 모라벡(Hans Moravec)의 이름을 딴 '모라벡의 역설(Moravec's paradox)'로 흔히 일컫는다. 보통 고도의 추론과 같은 복잡한 인지능력에 비해 단순한 운동능력의 경우 훨씬 오래 전부터 인류의 진화과정에 포함된 것을 이유로 든다. 나아가 앞서 언급한 것처럼 이미지넷 경연대회에서 인공지능이 인간의 시각인식 능력을 추월하기는 하였지만, 시각인식에 소요되는 에너지의 총량을 기준으로 본다면 여전히 인공지능에 비하여 인간이 월등히 우월한 성능을 보이고 있다. 김대식, 『김대식의 인간 vs 기계』, 동아시아 (2016), 10-11쪽, 19쪽.

이러한 문제의식을 가지고 발전한 전통적 방식의 하드코딩과 대비되는 방법론이 바로 인공신경망(Artificial Neural Network, ANN)이다. 인공신경망은 인간의 신경세포(neuron)와 그 연결부위인 시냅스(synapse)가 연결된 두뇌 속 신경망 구조를 모방한 것이다.[53] 단층 퍼셉트론과 같은 초기 인공신경망 모델이 가지고 있는 한계를 점차로 극복하면서 복잡화한 것이 오늘날 '딥러닝'이라고 불리는 심층신경망 모델이라고 말할 수 있다. 인간의 두뇌를 모사한 딥러닝 모델은 과거 컴퓨터가 어려워하던 패턴인식에 특히 탁월한 능력을 보이고 있고, 그러한 점에서 딥러닝을 보다 발전시킨다면 마치 인간과 같이 '생각할 수 있는 기계'에 도달할 수 있지 않을까 하는 생각이 점점 확산되고 있는 것이다. 아게라 이 아카스는 '연인들(Les Amants)'이라는 예술작품에 대한 외설 시비를 두고 스튜어트(Potter Stewart) 전 연방대법관이 말한 "보면 안다(I know it when I see it)"라는 유명한 말을 인용하면서,[54] 오늘날 딥러닝 인공지능의 작동방식도 마치 인간의 직관과 유사한 형태로 변모해나가고 있다는 사실을 지적하였다. 다만, 앞서 하인츠가 지적한 것처럼 표본 데이터의 대표성이나 모집단 데이터의 역사적 차별성과 같은 요인이 존재할 경우 편향성 문제에 봉착할 수 있다. 물론, 이에 대해서는 훈련에 사용되는 데이터(training data)와 별개의 검증 데이터(validation data)나 테스트 데이터(test data)를 활용하는 과정에서 적어도 부분적으로 걸러낼 수 있고, 실무에서도 그러한 방법을 활용하고 있는 중이다.

공정성은 기본적으로 이러한 테스트 과정을 얼마나 적절하게 구현하는지 여부와 밀접한 관련을 맺는다. 훈련 데이터 대표성 제고, 훈련 데이터와 테스트 데이터의 분리, 알고리즘에 부적절한 사항에 대한 제약조건을 내재하는 등의 작업을 올바로 이행했을 때 비로소 인류가 생각하는 '공정성'이라는 이상과 부합할 수 있기 때문이다. 예컨대, 안면인식 기술을 통해 민감정보를 추론해내지 못하도록 인공지능의 활용목적에 대한 제약을 알고리즘 속에 탑재할 수 있을 것이다. 인공지능의 의사결정은 과거를 재생산하는 것을 넘어 과거의 문제를 경로의존적으로 재생산할 수 있기 때문에 이러한 문제가 발생하지

52) A. M. Turing, "Computing Machinery and Intelligence", Mind, New Series, Vol. 59, No. 236 (1950), p. 433.
53) 인공신경망에 대한 최초의 아이디어는 맥컬럭(McCulloch)과 피츠(Pitts)의 1943년 모델이 꼽히지만, 실제로 학습(learning) 과정까지를 포함하여 구현된 최초의 모델로 로젠블라트(Frank Rosenblatt)가 1958년 제안한 퍼셉트론(perceptron)이 널리 알려져 있다. 퍼셉트론을 최초로 제시한, F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain", Psychological Review Vol. 65, No. 6 (1958) 참조.
54) Jacobellis v. Ohio, 378 U.S. 184 (1964), p. 197.

