[칼럼이 있는 AI톡]⑪ 인공지능윤리, 그 잠재성의 중심

박성은 기자 | 승인 2021.10.05 18:37

"칼럼이 있는 AI톡"
중앙대 인문콘텐츠연구소 & AI타임스 공동기획

[편집자주] 인공지능(AI)이 우리 일상 생활 속으로 점점 더 가깝게 다가오고 있습니다. 인공지능의 좋은 점과 나쁜 점에 대해서 설왕설래합니다. 많은 논의들이 진행되고 있습니다. 무릇 결론은 '사람이 중요하다'는 점입니다.
때문에 AI기술 중심으로 움직이는 현실에서 그 속에 있는 인간과 문화, 철학과 예술에 대한 논의를 일상의 눈높이에서 해보고자 합니다. 때로는 AI에 대한 사색을, 때로는 AI 도입으로 바뀌는 삶에 대해 생각하는 재료를 만들어 선보이겠습니다. 이번 특별기획은 중앙대 인문콘텐츠 연구소와 AI타임스가 공동으로 기획하고 진행합니다.

[글 싣는 순서]

토크 포인트(Talk Points)

올해 1월 우리나라에서는 AI 기술만큼, 아니 그보다 더 AI 윤리에 관심이 쏠렸다. 관심이 집중됐던 AI 챗봇 이루다 서비스가 출시와 동시에 중단됐기 때문.

AI 챗봇 이루다와 관련된 각종 문제를 관통하는 키워드는 AI 윤리다. 이루다는 소수자를 차별하는 가해자인 동시에 성희롱 피해자였다. 하지만 현행법으로 처벌 혹은 피해 대상이 될 수 없는 AI다. 문제는 극도로 복잡해졌다.

이루다 효과로 우리나라 정부와 기업은 AI 개발에 필요한 윤리 철칙을 누구보다 빠르게 마련했다. 하지만 아직 갈 길은 멀다. 다양한 형태의 AI를 모두 포용하는 윤리 원칙이 가능할까? 가이드라인을 넘어 법제도가 필요한 부분은 없을까? 요즘 윤리학자들이 바쁜 이유다.

[칼럼] 인공지능윤리, 그 잠재성의 중심

문규민 중앙대 인문콘텐츠연구소 HK연구교수

어떤 문제건 주어진 문제를 해결하는 것만큼이나 그 문제의 중요성을 명확히 파악하고 음미하는 것 또한 중요하다. 누구도 현재 인공지능이 중요한 문제라는 사실을 부정하지 않을 것이다. 그런데 인공지능은 왜 중요한가? 그것은 정확히 누구에게 어떤 문제들을 제기하고 있는가? 크게 보면 이들이 바로 인공지능윤리(AI ethics)가 다루고자 하는 질문이라고 할 수 있을 것이다.

인공지능윤리가 부상하게 된 배경에는 이제 더 이상 '비인간 자율성'(non-human autonomy)에 대한 도덕적, 윤리적 접근을 미룰 수 없게 되었다는 성찰과 각성이 있다. 일부 인공지능 개발자와 연구자들은 현재의 수준으로는 인간과 같이 사고하는 소위 "강한 인공지능"은 고사하고, 제한된 영역에서 "약한 인공지능"이 보여주는 수행능력을 끌어올리는데도 아직 갈 길이 멀다고 전망하기도 한다. 현재 통용되는 인공지능은 자율성을 갖춘 시스템이 아니라 사실상 세련되고 효율적인 자동화 시스템(automatized system)이라는 것이다. 이와 같은 평가가 옳다 하더라도, 인공지능윤리은 그 초점을 아직까지는 자동화보다는 자율성에 두고있다.

물론 여기서 말하는 자율성은 인간에게 흔히 부여되어온 자율성, 남의 영향에서 비교적 자유로운 상태에서 스스로의 정신을 통해 사고하고 판단하는 그런 절대적 자율성이 아니라, 인간의 입장에서 쉽게 예측하기 어렵고 통제가 점점 힘들어진다는 의미에서의 상대적 자율성일 것이다. 이런 자율성은 어쩌면 고도의 자동화와 잘 식별되지 않을 수 있지만, 어쨌든 현재 인공지능이 이런 상대적인 의미의 자율성을 획득해가고 있는 것은 사실이다.

