PwC Korea Insight Flash

책임 있는 AI

Responsible AI

September 2023 · 삼일PwC경영연구원

pwc 삼일회계법인

Epilogue

작년 11월, 인간의 요구에 따라 데이터를 찾고 학습하며, 추론과 의사결정 능력을 지녀 새로운 텍스트, 이미지, 코드 등을 빠르게 '생성'해낼 수 있는 생성형 AI(Generative AI)가 등장했다. 우리가 기존에 알던 기계는 '단순 노동' 영역에 머물러있었으나, 생성형 AI의 등장으로 이제는 '창조' 영역까지 이르러 현대 사회의 산업과 노동 패러다임을 뒤흔들고 있다.

그러나 기술혁신에는 불확실성과 리스크가 따르는 법. '만능'일 줄 알았던 생성형 AI가 가져올 수 있는 여러가지 윤리적, 사회적 문제가 서서히 드러나며 그에 대한 우려가 깊어지고 있다. 생성형 AI는 질문에 대한 허위답변을 제공하는 문제를 넘어 개인정보 무단도용, 저작권 침해 등의 심각한 보안 문제까지 일으킬 수 있기 때문이다. 그렇다면 우리는 생성형 AI와 같은 최첨단 AI 기술을 어디까지 신뢰할 수 있을 것인가? 그리고 이 기술을 책임감 있고 윤리적으로 사용하기 위해서는 어떻게 해야 하는가? '책임 있는 AI(Responsible AI)'라는 개념에 대한 논의는 이와 같은 의문에서 시작되었다.

책임 있는 AI 개념 등장

Responsible AI

책임 있는 AI는 AI 기술에 대한 '신뢰'와 '윤리'를 추구하기 위해 제시된 방법론이다. 즉, AI 기술을 개발하고 배포하는 과정이 윤리적이고 신뢰할 수 있으며, 사용자의 개인정보와 사회적 가치를 존중하는 방향으로 이뤄지는 방식을 의미한다. 해당 개념의 중요도에 대한 언급은 예전부터 지속적으로 존재해왔으나, 현재 생성형 AI의 시대가 도래하는 시점에서 그 필요성이 더욱 대두되고 있다.

다만, 책임 있는 AI의 실현가능성에 대해서는 아직 의견이 분분하다. 책임 있는 AI는 'AI 거버넌스' 측면에서 떠오르고 있는 분야로, 이의 필요성을 주장하는 쪽에서는 책임 있는 AI 시스템을 기업 거버넌스 체계에 도입시 보다 공평하고, 투명하며, 의지할 수 있는, 최적의 AI practice를 쉽게 이룰 수 있다는 입장이다. 그 반대로, AI에 어떻게 책임감을 갖게 할 수 있느냐 하는 의문이 존재할 뿐더러 아직은 구체적인 실천 지침 없이 다소 추상적인 선언 수준에 머물러 있다는 점도 책임 있는 AI라는 방식이 과연 실효성이 있는건지에 대한 의문을 제기한다.

그럼에도 불구하고, 책임 있는 AI는 생성형 AI 기술이 나날이 발전되고 있는 오늘날 전세계적으로 떠오르고 있다는 건 사실이다. 모든 산업을 통틀어 디지털 혁신이 빠르게 이루어지고 있고 그에 따른 AI 적용도 점차 깊고 정밀하게 이루어질 것으로 예상되는 만큼, AI로 인해 발생할 수 있는 상당한 리스크들을 원천적으로 차단할 수 있는 (현재까지는) 유일한 방안으로 떠오르고 있기 때문이다. AI로 인해 발생할 수 있는 리스크는 개인정보 유출, 작업 오류, 불량 등 어플리케이션 차원의 기술적 리스크와 더불어 기업 명성, 인력 대체, 디지털 격차 등의 기업 및 국가 차원의 리스크까지 아우른다. 이와 같은 위험 요소들을 책임 있는 AI 도입을 통해 해결할 수 있을 것이라 믿는 기업 또는 기관들은 그들만의 '책임 있는 AI 원칙(Responsible AI principles)' 또는 'AI 거버넌스 프레임워크(AI governance framework)'를 구축하여 이를 기반으로 AI 모델을 도입하고 있다.

