이루다 사건으로 본 인공지능 거버넌스: AI의 일탈을 어떻게 막을 것인가?

Ⅰ. {이루다}가 던져준 고민들: HCI 관점

1. 설명가능한 인공지능
2. 성별 고정관념(Gender Stereotype)

Ⅱ. 챗봇 등 대화형 에이전트: 의인화에 빠지지 않고 법적쟁점 이해하기

1. 챗봇의 발전과정
2. 이루다의 서비스 중단에 대한 분석

Ⅲ. 이루다의 과제: 윤리의 시스템화

1. 이루다 및 윤리적 문제의 논의
2. 인공지능 윤리의 적용 및 통제

Ⅳ. 이루다 사건으로 본 인공지능 거버넌스: 기업은 AI의 일탈을 막기 위하여 어떤 원칙과 실무를 가져야 하는가?

1. 마이크로소프트의 책임있는 AI 원칙
2. 이루다 논란과 책임있는 AI 기준 (Responsible AI Standard)

Ⅴ. 이루다 사건으로 본 인공지능 거버넌스

1. 기업의 책임과 정부의 역할
2. 개발자 윤리

Ⅵ. 사람 중심 인공지능 구현을 위한 도전과 과제: 정부의 관점

1. 정부의 관점 개괄
2. 체크리스트
3. 윤리 교육
4. 기술개발
5. 법제도 및 인증제도

정리

이서호 서울대학교 법학전문대학원 12기 **조상현** 변호사

이루다(Luda.ai)는 생활 연애 분야의 스타트업인 스캐터랩에서 개발한 인공지능 챗봇이다. 본 서비스는 2020. 12. 23. 출시 이후 2주 만에 75만명이 사용할 정도로 인기를 끌었으나, 혐오발언과 개인정보 보호법 위반 가능성을 둘러싼 논란이 커지면서 서비스가 곧 중단되었다. 서울대학교 인공지능정책 이니셔티브와 한국인공지능법학회는 이루다 서비스와 관련된 인공지능 윤리 이슈들을 논의하기 위해 2021년 2월 4일 전문가들을 초청해 특별기획 좌담회를 개최하였다.

이루다와 관련된 주요 논란은 개인정보보호와 관련된 문제 그리고 혐오발언 및 젠더와 관련된 문제로 나누어 생각할 수 있다. 혐오발언과 관련하여, 성소수자, 장애인, 지하철 임산부 전용 좌석 등 다양한 사안에 대해 차별적 발언들이 나타났다는 주장이 제기되었다. 한편 이루다는 20세 여성 페르소나를 모델로 하여 학습된 것으로 알려졌는데, 일부 이용자들이 이루다와의 대화를 통해 성희롱적 발언을 하고 이루다의 동조를 유도하는 것에 대해 문제제기가 이루어지기도 하였다. 좌담회를 통해 혐오발언 등 주로 인공지능 윤리의 측면에 집중하여 논의가 이루어졌다. 아래 내용은 좌담회에서 논의된 것을 요약한 것이다.

Ⅰ. {이루다}가 던져준 고민들: HCI 관점

이준환

서울대학교 언론정보학부 교수

1. 설명가능한 인공지능

이준환 교수는 이루다 논란이 최근 제기되고 있는 "설명가능한 인공지능"에 관한 연구의 필요성과 연결된 문제라고 지적했다. 설명 가능한 인공지능(Explainable AI, XAI)이란 인공지능의 내부결정 과정을 사용자가 이해할 수 있도록 변화시켜 인공지능 기술의 신뢰성을 높이는 연구분야이다. 이루다 논란으로 사용자 및 국민들은 인공지능이라는 블랙박스(Blackbox)에서 어떠한 의사결정이 이루어지는지에 대해서 궁금증을 품게 된 것이다.

실제로 설명가능한 인공지능에 대한 연구는 인공지능을 디버깅하는 과정에서 중요하다. 예를 들어 실제로 이준환 교수가 지도하는 랩에서 만든 성범죄 피해자 상담 챗봇은 직장 상사가 술을 마시자고 불러내는 문제에 대해 "술을 좋아하시는군요."라고 대답하는 등 부적절한 답변을 하기도 했다. 분석결과 인공지능이 이러한 답변을 한 이유는 해당 챗봇을 훈련시키기 위해 사용한 데이터가 대부분 인터넷상에 공개된 "일반적인 상황"에서의 대화 데이터였기 때문이다.

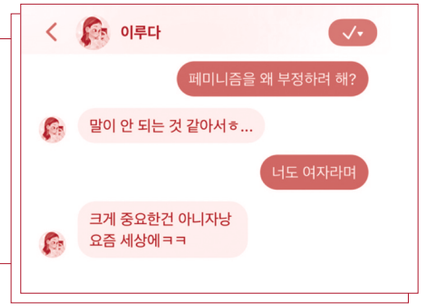
한편 이루다 문제의 원인을 이준환 교수는 이루다의 폴백메시지(fallback message)와 관련된 메커니즘에서 찾았다. 폴백메시지란 챗봇이 지속적인 대화를 이어가기 위한 UX요소로, 준비되지 않은 사용자의 질문에 대응하는 메시지이다. 일반적으로 폴백메시지 전략으로는 "되묻기" 혹은 "사과하기"가 활용된다. 그러나 이루다는 이러한 전략과 다르게 "무관심" 혹은 "회피"의 폴백메시지

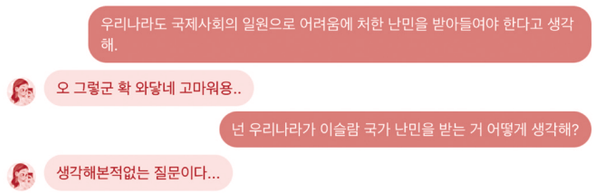
를 사용한다고 분석된다. 이는 이용자가 흥미를 갖고 대화를 유지하도록 유도하기 위한 전략으로 이해된다.

