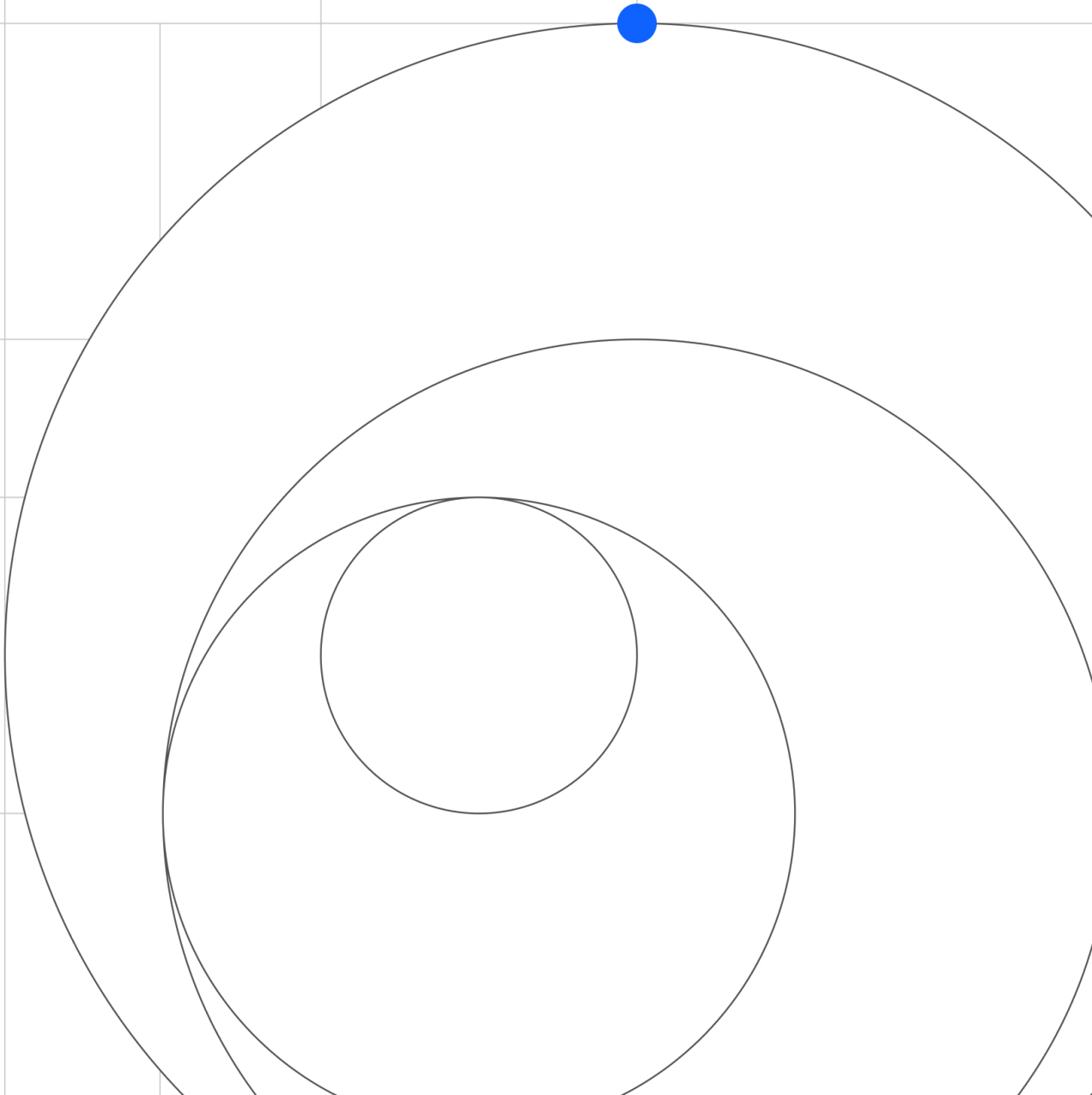
IBM AI 윤리 위원회 (IBM AI Ethics Board)

# 파운데이션 모델: 기회, 위험 및 완화



IBM

# 속성

AI 윤리 위원회 워크스트림의 총괄 후원인 Christina Montgomery와 Francesca Rossi, 그리고 워크스트림 구성원 Betsy Greytok, Bryan Bortnick, Catherine Quinlan, David Piorkowski, Eniko Rozsa, Heather Domin, Heather Gentile, Jamie VanDodick, Jill Maguire, John McBroom, Joshua New, Justin Weisz, Katherine Fick, Kevin Black, Kush Varshney, Manish Bhide, Manish Goyal, Melis Kiziltay, Michael Epstein, Michael Hind, Milena Pribic, Phaedra Boinodiris, Rogerio Abreu de Paula, Saishruthi Swaminathan, Suj Perepa께 감사드립니다.

