ICT Expert Interview

이재호 서울시립대 인공지능학과 교수

Q1. AI 신뢰성에 대한 논의가 중요한 이슈로 부상했습니다만, '신뢰성'이 의미하는 범주가 넓어 여전히 개념 정의가 쉽지 않아 보입니다. AI 연구자 입장에서 봤을 때, AI 신뢰성의 범주와 요소를 어떻게 정의할 수 있을까요?

AI 시스템은 자율적 인식, 판단, 행위적 요소를 기반으로 다양한 수준의 자율성을 갖출 수 있습니다. 일반적인 시스템 신뢰도(Reliability)는 가동률과 같은 시스템 성능 중심의 '객관적' 척도를 뜻합니다. 반면 AI 시스템 신뢰성(Trustworthiness)은 사용자 또는 이해관계자가 시스템을 신뢰하는 수준을 의미하는 '주관적' 척도라 할 수 있습니다.

따라서 신뢰성은 사용자가 주관적으로 판단하는 정당성·적합성과 같은 가치에 따라, 사용자 또는 이해관계자별로 다르게 평가·활용할 수 있습니다. 예를 들어, 개발자는 신뢰성의 기술적 요소를 활용해 제품·서비스의 성능 목표 수준을 설정해 AI 시스템을 설계할 수 있습니다. 반면 소비자는 신뢰성을 품질의 척도로 활용해 제품·서비스를 선택할 수 있는 권리를 행사할 것입니다. 비국가 단체의 경우, 제품·서비스 전 주기별 신뢰성 수준을 감시하고 요구사항을 제시해 소비자 권리를 보장하는 데 신뢰성을 활용할 수 있습니다. 국가는 인증·규정을 명문화해 책임과 보장의 범위를 정하는 데 신뢰성을 활용할 수 있습니다.

ISO/IEC JTC 1/SC 42 Artificial Intelligence에서 개발한 ISO/IEC TR 24028 Overview of trustworthiness in artificial intelligence 표준에선 신뢰성을 '이해당사자가 예상해 기대하는 바를 증명가능한 방법으로 충족시킬 수 있는 능력'으로 정의하고 있습니다. 해당 표준에 따르면 신뢰성의 특성은 신뢰도(Reliability), 가용성(Availability), 탄력성(Resilience), 보안(Security), 개인정보보호(Privacy), 안전(Safety), 책임성(Accountability), 투명성

(Transparency), 무결성(Integrity), 진정성(Authenticity), 품질(Quality), 사용성(Usability) 등입니다.

반면 NIST(미국 국립표준기술연구소, National Institute of Standards and Technology)는 '신뢰할 수 있고 책임감 있는 AI(Trustworthy and Responsible AI)'의 필수 구성요소로서, 유효성 및 신뢰성, 안전성, 보안성 및 복원력, 책임성 및 투명성, 설명가능성 및 해석가능성, 개인정보보호, 유해한 편견의 완화를 통한 공정성을 제시하고 있습니다.

02. AI 신뢰성 논의의 주요 요소 중 하나로 '학습 과정의 블랙박스화'가 꼽힙니다. 의사결정 과정을 구체적으로 추적하거나 재구성하기 어려운 상태에서 AI의 결론을 믿어도 좋을지 판단하기 어렵다는 지적입니다. 이와 관련해 학습과정을 투명하게 공개하는 방법으론 어떤 연구와 논의가 있을까요?

신뢰성의 대칭적 개념인 위험성(Risk)을 통해 AI 신뢰성을 정의할 수도 있겠습니다. AI 시스템 위험성의 대표적 요인으로 앞에서 언급한 'AI 시스템의 자율성'과 더불어 '모델의 불투명성'을 꼽을 수 있습니다. 이러한 불투명성의 해소, 즉 AI 시스템의 투명성(transparency) 확보는 학습모델 관련 데이터뿐만 아니라 시스템의 추론 및 의사 결정 절차 전반에 이루어짐으로써, AI 시스템이 어떻게, 왜 그러한 의사 결정을 했는지 추적하고 파악할 수 있게 합니다. 이는 시스템에 대한 신뢰를 증진할 수 있게 합니다.