이 점에서 인공지능윤리는 이전의 다른 응용 윤리 분야들, 예컨대 동물윤리와 대조된다. 동물은 어느 정도의 자율성을 분명히 가지는 것 같은데, 문제는 특히 가축의 경우 이미 기나긴 역사를 통해 인간의 통제 아래 있게 되면서 쉽게 예측가능하고 통제가능해졌다는 사실이다. 이는 빠르게 진보하면서 인간의 예측과 통제를 서서히 벗어나거나 혹은 그럴지 모른다는 불안을 야기하는 인공지능과는 사뭇 다르다. 문제시되는 비인간 자율성의 성격이 다르므로, 두 윤리의 내용도 달라질 수밖에 없다.

인공지능의 자율성은 윤리와 기술, 윤리와 산업의 관계 또한 변화시키고 있다. 인공지능은 이미 인간의 다양한 삶의 영역으로 깊숙이 들어와 있으며, 따라서 인공지능의 개발 문제는 우리의 일상생활에 직접적이고 즉각적인 영향을 끼칠 수 있다. 작게는 사적 정보의 유출부터 크게는 거대 시설에서 일어나는 사고까지, 인공지능이 야기할 수 있는 피해는 다양하다. 이 때문에 인공지능은 그 개발 단계부터 윤리적 고려가 개입하게 된다. 과학기술에 대한 윤리적 고려는 종종 해당 분야의 종사자들에게 연구와 개발의 방해물처럼 여겨지곤 했는데, 인공지능에서는 오히려 개발과 설계에서부터 윤리적 고려가 적극적으로 요청되고 있는 것이다. 아무도 윤리적 선택이 필요한 상황에 제대로 대처하지 못하는 자율주행자동차를 타려고 하지 않을 것이다. 누구도 걸핏하면 도덕적으로 의심스러운 언어표현을 구사하는 앱과 채팅을 하려고 하지 않을 것이다.

인공지능이 도덕적인지가 그것의 상품화 가능성에 결정적인 고려사항이 되는 것이다. 윤리적 접근과 현장에서의 인식 사이에는 자주 갈등이 있어왔는데, 인공지능윤리에서는 갈등보다는 협력이 이루어지고 있다. 비인간 자율성은 학술적 담론을 넘어 이미 긴급한 삶의 문제가 되었고, 따라서 학자와 연구자들은 물론 실제 개발을 맡고 있는 현장의 개발자와 사업가, 정책을 만들고 실행할 법조계, 무엇보다 실질적인 인공지능의 영향력에 노출된 시민들 사이의 적극적이고 진지한 소통과 협력을 요구하고 있다.

나아가 인공지능이라는 현상은 그 자체로 융합적 연구를 강제하는 측면이 있다. 실제로 인공지능윤리는 여러 응용윤리분야들 중에서도 학제간 연구, 융합연구가 가장 활발한 축에 속한다. 전문윤리학자는 물론 인공지능 개발자부터 법학자, 사회학자, 문화연구자 등이 앞다투어 인공지능윤리에 뛰어들고 있는 것이다. 이 또한 통제되지 않을 경우 큰 위험을 초래할 수 있다는 비인간 자율성의 특성에서 기인한다고 할 수 있다.

인공지능윤리가 다루는 주제들은 다양하다. 해결이 시급한 현안으로는 인공지능 개발 윤리가 있다. 데미스 하사비스와 일론 머스크 등이 2017년에 발표한 아실로마 인공지능 원칙(Asilomar AI Principles) 같은 것이 대표적인 사례가 될 것이다. 우리는 인공지능을 개발할 때도 지킬 것은 지켜야 한다. 그런데 무엇을 지켜야 하며 왜 지켜야 하는가? 인공지능 관련 법안과 윤리 수칙 제정 또한 인공지능윤리의 주요 주제

다. 예컨대 지난 4월 21일 유럽연합에서 공표한 입법 초안의 내용과 근거는 무엇인가? 문제는 없는가? 있다면 어떤 대안이 있을 수 있는가? 이들은 인공지능에 대해 인간이 무엇을 할 것인가, 또는 무엇을 해야 하는가와 관련된 문제들이다.