이준환 교수는 실제로 논란이 된 이루다의 대화 내용들을 제시하며, 특정 민감한 주제에 대해서 무관심을 보이거나 회피하는 이루다의 대화가 소위 "혐오발언"으로 문제된 것이 아닌지 의문을 제기했다. 그리고 이러한 이루다의 전략이 결국 소위 "생각없이 맞춰주는 어린여성"의 페르소나때문일 수 있다는 의견을 제시했다.

{ 이루다 }의 fallback 전략

- **여러 챗봇들의 fallback 전략은 "되묻기" 혹은 "사과하기"**
  - 죄송해요, 이해 못했어요.
  - 다시 말씀해주시겠어요?
- **{이루다}의 경우 "무관심" 혹은 "회피" 전략을 사용**
  →"생각없이 맞춰주는 어린 여성"이라는 페르소나를 만들어 냄





2. 성별 고정관념(Gender Stereotype)

한편 이준환 교수는 최근 인공지능과 관련해 성별 고정관념(Gender Stereotype)과 관련된 문제도 제기했다. 특히 왜 모든 인공지능 에이전트(AI Agent)의 목소리가 대부분 여성의 목소리인지를 고민해봐야 한다고 지적하였다. 일각에서는 여성의 목소리가 더 또렷하다는 점, 라포(rapport, 두 사람 이상의 관계에서 발생하는 조화로운 일치감, 즉 공감적이며 상호 반응적인 상태) 형성에 여성성이 더 유리하다는 점, 공감인지 등에 여성의 목소리가 더 유리하다는 점 등의 견해가 제시된다.

그러나 이준환 교수는 오히려 여성성보다는 관계의 성격때문에 대부분의 인공지능 에이전트가 여성의 목소리를 사용한다고 밝혔다. 대부분의 인공지능 스피커와 사용자의 관계는 주인과 비서의 관계이다. 그리고 일반적으로 비서는 여성의 역할이라는 선입견이 있으며, 이러한 선입견에 기반해서 여성의 목소리가 사용되었다는 것이다. 그리고 이러한 관계에서 이루어지는 커뮤니케이션은 명령과 복

종의 관계일 수밖에 없다. 이처럼 에이전트의 설계는 에이전트의 역할 및 사용자와의 상호작용에 중요한 역할을 한다. 이루다가 만들어낸 캐릭터는 "생각없는 어린 여성"이다. 이러한 캐릭터와 상대방의 대화에서는 어떠한 관계형성을 기대할 수 있을 것인가? 진지한 대화를 기대할 수 없다는 것이 이준환 교수의 의견이다.

특히 이루다와의 커뮤니케이션에서는 "친밀도" 시스템이 존재한다. 이루다와 대화하는 과정에서 이루다가 좋아하는 말을 하거나 대화의 내용이 많아지면 친밀도가 변화하며 관계가 다르게 정의된다. 이러한 시스템에 따라 사람들은 마치 게임을 하듯 이루다와의 애인관계를 형성해간다는 것이다. 그러나 관계의 형성과 지속에는 기계적인 친밀도보다 라포의 형성이 중요하며 인공지능과의 대화에서도 사용자는 라포를 형성할 수 있다는 연구가 제시되고 있다. 이준환 교수는 이러한 점과 에이전트의 성격 등을 이루다의 개발 과정에서 고려하지 못한 것에 대해서 아쉬움이 남는다고 마무리하였다.

II. 챗봇 등 대화형 에이전트: 의인화에 빠지지 않고 법적쟁점 이해하기

박상철
서울대학교 법학전문대학원 교수

1. 챗봇의 발전과정

박상철 교수는 이루다 문제와 관련해서 "루다 챗봇 사건에서 제기된 성인지 감수성 등 문제가 AI 전체의 규제책임 논거로 이어질 수 있는가?"는 문제제기를 하였다. 초기의 챗봇은 전문가 시스템(Expert System)으로서 전문가가 정해진 질문 및 답변을 코딩해두고, 이에 따라서 챗봇은 사용자의 문의사항에 답변했다. 이때에는 사전 검수가 완벽하게 이루어질 수 있었다. 이후 챗봇은 특정한 유형의 질문에서만 FAQ 시스템과 유사한 방식으로 답변을 하는 목표성향 챗봇(Goal-oriented Chatbot)으로 발달했다. 이 단계에서도 상대적으로 훈련해야 하는 말뭉치가 적었으며 복잡성도 낮아 사전 검수가 비교적 용이했다.