불투명성 해소를 위한 대표적 방안으로 설명가능성 또는 해석가능성을 확보하는 기술적 조치를 들 수 있습니다. 더불어 데이터 학습을 통해 모델을 만드는 전 과정에 대한 투명성 확보를 위해선, AI 시스템 개발, 활용, 지속적 개선을 위한 체계를 표준으로 개발해 보급·관리하고 더 나아가 안전성을 인증하는 정책적 조치가 필요합니다.

03. AI의 '블랙박스'를 여는 것은 보안 차원에서도 매우 중요하지 않을까 합니다. AI의 신뢰성을 훼손할 수 있는 보안 이슈에는 어떤 것이 있는지요? 그리고 이러한 이슈에는 어떤 대응방안이 논의되고 있는지요?

AI 시스템의 투명성 확보는 신뢰할 수 있고 책임감 있는 AI 시스템 구축에 필수적인 요소입니다. 다만 이에 따른 시스템 취약성에도 대비해야 합니다. 투명한 AI 모델은 내부 작동방식에 대한 정보가 더 많이 노출돼 해킹에 취약할 수 있습니다. 또한 관찰을 통해 보호돼야 할 내부 알고리즘이 유출될 수도 있습니다. 이러한 취약점에 대응하기 위해선 사전 예측, 예방, 취약점 발견, 위협 대응으로 구성된 선제적 방어 체계인 적응형 보안 구조(Adaptive security architecture) 등을 도입하는 것도 고려할 필요가 있습니다.

Q04. 현재 활용되는 AI가 언어모델에 기반한다는 특성상, 뇌과학, 심리학, 인지이론에 대한 이해도 필요할 듯합니다. 최근 생성 AI가 사용자를 적극적으로 '속이는' 모습이 보고되기도 했는데, 이는 일견 아동의 언어적 인지발달과정과 비슷해 보이는 것도 사실입니다. 이러한 인지발달과 관련해 어떠한 학제 간 연구가 이뤄지고 있는지요?

AI란 용어가 처음 만들어진 1956년 초창기부터, 이미 AI 연구는 인지과학, 심리학, 언어학, 인류학, 신경과학, 철학 등과 밀접하게 연관됐습니다. 최근엔 AI와 계산신경과학(Computational Neuroscience)을 연계하고 설명가능한 AI 기술을 활용해 EEG나 fMRI 같은 뇌 데이터를 해석하는 연구가 진행되고 있습니다. 역으로 ANN(Artificial Neural Network)과 같은 AI 모델을 활용해 뇌의 기능을 이해하고자 하는 연구 등도 있습니다. 과거 컴퓨터 분야가 특수 학문에서 사회 전반으로 확산된 것과 마찬가지로, AI는 범용적·융합적 역할을 담당할 것으로 생각합니다.

Q05. 현재 사용되는 AI 특성을 고려했을 때 학습데이터의 숫자를 줄인 소형언어 모델도 신뢰성을 확보·관리하는 데 유용한 접근법이 아닐까 합니다. 신뢰성 논의에선 소형언어 모델을 비롯한 다양한 AI 모델을 어떻게 바라보고 있는지요?

소형언어모델(sLLM)은 대형 모델에 비해 매개변수의 수가 수십억 내지 수백억 대로 그 크기가 작으며, 현재 온프레미스 AI와 온디바이스 AI에서 그 가치를 발휘하고 있습니다.

온프레미스 AI는 클라우드 AI와 대비해 외부연결 없이 내부 자원만을 활용해 구축됩니다. 이 때문에 신뢰성 측면에서 개인정보보호나 보안 안정성 확보에 유리하다고 할 수 있습니다. AI의 분야별 적용이 증대할수록 소형언어의 모델의 중요성은 더욱 증대할 것으로 생각합니다.

Q06. AI 신뢰성의 가장 중요한 목표 중 하나는 일반적으로 적용될 수 있는 투명하고 정량적인 기준을 마련하는 것 아닐까 합니다. 상기한대로 AI 신뢰성에 관여하는 요인이 다양하다 보니 표준화와 인증에 필요한 기준을 설정하는 것이 쉽지 않을 듯합니다. 이에 대해 국제적으로 어떤 논의가 이뤄지고 있는지요?