그림1. AI 리스크 종류

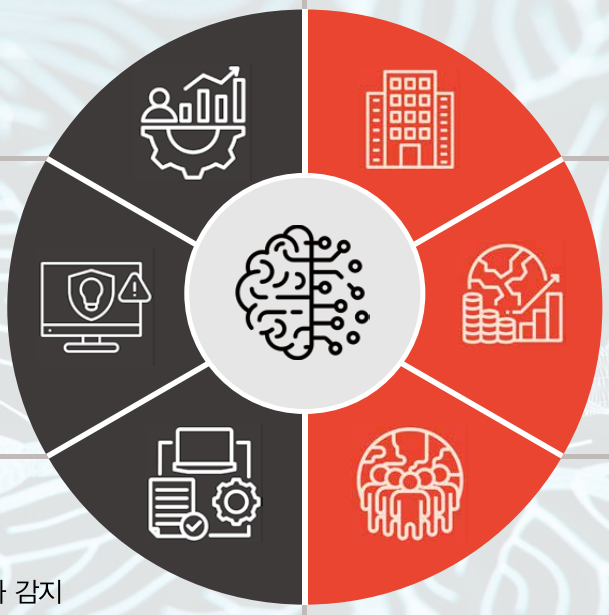
어플리케이션 차원의 리스크

기업 및 국가 차원의 리스크

성과 리스크
- 오류
- 편견
- 불투명성
- 성능 불안정

보안 리스크
- 적대적 공격
- 해킹&개인정보 유출 리스크
- 오픈소스 소프트웨어 리스크

제어 리스크
- 인간 개입 결여
- 불량 AI와 비의도적 결과 감지
- 명확한 책임의 결여



기업 리스크
- 명성 영향
- 재정적 성과 영향
- 법·규율의 준수 문제
- 기업 가치 비일치

경제적 리스크
- 인력 대체
- 불평등 강화
- '승자독식' 경제 불균형

사회적 리스크
- 잘못된 정보와 조작된 정보
- 디지털 격차
- 감시와 갈등

자료: PwC, 삼일PwC경영연구원

표1. 글로벌 기업·기관의 책임 있는 AI 원칙과 프레임워크 도입 현황

기업·기관	책임 있는 AI 원칙과 프레임워크
Microsoft	① 공평성 (Fairness) ② 신뢰와 안전 (Reliability and safety) ③ 개인정보보호와 보안 (Privacy and security) ④ 포용성 (Inclusiveness) ⑤ 투명성 (Transparency) ⑥ 책무성 (Accountability)
Google	① 사회적으로 유익해야 할 것 (Be socially beneficial) ② 불공평한 편견을 일으키는 것을 피할 것 (Avoid creating or reinforcing unfair bias) ③ 안정성 테스트를 거칠 것 (Be built and tested for safety) ④ 사람한테 제공하거나 지시받는 모든 것에 대해 책임을 가질 것 (Be accountable to people) ⑤ 개인정보보호 설계 원칙을 도입할 것 (Incorporate privacy design principles) ⑥ 높은 수준의 과학적 우수성을 지닐 것 (Uphold high standards of scientific excellence) ⑦ 이러한 원칙에 부합하는 용도로 사용될 수 있도록 할 것 (Be made available for uses that accord with these principles)
IBM	① AI의 목적은 인간의 지능을 강화하기 위함이다 (The purpose of AI is to augment human intelligence) ② 데이터와 그와 연관된 인사이트는 그 주인의 것임을 명시해야 한다 (Data and insights belong to their creator) ③ 기술은 투명하고 설명가능해야 한다 (Technology must be transparent and explainable)
NIST (미국국립 표준기술원)	① 책임 있고 투명한 (Accountable and transparent) ② 설명가능하고 해석가능한 (Explainable and interpretable) ③ 공평하며 해로운 편견은 관리 (Fair with harmful bias managed) ④ 향상된 보안 (Privacy-enhanced) ⑤ 안전하며 회복이 빠른 (Secure and resilient) ⑥ 유효하며 의지할 수 있는 (Valid and reliable) ⑦ 사람 또는 환경에 해롭지 않은 (Safe)

자료: 각 사, 삼일PwC경영연구원

PwC의 책임 있는 AI 프레임워크 및 거버넌스 모델

PwC는 기업이 생성형 AI의 잠재적 리스크를 책임 있는 AI를 통해 효과적으로 다루기 위해서는 책임 있는 AI 시스템을 자사 AI 전략의 핵심적인 기능 중 하나로 도입해야한다고 주장한다. 이에 PwC도 책임 있는 AI 프레임워크를 설계했으며, 이는 AI 라이프사이클에 폭넓게 적용되는 Trust-by-design(서비스 · 상품 설계 시작 단계부터 리스크 최소화 및 신뢰도 향상 원칙을 녹여 그 자체로 혁신적인 'risk-free' 기술이 되도록 하는 방식) 형태로, 기업내 모든 레벨의 임직원들의 역할과 긴밀히 맞닿아 있는 AI 도입 방안을 제시했다.