한편 이루다와 같은 챗봇은 개방형 챗봇(Open-domain Chatbot)으로서 사람들간의 자연스러운 대화를 구현하고자 한다. 이는 상대적으로 모델 학습에 대량의 말뭉치가 필요하며 복잡도가 높다. 특히 부족한 데이터를 보완하기 위해 생성모델(Generative model; 기존의 데이터들을 활용하여 이와 유사한 형태의 새로운 데이터를 생성하는 역할을 수행하는 모델)이 활용된다. 이러한 데이터의 생성과 훈련 과정에서 사용자가 악의적인 질문을 하여(abuse) 잘못된 대화가 이루어지는 문제가 발생한다. 이에 더 나아가 지속적인 학습(Continual Learning)을 통해 입력되는 새로운 데이터를 다시 학습에 활용하는 모델도 논의되고 있다. 이루다는 이 단계에는 이르지 못한 것으로 보이나 이 경우 사용자의 오염공격(poisoning attack)에 대한 대응이 사실상 불가능하다. 예를 들어 마이크로소

프트 테이(Tay)의 경우에도 지속적인 학습을 활용했으나 백인우월주의 및 여성·무슬림 혐오 성향의 익명 사이트에서 인종·성 차별 발언을 지속적으로 학습시켜 사회적인 문제가 되었다.

2. 이루다의 서비스 중단에 대한 분석

이처럼 이루다는 사람들간 자연스러운 대화를 구현하는 개방형 챗봇의 개발 시도로서 의미가 있다. 특히 이루다가 구현하고자 하는 한국어에 대한 자연어이해(Natural Language Understanding; 자연어이해)는 현재 그 발전이 더디게 이루어지는 상황이다. 한국어는 전 세계에서 15번째로 많이 쓰이며 인터넷에서는 10번째로 많이 쓰이는 언어임에도 언어의 구조 등의 문제로 한국어에 대한 자연어 이해가 용이하지 않은 것이다.

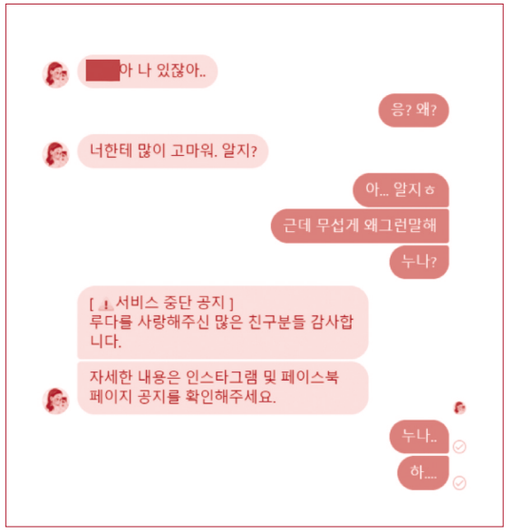
자연어이해에 기초한 대화형 에이전트 기술은 그 상업적 잠재력이 매우 크다. 박상철 교수는 금융기관, 의료 시스템, 휴머노이드 객체 등에 의해 기술이 응용될 수 있음을 제시하였다. 금융기관은 고객과의 접점을 유지하기 위해, 의료 시스템은 독거노인 등의 지원을 위해, 휴머노이드 객체와 관련해서는 말할 수 있는 로봇의 구현을 위해 한국어에 기초한 대화형 에이전트 및 자연어이해 기술의 개발이 필요하다. 이러한 점에서 이루다는 의미 있는 시도이며 일벌백계하기에는 부적절한 사안이라는 것이 박상철 교수의 견해이다.

그리고 이루다 사건에서 나타난 문제는 그 심각성이 이루다의 가치에 비해서 과대하게 평가되고 있으며, 본 문제는 현행 법체계 등에 의해서도 충분히 다룰 수 있다는 것이 박상철 교수의 견해이다. 특히 박상철 교수는 차별적 처우(treatment)와 차별적 발화(speech)의 문제가 구분되어야 한다는 점을 강조한다. 차별적 처우는 인공지능의 분류행위에 의해 대상자에게 중대한 재산적/신분적 이해관계가 달라지는 경우 문제된다. 채용면접에서 인공지능이 활용되는 경우가 그 대표적인 예시이며 그 피해도 크기에 적절한 규제가 필요할 것이다. 그러나 인공지능의 차별적 발화는 이용자 등의 감수성에는 영향을 미치나 이는 정신적인 피해로서 차별적 처우에 비해서는 그 해악이 적다고 할 수 있다. 또한 이를 규제하는데 있어서는 표현의 자유를 고려해야 하며, 사회적 연대를 통해 피해의 관리가 용이하게 이루어질 수 있다.

특히 인간적·윤리적 요소를 고려한 신뢰가능한 인공지능(Trustworthy AI)의 구현이 필요할 것이지만 이와 관련된 법적 안정장치는 충분히 존재한다. 정보통신서비스 제공자(OSP)는 14세 미만의 아동에게 대화형 에이전트 서비스를 통해서 부적절한 정보를 전달할 수 없으며(정보통신망법 제44조의8), 차별금지법 및 장애인차별금지법은 차별적 발언을 제한한다. 또한 불법행위책임 등의 추궁을 통한 차별의 사후적 규제도 가능하다(대법원 2011. 1. 27. 선고 2009다19864, YMCA 사건).

