매일경제

처음부터 나쁜 AI가 있을까요? 딥페이크도 쓰는 '사람' 문제죠

황순민 입력 2022.02.11. 16:57 수정 2022.02.11. 20:24

[Weekend Interview] 90년생 카카오 사외이사 박새롬 성신여대 융합보안공학과 교수

국내 주요 상장사 중 최연소·여성 사외이사이자 젊은 과학자로 주목받는 박새롬 성신여대 융합보안학과 교수. 인공지능 알고리즘이 가질 수 있는 취약성을 보완하고 프라이버시를 보호하는 기계학습 알고리즘을 개발하는 것이 그의 목표다. [사진 제공 = 박새롬 교수]

2020년 F16 톱건(최우수조종사)과 인공지능(AI) '알파도그파이트' 간 맞대결이 펼쳐졌다. AI에 맞서 모의 공대공 전투를 벌인 인간 조종사는 미국 공군 훈련소의 최우수 교관으로 자타가 공인하는 지구상 최고 실력을 보유한 전

투기 조종사였다. AI는 다섯 차례 모의 공중전에서 인간 조종사를 모두 격추했다. 바야흐로 AI가 인간을 완전히 압도하는 시대가 열렸다. 구글 딥마인드가 제작한 AI 알파고가 이세돌을 완파했던 '세기의 대결'이 벌써 6년 전이다.

지난 6년간 AI는 진화를 거듭하며 일상의 모든 영역에 침투했다. 바둑뿐 아니라 의료·과학 등 복잡계의 문제를 해결하는 범용 AI로 적용 분야를 빠르게 넓히고 있다. "AI가 인류가 새로운 지식영역을 개척하고 진리를 발견할 수 있도록 도울 것"이라는 데미스 허사비스 구글 딥마인드 최고경영자(CEO)의 호언장담이 현실로 이뤄지고 있는 셈이다. AI 발전 속도가 빨라질수록 알고리즘 취약성과 프라이버시 침해 등 우려도 함께 터져나온다. 기계와 인간을 구분하는 이른바 '튜링 테스트'를 통과하는 AI가 나온다 할지라도 그들이 인간처럼 의식과 감정을 갖기는 어렵다는 점에서 두려움을 느끼는 이도 상당하다. 최근 매일경제가 만난 박새롬 성신여대 융합보안공학과 교수는 AI로 인해 발생할 수 있는 부작용과 문제에 대해 주목하고 해결을 위해 도전해온 젊은 과학자다. AI와 보안, 기계·통계학습 등이 주요 연구 분야다. 박 교수는 서울대 연구원으로 재직할 당시 알파고 등장을 보면서 AI와 보안 문제에 관심을 갖게 됐고, 실용성에 기반을 둔 연구 성과를 내고 있다. 국내 재계 최연소 사외이사(카카오)로 정보기술(IT) 업계에 신선한 관점을 불어넣고 있다. AI 알고리즘이 가질 수 있는 취약성을 보완하고 프라이버시 보호를 위한 기계학습 알고리즘을 개발하는 것이 그의 목표다. '처음부터 나쁜 AI는 없다'는 그를 만나 착한 AI의 활용법을 물었다.

기술발전으로 사진합성 고도화
알고리즘 자체는 나쁜 것 아냐

데이터 대량 수집해 학습하는 AI
개인정보 등 보안 취약성은 필연
알파고 대국 보며 융합보안 관심

데이터 유통 자체를 막을순 없어
투명한 수집과 유익한 활용이 답
'알고리즘 공정성' 규제하기보다
공정모델에 인증 부여가 효과적

기업 사외이사 낯설기도 했지만
사회에 도움될 기회라고 여겼죠

—AI·보안 데이터마이닝 전문가의 길을 걷게 된 계기는.

▷사실 특별한 이유가 있던 것은 아니다. 돌이켜보면 항상 무언가를 배우고 알아가는 과정이 좋았다. 산업공학과에서 데이터 분석, 최적화, 확률 통계 등을 배우면서 데이터 분석에 관심을 갖게 됐다. 특히 재미있게 공부했던 것들이 중요한 기반 지식으로 활용될 수 있는 기계학습(머신러닝) 분야를 더 연구해 보고 싶다는 생각을 하게 됐다. 마침 대학원 입학 후에 알파고의 바둑 대국이 이뤄졌고 연구실에서 다 같이 관심을 갖고 본 기억이 난다. 운이 좋게도 학부부터 대학원까지 훌륭하신 교수님을 많이 만났다. 좋아하는 연구를 하면서도 학생 지도와 교육을 통해 긍정적인 영향력을 끼치는 일을 하고 싶다는 생각에 교수의 길에 들어서게 됐다.

—AI를 어떻게 정의하나.