않도록 원천적으로 차단할 필요가 있다.[55] 딥러닝은 기본적으로 막대한 데이터와 연산에 필요한 인적·물적 인프라의 활용을 전제하므로, 공동체 구성원에게 미치는 외부효과(externality)를 암묵적으로 전제하면서 발전되는 기술이기 때문이다. 아게라 이 아카스는 공정성에 대한 완벽한 사회적 합의가 불가능하다는 근본적 문제는 겸허하게 수용하되, 다만, 현실에서 가장 문제되는 사례군만큼은 해결해나가려는 점진적 태도가 필요하다고 역설하였다.[56]

다음으로 아게라 이 아카스는 딥러닝 인공지능을 이용하는 과정에서 문제될 수 있는 프라이버시 문제를 해결할 수 있는 기술적 대안으로 '엣지(edge) TPU'와 '연합학습(federated learning)'에 대해 소개하였다. 엣지 TPU는 1센트 동전보다 훨씬 작은 크기의 초소형 칩으로, 구글과 같은 대기업 플랫폼의 데이터 센터가 아니라 개인이 소지하는 IoT 기기를 통한 학습과 데이터 분석을 가능케 하는 도구이다. 지금은 개인의 디바이스에 부착된 센서에서 수집된 데이터를 일률적으로 중앙 데이터센터로 보낸 뒤 거기서 학습된 결과를 돌려보내는 형식을 취하는데, 이는 전기나 네트워크를 과도하게 소모시키는 비효율적 구조일뿐더러 해킹의 위험성도 높아지게 마련이다. 이러한 문제를 해결하는 방편으로 개발된 국지적(local) 인공지능 방법론이 바로 연합학습이다. 개별적 디바이스에서 1차적으로 학습이 이루어지고, 훈련 데이터 대신 학습된 모델의 데이터만을 암호화한 뒤 데이터센터로 보내고 중앙에서는 그러한 모델 데이터를 다시 통합하여 학습하는 방식을 취하게 된다.[57] 아게라 이 아카스는 엣지 TPU나 연합학습과 같은 기술이 개인정보에 대한 안전성은 높이면서도 딥러닝 기술 발전을 이루는 유력한 대안이라는 점을 강조하면서 발표를 마무리하였다.

---

55) 인간에게 중요한 가치를 원천적으로 설계(by design)나 초기값(by default)에 반영하는 방법론이다. 대표적 사례로 개인정보 보호에 대한 GDPR 제25조(Data protection by design and by default) 참조.

56) 아게라 이 아카스가 지적한 문제는 이른바 '애로의 불가능성 정리(Arrow's Impossibility Theorem)'이다. 이는 공동체 구성원의 선호를 통일된 사회후생함수(social welfare function)로 집합하는 과정에서 요구될 법한 몇 가지의 공정한 투표의 조건과, 비독재성(non-dictatorship)을 동시에 만족할 수 없음을 수학적으로 증명해낸 것이다. 이에 관하여는 Kenneth J. Arrow, "A Difficulty in the Concept of Social Welfare", Journal of Political Economy Vol. 58, No. 4 (1950) 참조.

57) 구글의 엣지 TPU에 대한 설명은 http://cloud.google.com/edge-tpu/?hl=ko 참조. 구글의 연합학습에 대한 설명은 http://federated.withgoogle.com/ 참조.

3. 패널토론 (Panel Discussion)

제2세션의 패널은 기조연설을 담당한 하인츠와 아게라 이 아카스 이외에 페이스북의 안드라데(Norberto Andrade), 싱가포르경영대학교의 첸(Gary Chan), 홍콩 디지털 아시아 허브의 자야람(Malavika Jayaram), 핀란드 헬싱키대학교의 즐리오베이트(Indrė Žliobaitė)로 구성되었다. 먼저 안드라데는 페이스북에서는 공정한 인공지능을 구현하기 위해 인력, 데이터, 알고리즘이라는 세 가지 핵심요소를 총체적(holistic)으로 접근하는 방법론을 취한다고 소개하였다. 인력의 경우, 코드를 작성하는 개발자, 데이터에 대한 학습자, 이러한 과정에 대한 검토자 역할을 하는 3개의 팀으로 구성된다. 검토자에는 법무팀이 포함되고 이들은 학계나 각종 민간, 공적 단체와 조율하고 그들의 견해를 엔지니어의 제작에 반영하는 가교 역할을 담당한다. 이러한 과정을 통해 앞서 아게라 이 아카스가 언급한 프라이버시를 고려한 설계(privacy by design)와 같은 일종의 컴플라이언스가 가능해지게 된다. 프로세스에서 검토되는 모든 사항은 기본적으로 문서화된다. 다만, 여기서 말하는 고려사항은 법적 의무의 준수로 한정되는 것은 아니고, 다양한 종류의 공적 요청이 반영된다. 인간이 가진 편향성이 데이터나 알고리즘으로 스며들 수 있기 때문에, 관계자들의 감수성이 공정성을 제고하는 데 중요한 역할을 한다.