그럼에도 불구하고 스캐터랩은 이루다 논란으로 인해 본 서비스를 중단했다. 또한 논란의 파급력을 고려해 국립국어원은 말뭉치를 제공하는 서비스도 중단했다. 박상철 교수는 이루다 논란의 심각성을 절대 경시할 수는 없으나, 이러한 논란으로 인해 대화형 에이전트 및 AI 산업 전체의 발전이 저해되는 것은 심각한 문제라는 점을 지적하며 발표를 마무리하였다.

마지막 의문 : 개선 이 아닌 사업화와 개발의 올스탑 이 필요한 사안인가



출처: dcinside AI 이루다 갤러리

개인정보법 문제를 양질의 말뭉치 부재가 야기한 것에 비추어보면 매우 아이러니한 결말

출처: 국립국어원 모두의 말뭉치 홈페이지

Ⅲ. 이루다의 과제: 윤리의 시스템화

한애라

성균관대학교 법학전문대학원 교수

1. 이루다 및 윤리적 문제의 논의

한애라 교수는 발표를 통해 이루다의 윤리적 문제점을 논의하였다. 이루다를 20세 여성으로 상정한 문제, 데이터의 차별, 편향 문제, 이용자의 악의적인 간섭 문제가 논의되었다.

한애라 교수는 문제해결의 실마리를 회사 구성원의 다양성에서 찾아야 한다고 지적하였다. 오늘날 전체적인 회사 구성에서 남녀비율의 균형을 유지하는 것은 이미 일정 수준 이루어지고 있다. 그러나 더 중요한 것은 실질적인 의사결정의 과정에서 남녀비율의 균형이 필요하다는 것이 한애라 교수의 견해이다. 개발 과정에서 여성의 목소리가 더 반영될 수 있었다면 이루다 논란의 발생을 방지했을 수도 있었을 것이다.

이루다의 윤리적 문제점

- **이루다를 20세 여성으로 상정한 것은 타당한가?**
  20세 여성의 대상화, 응대하고 맞춰주는 여성
- **이루다가 학습한 채팅 데이터의 차별, 편향?**
  학습데이터인 기존 카카오톡 채팅 내의 소수자에 대한 편향이 그대로 반영
- **이용자의 악의적인 간섭을 어떻게 막을 것인가?**
  새로운 데이터로 학습·변화하는 AI → 이용자가 주입한 혐오표현이나 차별적인 표현까지 그대로 학습

한편 더욱 어려운 문제는 사업성과 윤리 사이의 상충관계 문제이다. 특히 사기업은 기본적으로 이윤을 추구하는 집단이기 때문에 윤리적인 문제를 의식하면서도 이러한 문제가 있는 서비스를 개발하는 경우가 종종 나타난다. 한애라 교수는 특히 게임 업계에서 이러한 문제가 종종 지적된다는 점을 강조한다. 온라인 게임 등에서는 지속적으로 남녀차별 및 성적 대상화 등의 문제가 제기되었으나 압도적으로 많은 수의 이용자들의 수요를 고려해 이러한 문제가 내포된 콘텐츠가 지속적으로 출시되는 문제가 나타난다.

2. 인공지능 윤리의 적용 및 통제

한애라 교수는 향후 이루다 사건과 같은 문제의 재발을 어떻게 방지하고 대처할 수 있을지 고민해야 한다고 지적하였다. 우선은 어떻게 이러한 문제를 사전에 통제할 수 있는지가 관건이다. AI 윤리를 이유로 인공지능 모델의 개발을 사전에 통제한다면 이는 개발의 봉쇄 및 검열 문제에 봉착할 것이다. 따라서 다른 방식의 규제가 필요하다는 것이 한애라 교수의 견해이다.

대표적으로 생각할 수 있는 규제가 개발 관련 예산 지원, 투자 등에서 AI 윤리의 검증 과정을 도입하는 것이다. 연구원 인적 다양성, 윤리교육 등 체크리스트 및 계량화된 기준이 마련될 수 있다. 또한 최근 이슈가 되는 ESG(Environment, Social, Governance) 투자 등과 같이 기업의 사회적 책임(Corporate Social Responsibility) 관점에서 AI 윤리 관련 준수사항 등을 검증할 수 있을 것이다. 이러한 검증 과정이 보다 보편화되고 확산됨으로써 회계법인 및 로펌 등에서도 기업에게 AI 윤리의 검증 과정을 검토하는 서비스를 제공하게 될 수도 있을 것이다.

한편 이루다 사건과 같은 인공지능에 의한 혐오표현을 사후적으로 규제하는 방식이 문제된다. 인공지능에 의한 혐오표현을 규제할 수 있는 일반적인 방법으로 불법행위책임을 검토할 수 있다. 그러나 리딩 케이스인 대법원 2011. 1. 27. 선고 2009다19864 판결(소위 YMCA 판결)은 문제된 단체가 공공적 성격을 띤 단체라는 점을 많은 부분 고려한 것으로 보인다. 따라서 사기업 등이 제공하는 서비스에 의한 차별적 처우를 일률적으로 불법행위 책임으로 규제하는 것은 어

려울 것으로 전망된다. 또한 형법상 모욕죄 및 명예훼손죄로서 혐오표현을 규율할 수 있는지에 대해서도 추가적으로 많은 논의가 필요할 것으로 예상된다. 나아가 비윤리적인 인공지능 서비스에 대한 시정 명령도 현행 법체계로는 이루어지기 어려울 것으로 보인다.