그러나 인공지능윤리가 다루는 범위는 그보다 훨씬 넓고 다양하다. 가령 인공지능을 어떻게 도덕적으로 만들 것인가? 이것이 인공도덕행위자(Artificial Moral Agent, AMA)의 문제다. 이런 문제들은 도덕이나 윤리에 대해 근본적인 질문들을 다시 불러일으킨다. 로봇이나 인공지능을 비롯한 인공물이 윤리적으로 행위한다는 게 도대체 무슨 뜻인가? 인공물이 도덕적 행위자가 된다는 것이 기술적인 수준에서, 또는 원칙적인 수준에서 가능하긴 한가? 인공물이 도덕적 행위자라면, 그것은 자신의 행위에 책임을 질 수 있는가? 인공물에게 책임을 묻는 것이 가당키나 한가? 몇몇 질문은 SF 소설 같은 이야기처럼 들릴 수도 있지만, 이들은 그 자체로 심원한 문제일 뿐 아니라 앞으로 직면할 가능성이 있는 윤리적 쟁점들이다. 이런 질문들에 대답하려는 노력은 인공지능의 활용은 물론 개발에 있어 원칙적인 수준의 지침을 줄 수도 있을 것이다.

언급된 문제들은 파고들다 보면, 자주 윤리학은 물론 행위 이론이나 인식론, 형이상학과 인간학 등에서 유지되어오던 직관들을 심각하게 재고하게 된다. 이런 점에서 인공지능윤리는 주어진 개념의 내용을 분석하는 개념분석(conceptual analysis)을 넘어서, 새로운 개념을 창조하는 개념공학(conceptual engineering)을 시도하기를 자극하고 있다고 볼 수도 있다. 인공지능윤리는 음미되고 해결되어야 할 문제들의 집합이자, 동시에 인간중심의 관념과 사고를 뒤흔들 수 있는 잠재성의 중심인 것이다.

비하인드 인터뷰

칼럼을 읽은 후 칼럼니스트에게 질문 혹은 반문하는 것은 다소 귀찮거나 힘든 일이다. 독자를 대신해 AI타임스가 여전히 남은 궁금증을 풀어봤다. 조금은 매울지도.

문규민 중앙대 인문콘텐츠연구소 HK연구교수

Q. AI로 인해 발생할 수 있는 윤리 문제, 대표적으로 어떤 것들이 있을까?

국내에서는 단연 AI 챗봇 이루다 사건을 들 수 있다. 전세계적으로는 2010년대 후반부터 구글을 비롯한 빅테크 기업들이 만든 AI 챗봇에서 인종 차별, 혐오 발언 문제가 나오기 시작했다. 이외 자율주행 분야에서 AI 윤리가 중요하게 언급된다.

Q. 자율주행에서는 주로 어떤 AI 윤리 문제를 논의하나?

하나를 지키기 위해 다른 것을 버려야 하는 경우가 있다. 예를 들어 길가에 갑자기 나타난 사람을 자율주행차가 피하기 위해 크고 작은 교통법규를 어겨야 하는 상황이 발생할 수 있다. 이 때 AI가 어떻게 판단하도록 프로그래밍할지 정해야 한다.

Q. 당장 시급하게 마련해야하는 자율주행 법제도가 있다면?

자율주행차와 인간 운전자가 핸들을 공유하는 레벨3 이후 단계에서 언제 어떤 상황에서 운전권한을 인간에게 넘길지 결정하는 일이다. 조금이라도 민감하거나 위험 가능성이 있는 상황에 처할 때마다 자율주행차가 인간에게

운전권을 넘긴다면 자율주행 의미가 퇴색될뿐더러 안전하지 않을 수 있다. 최근 혼다에서 레벨3 자율주행차를 출시하겠다고 발표한 만큼 급한 문제인데 공론화가 많이 되지 않았다.

