이준기의
빅데이터

〈중앙 SUNDAY 제공〉

내편 네편 불신의 사회 … AI에게 '판단' 맡기면 공정할까

이준기
연세대 정보대학원 교수

외국 대학에서 교수로 근무하다가 국내 대학으로 옮긴 뒤, 얼마 지나지 않아 교무처 보직을 맡은 적이 있다. 그곳에서의 핵심 업무는 교수 채용과 승진, 승봉(호봉 상승) 심사였다. 즉, '평가'하는 일이었다. 그런데 일을 시작하자마자 깨달은 사실이 있다. 평가 자체보다 평가 기준을 만드는 것이 더 어렵다는 점이다. 학문은 늘 움직인다. 새 저널이 생기고 새 분야가 뜬다. 그때마다 기준을 다시 짜고, 설명하고, 설득해야 한다.

해외 대학의 평가는 대체로 정성평가 중심이다. 그에 비해 국내에선 점수제가 강하다. 이유를 곰곰이 생각해 보면 한 가지 결론에 닿는다. 우리는 학연·지연 같은 관계망이 촘촘하다. 심사 과정에 이해관계가 섞일 위험도 크다. 그러니 사람의 판단을 노출시키는 방식보다, '객관성'의 외관을 갖춘 점수제가 안전장치처럼 작동해 왔다. 점수는 중립이어서가 아니라, 갈등을 덜 보이게 해서 편리하다.

행정·복지·치안 결정 기계화의 명과 암

하지만 이는 뒤집어 말하면 평가에 대한 사회적 신뢰가 약하다는 뜻이기도 하다. 신뢰의 결핍은 곧 비용으로 이어진다. 대학 평가는 새 기준이 나올 때마다 논쟁이 반복된다. 최근 우리 사회는 판결에 대해서도, 심지어 연예인의 은퇴조차 '사실'보다 '우리 편인가'가 먼저 거론되곤 한다. 요즘 '리더의 부재'라는 말이 나올 정도로 전문가의 말도 쉽게 진영의 논리에 파묻힌다. 그렇다면 공정한 판단은 어떻게 가능할까. 이 질문이 무거워질수록, 마음 한쪽에서 엉뚱한 상상이 튀어나온다. 플라톤이 말한 '철인'처럼, 이해관계에서 자유로운 누군가가 나타나 갈등을 정리해 주면 어떨까.

인공지능(AI)의 고도화는 이 상상에 현실감을 부여한다. '보조 도구'가 아니라 판단의 주체로 AI를 세우자는 논의가 커지는 이유다. 이미 행정, 복지, 교육, 노동, 치안, 사법 시스템에서 알고리즘이 사람의 삶을 좌우하는 결정을 내려온 역사가 있다. 기술이 좋아질수록, 그리고 판단 비용이 커질수록, "그냥 AI가 결정하게 하자"는 유인은 강해진다.

AI 판단의 장점은 분명하다. 적어도 인간관계의 압력에서 상대적으로 자유롭다. 친분이나 청탁에 흔들릴 감정도 없다. 일관된 규칙으로 빠르게 처리할 수 있다. 판사의 판결이 점심 후에 좀 더 관대해진다는 연구도 있지만, AI 판단은 컨디션이나 그날의 기분 등에 좌우되지 않는다.

사실 우리는 이미 '판단의 자동화' 속에서 산다. 신용 점수는 대출 가능 여부와 금리를 갈라놓고, 보험료 산정은 병력과 생활 습관을 숫자로 환산한다. 기업의 채용 전형은 지원자의 이력서를 필터링하고, 온라인 플랫폼은 글과 영상의 노출을 조정한다. 이 판단들은 대부분 사람의 얼굴을 갖지 않는다. '시스템이 그렇게 정했다'는 말로 끝난다. 생성형 AI는 여기서 한발 더 나아가, 근거를 말로 포장하며 '그럴듯한 판정문'까지 써낼 수 있다. 그래서 더 위험하고, 그래서 더 매력적이다.