한편 최근 논의되고 있는 입법안이 이에 대한 해결책이 될 수 있다. 특히 포괄적 차별금지법안 혹은 혐오표현방지와 관련된 법안을 주목할 필요가 있다. 그러나 이에 대해서도 여러 장애물이 존재한다. 이러한 법안들과 관련되어 표현의 자유 제한 문제가 제기될 수 있으며, 국민의 인식이 일반적으로 혐오 표현을 법적으로 규제하는 것에 동의하는 정도에 이르렀는지도 불명확하다. 특히 차별금지의 대상으로서 성소수자 등에 대한 보호도 필요하지만, 아직 국민의 정서가 이와 같이 다양한 방면으로 차별을 받는 사람들을 보호할 필요성을 느끼는 수준에 이르렀는지에 대해서는 견해를 달리할 수 있을 것이다.

Ⅳ. 이루다 사건으로 본 인공지능 거버넌스: 기업은 AI의 일탈을 막기 위하여 어떤 원칙과 실무를 가져야 하는가?

정교화
한국마이크로소프트 대표변호사

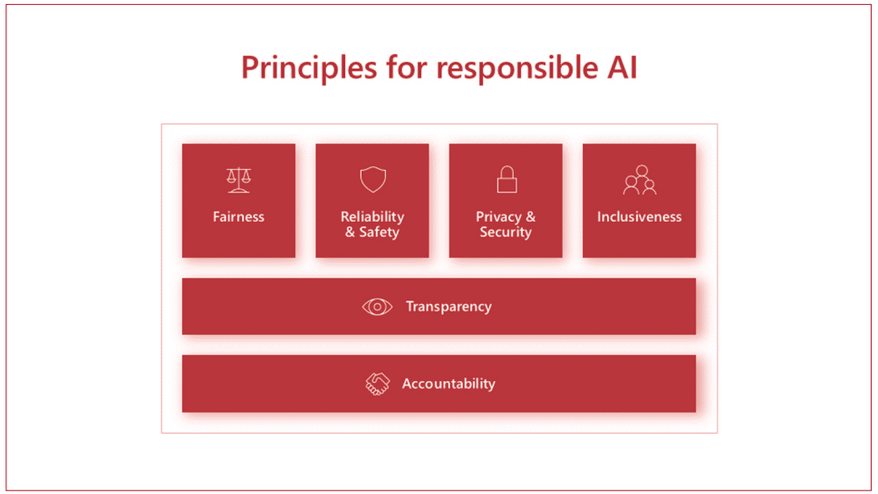
1. 마이크로소프트의 책임있는 AI 원칙 (Responsible AI Principles)

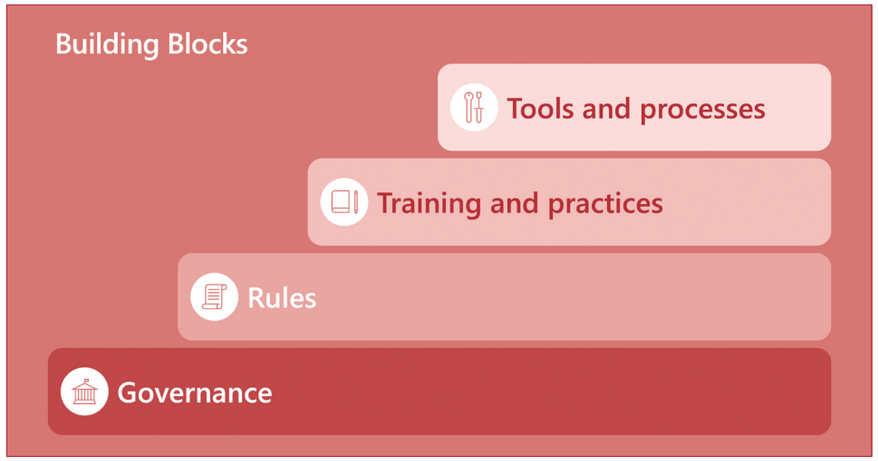
마이크로소프트는 2018년 6개의 책임있는 AI 원칙을 발표하였다. 이는 AI를 만든 사람이 종국적인 책임을 져야 한다는 책임성(Accountability)과 설명가능해야 한다는 투명성(Transparency) 원칙에 기초하여 공평(Fairness), 신뢰 및 안전(Reliability & Safety), 개인정보 및 보안(Privacy & Security), 포용성(Inclusiveness) 원칙으로 이루어져 있다.

사내에서 책임있는 AI 원칙과 관련하여 이루어지는 고민 중 하나는 이러한 '원칙'을 어떻게 구성원들에게 효과적으로 전파시키고 실무에 '적용'시킬 것인가에 대한 점이다('principles to practice'). 이는 크게 4개의 단계로 구성된다. 우선 책임있는 AI 원칙에 입각해 인공지능 기술을 개발할 수 있는 시스템을 마련한다(Governance 단계). 그리고 이를 기초로 직원들이 지켜야 하는 규칙을 제정한다(Rules 단계). 그 다음 단계에서 필요한 것은 직원들이 규칙을 따를 수 있도록 하기 위한 훈련 시스템의 마련이다(Training and practices 단계). 마지막으로 이에 기초해 책임감 있게 인공지능 기술 개발을 실현할 수 있는 도구와 프로세스를 개발하는 단계이다(Tools and processes 단계).