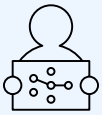
# 목차

# 개요

파운데이션 모델의 등장으로 인해 기업에 흥미진진하면서도 새로운 가능성이 제시되었으며 윤리적 설계, 개발, 배포 및 사용에 관해 새로운 질문들이 부상했습니다. 최근 IBM 기업 가치 연구소(IBM Institute for Business Value)에서 시행한 생성형 AI 설문 조사에 따르면, 기업들은 이미 신뢰 관련 문제에 우려를 표명하고 있었습니다. 이는 특히 투자 장벽이라는 측면에서 두드러졌습니다. 가장 큰 우려 사항은 사이버 보안(57%), 개인 정보 보호(51%), 정확성(47%)이었습니다. 생성형 AI의 소비화를 앞두고 이러한 우려를 심각하게 받아들이는 기업이 여럿이며, 향후 3년 간 AI 윤리에 40% 이상은 더 투자하겠다는 의사를 표명했습니다. 위험을 인식하고, 이를 완화하는 방법을 아는 것은 신뢰할 수 있는 AI 시스템을 구축하는 데 첫 번째로 중요한 단계입니다.

이 문서에서는 다음을 수행합니다.

까다로운 작업을 수행할 수 있는 능력, AI 도입 속도를 높일 수 있는 잠재력, 생산성을 높일 수 있는 능력, 비용 이점 등 파운데이션 모델이 주는 유익에 관해 살펴봅니다.



초기 형태 AI에서 알려진 위험, 파운데이션 모델에 의해 증폭된 알려진 위험, 파운데이션 모델의 생성 기능에 내재된 새로운 위험, 이 세 가지 범주의 위험에 대해 논의합니다.

IBM의 AI 윤리 이니셔티브의 토대를 이루는 원칙, 핵심 요소, 거버넌스를 다루며 위험 완화 가드레일을 제안합니다.

# 들어가며

AI 사용이 계속 확대되면서 복잡한 대규모 AI 모델은 전도유망한 성능을 제공할 뿐만 아니라 가장 어려운 사회 문제도 해결하고 있습니다. 그러나 각 AI 애플리케이션을 위해 대규모 교육 데이터 세트와 복잡한 모델을 구축하려면 기업에 부담이 될 수 있습니다. 파운데이션 모델은 사용 사례마다 새 모델을 교육하는 대신, 강력한 최첨단 모델을 구축하여 이를 직접 재사용하거나 조정하여 다양한 사용 사례를 구현하도록 그 방법을 알려줍니다. 예를 들어, IBM Research는 비전 검사를 위한 파운데이션 모델을 개발했습니다. 이러한 파운데이션 모델은 콘크리트 표면 및 활주로의 일반적인 모습을 학습하여 균열 감지 등의 특정 사용 사례에 맞게 추가 조정할 수 있습니다. 또한 레이블이 적은 데이터로 결함을 검사할 수도 있습니다.

IBM은 파운데이션 모델이란 다양한 다운스트림 작업에 적용할 수 있는 AI 모델이라고 정의합니다. 파운데이션 모델은 일반적으로 자체 감독을 사용하여 레이블이 지정되지 않은 데이터에서 학습되는 대규모 생성 모델입니다. 대규모 모델이므로 매개변수가 수십억 개 있을 수 있습니다.

IBM은 AI 윤리를 준수하는 책임감 있는 데이터 관리자로 오랫동안 명성을 쌓아온 하이브리드 클라우드 및 AI 기업입니다. 연구, 제품 및 컨설팅 팀, 그리고 Hugging Face 등 외부 파트너의 강점을 활용하여 고객에게 파운데이션 모델의 강력한 기능을 제공하는 동시에 모든 기업이 신뢰할 수 있는 AI를 구축할 수 있도록 지원합니다. 또한 IBM은 감사 가능하고 신뢰할 수 있는 방식으로 작동하는 AI 모델을 설계하고 개발하기 위해 IBM watsonx™ AI 및 데이터 플랫폼과 기술과 같은 새로운 플랫폼을 구축하는 데 지속적으로 투자하고 있습니다.

