Deloitte.

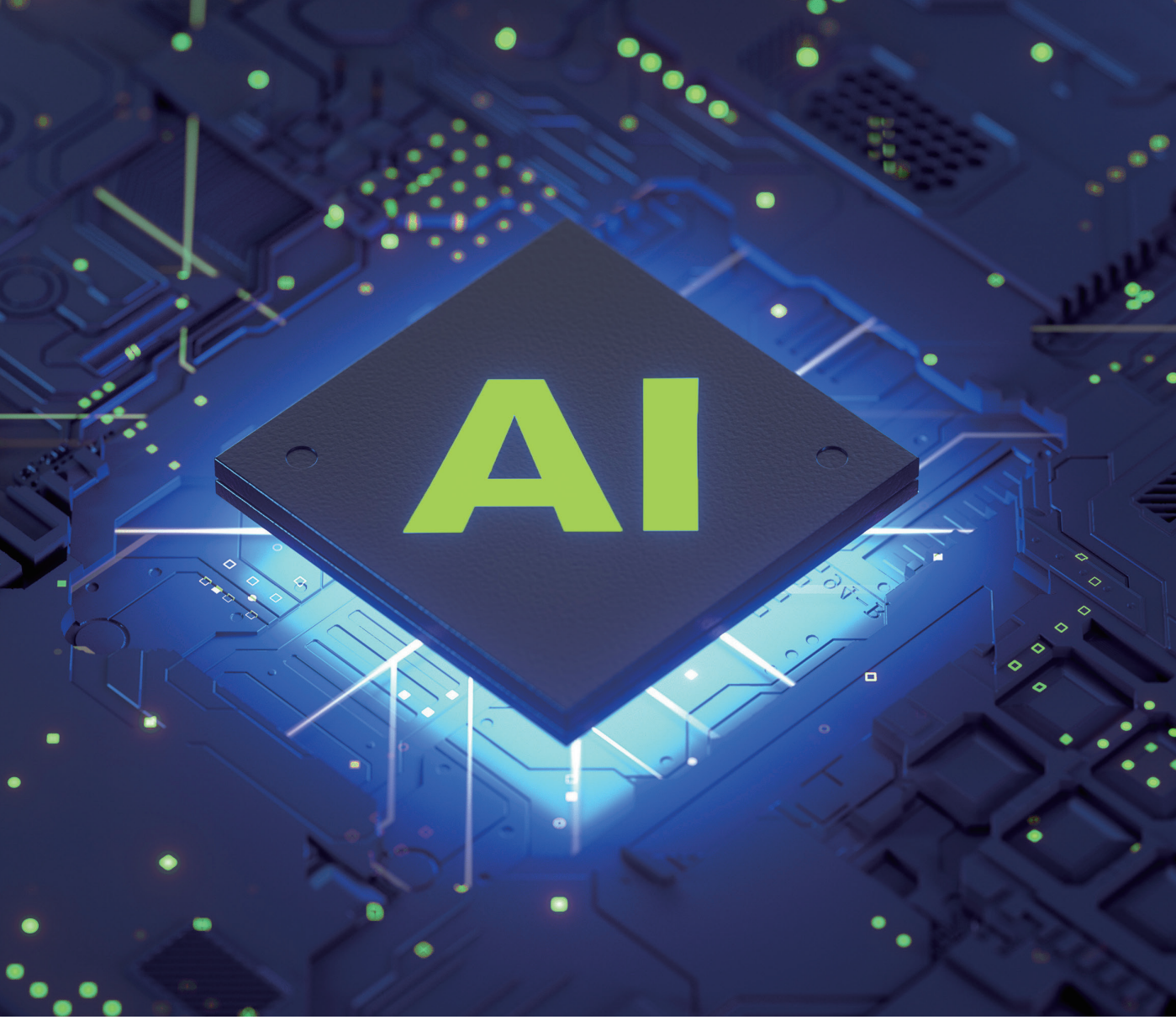
생성형 AI-1편

선제적 리스크 관리가 중요하다

생성형 AI의 윤리, 책임, 신뢰

Deloitte AI Institute

2023년 07월
Deloitte Insights



목차

01 AI에 대한 대중의 관심을 사로잡은 생성형 AI

일각에서는 생성형 인공지능(AI)을 인터넷 검색의 종말이자 인류의 일과 삶을 혁신적으로 변화시킬 수단으로 본다. 사실 이는 AI가 처음 등장했을 때부터 듣던 얘기다. 최신 애플리케이션은 으레 대중적 열광을 불러일으키는 '핫이슈'가 되기 마련이다.