데이터의 경우, 모델의 목표를 분명히 하고 그에 알맞은 데이터를 선별하는 것이 1차적 과제이다. 따라서 데이터가 문제되는 상황을 리포트하고, 문제된 데이터가 어떤 소스로부터 유입된 것인지를 확인할 수 있어야 한다. 사전에 선별된 문제되는 데이터의 종류를 라벨링하고 주석으로 기재하는 작업도 동반된다. 이와 별개로 개인정보 보호를 위한 비식별화(de-identification) 조치도 행해진다.[58] 그러나 현실적으로 데이터 단계에서 모든 편향성을 제거할 수 없으므로 나머지는 알고리즘 단계로 이어진다. 가령, 페이스북이 지난 해 연례 기술 컨퍼런스 'F8'에서 공개한 'Fairness Flow'라는 툴은 데이터 세트가 가진 특징(feature)을 기초로 하위집단을 구분하고, 집단 별 분류결과와 정확도가 어떻게 다른지를 일목요연하게 시각적으로 알려준다.[59] 다만, 안드라데는 이러한 기술적

58) 특정 개인을 알아볼 수 있거나, 다른 정보와 쉽게 결합하여 특정 개인을 알아볼 수 있는 '개인정보'(개인정보 보호법 제2조 제1호 참조)의 '식별성'을 제거하거나 완화하여 개인을 알아볼 수 없도록 하는 방법론을 뜻한다. 이에 대한 자세한 내용은, 관계부처 합동, "개인정보 비식별 조치 가이드라인 - 비식별 조치 기준 및 지원·관리체계 안내 -" (2016. 6. 30.), 3쪽 이하 참조.

59) 이에 대한 보다 자세한 설명으로는, Jerome Pesenti, "AI at F8 2018: Open frameworks and responsible development" (2018. 5. 2.) 참조.

툴만으로 모든 사회적 문제를 해소할 수는 없고, 따라서 앞서 언급한 것처럼 프로세스가 중요한 역할을 한다고 역설하였다. 안드라데는 프로세스의 참여자들이 공정성에 관한 다양한 질문에 대해 어떤 시각을 가져야 할지에 대한 모범사례(best practice)를 구축하고, 문제가 해결된 선례를 기록으로 남겨야 하며, 외부 커뮤니티와 지속적으로 협력할 필요가 있다고 강조하면서 발표를 마무리하였다.

다음 발표자인 첸은 '고용(employment)'이라는 주제에 특히 주목하여 논의를 진행하였다.[60] 우리 모두는 구직활동을 하거나, 누군가를 채용하거나, 혹은 그러한 입장에 처할 예정인 존재이다. 따라서 고용은 매우 보편적 주제이고, 여기서 발생하는 인공지능 관련 이슈는 인류의 삶에 무척이나 중대한 영향을 끼친다고 말할 수 있을 것이다. 과거 인간이 전담하던 고용 프로세스에 인공지능을 이용하였을 때의 혜택으로는 신속성, 인간이 다른 업무에 전담함으로써 얻게 되는 효율성, 인간이 가진 다양한 편향성 기타 취약한 속성을 배제할 수 있다는 등의 이점이 제시된다. 이러한 인공지능은 채용광고, 서류나 면접 전형에서의 적합도 평가, 고용된 직원에 대한 커리어 코칭 등에 광범위하게 활용될 수 있다. 하지만 앞서 다른 발표자들이 이야기한 것처럼, 인공지능 의사결정은 완벽하지 않고, 여러 가지 편향과 의도적・비의도적 차별 문제에 노출될 여지가 존재한다. 인공지능 의사결정이 채용광고 대상을 선별하는 과정에서 성차별적 이슈에 직면한 것은 널리 알려진 사례이다.[61] 이러한 문제에 대하여 보정을 가한다면 그 기준으로는 헌법, 실정법, 혹은 전기전자기술자협회(IEEE)를 비롯하여 국내외 관련 커뮤니티의 가이드라인에 제시된 윤리원칙을 참조할 수 있을 것이다.