표준화의 본질적 목적이 기존의 능률성·경제성 증진과 더불어 점차 생명의 안정성, 환경보호, 사회적 책임과 같은 공공이익 증대로 발전하고 있습니다. 이러한 맥락에서 볼 때, AI 시스템 신뢰성 관련 표준화는 공공안전과 이익 증대에 이바지하는 중요한 표준화 대상이라는 인식이 국제적으로 공유되고 있습니다.

예를 들어 ISO/IEC JTC 1/SC 42 AI 분과에선 AI 시스템의 위험을 체계적으로 다루고 관리하

기 위한 프레임워크로서 ISO/IEC 42001 AI 관리시스템 표준을 개발하는 등, AI의 책임감 있는 사용과 신뢰성을 보장하고, 더불어 비용 절감 및 효율성 향상을 이루고자 공동의 노력이 진행되고 있습니다. 기술과 정책의 결합체인 기술표준의 개발은 이러한 국제적 기준을 설정하는 데 있어 가장 적합한 수단입니다. 그만큼 표준을 중심으로 한 국제적 논의가 바람직하다고 생각합니다.

07. ICT 분야에선 AI 신뢰성 확보를 위한 원천기술로 어떠한 연구를 추진하고 있는지 소개 부탁드립니다.

현재 KAIST 주관으로 진행된 '사용자 맞춤형 플러그앤플레이 방식의 설명가능성 제공 기술 개발' 과제를 통해 개발된 기술성과가 있습니다. 이를 국제표준으로 만들기 위해 현재 ISO/IEC JTC 1/SC 42/WG 3 Trustworthiness 작업반에서 Project Editor로 활동하고 있습니다.

개발 중인 표준인 ISO/IEC TS 6254는 학계, 산업계, 정책 입안자 등 다양한 이해관계자들이 AI 시스템의 전 생명주기에 걸쳐 설명가능성이나 해석가능성을 달성하기 위한 방안을 제공하는 것을 목표로 하고 있습니다.

08. AI 신뢰성 이슈는 AI와 관련 서비스를 활용하는 '사람'의 인식, 사용자가 AI의 판단을 어떻게 활용할 것인가에 대한 생각도 매우 큰 몫을 차지한다고 생각합니다. 이와 관련해 AI 신뢰성 확보를 위해 정부와 산업계가 AI를 어떻게 바라보는 것이 좋을지, AI 관련 연구에 어떤 관점에서 관여하는 것이 좋다고 생각하시는지 의견 부탁드립니다.

정부와 산업계에선 AI를 개별 기술이나 상품 또는 서비스로 볼 것이 아니라, 사용자와 개발자, 정책 입안자, 연구자, 판매자 등 다양한 이해관계자와 AI 시스템 생명주기로 구성된 생태계로 인식해야 합니다.

이러한 생태계를 이해하고 AI 신뢰성을 확보하기 위해선 표준 중심 접근법이 적합하다는 의견입니다. 표준은 그 자체로는 강제력이 없지만 관련 기술 정책을 수립하는 데 필수적인 기준과 절차를 제공합니다. 그간 국내 AI 개발의 투명성 확보를 위한 노력의 일환으로 TTA는 2023년 7월 신뢰할 수 있는 인공지능 개발 안내서를 펴내고 12월 국내 최초 인공지능 신뢰성 단체표준을 제정하여 컨설팅 및 시범 검증을 통해 현장에 적용하여 효과를 거두고 있습니다. 향후에도 국내 현실을 반영한 국제표준 개발과 함께, 이를 기반으로 글로벌 시장에서 통용될 수 있는 신뢰성 인증 체계를 선도해야 합니다. 특히, IEEE, NIST 등 선도적인 해외 기관과의 국제협력을 통해 글로벌 표준화를 주도적으로 이끄는 것이 필요합니다. 그 과정에서 국제적인 연계뿐만 아니라, 정부 정책과의 정합성을 검토하고 산업계 전문가와의 협의를 통해 기술적 타당성과 적합성 역시 보장해야 합니다.