

AI를 위한 데이터, 데이터를 위한 AI

인공지능(AI)에게 질문을 던져 봤다. “2000년 이후 한국의 연간 실업률이 가장 높은 해는 언제일까.” AI는 모든 연도의 실업률을 차례로 보여 줬고 출처까지 제시하며 2008년이 가장 높다고 답했다. 글로벌 금융위기가 원인이란 분석도 덧붙였다. 하지만 이는 사실과 다르다. 통계청 경제활동인구조사에 따르면 정답은 2000년이다. 다시 물었더니 AI는 자신이 틀렸다는 것을 바로 인정한 뒤 2019년이란 새로운 답변을 내놓았다. 이유는 코로나19로 촉발된 경제위기 때문이라고 했다.

AI의 등장으로 누구나 쉽게 통계를 검색하고 이용할 수 있는 시대가 열렸다. 하지만 AI에만 의존해 데이터를 활용했을 때 발생할 수 있는 위험성도 있다. 그럴싸한 설명과 함께 제시하는 AI의 답변이 사실과 다를 때가 생각보다 많다. 부정확한 데이터에 기반한 의사결정의 파장은 개인, 기업 등 우리 모두에게 결코 가볍지 않다.

왜 이런 일이 발생할까. 대부분 AI는 텍스트를 기반으로 학습하고 검색한다. 공인 기관이 제공하는 ‘통계표’ 형식의 정형화된 수치 자료는 직접 활용하지 못한다. 블로그나 오래된 기사, 영상 속 데이터를 주로 참고하는데 출처가 불확실하고 자료의 신뢰성을 보장하기 어렵다는 것이 문제의 핵심이다. AI는 자신이 아는 경로에서 답을 찾지

공직자의 창



안형준
통계청장

못하면 기존에 학습한 패턴에 따라 데이터를 생성하기도 한다. 안타깝게도 결과가 항상 옳다고 말하기 어렵다.

해법은 결국 데이터에 있다. AI가 올바르게 작동하려면 신뢰할 수 있는 데이터가 학습 자료로 제공돼야 한다. 특히 AI가 공식 통계 데이터베이스(DB)를 활용할 수 있도록 가르쳐야 한다. 각 통계 수치의 기준, 의미, 작성 방법을 알려 주는 설명 자료가 필요하다. 이를 ‘통계 메타데이터’라고 한다. 정확한 길을 찾기 위한 ‘지도’에 해당한다.

통계청은 AI가 이용자 요구에 따라 ‘진짜’ 데이터를 찾아갈 수 있는 기반을 마련하고자 내년 ‘AI 친화적 통계 메타데이터 구축’ 예산을 확보했다. 내년에 표준을 정비한 뒤 단계적으로 국가가 승인한 메타데이터를 제공할 계획이다. 이 사업은 전 국민을 대상으로 AI를 확산하기 위해 새 정부가 추진하는 ‘모두의 AI’ 정책을 실현하는데 기여할 것으로 기대된다. 민간과 공공기관은 믿

을 수 있는 자료를 기반으로 AI 서비스를 개발하게 되고(Data for AI), 국민은 보다 똑똑해진 AI를 통해 정확한 데이터를 쉽게 접하게 될 것이다(AI for Data).

이런 비전은 9일부터 이틀간 ‘AI로 만들어 가는 국가통계의 미래’라는 주제로 서울에서 열리는 통계청·경제협력개발기구(OECD) 공동 국제회의에서 공유될 예정이다. AI가 각국 통계청의 데이터베이스에 담겨 있는 공식 통계에 접근하지 못하는 것이 비단 우리나라만의 문제는 아니기에 국제사회도 한국 통계청의 시도에 주목하고 있다.

OECD는 이번 회의에서 향후 각국의 AI 공동 대응과 국제협력 행동계획에 관한 선언문을 한국 통계청과 공동으로 발표할 계획이다. 지금까지 국제사회가 단순히 AI 기술을 어떻게 활용할지를 주로 논의해 왔다면 앞으로는 국민이 안심하고 활용할 수 있는 AI 데이터 환경을 만들자는 데 공감대를 형성하고 머리를 맞댈 것으로 보인다.

AI 대전환 시대 속에서도 변하지 않는 진리가 있다. 아무리 뛰어난 기술도 신뢰할 수 있는 데이터 기반이 없으면 활용에 한계가 있다는 사실이다. 한국형 ‘AI 메타데이터 표준’이 향후 국제통계사회의 AI 대응을 선도해 나갈 열쇠가 될 것으로 믿는다.