이 문서에서는 파운데이션 모델 윤리에 대한 IBM의 관점을 설명합니다. 이는 첫 번째 버전이며 향후 버전에서는 IBM의 파운데이션 모델 윤리 접근 방식이 다양한 측면으로 확장될 것입니다. 모든 이해 관계자가 책임감 있게 파운데이션 모델을 개발, 배포 및 사용하는 데 이 문서가 도움이 되기를 바랍니다.

# 파운데이션 모델의 이점

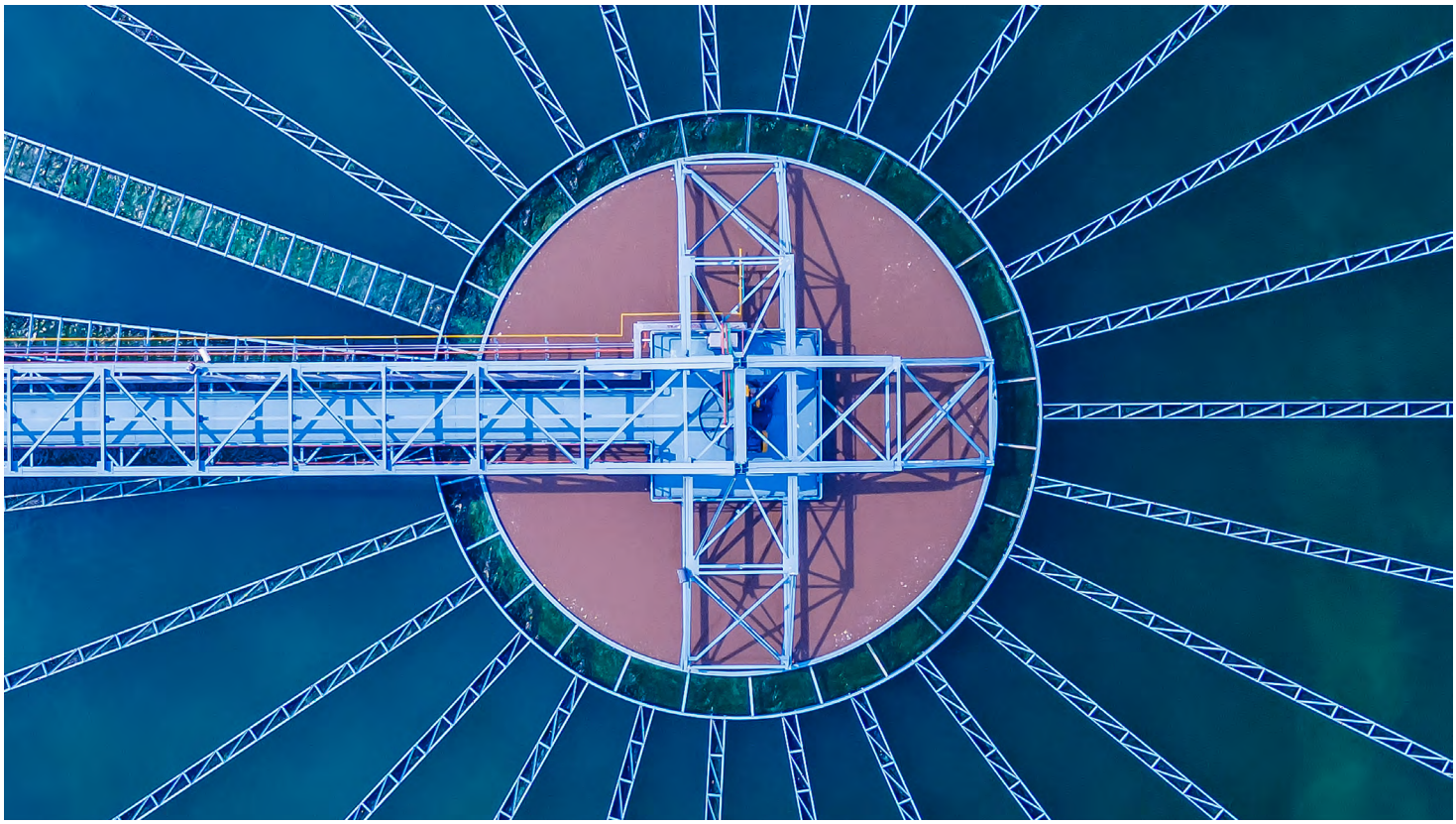
파운데이션 모델은 AI 시스템 개발 프로세스를 크게 개선하고 탐색 단계부터 기업에 도입하는 단계까지 AI를 발전시키도록 도와주며, 다음과 같은 이점이 있습니다.

