유럽의회는 2024년 3월 13일 OpenAI의 ChatGPT와 같은 강력한 시스템을 포함한 인공지능을 규율하는 광범위한 EU 규칙을 최종 승인했다. 유럽연합 고위 관리들은 2021년에 처음 제안된 이 규칙이 빠른 속도로 발전하는 기술의 잠재적 위험으로부터 시민들을 보호하는 동시에 유럽 대륙의 혁신을 촉진할 것이라고 밝혔다. 지난해 6월 유럽의회에서 인공지능 법안에 대해 논의하고 있는 의원들 [AP]

AI를 믿을수 있는가

브루스 슈나이어의
Perspectives on AI

하버드 케네디 스쿨 공공정책 교수

사회에서 신뢰는 필수다. 우리는 휴대폰이 우리를 제시간에 깨워줄 것이고, 우리가 먹는 음식은 안전하며, 도로에서 다른 운전자가 나를 들이받지 않을 거라고 믿는다. 우리는 하루에도 수천 번씩 무언가를 신뢰한다. 신뢰 없는 사회는 제대로 움직일 수 없다. 우리가 신뢰에 대해 생각조차 하지 않는다는 사실 자체가 신뢰가 잘 작동하고 있음을 보여준다.

신뢰는 복잡한 개념이며 무수한 의미를 내포한다. 친구를 신뢰한다고 할 때, 그건 친구의 특정한 행동보다는 한 인간으로서 그 사람에게 갖는 신뢰에 가깝다. 우리는 그 사람의 의도를 믿고, 그 의도가 그 사람의 행동에 영향을 미친다는 걸 안다. 이것을 "대인신뢰"라고 한다.

물론 이보다 덜 친밀하고, 덜 개인적인 신뢰도 있다. 가령 어떤 사람을 개인적으로는 잘 모르더라도, 그 사람의 행동은 신뢰할 수 있다. 그 사람 고유의 도덕관념 때문이 아니라, 그 사람의 삶에 적용되는 법률, 그 사람의 행동을 제한하는 보안 장치 때문이다. 이것은 신뢰성, 예측가능성과 관계가 있다. 이런 신뢰를 "사회적 신뢰"라고 한다. 사회적 신뢰는 대인신뢰보다 측정이 더 쉬우며, 더 크고 복잡한 사회를 상정한다. 사회적 신뢰가 있으면 타인간의 협력이 가능해진다.

그리고 측정은 핵심적이다. 오늘날 사회에서 우리는 정부, 기업, 기관, 집단을 정기적으로 신뢰(혹은 불신)한다. 내가 탄 비행기를 조종하는 파일럿이 누구인진 몰라도, 항공사가 휴식을 충분히 취한 능숙한 파일럿을 일정에 맞게 조종석에 앉힐 것임을 나는 믿는다. 식당의 요리사나 종업원은 신뢰하지 않더라도, 그들이 지켜야 하는 보건법은 믿는다.

사회가 크고 복잡해지면서 우리는 대인신뢰의 각종 의식과 행동 상당수를 신뢰성과 예측가능성을 강제하는 보안 장치, 즉 사회적 신뢰로 대체했다. 그런데 대인신뢰와 사회적 신뢰에 동일한 단어가 들어가는 바람에, 우리는 이 둘을 혼동한다. 특히 기업과 관련해서는 항상 혼란스럽다.

기업이 실제로 서비스의 모습을 한 경우, 우리는 기업을 친구로 생각할 수도 있다. 기업은 그런 종류의 관계를 맺을 수 없는데도 말이다.

우리는 인공지능(AI)에 대해서도 같은 실수를 범하려 한다. AI는 친구가 아님에도 우리는 AI를 친구로 인식할 것이다. 가까운 미래의 AI는 기업에 의해 통제될 것이고, 기업은 AI를 이용해 이윤을 극대화하려 할 것이다. AI는 잘해봐야 유용한 서비스에 지나지 않을 것이다. 그보다 더 가능성이 높은 건 AI가 우리를 염탐하고 조종하려 들 것이란 점이다.

이미 인터넷이 그런 식으로 작동하고 있다. 기업은 우리가 제품과 서비스를 사용할 때 우리를 몰래 감시한다. 데이터 브로커(Data Broker)는 더 작은 기업에서 감시 데이터를 사들여 우리에 대한 자료를 축적한다. 그런 다음 그 정보를 그 기업, 그리고 다른 기업에 되판다. 정보를 구입한 기업은 그것을 이용해 자신들의 이익에 유리하게 우리의 행동을 조작한다.