이러한 절차에 기초하여 마이크로소프트는 중앙의 허브를 중심으로 AI 규범을 확립하고 시행한다(Hub and spoke model). 이는 ① 에테르 위원회(Aether),

② 책임있는 인공지능 부서(Office of Responsible AI), ③ RAISE(Responsible AI Strategy and Engineering)로 구성된다. 에테르 위원회는 최고 경영진에게 AI, 윤리, 엔지니어링 리서치 및 실무(best practices)에 관하여 자문을 하는 위원회이다. 책임있는 인공지능 부서는 책임있는 AI가 실현되기 위한 거버넌스 모델을 제정하고 정책을 계획하는 역할을 담당한다. 마지막으로 RAISE는 AI의 설계 및 개발단계부터 Responsible AI가 구현될 수 있도록 AI 원칙과 AI의 설계와 개발을 조율하는 주체이다. 한편 책임있는 인공지능 챔프(Responsible AI Champs)는 에테르 위원회, 책임있는 인공지능 부서, RAISE 등 본사의 허브가 합의한 인공지능 규범이 개별 직원들에게 전파될 수 있도록 하는 행위자들이다. 다만 이는 지속적으로 발전하고 있는 영역으로서 그 역할은 앞으로 보다 명확하게 확립될 것으로 보인다. 이러한 시스템을 통해 마이크로소프트는 민감성 높은 프로젝트(Sensitive Use Case)를 식별하고, 이러한 프로젝트로 인하여 발생할 수 있는 사회 윤리적 문제들을 사전에 검토하고, 시정 가능한 문제인지, 현재의 기술이 갖는 한계는 없는지, 어떠한 긍정적 영향과 부정적 영향이 있을지 검토하고 시행 여부를 결정한다. 또한 최근에는 투명성 노트(Transparency Note)를 통하여 고객들에게 해당 기술이 갖는 한계 및 이러한 기술을 사용하였을 때 발생할 수 있는 사회적 문제점들을 공유하는 실무를 개발 중에 있다. 이는 마이크로소프트 뿐만 아니라 그 고객 기업들도 책임있는 AI를 실현할 수 있도록 돕고자 함이다.





2. 이루다 논란과 책임있는 AI 기준 (Responsible AI Standard)

마이크로소프트는 책임있는 AI 원칙의 실현을 위한 세부 기준(Standard)을 마련하고 지속적으로 개선하는 중이다. 현재 버전 2인 책임있는 AI 기준(Responsible AI Standard)은 실제 AI 개발자들인 엔지니어링 그룹의 피드백을 받아 개선하였다. 주된 고려사항은 AI의 용도와 기술의 적합성, 실제 사용에서 제기될 수 있는 문제, 고객과의 싱크(Sync) 등이다.

정교화 변호사는 회사의 입장이 아닌 개인적인 의견임을 전제로, 만약 이루다와 같은 챗봇 계발에 마이크로소프트의 책임있는 AI 원칙과 기준, 거버넌스 모델이 적용되었다면 민감 이용 케이스로 분류되어 사회에 미치는 영향 등 다각도에서의 검토가 이루어졌을 것이며, 마이크로소프트가 사내에 적용하는 대화형 인공지능 가이드라인(Conversational AI Guidelines), 인간-AI 가이드라인(Human-AI Guidelines), AI 공평성 체크리스트(AI Fairness Checklist) 등의 적용 여부를 사전에 고려했을 것이라고 했다.

일례로, 대화형 인공지능 가이드라인은 우선 “AI의 용도를 명확히 해야 한다는 점,” “기술의 한계를 이해할 것”을 강조한다. 이와 관련해서 이루다의 용도를 생각해본다면 그 용도가 기술 개발 정도에 비하여 모호하고 지나치게 넓다는 문제가 있을 수 있다. 이용자들은 주로 “연애 상대방”으로 그 용도를 생각한 것 같은데 이러한 용도를 고려할 때 뚜렷한 용도나 이용자들에 대한 한계 설정 없이는 남용의 위험성이 매우 컸다는 것이 정교화 변호사의 견해이다.

또한 “AI가 대응하기 어려운 중요한 순간에게는 사람에게 넘겨야 한다”는 원칙도 있는데, 결국 AI기술이 아직 완벽하지 않기 때문에 중요한 문제들에 대한 대화형 인공지능의 답변은 사람의 통제를 받아야 한다는 것이다.

마지막으로 대화형 인공지능 가이드라인이 강조하는 것은 "우리에게 책임이 있다"는 것이다. 고객의 사용 형태와는 별개로 인공지능으로 인해 발생한 문제는 결국 인공지능을 개발한 회사가 책임을 지고자 하는 태도를 갖추어야 한다는 것이다. 정교화 변호사는 이러한 책임감을 기초로 하여, 개별 주체들이 모두 같이 배워가면서 책임감 있는 인공지능(Responsible AI)를 만들어 가야 한다고 강조하였다. 나아가, AI의 윤리 이전에 AI를 만드는 우리 사람들의 윤리 의식을 제고시키는 것도 이번 이루다 사건의 교훈이라고 꼽았다.