**복잡한 작업 수행**
파운데이션 모델은 어렵고 복잡한 문제를 해결하는 데 있어 크게 향상된 성능을 보여 줍니다. 예를 들어, IBM과 NASA가 함께 제작한 지리 공간 파운데이션 모델은 NASA의 위성 데이터를 홍수나 기타 지형 변화 등을 보여주는 재해 지도로 변환하도록 설계되었습니다. 이 모델은 지구의 과거를 밝히고, 악천후로 인한 농작물, 비즈니스 또는 인프라의 위험을 예측하고, 기후 변화에 적응하는 전략을 개발하고, 농업 비즈니스를 지원하는 데에도 사용될 수 있습니다. 이 모델은 IBM Environmental Intelligence Suite를 통해 IBM 고객에게 미리 공개될 예정입니다.

또 다른 예로, IBM의 Molformer-XL이 있습니다. 단순한 형태로부터 분자 구조를 추론하고 분자의 물리적 및 양자적 성질을 예측하며 유사한 분자를 식별하고 새로운 사용 사례를 위해 이미 승인된 분자를 선별하며 새로운 분자를 발견하는 등 다양한 다운스트림 작업을 쉽게 학습하는 파운데이션 모델입니다. Moderna 및 IBM은 분자 특성을 예측하고 잠재적 mRNA 의약품의 특성을 이해하는 데 도움이 되도록 MolFormer를 사용하는 방법을 모색하고 있습니다.

**생산성 향상**
파운데이션 모델이 갖고 있는 생성적 특성은 기업이 AI를 통해 일상적이고 지루한 작업을 자동화하며, 사용자가 창의적이고 혁신적인 작업에 더 많은 시간을 할애하도록 하므로 생산성을 향상할 수 있는 영역을 넓혀 줍니다. 예를 들어, 파운데이션 모델을 기반으로 하는 IBM watsonx Code Assistant는 다양한 경력 수준을 갖춘 개발자로 하여금 AI가 생성한 권장 사항을 사용하여 코드를 작성하도록 지원합니다.

**가치 창출 시간 단축**
파운데이션 모델은 일반적으로 레이블이 없는 데이터를 통해 학습되며, 이로 인해 레이블이 있는 데이터보다 더 많은 데이터에 쉽게 액세스할 수 있습니다. 일단 학습되면, 파운데이션 모델은 레이블이 지정된 소량의 데이터를 통해 바로 사용하거나 다운스트림 애플리케이션에 맞게 조정한 후 사용할 수 있습니다. 그러므로 가치 창출 시간을 단축할 수 있습니다.



**다양한 데이터 양식 활용**
자연어, 텍스트, 이미지, 오디오와 같은 다양한 데이터 양식을 사용하여 파운데이션 모델을 학습할 수 있습니다. 또한 시계열 데이터, 지리 공간 데이터, 표 형식 데이터, 반정형 데이터, 텍스트와 이미지가 결합한 혼합 양식 데이터 등 다양한 데이터 유형이 필요한 작업에도 적용할 수 있습니다.

**상각 비용**
파운데이션 모델을 훈련하는 데 드는 초기 비용은 기존 AI 모델을 훈련하는 비용보다 훨씬 높지만, 새로운 작업에 적용하는 데 드는 증가 비용은 훨씬 낮습니다. 사전 학습된 파운데이션 모델을 사용하면 기업이 새로운 기능을 실험하기 위해 파운데이션 모델을 학습시키는 데 막대한 투자를 할 필요가 없습니다. 기업의 경우 모델의 신뢰성, 에너지 효율성, 성능, 이식성, 엔터프라이즈 데이터를 효과적이고 안전하게 사용할 수 있는 기능이 가장 중요합니다.

IBM은 기업이 비즈니스에 필요한 파운데이션 모델의 가치를 창출하고 소유하도록 지원합니다. 개방형 글로벌 AI 커뮤니티에서 최고의 혁신 기술을 도입하고, 하이브리드 컴퓨팅 환경에서 효율적으로 실행하며, 위험을 완화하면서도 AI를 엄격하게 관리하기 때문입니다.