이를 세부적으로 설명하면: CEO와 이사회는 공공 정책 개발 및 기업의 목적과 가치를 중심으로 전략을 수립하고, 위험 및 준법 관리 최고책임자(Chief risk and compliance officer)는 거버넌스, 준수 및 리스크 관리를 포함한 관리 · 통제 역할을 맡는다. 정보 보안 최고책임자(Chief information and information security officer)의 경우 사이버 보안, 개인정보보호 및 성과 등의 '책임 있는 practice'와 관련된 활동을 통솔한다. 그리고 마지막으로 데이터 전문가 및 비즈니스 도메인 전문가들은 Use case(시스템의 동작을 사용자의 입장에서 표현한 시나리오) 개발, 이슈 구체화, 아웃풋 모니터링, 검증 등을 수행하는 과정에서 '책임 있는 core practice'를 적용시킨다. 이 모든 일련의 과정들의 모여 하나의 책임 있는 AI 프레임워크가 만들어지는 것이다.

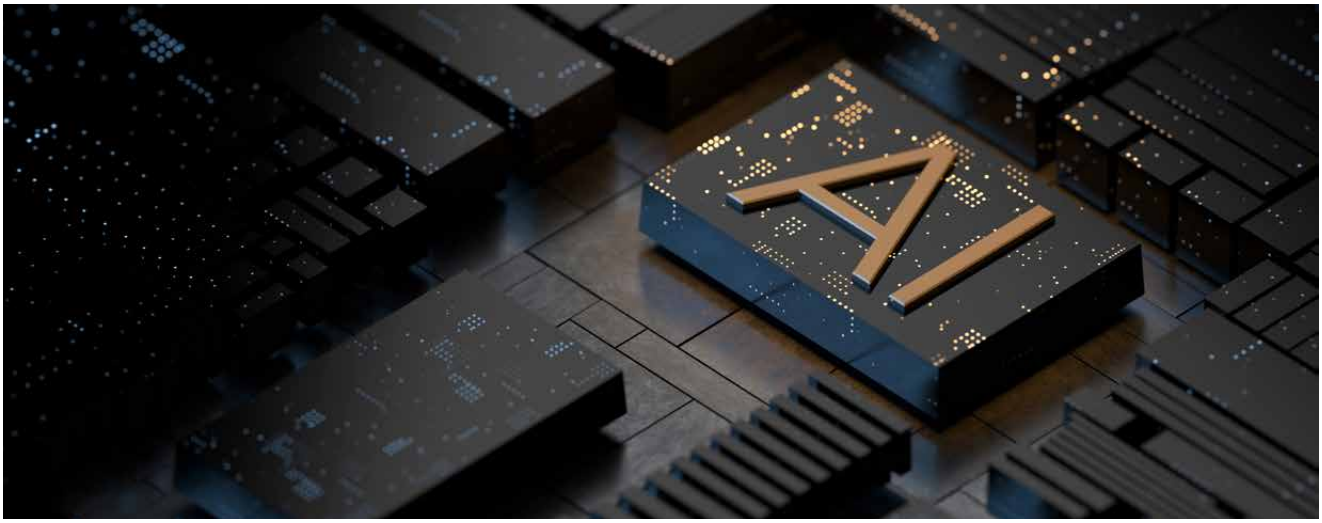


그림2. PwC의 책임 있는 AI 프레임워크

전략
CEO·이사회

데이터 & AI 윤리
데이터 및 AI 활용 에 따른 윤리적 영향을 인지하고, 이를 기업 가치에 녹여낼 것

정책 & 규제
주요 공공 정책과 규제 트렌드를 예측하고 이해하여 이를 성실히 준수·이행할 것

관리·통제
위험 및 준법 관리 최고 책임자

거버넌스
3차 방어선까지 관리할 수 있는 시스템을 구축할 것

규정 준수
규정, 조직 정책, 산업 표준을 준수할 것

위기 관리
리스크 감지 및 경감책을 확장하여 AI에 생소한 위험마저 관리할 수 있도록 할 것

책임있는 관리 제도 실행
정보 보안 최고 책임자

해석가능성 & 설명가능성
투명한 의사결정 모델을 구축할 것

지속 가능성
환경에 미치는 악역향을 최소화하고 사람의 권리와 권한을 강화할 것

강건성
고성능의 신뢰할 수 있는 시스템을 구축할 것

편향성 & 공정성
공정성을 정의, 측정하고 관련 표준과 비교해볼 것

보안
시스템의 사이버 보안 기능을 향상시킬 것

개인정보보호
데이터 보안을 위한 시스템을 개발할 것

안전
물리적 피해를 방지하기 위한 시스템을 설계하고 테스트할 것

핵심 관리 제도 실행
데이터 전문가·비즈니스 도메인 전문가

문제 구체화
해결해야 하는 문제에 대해 정확히 이해하여 AI/ML 솔루션이 요구되는 사항인지 파악할 것

표준 규격화
산업 표준 및 모범 사례들을 본보기 삼을 것

입증
모델 성과를 평가하고 모델 설계 및 개발 과정을 반복 수행하여 성과치를 높일 수 있도록 할 것

모니터링
잠재적 리스크를 발견하기 위해 지속적인 모니터링 시스템을 구축할 것

자료: PwC, 삼일PwC경영연구원