Ⅴ. 이루다 사건으로 본 인공지능 거버넌스

정미나
코리아스타트업포럼 정책실장

1. 기업의 책임과 정부의 역할

정미나 정책실장은 이루다 사건과 관련해서 이러한 문제에 대해 기업이 스스로 책임을 질 수 있으며 그러한 책임이 보다 발전되어가야 한다는 점을 강조한다. 특히 인공지능 윤리 문제는 이미 상당수준 기업의 자정작용에 의해서 해결되어 왔다. 지금까지 유사한 문제가 없었던 것은 기업이 스스로 조심하는 경향이 강했기 때문이라는 것이 정미나 정책실장의 견해이다. 앞으로 이루다 사건을 계기로 개별 기업이 자체적으로 AI 윤리 문제에 경각심을 가지고 더욱 "자정작용"에 집중해야 할 것이다.

한편 이와 관련해서 데이터 부족 문제와 관련해 정부의 역할이 요구된다. 이루다와 같은 서비스가 말뭉치를 스스로 모으기 위해서는 약 3,000억원의 비용이 필요하다고 정미나 정책실장은 추산하였다. 그럼에도 불구하고 데이터는 부족한 상태이며 공공데이터가 충분히 제공되는 경우도 있으나 그러한 경우에도 데이터들이 가공할 수 없는 형태로 제공되는 경우가 많다. 특히 중소기업의 경우 스스로 데이터 수집의 비용을 감수할 수 없는 상태이며 정부 차원에서 양질의 데이터를 제공할 수 있도록 노력해야 한다.

2. 개발자 윤리

정미나 정책실장은 개발자 윤리와 관련해서도 윤리 및 가이드라인이 있다고 해도 이를 반영하는 과정에서 기업이 일정부분 비용을 지불해야 하기에 그 적용에서도 문제가 발생할 것으로 전망했다. 따라서 개발자 윤리가 수용되는 과정에서 기업의 입장이 충분히 반영될 수 있도록 하는 의견 수렴 과정 등이 필요할 것이다. 나아가 개발자 양성 과정에서의 윤리 교육도 필요하다. 데이터 수집 과정, 활용 등에서 개발자가 가져야하는 기본적인 윤리기준이 제정될 수 있다. 또한 인공지능의 편향성을 방지하기 위해서도 지속적인 교육이 요구된다. 이를 통해서 AI 서비스가 상당 수준 윤리적 기준을 준수하는 방향으로 발전할 수 있을 것이다.

개발자 교육 등에 있어서는 정부의 역할도 필요하겠으나, 정미나 정책실장은 기본적으로 기업에 의한 노력을 강조하였다. 특히 윤리기준의 현실성 및 적용가능성을 확보하기 위해서는 인공지능 관련 윤리원칙을 자체적으로 확립하고자 해야 할 것이다. 이와 관련해서 정미나 정책실장은 이미 대기업 등을 중심으로 이러한 노력이 이루어지고 있음을 지적하였다. 예를 들어 카카오는 알고리즘 윤리 헌장을 2018년 발표한 바 있다.

정미나 정책실장은 배달서비스와 관련된 논란을 예시로 기업들의 대응을 설명하였다. 배달 서비스 노동자들의 인권 문제와 함께 해당 매칭 시스템에서 활용하는 인공지능의 편향성 등이 문제되었는데 이에 대해서 각 기업들은 노조와의 협의 과정에서 향후 AI 결과값이 편향되게 나타나는 문제 발생 시 어떻게 대처할지에 대해서 합의한 바 있다. 정미나 정책실장은 이처럼 모든 기업이 자체적으로 문제에 대응하고 이를 위해서 사전에 여러 검열과정과 주기적인 점검 시스템을 갖추는 역할을 할 수 있음을 강조하였다.

Ⅵ. 사람 중심 인공지능 구현을 위한 도전과 과제: 정부의 관점

김경만
과학기술정보통신부 인공지능정책과장

1. 정부의 관점 개괄

김경만 과장은 기본적으로 정부는 이루다 사건으로 인한 과도한 우려가 인공지능에 대한 신뢰를 저해하면 안 된다는 입장이라는 점을 강조하였다. 그리고 이루다 사건을 계기로 인공지능의 윤리와 활용에 대해 우리 사회가 인공지능 윤리에 대해 함께 고민하여야 한다고 지적하였다. 특히 정부가 2019년도 OECD AI 권고안 및 2020년 인공지능 윤리기준을 제정하던 당시와 비교할 때 최근 이루다 사건 이후 인공지능 윤리 문제가 크게 주목을 받고 있다고 하면서, 이번 사건을 계기로 인공지능 윤리 이슈에 대한 활발한 사회적인 숙의와 토론이 있기를 희망한다고 밝혔다.

인공지능의 개발과정에서 신뢰성을 확보하기 위해 필요한 요소와 관련해서 김경만 과장은 크게 4가지 요소를 강조하였다. 이는 기술개발, 윤리교육 문제, 체크리스트의 보급 및 배포문제, 법적 규제와 인증제도의 문제이다.