문제는 AI에도 보이지 않는 변수가 있다는 것이다. 첫째는 편견이다. AI는 과거의 데이터를 먹고 자란다. 과거가 불공정했다면, 그 불공정이 '정답'처럼 재학습될 수 있다. 두 번째는 AI가 어떤 가치관을 갖고 있는가다. 지난 글에서 필자는 AI도 각각의 MBTI를 갖고 있다는 글을 쓴 적이 있는데, 이는 단적인 예다. 어떤 데이터를 주로 학습하였는가에 따라 AI는 다른 성향을 보여준다. 실제로 기업의 사회적 책임, 낙태의 적법성 등의 문제에 대하여 AI들은 서로 다른 견해를 보여 주곤 한다. 결국 어떤 성향의 AI에게 판단을 맡기느냐에 따라 결과가 달라진다면, 이는 또다시 '우리 편'을 가르는 인간의 문제로 귀결될 것이다. 세 번째이자 가장 큰 문제는 '맥락의 부재'다. 예를 들어 우리가 정의

친분·청탁에 흔들릴 '감정' 없지만
AI도 학습 데이터 따라 편견 가져

인간의 동기·맥락 모르는 AI 판사
판결 정확해도 정의롭진 않을 수도

투명성·이의제기권·책임주체 등
AI 판단 시스템에 안전장치 둬야

(justice)를 어떻게 규정하느냐에 따라 다르겠지만, 동기와 맥락이 있는 평가는 단순한 결과의 평가와는 다르다. 절도는 맞으나, 생존을 위한 불가피한 선택이라는 맥락과 정상 참작이라는 주관성, 인간의 고통, 절박함, 후회와 같은 감정적 경험의 질(qualia)을 이해하지 못하는 AI라면, 이때 AI의 판정은 '정확'할지는 몰라도 '정의롭지' 않을 수 있다는 철학적 난제가 남을 수 있다.

그래서 많은 사람은 'AI는 보조, 최종 판단은 인간'이라는 절충안을 떠올린다. 그런데 현실은 종종 반대로 간다. 결정 비용을 줄이고 속도를 올리기 위해, 공정성을 높이기 위해 AI가 1차가 아니라 '최종'을 맡는 경우가 있다. 2020년 영국에서는 코로나 사태로 시험을 치르지 못하게 되자 성적을 표준화 알고리즘으로 산출했고, 그 결과가 학생의 진학을 사실상 결정했다. 이에 대한 신뢰성이 흔들리고 사회적 반발이 커지자 정책은 결국 철회됐다.

호주의 '로보데트(Robodebt)'는 더 극단적이다. 정부가 소득 자료를 자동 산출해 복지 수급자에게 채무 통지를 보냈고, 많은 사람이 '설명할 수 없는 빚'에 시달렸다. 네덜란드의 보육수당 사태 역시 자동화된 부정수급 판정이 과도한 환수와 낙인으로 이어져 사회를 뒤흔들었다. 이들 사례에서 공통점은 하나다. 결정이 자동으로 내려진 뒤에는, 개인이 뒤늦게 이의를 제기해도 이미 피해가 누적된다는 점이다.

알고리즘 숨기면 사후 구제 쉽지 않아

플랫폼 노동 영역에서도 '결정'은 기계화된다. 배달·이동 서비스에서 계정 비활성화, 배차 우선순위, 근무 슬롯 배정 같은 처분이 알고리즘으로 실행된다. 문제는 당사자가 이유를 알기 어렵고, 소명 절차도 제한된다는 것이다. 더 큰 문제는 '정당성의 착시'다. 사람의 판정은 불완전해도, 질문하면 이유를 들을 수 있다. 그러나 알고리즘은 종종 이유를 숨긴다. 규칙을 공개하면 악용될 수 있다는 주장, 기업의 영업 비밀, 개인정보 보호가 뒤엉키면서 결정 과정이 닫힌다. 닫힌 결정은 반박하기 어렵다. 그 위에 생성형 AI의 문장력이 얹히면, 불투명한 결론이 마치 합리적 숙고의 산물처럼 보일 위험이 커진다.