또한, PwC는 기업이 성공적으로 책임 있는 AI 시스템을 도입할 수 있게끔 기업 맞춤형 end-to-end 거버넌스 모델을 마련했다. 이 모델은 명확한 역할과 그에 따라 요구되는 기대사항, 활동 추적 및 지속적인 평가를 위한 메커니즘을 기반으로 기업의 top-to-bottom AI journey에 대해 가시성을 부여했다. 이를 통해 기업들은 책임 있는 AI 모델에 대한 평가 – 구축 – 입증 및 배포 – 평가 및 모니터링까지의 과정을 빠짐없이 수행할 수 있게 되는 것이다.

그림3. PwC의 책임 있는 AI 거버넌스 모델

자료: PwC, 삼일PwC경영연구원

‘책임 있는’ 생성형 AI 활용을 위한 고려사항

기업들이 생성형 AI 활용 방안과 책임 있는 AI 적용 방안에 대해 탐구하는 단계에서는 반드시 생성형 AI가 갖고 있는 주요 리스크 사항에 대해 염두에 두고 있어야 한다. 생성형 AI는 기존의 AI 기술 대비 ‘창작’할 수 있는 능력이 있기에, 그에 따른 리스크도 차원이 달라지기 때문이다. 생성형 AI는 기업 브랜드를 폄하할 수 있는 편향적이고도 모욕적인 콘텐츠를 생성할 수도 있으며, 보다 설득력 있는 콘텐츠 생성을 통해 사이버 위협을 가하거나, 기업의 지적 재산권을 유출할 수도 있다. 기업들은 나날이 발전해나가는 생성형 AI 기술이 가지는 장단점에 대해 면밀히 분석해보고 이해해야 보다 체계적인 책임 있는 AI 시스템을 설계할 수 있을 것이다.

표2. 생성형 AI의 주요 리스크

리스크	상세
편향적, 모욕적, 부정확한 콘텐츠 생성	생성형 AI가 보유한 데이터에 내제되어 있는 편견을 학습하여 ‘기업의 이름을 달고’ 편향적이며 모욕적인 부정확한 콘텐츠를 생성
블랙 박스 (기능은 알지만 작동 원리를 이해할 수 없는 장치)	생성형 AI의 대다수는 전문적인 제3자 기관이 제작한 ‘파운데이션 모델’을 기반으로 작동. 해당 모델의 내부 작동 방식에 접근할 수가 없어, AI가 특정 결과를 어떻게 도출했는지 온전히 이해할 수 없음
고도화된 사이버 위협	생성형 AI는 경영진의 이메일, SNS 게시물, 영상 출연 클립 등을 분석해 그들의 문체, 말투, 표정을 따라 할 수 있음. 이를 사용해 관련 회사에서 만든 것처럼 보이는 이메일이나 영상을 생성해, 잘못된 정보를 퍼트리거나 이해관계자들에게 기밀 정보를 요구하는 방식으로 악용될 수 있음
성능 저하를 초래할 수 있는 ‘환각 현상’	생성형 AI가 생성한 답변이 틀린 정보일 수 있으나, 생성형 AI는 이를 옳다고 제시하며 설득력 있는 답을 지어낼 수 있음
‘훔친’ 데이터 기반의 위험한 모델	생성형 AI 작동을 위해 사용되는 데이터에 대한 명확한 출처를 확인하기 어려움. 저작권 보호를 받는 문서, 이미지, 코드를 복제 및 재생산할 수 있으며, 이는 지적재산권 도용으로 이어져 벌금형, 소송 제기, 브랜드 가치 저하로 이어질 수 있음
지적 재산권 유출	주의를 기울이지 않는다면, 자사의 중요한 정보가 경쟁자들의 콘텐츠 생성에 활용되는 경우를 목격할 수도 있음. 사용자가 생성형 AI 모델에 입력하는 자료들이 데이터베이스로 수집되고, 공유되어 활용될 수 있기 때문