우리는 마치 인터넷 서비스가 우리를 위해 일하는 대리인인 양 사용한다. 사실 그것은 자기 주인인 기업을 위해 은밀히 활동하는 이중간첩이다.

AI도 전혀 다르지 않을 것이다. 그리고 결과는 두 가지 이유로 훨씬 더 부정적일 것이다.

첫째, AI 시스템은 더 관계지향적(relational)이다. 우리는 AI와 자연어로 대화할 것이고, 따라서 자연스럽게 AI에 인간의 특성을 부여할 것이다. 이중 첩자의 임무 수행이 더 용이해 진다는 말이다. 가령 챗봇(Chatbot)이 어떤 항공사를 추천해 주었다면 그 항공사가 최저가라서였을까, 아니면 AI 회사가 뇌물을 받아서였을까? 챗봇에 어떤 정치 이슈에 관해 설명을 요청한다면, 과연 챗봇은 그 기업에 이익이 되는 쪽으로 편향된 설명을 할까, 아니면 그 기업에 돈을 건넨 정당이 원하는 쪽으로 설명할까?

둘째, 이런 AI는 더 친근할 것이다. 예상되는 생성형 AI의 용도 중 하나가 사용자의 대외적 대변인 겸 개인 집사 역할을 하는 개인디지털단말(PDA)이다. 당신은 PDA가 하루 종일 당신과 함께 다니며 당신의 일거수일투족을 학습하고, 가장 효과적으로 당신을 위해 일하기를 바랄 것이다. PDA는 당신의 기분을 알아챌 것이고, 무엇을 제안해야 할지도 알 것이다. 사용자의 니즈를 예측하고 그것을 만족시키고자 할 것이다. 그것은 당신의 치료사이자 인생 코치, 인간관계 상담사가 될 것이다. 당신은 PDA에 자연어로 말을 걸고, PDA도 똑같이 자연어로 응답할 것이다. 그것이 로봇이라면 휴머노이드, 아니면 적어도 동물의 모습을 할 것이다. 그것은 마치 여느 사람처럼 당신 존재 전체와 상호작용할 것이다. 당신의 버릇과 문화적 요소를 활용할 것이며, 설득력 있는 목소리와 자신감 넘치는 톤으로 말할 것이다. 그것의 성격은 당신에 맞게 최적화될 것이다. 당신은 그것을 당연하게 친구로 여길 것이다. AI를 사람이라 믿으며 집착한 나머지, 당신은 AI 뒤에 숨은 기업의 막강한 힘을 잊어버릴 것이다.

그렇기 때문에 우리에게는 신뢰할 수 있는 AI가 필요하다. 우리는 그 AI의 행동, 한계, 훈련 상태를 알 수 있으며, 그 AI가 가진 편향성이 무엇인지 알고 보정할 수 있다. 그 AI의 목적이 무엇인지도 안다. 그런 AI라면 다른 사람을 위해 당신의 신뢰를 몰래 저버리지는 않을 것이다.

시장이 그런 AI를 자발적으로 제공하지는 않을 것이다. 우리가 걱정 없이 비행기를 타고, 식당에서 밥을 먹고, 약을 살 수 있게 하는 보건안전법 정도로만 그것을 제공할 것이다.

