신뢰할 수 있는 인공지능

"인간이 느끼는 가장 강력하고 오래된 공포는 미지에 대한 공포다." 영국의 호러소설 작가, 하워드 필립 러브크래프트가 남긴 말이다. 사람을 포함한 모든 동물은 자신이 알 수 없는 것에 본능적으로 공포를 느낀다고 한다. 생명이 오랜 시간 동안 진화하면서 축적해 온, 혹시 모를 위험을 피하려는 생존 메커니즘이다.

그래서 생소한 상황에 맞닥뜨리면 사람들은 늘 경계하고 두려워한다. 상대방이 능동적일수록 더 그렇다. 이러한 두려움을 해소하기 위해 인류는 '모르는 것'을 '아는 것'으로 바꾸어서 타인에 대한 벽을 낮추고자 예절과 문화를 만들었다. 이처럼 일정한 프로토콜을 공유함으로써 상대방을 알고 나면 신뢰가 가능해진다.

4차산업혁명 시대를 맞아 인공지능(AI)이 널리 사용되면서 비슷한 문제가 나타났다. 심층신경망에 의한 머신러닝을 핵심 메커니즘으로 활용하는 현대의 AI는 종종 '블랙박스'로 표현되곤 한다. 심층신경망의 특성상 입출력 구조를 구체적으로 예측하거나 분석하기 어려워서 AI의 판단 근거가 불투명하기 때문이다. 자연히 대중은 물론, 연구자 사이에서도 'AI의 판단을 과연 믿을 수 있는가'라는 문제가 제기되기 시작했다.

AI의 알고리즘 자체의 불투명성과 함께 데이터 역시 중요한 이슈로 떠올랐다. 수집되는 데이터의 양이 급증하면서 딥페이크와 같은 가짜 데이터나 적극적으로 속일 목적으로 만든 공격용 데이터, 기계나 시스템의 한계로 불완전하게 확보된 데이터처럼 AI의 판단에 악영향을 줄 수 있는 데이터가 학습체계에 유입될 가능성이 커지고 있다.

따라서 AI를 유용하게 활용하려면 예측불가능성을 최소화하는 한편, AI가 누구나 수용할 수 있는 결과값을 도출해야만 한다. 이러한 원칙을 구체적으로 표현한 개념이 바로 AI의 신뢰성(trustworthiness)이다. 앞서 비유했듯 AI의 신뢰성은 사람이 AI와 함께 일하기 위해 반드시 준수해야 하는 최소한의 규범이라고 할 수 있다. 신뢰성을 확보함으로써 AI 사용에 따른 위험요소를 줄이고 불투

명성, 편향성과 같은 기술적 한계를 해결할 수 있다.

AI 연구를 이끄는 주요국은 신뢰성 문제를 중요하게 여기고 관련 규범과 제도를 마련해 왔다. 미국은 2016년 발표된 '국가 AI R&D 전략'에서 AI 신뢰성 확보의 필요성을 제기하고 2020년 1월 연방정부 규제 가이드라인에 AI 윤리 및 안전, 신뢰성 제고를 위한 원칙을 포함시켰다. 유럽연합(EU)은 2019년 4월 '신뢰할 만한 AI 윤리 가이드라인'을 발표하고 2021년 4월 세계 최초로 AI 법안을 제안하여 고위험 AI를 체계적으로 규제하는 제도적 장치를 마련했다. 한국에서도 2020년 12월 'AI 법·제도·규제 정비 로드맵'을 통해 정부 차원에서 과제를 발굴하는 한편 'AI 윤리 기준'을 발표해 민간 중심으로 AI 윤리 기준을 정립하도록 촉진하고 있다.

학계에서는 AI 관련 기술이 빠르게 성장하기 시작한 2000년대 중반 이후 인간이 신뢰할 수 있는 인공지능의 조건을 본격적으로 탐색하기 시작했다. AI의 신뢰성은 ICT뿐 아니라 사회학부터 윤리철학까지 포괄하는 광범위한 주제인 데다 AI의 파급효과가 인류 전체에 미치기에 학계 전반의 광범위한 협력과 합의가 필요했다.

2004년 발표된 '후쿠오카 세계 로봇 선언'은 AI를 타자화하여 위험성을 과대해석하지 않고 인간의 관리 책임을 명시했다는 점에서 현대적인 AI 신뢰성 논의의 출발점이라고 할 만하다. AI의 '믿을 수 없는 행동'의 책임 소재를 기술 자체에 묻기보다 시스템을 설계하고 관리하는 인간에게 둔 것이다.

후쿠오카 선언에 나타난 문제의식은 20년 가까이 관련 논의가 활발하게 이루어지면서 2017년 발표된 '아실로마 원칙'을 통해 대중적인 호소력을 지니기 시작했다. 아실로마 원칙은 연구, 윤리와 가치, 장기적 이슈를 폭넓게 다뤘으며, AI 시대의 윤리적 논점을 공개적으로 논의함으로써 AI 기술이 인류의 공동선을 위해 활용되는 데 목표를 둔다.

AI에 대한 윤리적, 사회적 담론은 기술에도 분명한 영향을 주고 있다. 최근 AI 연구는 효율적인 학습 체계를 구축하는 수준을 넘어서서 '설명 가능한 인공지능(EXplainable Artificial Intelligence)'을 목표로 추진되고 있다. 이는 AI가 판단의 근거를 사람이 이해할 수 있는 형태로 제시해야 한다는 뜻으로, AI의 불확실성을 줄임으로써 인류 공동의 가치관에 부합하도록 이끄는 데 목표가 있다. 이미 미국 방위고등연구계획국(DARPA)을 위시한 주요 연구기관에서는 인공지능이 문제를 어떻게 해결했는지 스스로 설명하게 하는 프로젝트를 진행하고 있다.

AI 기술의 발전이 우리에게 제시하는 새로운 도전, 신뢰성의 문제를 ICT 연구에서는 어떻게 해결하고 있을까? 다양한 분야에서 펼쳐지는 '인간이 이해할 수 있는 인공지능을 위한 여정'을 TTA 저널 201호에서 살펴본다.