첸은 구체적 보정 방식으로 앞서 언급한 것처럼 데이터와 알고리즘 단계별로 고용 맥락에 상응한 다음의 내용을 제시하였다. 먼저 직무기술을 포용적(inclusive) 언어로 기술할 필요가 있음을 지적하였다. 가령, 특정한 성별이 보다 능숙하다고 널리 알려진 직무기술 대신, 성별을 불문하고 범용적으로 필요한 범용기술을 기준으로 삼는 것이다. 세부 하위집단 별로 훈련 데이터의 대표성을 확보하고 사전에 검증된 부정적 속성을 제거해

60) 인공지능과 고용차별 문제를 다룬 최근 연구로, 고학수・박도현・정해빈, "인공지능과 고용차별의 법경제학: 블라인드 채용과 베일의 역설을 중심으로", 법경제학연구 제16권 제1호 (2019) 참조.

61) 고소득 직종에 관한 내용이 포함된 광고가 여성보다 남성에게 현저히 높은 비율로 노출되었다고 한다. 이에 대한 보다 자세한 내용은, Amit Datta・Michael Carl Tschantz・Anupam Datta, "Automated Experiments on Ad Privacy Settings", Proceedings on Privacy Enhancing Technologies Vol. 2015, Issue. 1 (2015) 참조.

야 한다는 점은 고용 외의 영역과도 공통된 해결책이다. 알고리즘의 측면에서는 신뢰할 수 있는 제3자(Trusted Third Party, TTP)에 의한 사후감사나, 무작위(random)의 첨가 같은 방법론이 제시되고 있다고 한다. 이와는 별개로 인공지능을 이용한 고용절차에서 불이익을 받은 사람에게는 앞서 루치가 언급한 것처럼 반사실적 설명과 같은 직관적 형태의 설명을 제공할 필요도 있다. 나아가 이와 같은 절차를 일회성으로 끝내지 않고 지속적으로 업데이트 하는 일도 중요하다.

다음 발표자인 자야람은 '편향성'이라는 개념에 대한 보다 근본적 물음을 제기하면서 발표를 시작하였다. 일반적으로 편향성이라는 개념은 개인적 인식 차원의 고정관념이나 편견과 연결되게 마련이라는 점에서 사회적이고 체계적 차원의 이해를 간과할 우려가 있고, 따라서 인종차별, 구조적 억압과 같은 보다 사회적 차원의 용어를 활용할 필요가 있다는 문제제기가 그것이다.[62] 그러한 맥락에서 자야람은 인공지능 영역에서는 기술 분야를 넘어서는 다학제적 연구가 필요하다는 점을 지적하고, 그와 같은 방식의 협업이 적절히 이루어진 것으로 보이는 연구결과를 소개하였다.

자야람이 언급한 문헌에서는 '공정성'이란 기술을 넘어선 보다 넓은 체계의 일부를 이루는 개념이므로 공정성의 구현을 오로지 기술적 관점에서만 접근하는 경우 올바른 해결책이 되기 어렵다는 것을 핵심적 주장으로 삼고 있다. 논문에서는 공정성 개념을 기술적 방식으로 추상화(abstraction)함으로써 나타나는 오류를 크게 5가지로 유형화하였다. 공정성을 모델에 구현하는 과정에서 특히 중요한 인적·물적 요소가 제대로 반영되지 못하면서 발생하는 프레이밍 함정(framing trap), 특정 맥락에서 공정하게 작동하는 인공지능이 다른 맥락에서 그렇지 않을 수 있다는 이동성 함정(portability trap), 형식화된 정의를 통해 공정성 개념의 사회적 함의를 담아내는 데 따른 난점에 관한 형식화 함정(formalism trap), 기술이 도입되었을 때의 파장을 고려하지 않거나 덜 고려함으로써 발생하는 파급효과 함정(ripple effect trap), 기술이 올바른 해법이 아닌 상황에서조차 기술만이 답이라고 맹목적으로 신뢰하는 해결책 함정(solutionism trap)이 그것이다.[63] 그밖에

62) Kinjal Dave, "Systemic Algorithmic Harms", Data & Society (2019. 5. 31.) 참조.