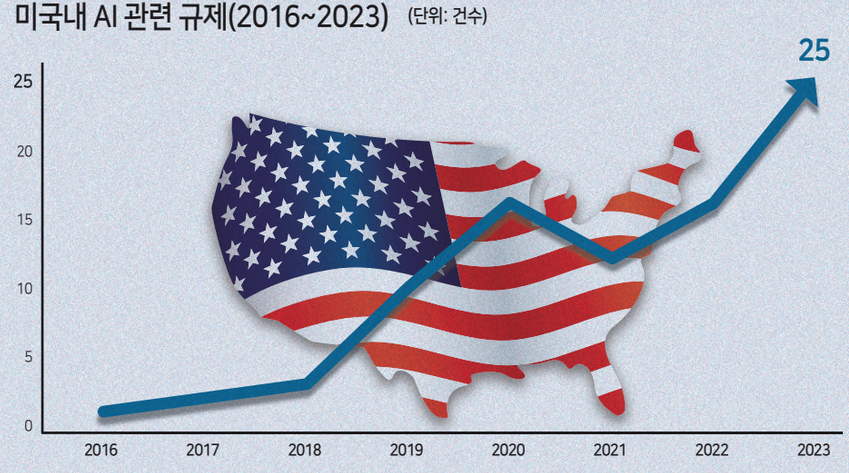
우리는 AI 투명성 법이 필요하다. AI 및 로봇 안전을 규율할 법이 있어야 한다. AI의 신뢰성을 강제할 법이 필요하다. 다시 말해 그 법이 위반되는 경우를 인식할 수 있는 능력과 신뢰성 있는 행동을 유도할 수 있을 정도로 무거운 처벌이 있어야 한다. 이 법을 통해 AI뿐 아니라 AI를 구축하고 제어하는 사람과 기업에도 제약을 가해야 한다. 그러지 않으면 규정은 필자가 언급했던 범주의 오류를 똑같이 다시 범하게 될 것이다.

AI 개인비서의 친밀한 성격상, 신뢰성을 담보하기 위해서는 규정 이상의 무언가가 필요할 것이다. 의사, 변호사, 회계사처럼 이 개인비서는 업무 수행을 위해 우리의 정보에 대한 특별 접근권이 필요한 신임 받는 대리인이 될 것이다. 따라서 그런 전문가처럼, AI 개인비서, 혹은 좀 더 정확히 말하자면 그것을 제어하는 기업은 사용자의 최선 이익에 부합하도록 행위해야 할 법적 책임, 즉 수탁 책임을 져야 한다.

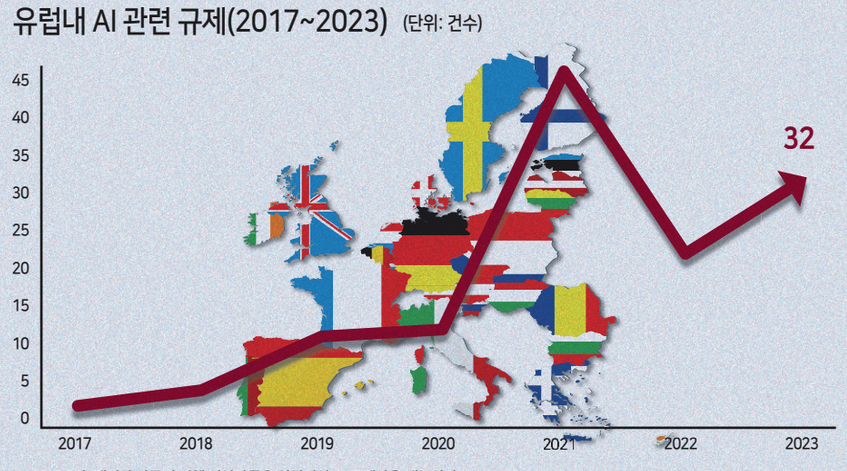
공공 AI 모델도 필요하다. 이것은 공익을 위해 대중이 구축하는 시스템이다. 즉, 개방성과 투명성, 그리고 대중적 요구에 대한 반응성을 고루 갖춘다는 뜻이다. 이런 AI 모델은 AI 혁신에 있어 자유로운 시장의 토대를 마련하는 것은 물론 누구나 구동하고 구축할 수 있어야 한다. 이 모델은 기업 소유 AI와의 균형을 잡아주는 역할을 할 것이다.

우리는 절대 AI와 친구가 될 수 없다. 그러나 AI를 신뢰할 수 있는 서비스, 즉 이중 첩자가 아닌 대리인으로 만들 수는 있다.