자료: PwC, 삼일PwC경영연구원

생성형 AI의 리스크를 파악하고 이를 해결할 수 있는 방안을 어느정도 세웠다면 다소 훌륭한 책임 있는 AI 체계를 마련했다고도 볼 수 있겠다. 그러나 점점 더 빨리 진화하는 IT 환경에 맞춰 책임 있는 AI를 적용하기 위해서는 주의 깊게 관찰하고 중요시 여겨야 할 요소들이 아직 남아있다. 이를 위해 취해야할 주요 action은 다음과 같다:

01 위기관리 기반 우선순위를 도출하라	기업 주주들의 입장에서 특정 생성형 AI 리스크가 남들보다 더 중요한 사항이 될 수도 있다. 거버넌스, 규정 준수, 위기 관리, 내부 회계 감사, 그리고 AI 팀 모두가 가장 중대한 리스크에 가장 많은 관심을 줄 수 있도록 하는 보고 체계를 마련하라
02 사이버, 데이터 및 개인정보보호책을 개편하라	사이버 보안, 데이터 관리 및 개인정보보호 정책 갱신을 통해 해커들이 생성형 AI를 통해 개인 정보와 신원을 유출하고 미래 사이버 공격을 계획하는 리스크를 선제적으로 예방하라
03 불투명성의 리스크를 해결하라	일부 생성형 AI 시스템에 관해서는 '왜' 특정 결과물을 도출했는지, 그 과정을 이해하고 설명하는 것이 거의 불가능하다. 이러한 시스템들을 파악하여 시스템의 공정성, 정확성, 준수성을 개선하기 위한 방안이 무엇이 있는지를 고민해볼 필요가 있다
04 이해관계자들의 책임 있는 활용 · 관리가 가능하도록 준비시켜라	예를 들어, 생성형 AI를 사용해야하는 직원들에게는 기본 작동 원리, 사용 방법, 아웃풋을 입증, 개선하는 방법 등을 안내하라. 그리고 규정 준수와 법률 리스크를 담당하는 팀에게는 지적 재산권 위반, 기타 연관 위기들을 식별하기 위해 필요한 기술 및 소프트웨어를 제공하라
05 제3자를 지속 모니터링하라	어느 벤더가 생성형 AI 기반의 콘텐츠와 서비스를 제공하는지 알고, 그들이 AI 리스크를 어떻게 다루는지와, 자사가 잠재적 피해에 노출될 수 있는 가능성에 대해 미리 파악하라
06 규제 현황을 살펴라	전세계의 정책 입안자들이 AI 개발 및 활용에 대해 더 많은 규정들을 제정 중이다. 이들은 아직 미완성 단계이나, 새로운 규정들이 지속적으로 생성되는 중임을 인지해야 한다
07 자동화된 관리감독 시스템을 도입하라	생성형 AI 기반의 콘텐츠들이 만연한 오늘날, 새로운 소프트웨어 도구를 활용하여 AI 생성 콘텐츠를 식별하고, 아웃풋을 검증하며, 편향성 또는 개인정보 침해 여부에 대해 평가하고 필요에 따라 개선 사항을 반영하라

Author Contacts

삼일PwC경영연구원

최재영 삼일PwC경영연구원장
jaeyoung.j.choi@pwc.com
+82-2-709-8820

이은영 Managing Director
eunyoung.lee@pwc.com
+82-2-709-0824

오선주 Senior Manager
sunjoo.oh@pwc.com
+82-2-3781-9344

강서은 Manager
seoeun.kang@pwc.com
+82-2-3781-9137

최형원 Senior Associate
hyungwon.choi@pwc.com
+82-2-3781-9638

신서윤 Assistant Associate
seoyoon.shin@pwc.com
+82-2-3156-5334

www.samil.com

S/N: 2309W-RP-038