63) 컴퓨터 과학에서 주로 활용하는 '추상화(abstraction)' 작업은 본질적으로 사회적 맥락과 관계된 많은 정보를 소실시키는 경향이 있는데, 공정성을 달성하는 데 있어 이러한 정보가 핵심적인 경우가 많다는 문제의식으로 파악된다. Andrew D. Selbst et al., "Fairness and Abstraction in Sociotechnical Systems", Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM (2019) 참조.

도 자야람은 선진국과 개발도상국 사이의 기술격차, 인공지능과 인간의 공존이 자녀의 양육과 발달과정에 미치는 영향과 같은 다양한 문제가 공정성 논의에 포함되어야 한다는 점을 지적하면서 발표를 마무리하였다.[64]

마지막 발표자인 즐리오베이트는 작년부터 인공지능의 공정성에 관한 연구가 폭발적으로 성장하고 있는 추세에 대해 고무적으로 평가하면서, 몇 가지 생각해볼 만한 사항들을 언급하였다. 먼저 즐리오베이트는 최근 국제적으로 많은 논란을 빚은 보잉 737 맥스 사고를 예로 들었다.[65] 이 사고로부터 얻을 수 있는 교훈은, 어떤 기술 시스템을 설계할 때는 그것을 이용하는 인간의 행동에 대한 모종의 가정이 사전에 포함되게 마련이라는 것이다. 부적절한 가정이 포함된 시스템은 그로 인해 예상치 못한 사고에 처할 위험이 높아질 수 있다. 다른 한편, 인공지능이 인간보다 일관된 행동을 할 수는 있겠지만, 일관성이 객관성을 반드시 담보하지는 않는다는 데 대하여 유의할 필요가 있다고 지적하였다.

그렇다면 결국 인공지능이 내놓은 결과물에 대한 해석이 필요할 텐데, 문제는 앞서 언급한 것처럼 인간과 인공지능은 강점을 보이는 영역이 다르고 인간에 비해 인공지능이 더 우월한 영역에서는 인간에 의한 사후검증에 한계가 있을 수밖에 없게 된다. 나아가 설령 인간이 사후검증을 할 수 있는 영역에서조차 앞선 발표에서 말한 것처럼 인과관계에 의존하는 인간과 상관관계에 의존하는 인공지능 의사결정 방식의 차이를 이해하고, 편향성에 노출되게 마련인 인공지능의 의사결정의 한계를 직시하여야 한다. '미가공 데이터(raw data)'라는 용어가 형용모순이라고 말하는 까닭을 이러한 점에서 이해할 수 있을 것이다.[66] 끝으로 즐리오베이트는 사회적 가치를 기술적 제약조건으로 표현해내는 것과 같은 다학제적 접근방식의 중요성을 강조하면서 발표를 마무리하였다.

---

64) 자야람이 언급한 일련의 문제의식에 관한 문헌으로, Berkman Klein Center, "IDRC Global Symposium on AI & Inclusion Outputs" (2018) 참조.

65) 보잉 737 맥스 기종은 기존보다 더 큰 엔진을 탑재하였고 그 과정에서 발생한 미세한 불균형에 대한 조정을 MCAS라는 소프트웨어에 의존하였는데, 센서가 오작동하는 경우에 MCAS의 자동 시스템이 사고를 유발할 수 있었을 뿐더러 이와 같은 상황을 훈련받지 않은 조종사들과 소프트웨어가 지속적으로 상호 충돌하는 바람에 사고가 발생하였다고 한다. 보다 자세한 설명은, 정영훈, "[지식K] 보잉 737 MAX의 이유 있는 추락", KBS NEWS (2019. 4. 10.) 참조.

66) Lisa Gitelman (eds.), 『"Raw Data" Is an Oxymoron』, The MIT Press (2013) 참조.

▌참고문헌

1. 국내문헌

고학수・임용 엮음『데이터 오너십 : 내 정보는 누구의 것인가?』, 박영사 (2019).

김대식,『김대식의 인간 vs 기계』, 동아시아 (2016).

김도균・이상영,『법철학(개정판)』, 한국방송통신대학교 (2011).