이처럼 행정·교육·노동·치안의 여러 영역에서 '결정'이 기계화되면, 사후 구제만으로 신뢰를 담보하기 어렵다. 처음부터 어떤 데이터와 규칙이 쓰였는지, 예외는 어떻게 처리되는지, 이의 제기는 어떤 경로로 가능한지, 그리고 최종 책임은 누가 지는지까지 제도 설계로 들어가야 한다. 판단이 곧 권력이라면, AI 판단은 기술이 아니라 제도다.

자칫 이러한 우려가 'AI 도입을 반대하자'는 주장으로 읽힐지 모르겠다. 하지만 필자의 생각은 '그럼에도 불구하고 도입해야 한다'는 쪽에 가깝다. 객관적 신뢰가 중요하고 사회적 비용이 큰 정부 예산 배정이나 입찰, 연구개발(R&D) 심사, 자동차 보험손해율 산정, 복지 수급자 배정 등은 오히려 AI를 적극적으로 실험해봐야 할 분야다.

이는 먼 미래의 이야기만도 아니다. 최근 서울시교육청이 2033학년도 수능의 논·서술형 전환을 제안하며, 채점 부담을 줄이기 위한 AI 활용 가능성이 거론됐다. 논·서술형의 취지는 좋다. 그러나 채점이 'AI의 판정'으로 넘어가는 순간, 투명성과 책임의 요구는 지금보다 훨씬 커진다. AI가 점수를 매기는 사회는 편해질지 몰라도, 설명할 수 없는 사회가 되기 쉽다.

그래서 'AI가 맞느냐 틀리냐'의 싸움으로는 답이 나오지 않는다. 핵심은 절차다. 결정 기준을 문서화하고, 데이터의 품질을 점검하고, 편향을 정기적으로 감사해야 한다. 자동 결정이 생활을 직접 좌우하는 영역이라면, 설명 의무와 이의 제기권은 최소한의 안전장치가 된다. 무엇보다 책임의 주체를 비워두면 안 된다. 시스템이 결론을 내렸어도, 책임질 기관과 사람이 있어야 민주주의가 작동한다.

결국 질문은 단순하다. 우리는 어떤 판단을 '기계에 맡길' 준비가 되어 있는가. 그리고 맡기더라도 어떤 안전장치를 둘 것이냐다. AI는 철인이 아니다. 때로는 과거를 그대로 비추는 거울이다. 그렇다고 거울을 부숴서 문제가 사라지지도 않는다. 필요한 것은 기술의 찬반이 아니라, 판단 시스템을 설계하는 감각이다. 투명하게 만들고, 이의 제기를 가능하게 하고, 책임을 명확히 해야 한다. 그래야만 AI 판단이 신뢰의 회복으로 이어질 수 있다.

〈광주일보와 중앙 SUNDAY 제휴 기사 입니다〉

이준기 연세대 정보대학원 교수. 서울대 계산통계학과 졸업 후, 카네기멜론대 사회심리학 석사, 남가주대 경영학 박사를 받았다. 인공지능의 기업 활용에 대해 여러 회사에 자문을 하고 있다. 저서로는 'AI로 경영하라' '오픈 콜라보레이션' '웹 2.0과 비즈니스 전략' 등이 있다.

KSA 한국표준협회
KOREAN STANDARDS ASSOCIATION
ISO 21388
보청기적합관리 인증센터

"고객에게는 신뢰와 만족"

국제보청기 since 1982

✓ 필요한 소리만 **똑똑히 들립니다.**
✓ 작은 사이즈로 착용시 **거부감이 없습니다.**
✓ 정직한 우수상품 **가격부담이 없습니다.**

본 점 서석동 남동성당 옆 062) 227-9940 / 062) 227-9970
서울점 종로 5가역 1층 02) 765-9940
순천점 중앙시장 앞 061) 752-9940