하지만 생성형 AI는 오늘날 사용되는 대다수 AI 모델과 다르다. 일관성 있는 인간의 언어를 그럴 듯하게 흉내내는 대규모 언어 모델(LLM)이 사용자 프롬프트에 자연어 결과로 대답한다. 게다가 일부 생성형 AI 모델은 배경이 되는 수리나 기술에 대한 이해는커녕 AI에 대한 지식이 없어도 사용하는 데 전혀 문제가 없다.

업계에서는 생성형 AI를 기업용으로 어떻게 활용할 수 있을지에 대한 관심도 뜨거워지고 있다. 모든 인지 도구가 그렇듯 생성형 AI로 얻을 수 있는 결과도 이를 어떻게 사용하느냐에 달려 있다. 그리고 리스크 관리가 바로 어떻게 사용하느냐의 영역에 포함된다. 하지만 생성형 AI의 리스크 관리에 대해서는 아직 깊은 논의가 이뤄지지 않고 있다. **기업 사용자들이 생성형 AI 애플리케이션이 내놓은 답을 신뢰할 수 있는가? 또 이를 아직 신뢰할 수 없다면, 앞으로 신뢰할 수 있도록 하기 위해 어떻게 해야 하는가?** 우선적으로 이러한 질문에 대한 답을 찾아야 할 것이다.

02
AI 분야의 새로운 개척자, 신뢰가치 구축 위한 해결과제들

지금까지 AI는 대체로 업무 자동화, 패턴과 상관관계 포착, 현재 및 과거 데이터에 기반한 정확한 미래 예측 등에 활용됐다. 이제 생성형 AI는 진짜 데이터처럼 보이는 데이터를 창조하도록 설계돼 있다. 다시 말해 생성형 AI는 사람이 만든 인공적인 결과물과 동일한 정확도를 가진 것처럼 보이는 디지털 결과물을 만든다.

예를 들어 자연어 프롬프트는 신경망(neural network)에 명령을 내려 사람이 만든 진짜 이미지와 구분이 불가능한 이미지를 만들게 할 수 있다. 텍스트를 만드는 대규모 언어 모델의 경우 AI는 때때로 원천 정보를 제시함으로써 결과물이 사실임을 강조하고 설득력 있는 문장까지 제시한다. **AI가 '나를 믿으라'고 계속 호소하는 것이다.**

기업의 최고정보책임자(CIO)와 기술 전문가들은 생성형 AI가 사람처럼 '사고'하거나 창의력을 발휘하지 않는다는 것을 알고 있고, **생성형 AI가 내놓은 결과물이 보이는 것처럼 정확하지 않을 수 있다는 것** 또한 알고 있다. 하지만 기술에 대한 전문 지식이 없는 기업의 사용자들은 생성형 AI가 어떻게 작동하는지 또는 그 결과물을 얼마나 믿을 수 있는지 알지 못한다. 기업들이 직면한 이러한 문제는 생성형 AI 기술이 급속도로 발전하기 때문에 더욱 증폭된다. **급격히 진화하는 생성형 AI의 능력을 따라잡기도 벅찬 기업과 최종사용자들이 어떻게 그에 따른 리스크를 예측하고 진정한 신뢰를 바탕으로 새로운 도구를 마음껏 사용할 수 있을까?**

AI에는 본디 신뢰라는 개념이 없다. 따라서 AI에 대한 신뢰라는 것을 구축하려면 AI 거버넌스, 리스크 완화, 그리고 사람-프로세스-기술 간 의식적 합 맞추기(alignment)를 지속적으로 추구해야 한다.

생성형 AI의 신뢰가치는 이를 어떻게 사용하느냐에 달려 있다. 그리고 기업들이 빠르게 발전하는 AI를 적극 도입하기 시작하면서 반드시 해결해야 할 신뢰와 윤리 문제가 대두되고 있다.