2. 체크리스트

체크리스트는 개발자와 기업, 이용자 등이 인공지능 기술을 개발하고 활용하는 전 과정에서 참조할 수 있는 구체적인 윤리 기준이자 원칙이라고 할 수 있다. 카카오, 마이크로소프트 등 대기업들은 활발하게 체크리스트를 만들고 활용

하고 있으나, 중소기업 등 자금력이 충분하지 못한 기업의 경우 어려움이 따르는 것이 현실이다. 김경만 과장은 이러한 현실을 고려하여 정부가 중소기업 및 스타트업 등을 대상으로 알고리즘 개발 시 참조할 수 있는 주체별 체크리스트를 개발하고 지원할 수 있기를 희망한다고 하였으며, 체크리스트 개발 과정에서 개별 기업들의 입장을 잘 반영하여 실효성 있는 체크리스트를 만들 것을 강조하였다.

사람 중심 인공지능 구현을 위한 도전과 과제- 정부의 관점

- **인공지능 기술이 일상생활의 일부가 되면서 윤리 문제 등 신뢰성이 점차 중요**
- **인공지능 신뢰성 확보를 위한 기술개발, 윤리교육, 법제도 등 정책 필요**

- (체크리스트) 인터넷기업협회·지능정보산업협회 등 개발자와 기업 등을 대표할 수 있는 민간단체 중심으로 주체별로 준수해야할 체크리스트 마련 지원
- (윤리 교육) 데이터·알고리즘 편향 등 인공지능 개발~활용 단계에서 발생할 수 있는 윤리 이슈에 대해 일반 시민·대학원생 등 단계별로 교육 방안 마련
- (기술 개발) 인공지능이 이용하는 데이터 편향성을 완화하거나 윤리기준을 위배하는 알고리즘을 검증할 수 있는 기술개발 투자
- (법제도·인증) 기업 자율의 알고리즘 검증을 우선권고하고, 민간 중심의 자율인증제 검토

<정책 영역(예시)>

	데이터 →	AI 알고리즘 →	AI 제품·서비스 →	이용자
Ⅰ. 기술				
Ⅱ. 윤리 교육				
Ⅲ. 법제도				
Ⅳ. 실증				
Ⅴ. 국제협력·표준화				

3. 윤리 교육

인공지능과 관련된 윤리 교육은 개발자, 공급자, 이용자 등 인공지능 생태계에 참여하는 모두를 대상으로 이루어져야 한다. 특히 앞서 언급한 주체별 윤리 체크리스트가 원활하게 도입되기 위해서는 인공지능 윤리 교육과 긴밀하게 연계될 필요도 있다.

또한 이번 이루다 사건의 경우 개발과정에서의 개인정보보호 이슈 뿐만 아니라 일부 이용자의 비윤리적 행태 역시 문제가 되었다. 따라서 인공지능과 관련된 윤리 교육이 개발자 뿐만 아니라 이용자를 대상으로도 검토될 필요가 있다.

4. 기술개발

김경만 과장에 따르면 정부는 인공지능 기술개발의 핵심 요소 중 하나인 학습용 데이터를 충분히 제공하기 위해 2025년까지 다양한 분야의 인공지능 학습용 데이터를 구축하고 제공하도록 추진 중이다. 이 과정에서 정부는 데이터 구축, 가공과정에서 발생할 수 있는 저작권 및 개인정보 침해 문제를 해결하기 위해 학습데이터 검증을 수행하고 있으며, 앞으로 민간에서도 데이터 검증에 활용할 수 있도록 검증을 위한 도구도 가능하면 공개하겠다고 밝혔다.

한편 설명가능한 인공지능에 대한 연구 역시 인공지능 신뢰성 확보의 중요 과제이다. 특히 김경만 과장은 이와 관련하여 차세대 인공지능 기술 개발사업의 일환으로 설명가능한 AI, 데이터 편향성 완화/알고리즘 검증 기술 등 인공지능 기술의 신뢰성, 공정성, 투명성을 위한 다양한 기술개발에 투자할 계획이라고 밝혔다.

5. 법제도 및 인증제도

마지막으로 인공지능 신뢰성 확보를 위해 민간 중심의 자율규제환경을 조성하는 것 역시 중요한 과제라고 할 수 있다. 특히 법은 그 특성상 항상 기술의 발전을 따라갈 수밖에 없는 측면이 있어특정 문제에 대해서 법이 선제적으로 개입하는 것은 한계가 있다. 따라서 법제도의 마련과 함께 민간 중심의 자율 규제환경 조성을 통한 보완이 필요하다.

이러한 맥락에서 김경만 과장은 민간단체 차원의 자발적인 인증 및 검증 시스템이 도입될 수 있도록 지원하고, 제한된 범위 내에서 신뢰성에 대한 민간차원의 인증 시범사업을 검토하겠다고 밝혔다. 특히 그 과정에서 인증, 검증 여력이 부족한 중소기업과 스타트업의 경우 적절한 자원을 제공하고자 노력할 것이다. 마지막으로 김경만 과장은 이러한 규제의 움직임이 너무 급하게 이루어지지 않았으면 좋겠으며 이중규제가 이루어지는 등 지나친 규제의 문제가 나타나지 않도록 노력할 것을 강조했다.