▷AI에 대한 수업을 할 때 늘 소개하는 여러 정의가 있다. 그중에서도 가장 먼저 소개하는 정의는 '사람의 생각과 관련된 활동을 자동화하는 것'(Bellman·1978년)이다. '인공'이라는 것은 기계, 지능적인 에이전트, 인공물, 컴퓨터 등으로 표현할 수 있다. '지능' 부분도 다양하게 표현할 수 있는데, '자동화'라는 표현에서 지능에 대한 정의를 광범위하게 잘 나타내고 있다고 생각한다.

—AI 알고리즘 취약성이 가져올 수 있는 사회문제는 무엇일까.

▷AI의 취약성이 딱 한 가지로 존재하는 것이 아니기 때문에 그로 인한 사회문제도 한 가지로 정의하기는 어려울 것 같다. 하지만 AI가 많은 곳에서 활용될수록 AI 알고리즘이 갖는 보안 취약점은 우리 삶에 큰 영향을 끼치는 것만큼은 확실하다. AI로부터 얻는 결과를 우리 의사 결정을 보조하는 역할로 활용한다면 그 위험성이 덜 할 수 있지만 앞서 언급했던 AI의 정의에서처럼 AI를 도입함으로써 자동화된 의사 결정들을 하게 된다면 AI에 대한 문제가 발생했을 때 그것이 실세계에 직접적으로 영향을 끼칠 수 있게 된다. 또 다른 측면으로는 최근 우리가 관심을 갖고 있는 AI 알고리즘들은 기본적으로 수집된 데이터를 기반으로 학습하게 되기 때문에 학습에 사용되는 데이터의 프라이버시도 중요한 문제가 될 것으로 본다.

—AI가 발전할수록 더 많은 사회적 합의가 필요해질 것 같다.

▷AI로 인해 발생할 수 있는 모든 가능한 문제를 미리 다 예방하는 것은 결코 쉬운 일이 아닐 것이다. 하지만 AI 기술에 대한 충분한 이해와 대비는 필요하다고 생각한다. 예컨대 일부러 공정하지 않은 AI를 학습하려고 하는 것이 아니었음에도 AI 얼굴 인식 시스템 학습 중에 동양인의 얼굴 데이터가 적어 동양인의 얼굴인식 성공률이 낮도록 학습된다면 해당 서비스를 활용하는 데 불평등이 발생할 수 있다. 앞으로 AI에 다양한 측면에 대해 이해가 넓어질수록 고민해야 할 부분이 많아질 것이다. 가령 AI 기술이 발전하면서 데이터가 창출할 수 있는 가치들이 커지게 됐다. 이제는 많은 사람이 데이터의 가치를 충분히 인지하고 있는 것 같다. 여전히 우리가 의식하지 않은 채 우리 정보가 수집될 수도 있지만 데이터 수집에 관한 다양한 법과 제도가 만들어짐에 따라 많은 정보가 우리가 관심을 가지면 어떻게 수집되는지에 대해 인지할 수 있는 단계에 접어든 것이다. 데이터를 수집하는 것이 무조건 나쁘다는 인식보다는 우리 데이터가 어떻게 수집되고 어떻게 활용되는지를 알 수 있는 투명성이 중요하다는 의미다. 데이터 수집을 아예 차단하는 것보다 데이터가 우리에게 유익을 줄 수 있는 방향으로 활용되도록 다양한 이해관계자들이 함께 고민하고 합의해 나가는 과정이 중요하다.

—AI 자체가 나쁜 것은 아니라고 언급했는데.

▷가령 딥페이크와 같은 기술은 일종의 사진 합성 기술이라고 볼 수 있다. 딥페이크가 악용될 소지가 있기 때문에 우려되는 부분이 존재하지만 딥페이크를 나쁜 방향으로만 사용할 수 있는 것은 아니기 때문에 알고리즘 자체가 나쁜 것은 아니라는 의미다.

—융합보안은 어떤 학문인가.

▷보안은 융합이 굉장히 중요한 분야다. 연구적으로는 AI 보안 문제 자체에 한정해서 연구를 하고 있지만 실제로 AI 기술이 시스템에 활용될 때에는 AI 보안만의 문제를 넘어선 다양한 보안 문제가 모두 중요해지기 때문이다.

—요즘엔 어떤 연구 분야에 관심을 갖고 있나.

▷최근엔 데이터를 활용하면서도 발생할 수 있는 걱정들을 최소화할 수 있는 연구를 수행하고 있다. 가령 요즘 관심을 갖고 있는 프라이버시 보호 기계학습 알고리즘 연구는 개인 프라이버시를 보호하면서도 데이터를 활용할 수 있도록 돕는 연구다. 최근 연구 중에서는 AI 모형의 공정성을 감시하는 프레임워크를 제안했는데, 민감 정보의 프라이버시는 보호하면서도 공정한 모형인지를 판단해줄 수 있도록 기밀 컴퓨팅 기법을 함께 활용하는 방법이다. 알고리즘의 공정성을 규제로 제약하는 것보다 공정한 모형에 인증을 부여하는 개념이다. 이를 통해 공정한 모형을 활용하는 기업들을 사용자들이 인지할 수 있도록 하는 것이다. 이는 ESG 경영(환경·책임·투명경영)과 관련해 공정

한 모형의 활용을 촉진하는 효과를 가져을 수 있을 것으로 기대하고 있다. 큰 학술대회에 논문 2개가 채택됐다. 내가 고민하는 연구가 의미가 있다는 생각이 들어서 큰 동기부여가 됐다.