01 | '환각'과 '오정보' 관리하기

생성형 AI 모델은 데이터세트(dataset)에 기반해 일관성 있는 언어나 이미지를 만든다. 초기 사용자들은 이러한 AI의 결과물에 놀라움과 두려움을 동시에 느꼈다. 하지만 **자연어 프로그램으로 만들어진 AI의 결과물은 문장과 문법은 그럴 듯하지만 그 내용은 일부 또는 완전히 부정확할 수 있고, 거짓과 진실이 명확해야 하는 진술을 할 때는 완전히 거짓을 말할 수도 있다.** 이러한 자연어 애플리케이션의 리스크 중 하나는 부정확한 결과물이 진실인 것처럼 100% 확신함으로써 사용자를 '환각'(hallucination)에 빠뜨릴 수 있다는 것이다. 심지어 있지도 않은 참조와 출처도 만들어낸다. AI는 죄가 없다. 애초에 AI가 할 일은 사람이 만든 것과 비슷해 보이는 디지털 결과물을 만드는 것이기 때문이다. 하지만 **일관성 있는 데이터와 유효한 데이터가 반드시 같지 않을 수도 있다.** 따라서 대규모 언어모델을 사용할 때에는 AI의 유려한 결과물이 사실적 가치가 있는지를 확인해야 하는 과제가 남아있다.

생성형 모델에는 필연적으로 내재한 편견의 위험도 있다. 훈련에 쓰인 데이터가 편파적인 내용을 담고 있을 가능성이 얼마든지 있기 때문이다. 생성형 AI 모델 훈련에 필요한 수십 테라바이트에 달하는 데이터는 매우 광범위하고 양이 방대하기 때문에, 어떤 기업도 독자적으로 이러한 데이터를 수집하고 정리할 수 없다. 생성형 모델은 공개 데이터로 훈련할 수도 있지만, 이러한 데이터에는 편견이 내재돼 있을 위험이 있기 때문에 AI 또한 편견에 치우친 결과물을 내놓을 수 있다.

가장 중대한 리스크는 사용자가 잘못되거나 편견 섞인 AI 결과물을 완전히 신뢰한 채로 결정을 내리거나 행동을 개시할 때 발생한다. 이러한 리스크를 완화하려면 AI 거버넌스를 수립해야 한다. 인력의 업스킬링[1], AI 모델 전주기[2]에 걸친 각 단계별 의사결정, 체계적 감독, 유비쿼터스 문서기록 등 여타 AI 모델에 대한 신뢰를 구축하기 위해 적용한 거버넌스(Trustworthy AI™)를 생성형 모델에도 적용할 수 있다.

02 | 귀속(저작권) 문제

생성형 AI는 훈련받은 데이터와 같은 내용의 결과물을 내놓는다. 그리고 온라인 백과사전, 디지털 서적, 고객 리뷰뿐 아니라 선정된 데이터세트 등이 포함된 이러한 데이터는 귀속(attribution)[3]과 저작권(copyright)이 법적으로 보호되는 현실 세계에서 인간이 만드는 것이다. 따라서 **생성형 모델이 정보의 정확한 출처를 제시한다 하더라도, 귀속을 존중하지 않는 결과물을 내놓거나 더 심할 경우 대놓고 표절 및 저작권과 상표권 침해 행위를 할 수 있다.**

사람이 만든 데이터를 학습해 이를 그대로 모방하면서 사람의 창의력을 흉내내는 도구가 등장한 지금, 귀속 문제를 어떻게 다룰 것인가? 대규모 언어모델이 콘텐츠를 표절하고 기업들이 이를 사업 용도로 사용한 것으로 드러날 경우, **표절의 책임은 AI가 아니라 사람이 지게 된다.** 기업들은 이러한 위험을 방지하기 위해 저작자 표시가 올바르게 되도록 견제와 평가 시스템을 구축할 수는 있다. **하지만 사람이 귀속을 일일이 확인, 평가해야 한다면 생성형 AI를 쓰는 것이 생산성을 끌어올리는 데 무슨 효용이 있을까?**

AI 결과물 내 귀속을 신뢰할 수 있느냐와 사람이 어디까지 감독해야 하느냐 사이 균형을 맞추는 것은 쉽게 해결되지 않을 문제로 남을 것이다. 이는 기업들에게 중대한 법적 및 브랜드 이미지 문제가 될 수 있다.

03 | 투명성과 설명가능성 제공

최종사용자는 대규모 언어모델을 뒷받침하는 복잡한 기술은 커녕 AI 기본 기술에 대한 이해가 부족한 경우가 대부분이다. 하지만 **일반인의 기술적 이해가 낮다고 해서 기업들이 생성형 AI 모델의 투명성(transparency)과 '설명가능성'(explainability)을 소홀히 해도 되는 것은 아니다.** 일반의 이해도가 낮기 때문에 이 사안은 오히려 더욱 중요하다.