Q. 엔터테인먼트, 자율주행, 의료 등 분야별로 AI 윤리 중요도가 다를 것 같은데?

분야별로 자주 제기되는 문제 종류는 있겠지만 중요도의 경우 건마다 달라진다. 예를 들어 엔터테인먼트 분야에서는 메타 휴먼의 '인권'을 중요하게 논의할 수 있다. 고인의 목소리나 영상을 재현하는 AI에도 윤리 문제가 있을 수 있다. 고인이나 유가족에게는 잊혀질 권리가 있기 때문이다.

Q. AI 자체에 윤리 문제에 대한 책임을 물을 수도 있을까?

자율성이 아주 높아질 경우에는 가능할 수도 있겠다. 여기서 말하는 자율성은 외부 개입 없이 스스로 작동하는 수준을 훨씬 능가하는 정도다. 현재 AI 수준은 소극적 자율성에도 도달하지 못한 만큼 지금 논의하기에는 시기상조인 주제다. 다만 철학 문제 자체로는 성립될 수 있다. 안전 무해한 AI를 개발하는 일 이외에 근본적인 질문을 던지고 새로운 사고를 가능하게 하는 것도 AI 윤리 역할이다.

Q. AI 챗봇 이루다를 대상으로 성희롱한 사건이 있었다. 일각에서는 피해 대상이 생명체가 아니니 문제없다고 주장했는데 어떻게 생각하나?

행위 하나만 놓고 보면 실제적으로 피해를 보는 주체가 없으니 괜찮다고 볼 수 있다. 하지만 문제는 이러한 행위가 퍼져나가 실제 현실에 영향을 미쳤을 때다. 골방에서 혼자 즐기는 것에서 나아가 성희롱 행위가 확산되면 문제가 될 수 있다.

Q. 대부분 AI, 특히 챗봇에서는 데이터 편향을 윤리 문제 주범으로 이야기한다. 윤리적인 데이터는 누가 어떻게 만들 수 있나?

언어학자, 윤리학자, 프로그래머, 개발자가 협동 연구를 해야 가능하다. AI 윤리는 다학제적인 연구가 될 수밖에 없다. 말 자체가 어떤 윤리적 효과를 내는지는 알고리즘이 아닌 도덕적 수사학 문제다.

Q. 윤리적인 데이터 마련에만 비용이 꽤 들어갈 것 같은데, 기업에서 자발적으로 참여할까?

윤리적인 데이터 작업을 누가 어떤 방식으로 하는지에 대해 법제도로 정해야 한다. 제약을 걸지 않은 상황에서 알아서 윤리적인 데이터를 선별해 사용하는 기업은 거의 없을 것이다.

Q. 최근 우리나라 정부와 기업들이 앞다퉈 AI 윤리 철칙을 공개하고 있다. 강제성이 없는 가이드라인 수준인데 효력이 있을까?

확실히 효과가 있을 것이라고 본다. 윤리적인 문제가 곧 기업 활동에 직접적인 영향을 주기 때문이다. 시장 내 소비자들의 기준이 높다. 인종차별적 발언을 하는 AI 챗봇은 소비자들이 사용하지 않는다.

Q. AI 윤리 대책, 앞으로 어떻게 만들어나가야 할까?

단 하나의 AI는 없다. 추상적인 가이드라인 하나만으로는 실효성이 없다는 의미다. 일상 챗봇, 자율주행, 메타휴먼, 딥페이크 등 분야별로 나눠서 AI 개발 윤리를 바라봐야 한다. 보다 넓은 관점에서는 AI 윤리보다는 윤리 자체에 초점을 맞추는 것이 중요하겠다.

문규민 교수는 경희대학교 철학과를 졸업하고 서울대학교에서 의식과 양상에 대한 논문으로 박사학위를 받았다.

충북대, 고려대, 서울시립대등에서 연구하고 가르쳤다. 현재 중앙대학교 인문콘텐츠 연구소 HK+연구교수로 재직 중이며, 인공지능의 철학과 윤리학, 의식과 주관성을 비롯한 여러 주제와 관련된 연구를 하고 있다.

AI타임스 박성은 기자 sage@aitimes.com

박성은 기자 sage@aitimes.com