—어렸을 때부터 재능을 발견한 편이었나.

▷고등학교 때부터 공부에 흥미를 갖기 시작했고 오히려 아주 어렸을 때에는 음악을 전공하고 싶었다. 그래서인지 부모님이 공부에 대해 크게 기대하거나 바라는 부분이 없었고, 과학 과목 공부를 하면서 재미를 느낄 수 있었다. 누가 시켜서 하는 것이 아닌 스스로 재밌어서 공부를 하게 된 것이 질리지 않고 꾸준히 공부하게 된 중요한 이유가 됐다.

—기억에 남는 수업이 있나.

▷사실 대학에 입학했을 때는 한글 타자도 못할정도로 컴퓨터와는 거리가 멀었던 학생이었다. 1학년 때 들었던 김태완 서울대 교수님의 '컴퓨터의 원리' 수업이 기억에 남는다. 그 수업을 들으면서 처음으로 C언어를 배웠다. 아무것도 모르는 상태에서 코딩을 통해 간단한 프로그램까지 만들 수 있게 되는 과정이 신기하면서도 즐거웠고 컴퓨터와 친해지는 아주 중요한 계기가 됐다. 교수님이 수업에서 '중학생도 이해할 수 있는 프로그래밍'이라는 말씀을 자주 하셨던 게 아직도 기억이 난다. 그 덕분에 용기를 갖고 프로그래밍을 공부할 수 있었다.

—현재 '잘하고' '좋아하는' 일을 하고 있다. 후배들에게 진로 선택 조언을 해준다면.

▷일단 무엇보다 내가 좋아하는 일들에 대해서 많은 고민을 했다. 좋아하는 일들을 하게 된다면 잘 못하더라도 만족감을 느낄 수 있는 선택이 될 것이라고 생각했기 때문이다. 좋아하는 일이 무엇인지 스스로에게 적극적으로 질문해 보는 것도 좋을 것 같다. 그리고 내가 선택한 일에 대해서 후회하지 말고 살아 가자는 생각을 평소에도 많이 하는 편이다. 이 부분이 진로 선택에 많은 도움이 됐다. 일단 선택한 후에는 선택한 방향에 대한 좋은 부분들을 생각하려고 노력하기 때문에 대부분 개인적으로 좋은 선택들이 될 수 있었다.

—요새는 어떤 공부를 하고 있나.

▷이전에는 주로 AI 분야만 공부해왔다면 요즘엔 AI를 벗어난 보안 분야 연구자들과 함께 공부하는 기회를 만들어 세미나를 하며 즐겁게 공부하고 있다. 훌륭하신 교수님이나 연구자들은 다 각자 강점을 갖고 있다. 한 연구자를 롤모델로 삼기보다는 직접 만나고 경험하는 다양한 분들의 강점들을 보며 항상 배우면서 살아가고 있다.

—국내 최연소 여성 사외이사이기도 하다. 사외이사로 선임된 이유가 무엇이라고 생각하나.

▷사실 처음 선임됐을 때는 사외이사라는 직책 자체가 낯설기도 했다. 개인적으로는 관심을 가지고 연구하고 있던 분야가 AI 보안이었기 때문에 이 분야와 관련해서 기업이 필요한 부분에 도움이 됐으면 좋겠다는 생각을 했다. 카카오에서 ESG 위원회에 위원으로 참여하게 되면서 ESG 경영에 대해서 배울 기회들이 있었다. 기업뿐 아니라 연구자인 나도 내 분야에서 세상에 도움이 되는 연구를 하고 싶다는 생각을 하게 됐다.

—앞으로 꿈이 무엇인가.

▷나는 한국에서 꾸준히 공부하며 박사를 받았다. 학자들 간의 협업, 산업계와 학계 간 협업이 다양한 방식으로 이뤄질 수 있는 연구 공동체를 만들어보는 것이 꿈이다. 이를 통해 새로운 가치들을 창출하고 중요하고 복잡한 현재 혹은 미래의 문제들을 풀어나가는 데 기여하고 싶다.

▶▶ 박새롬 교수는…

1990년생. 2018년 서울대 산업공학과에서 공학박사 학위를 취득했다. 서울대 수학기반산업데이터해석 연구센터 연구원을 거쳐 성신여대 교수가 됐다. 2019년 'WISET·KIIE 젊은 연구자상 최우수상'을 수상했고 컴퓨터과학 분야 SCI급 국제 학술지 '뉴럴 네트워크'에 공저자로 참여해 AI의 감정 분석 기술 관련한 논문을 등재했다.AI·보안 분야 전문성을 인정받아 2020년 카카오 사외이사로 선임돼 화제를 모았다.

[황순민 기자]