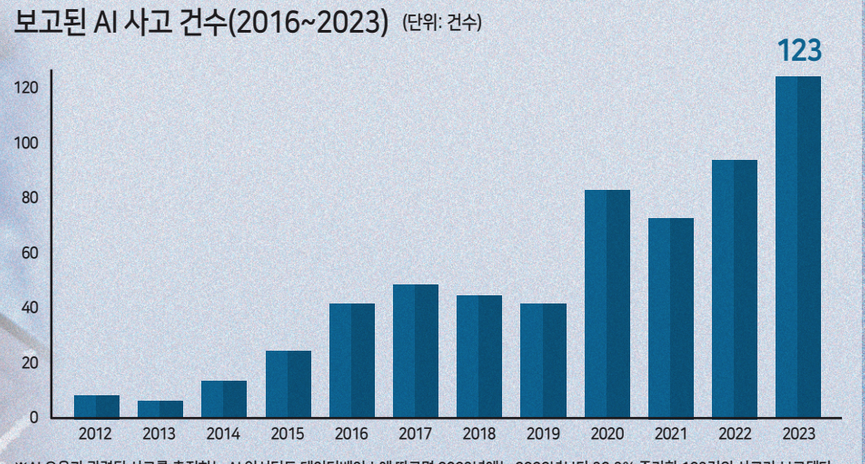
그것이 바로 번영하는 사회를 만드는데 필요한 사회적 신뢰를 쌓는 방법이다.



※미국의 AI 관련 규제는 지난해와 지난 5년 동안 크게 증가했다. 2016년 한 개에 불과했던 AI 관련 규제는 2023년에 25개로 증가했다. 작년에만 AI 관련 규제의 총 수가 56.3% 증가했다.

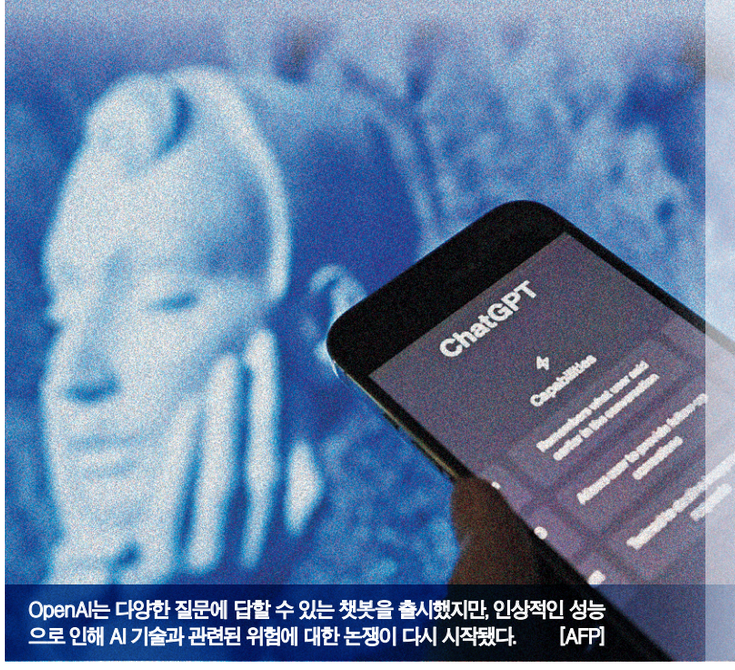


※2023년, 대서양 양쪽의 정책 입안자들은 실질적인 AI 규제안을 내놓았다.
유럽연합은 2024년에 제정된 획기적인 법안인 AI 법의 조건에 대한 합의에 도달했다.
한편, 바이든 대통령은 그해 미국에서 가장 주목할 만한 AI 정책 이니셔티브인 AI에 관한 행정명령에 서명했다.



※AI 오용과 관련된 사고를 추적하는 AI 인시던트 데이터베이스에 따르면 2023년에는 2022년보다 32.3% 증가한 123건의 사고가 보고됐다.
2013년 이후 AI 사고는 20배 이상 증가했다.
주목할 만한 사례로는 온라인에서 널리 공유된 테일러 스위프트의 성적으로 노골적인 딥페이크가 AI로 생성된 것이 있다.

[스탠포드대학교 HAI AI 인덱스 리포트]



OpenAI는 다양한 질문에 답할 수 있는 챗봇을 출시했지만, 인상적인 성능으로 인해 AI 기술과 관련된 위험에 대한 논쟁이 다시 시작됐다. [AFP]

브루스 슈나이어 프로필

'보안 구루'로 불리는 국제적으로 저명한 보안 테크놀로지스트
하버드대학교 버크만 클라인 인터넷 및 사회 센터의 연구원
하버드 케네디 스쿨 공공 정책 강사, 전자 프론티어 재단과 AccessNow 이사
전자 개인정보 정보 센터와 VerifiedVoting.org 자문위원
Inrupt, Inc.의 보안 아키텍처 책임자

편집디자인 전혜경