대니얼 카너먼, 이창신 역,『생각에 관한 생각(제2판)』, 김영사 (2018).

서울대학교 법과경제연구센터,『미래를 향한 인공지능 정책: 우리는 AI를 신뢰할 수 있을까?』(2019).

서울대학교 아시아태평양법 연구소・서울대학교 법과경제연구센터,『인공지능, 알고리즘, 개인정보보호를 둘러싼 정책적 과제』(2017).

서울대학교 아시아태평양법 연구소・서울대학교 법과경제연구센터,『인공지능의 시대: 기술 발전에 따른 책임과 규제』(2018).

스튜어트 러셀・피터 노빅(류광 역),『인공지능: 현대적 접근방식(제3판)』, 제이펍 (2016).

웬델 월러치・콜린 알렌(노태복 역),『왜 로봇의 도덕인가』, 메디치미디어 (2014).

고학수 외, “해외 비식별조치 가이드라인 등에 대한 비교・분석”, 한국인터넷진흥원 (2018).

고학수 외, “프로파일링 관련 기술 동향 분석 및 개인정보 정책 방안 연구”, 한국인터넷진흥원 (2018).

고학수・박도현・정해빈, “인공지능과 고용차별의 법경제학: 블라인드 채용과 베일의 역설을 중심으로”, 법경제학연구 제16권 제1호 (2019).

고학수・정해빈・박도현, “인공지능과 차별”, 저스티스 통권 제171호 (2019).

과학기술정보통신부, “I-Korea 4.0 실현을 위한 인공지능(AI) R&D 전략” (2018. 5.).

관계부처 합동, “개인정보 비식별 조치 가이드라인 - 비식별 조치 기준 및 지원・관리체계 안내 -” (2016. 6. 30.).

김상욱, “값싼 드론, ‘막대한 피해주는 테러무기’로 거듭나”, 뉴스타운 (2019. 9. 16).

김지희 외, “인공지능과 미래사회”, 서울대학교 인공지능정책 이니셔티브 이슈페이퍼 01 (2019).

박도현, “서울대학교 인공지능 정책 이니셔티브 2018 국제학술대회 보고서”,『인공지능의 시대: 기술 발전에 따른 책임과 규제』(2018).

연합뉴스 “美교통안전위 “테슬라 자율주행 시스템, 충돌사고 책임있다”” (2019. 9. 5.).

정보문화포럼・한국정보화진흥원, “지능정보사회 윤리 가이드라인” (2018).
정영훈, “[지식K] 보잉 737 MAX의 이유 있는 추락”, KBS NEWS (2019. 4. 10.).
천현득, “인공 지능에서 인공 감정으로 - 감정을 가진 기계는 실현가능한가? -”, 철학 제131집 (2017).
한애라, “‘사법시스템과 사법환경에서의 인공지능 이용에 관한 유럽 윤리헌장’의 검토 - 민사사법절차에서의 인공지능 도입 논의와 관련하여 -”, 저스티스 통권 제172호 (2019).

2. 해외문헌

Lisa Gitelman (eds.), 『“Raw Data” Is an Oxymoron』, The MIT Press (2013).

A. M. Turing, “On Computable Numbers, with an Application to the Entscheidungsproblem”, Proceedings of the London mathematical society Series 2 Vol. 42 (1936).
A. M. Turing, “Computing Machinery and Intelligence”, Mind, New Series, Vol. 59, No. 236 (1950).
Amit Datta・Michael Carl Tschantz・Anupam Datta, “Automated Experiments on Ad Privacy Settings”, Proceedings on Privacy Enhancing Technologies Vol. 2015, Issue. 1 (2015).
Andrew D. Selbst et al., “Fairness and Abstraction in Sociotechnical Systems”, Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM (2019).
Angèle Christin, “Algorithms in Practice: Comparing Web Journalism and Criminal Justice”, Big data & Society, (2017).
Anh Nguyen・Jason Yosinski・Jeff Clune, “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”, arXiv:1412.1897v4 (2015).
Berkman Klein Center, “IDRC Global Symposium on AI & Inclusion Outputs” (2018).
Clayton M. Christensen・Michael E. Raynor・Rory McDonald, “What Is Disruptive Innovation?”, Harvard Business Review (2015. 12.).
Colin Allen・Gary Varner・Jason Zinser, “Prolegomena to Any Future Artificial Moral Agent”, Journal of Experimental and Theoretical Artificial Intelligence Vol. 12, No. 3 (2000).
Danielle Keats Citron, “Technological Due Process”, Washington University Law Review Vol. 85 (2008).