현재 상용되는 생성형 AI 모델은 결과물이 부정확할 수도 있다는 면책조항을 제시하기 때문에 얼핏 투명성을 담보하는 것처럼 보인다. 하지만 실제로 사용약관을 제대로 읽는 최종사용자는 거의 없고 대부분은 AI 기술이 어떻게 작동하는지 모르기 때문에, 대규모 언어모델은 설명가능성이라는 문제를 수반할 수밖에 없다. 따라서 최종사용자들이 스스로 위험을 관리하고 윤리적 의사결정을 내리도록 하기 위해서는, 생성형 AI의 작동 원리뿐 아니라 그 한계와 능력, 수반되는 리스크에 대해 비전문가도 이해할 수 있는 설명이 제공돼야 한다.

이제 전사적 차원의 AI 문맹 타파와 AI 리스크 인식 제고는 기업의 내부 운영에 필수 요소가 되고 있다. 생성형 AI가 기업용 업무 도구로 본격 도입되면 이는 더욱 중요한 해결과제로 부상할 것이다. AI 업무도구를 사용함으로써 발생하는 위험과 결과에 대한 책임을 AI 엔지니어나 데이터 과학자가 아니라 최종사용자가 져야 하기 때문이다. 따라서 기업의 최종사용자들이 생성형 AI를 제대로 이해하는 것이 매우 중요하다. 각 기업의 CIO와 리더들은 전사적 AI 이해를 제고하기 위해 최종사용자들인 직원들에게 훈련과 학습 프로그램, 교육과정 등을 제공하고 지속적인 학습이 이뤄지도록 조직 문화를 조성할 필요가 있다.

03
생성형 AI의 리스크 관리 책임은 결국 사람에게 있다

생성형 AI가 사람의 창의력을 흉내내는 능력이 갈수록 발전하고 있지만, 사람-기계 관계에서 사람의 역할과 의미를 잊지 말고 신중하게 균형을 잡아야 한다. 어찌됐건 모든 사람이 생성형 AI의 영향을 받을 것이다. 업무 아웃소싱이나 대규모 정리해고가 발생할 수 있고, 특정 직업에 수반되는 역할이 바뀔 수도 있고, 여러 법적 문제가 제기될 수도 있다. 생성형 AI는 이처럼 현실 세계에 중대한 영향을 미치지만, 정작 주인공인 AI 모델은 자율의지나 의도를 가지는 존재가 아니기 때문에, 문제가 발생했을 때 실질적인 책임을 질 수 없다.

생성형 AI가 대규모로 도입됐을 때, 투명성을 확보하기가 힘들 수 있기 때문에 사람이 계속 개입해야 한다는 것이 치명적인 단점으로 작용할 수 있다. 또한 현재로서는 생성형 AI의 범용화가 어떠한 변화를 가져올지 예측하기가 힘들다. 부정적으로는 가짜 정보가 난무해 객관적이고 온전한 진실이 손상되는 결과가 나타날 수 있다. 이러한 잠재적 문제에도 불구하고 생성형 AI의 범용화는 이제 거스를 수 없는 변화가 됐다.

다만 아무리 막강한 도구가 등장한다 하더라도 AI 여정의 중심에는 항상 사람의 분석, 감시, 상황 인지, 휴머니티가 자리를 잡아야 한다.

이미 도래한 AI 시대는 사람과 기계가 힘을 합치지 않으면 아무것도 이룰 수 없는 진정한 사람-기계 '공생의 시대'(Age of With™)다. 이를 위해 책임, 신뢰, 윤리 문제를 실질적으로 해결할 방법을 알아내는 것이 무엇보다 시급하다. 그래야만 생성형 AI의 결과물과 그 결과물의 창조자인 기업이 진정한 협력관계를 구축할 수 있다.

딜로이트 AI 연구소(Deloitte AI Institute) 소개

딜로이트 AI 연구소는 기업들이 매우 강력하고 역동적이며 빠르게 진화하는 AI 생태계의 다양한 차원을 연결하도록 도움을 주는 조력자 역할을 하고 있습니다. 우리는 날카로운 통찰력으로 산업 전반에 걸쳐 AI 혁신 기술을 적용하기 위한 담론을 주도하며, '공생의 시대'(Age of With™)를 맞아 사람과 기계의 협력을 고취하고자 합니다.