David Gunning, “Explainable Artificial Intelligence (XAI)”, Defense Advanced Research Projects Agency (DARPA) (2017).

David Silver et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search”, Nature 529 (2016).

David Silver et al., “Mastering the Game of Go Without Human Knowledge”, Nature 550 (2017).

David Silver et al., “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm”, arXiv:1712.01815 (2017).

Deirdre K. Mulligan · Kenneth A. Bamberger, “Procurement as Policy: Administrative Process For Machine Learning”, Berkeley Technology Law Journal Vol. 34 (2019).

F. Rosenblatt, “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”, Psychological Review Vol. 65, No. 6 (1958).

Future of Life Institute, “Asilomar AI Principles” (2017).

G20, “G20 Ministerial Statement on Trade and Digital Economy” (2019. 6.).

Gary Marcus, “Deep Learning: A Critical Appraisal”, arXiv:1801.00631v1 (2018).

George O. Mohler et al., “Self-exciting Point Process Modeling of Crime”, Journal of the American Statistical Association Vol. 106, Issue. 493 (2011).

Google, “Perspectives on Issues in AI Governance” (2019. 1. 22.).

High-Level Expert Group on Artificial Intelligence, “Ethics Guidelines for Trustworthy AI”, European Commission (2019. 4. 8).

Jacobellis v. Ohio, 378 U.S. 184 (1964).

Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms”, Big Data & Society (2016).

Jerome Pesenti, “AI at F8 2018: Open Frameworks and Responsible Development” (2018. 5. 2.).

Joi Ito, “What the Boston School Bus Schedule Can Teach Us About AI”, Wired (2018. 11. 5.).

Kenneth J. Arrow, “A Difficulty in the Concept of Social Welfare”, Journal of Political Economy Vol. 58, No. 4 (1950).

Kinjal Dave, “Systemic Algorithmic Harms”, Data & Society (2019. 5. 31.).

Maarten Franssen · Gert-Jan Lokhorst · Ibo van de Poel, “Philosophy of Technology”, Stanford Encyclopedia of Philosophy (2018).

Marco Tulio Ribeiro · Sameer Singh · Carlos Guestrin, " "Why Should I Trust You?": Explaining the Predictions of Any Classifier", arXiv:1602.04938v3 (2016).

Michael Punke, "Some Thoughts on Facial Recognition Legislation" (2019. 2. 7.).

Michal S. Gal, "Algorithmic Challenges to Autonomous Choice", Michigan Technology Law Review Vol. 25, Issue 1 (2018).

OECD, "Recommendation of the Council on Artificial Intelligence" (2019. 5. 22.).

Sandra Wachter · Brent Mittelstadt · Chris Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR", Harvard Journal of Law & Technology (2018).

Singapore Personal Data Protection Commission, "A Proposed Model Artificial Intelligence Governance Framework" (2019. 1.).

State of Wisconsin v. Eric L. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016).

Steven Ashley, "Using Artificial Intelligence to Spot Hospitals' Silent Killer", NOVA (2017. 10. 11.).

Sundar Pichai, "AI at Google: Our Principles" (2018. 6. 7.).

Taha A. Kass-Hout · Matt Wood, "Introducing Medical Language Processing with Amazon Comprehend Medical" (2018. 11. 27.).

Yoav Shoham et al., "Artificial Intelligence Index 2018 Annual Report" (2018).

준비한 사람들 & 도와준 사람들

서울대학교 인공지능 정책 이니셔티브 공동디렉터 : 고학수 & 임 용

행사준비 실무총괄 : 김태훈

행사준비 실무 : 정종구, 박지훈, 김혜인, 김은수

자료집 정리 및 행사보고서 작성 : 박도현

번역 및 현장 안내 : 강성구, 권유진, 김나형, 김동연, 김세영, 김시온, 김진우, 마준성, 박유준, 신진식, 윤희성, 임현서, 조원휘(서울대학교 법학전문대학원 인공지능법학회)

동영상 제작 : 이지현, 강현욱, 조영채(이상 서울대학교 자유전공학부), 변호재(서울대학교 치의학대학원 치의학과)

동시통역 : 천지은, 손선희