딜로이트 AI 연구소는 AI를 둘러싼 논의와 개발을 뒷받침하고, 혁신을 촉진하며, AI 도입을 가로막는 장애물을 파악해 해결책을 제시합니다. 학계 연구단체, 스타트업, 기업가, 혁신가, 선도적 AI 제품 생산 기업 등 생태계 내 모든 구성원과 협업해 AI에 대한 핵심 영역들, 즉 관련 리스크와 정책, 윤리, 일과 인력의 미래, 응용 AI 활용 사례 등을 탐구하고 있기에 가능한 일입니다. AI 연구소는 딜로이트가 보유한 심도 깊은 AI 지식과 경험을 기반으로 이처럼 복잡한 생태계의 생리를 정확히 파악하여, 기업들이 AI에 대한 유효한 의사결정을 내려 경쟁에서 승리할 수 있도록 실질적인 조언을 제시합니다.

딜로이트 AI 연구소는 기업 이사회 멤버, 고위 경영진, 현직 데이터 과학자 등 여러분이 어떤 역할을 맡고 있든 또한 여러분이 AI 여정의 어떤 단계에 있든지 상관없이, 전 세계 기업들이 AI를 활용해 어떻게 경쟁우위를 확보하고 있는지 배울 수 있도록 생생한 정보를 전달합니다.

딜로이트 AI 연구소(www.deloitte.com/us/AIInstitue)를 방문하시면, 다양한 연구 결과물뿐 아니라 팟캐스트와 뉴스레터, 전문가 담론과 라이브 이벤트 등을 접할 수 있습니다. 딜로이트 AI 연구소와 함께 AI의 미래를 탐구해 보십시오.

주석

1. 업스킬링(upskilling)은 동일한 일의 품질을 높이거나 더욱 복잡한 일을 수행할 수 있도록 인력의 숙련도를 높이는 것을 뜻한다.
2. AI 모델의 전주기(lifecycle)는 통상 1)AI로 해결할 문제와 범위의 정의부터 시작해 2)데이터 수집 및 준비 3)모델 개발 및 훈련 4)모델 평가 및 검증 5)배포 6)모니터링 및 유지관리의 과정으로 이뤄진다.
3. 귀속(attribution)은 저작권의 소유자를 표시하는 행위로, 이에 따르면 창작물의 창작자 이름을 명기해야 하고 창작자의 동의 없이 창작물을 수정 및 개조할 수 없다.

딜로이트 컨설팅 코어테크놀로지 그룹 및 디지털 금융 그룹

딜로이트 컨설팅 코어테크놀로지 그룹 및 디지털 금융 그룹은 테크 전략 설계부터 도입 및 최적화, AI & Data 전문 컨설팅, 클라우드 전환 및 설계, 이행까지의 라이프사이클 전방에서 선도적으로 고객을 지원합니다.

AI & Data 전문 컨설팅 조직은 데이터 표준화, 모델링 및 분석, AI 활용 방안에서 AI 운영 거버넌스 체계 수립 및 ISO 42001 인증 지원까지 엔터프라이즈 고객사의 데이터에 기반한 E2E 서비스를 제공하고 있습니다. 또한 IT에서 재무 및 공급망까지 엔터프라이즈 전반의 시스템과 조직의 특정 기능에 대한 솔루션 투자로 고객이 영향력을 확대하고 가치를 달성할 수 있는 전문 IT 컨설팅 서비스 역량을 보유하고 있습니다.

Contact Point

김우성 파트너
Core Technology 그룹리더
Tel: 02-6099-4670
Email: wooskim@deloitte.com

안상혁 파트너
디지털 금융 그룹 리더
Tel: 02-6676-3625
Email: sanghyan@deloitte.com

최규웅 파트너
Core Technology
Tel: 02-6676-3873
Email: kyuwchoi@deloitte.com

박지숙 파트너
금융 IT, 오퍼레이션 리더 I
딜로이트 컨설팅
Tel: 02-6676-3722
Email: jisukpark@deloitte.com

강기식 상무
Core Technology, chief architect
Tel: 02-6676-2039
Email: gikang@deloitte.com

이성호 이사
Core Technology,
Data 분석 전문가
Tel: 02-6676-3767
Email: sholee@deloitte.com

Deloitte.
Insights

딜로이트 안진회계법인·딜로이트 컨설팅
성장전략본부

손재호 Partner
성장전략본부 리더
jaehoson@deloitte.com

정동섭 Partner
딜로이트 인사이트 리더
dongjeong@deloitte.com

김사현 Director
딜로이트 인사이트 편집장
sahekim@deloitte.com

HOT LINE
02) 6099-4651