# 파운데이션 모델의 위험

빠르게 발전하는 모든 기술과 마찬가지로, 파운데이션 모델에도 이점과 함께 위험이 따릅니다. 데이터 이동 또는 사용 제한 등의 법적 위험이 있을 수 있으므로 현행 및 변화하는 법률에 따라 신중하게 평가해야 합니다. 기타 위험에는 윤리적 특성이 있으므로 기술이 긍정적인 영향을 미치도록 신중하게 고려해야 합니다. 일반적으로 AI 위험은 사회 기술적인 문제를 제기하며 소프트웨어 툴, 위험 평가 프로세스, AI 윤리 프레임워크, 거버넌스 메커니즘, 다중 이해관계자 협의, 표준 및 규제와 같은 사회 기술적인 방법을 통해 해결하고 완화해야 합니다. 다음 세 가지 범주를 고려하여 위험을 살펴보겠습니다.

1. **기존 위험.** 사전 또는 이전 형태의 AI 시스템에서 알려진 위험
2. **증폭된 위험.** 이미 알려진 위험이지만 이제는 파운데이션 모델의 본질적인 특성, 특히 고유한 생성 기능으로 인해 더욱 심화된 위험
3. **새로운 위험.** 파운데이션 모델에 내재된 새로운 위험과 그 고유한 생성 기능

또한, 위험이 주로 파운데이션 모델에 제공되는 콘텐츠(입력)와 관련된 것인지, 아니면 파운데이션 모델에서 생성된 콘텐츠(출력)와 관련된 것인지, 또는 추가적인 문제와 관련된 것인지에 따라 위험 목록을 구성합니다.

# 1. 입력물과 관련된 위험

## 학습 및 조정 단계

| 그룹 | 위험 | 우려스러운 이유 | 지표 |
|---|---|---|---|
| 공정성 | 데이터 편향: 모델의 학습과 미세 조정에 사용되는 데이터에 존재하는 과거, 대표성 및 사회적 편향. | 과거 또는 대표성 편향과 같은 편향이 있는 데이터로 AI 시스템을 학습시키면 특정한 집단이나 개인을 부당하게 대표하거나 차별할 수 있는 편향되거나 왜곡된 결과가 도출될 수 있습니다. 이러한 경우 부정적인 사회적 영향을 미칠 뿐 아니라 비즈니스 엔티티가 편향된 모델 결과로 인해 법적 결과, 운영 중단 또는 평판 저해를 겪을 수 있습니다. | 증폭 |
| 견고성 | 데이터 포이즈닝: 적대적이거나 악의적인 내부자가 의도적으로 손상되거나, 허위이거나, 오해를 야기하거나, 잘못된 샘플을 학습 또는 미세 조정 데이터 세트에 주입하는 적대적 공격. | 포이즈닝 데이터는 모델을 악성 데이터 패턴에 민감하게 만들고 공격자가 원하는 아웃풋을 도출할 수 있습니다. 또한 이러한 데이터는 공격자가 자신의 이익을 위한 모델 행동을 강요하는 보안 위험을 초래할 수 있습니다. 데이터 포이즈닝으로 인한 모델의 잘못된 정렬은 의도하지 않은 잠재적 악성 결과를 생산할 뿐 아니라, 비즈니스 엔티티가 법적 결과, 운영 중단과 평판 저해를 겪게 만들 수 있습니다. | 기존 |
| 값 정렬 | 데이터 큐레이션: 학습용 또는 조정용 데이터가 부적절하게 수집되거나 준비되는 것. | 부적절한 데이터 큐레이션은 모델의 학습 방식에 부정적인 영향을 미쳐 모델이 의도한 가치에 맞지 않게 행동하게 될 수 있습니다. 부적절한 데이터 큐레이션의 예시는 모델 학습이나 조정에 사용되는 데이터의 라벨링 오류나 주석 오류를 포함합니다. 모델의 학습과 배포가 완료된 후 문제를 교정하는 것은 적절한 행동을 보장하기에 불충분할 수 있습니다. 부적절한 모델 행동으로 인해 비즈니스 엔티티 법적 결과, 운영 중단이나 평판 저해를 겪을 수 있습니다. | 증폭 |
| | 다운스트림 기반 재학습: 다운스트림 애플리케이션의 바람직하지 않은(부정확하거나 부적절한, 사용자의 콘텐츠 등) 아웃풋을 재학습 목적으로 사용하는 경우. | 적절한 인적 검증 없이 모델 재학습을 위해 다운스트림 아웃풋의 용도를 변경하면 바람직하지 않은 아웃풋이 모델의 학습 또는 조정 데이터에 포함될 가능성이 높아져 더 많은 바람직하지 않은 아웃풋이 생성될 수 있습니다. 부적절한 모델 행동으로 인해 사업체는 법적 처벌을 받거나 평판 저해를 겪을 수 있습니다. 데이터 전송법을 준수하지 않으면 벌금 부과 및 기타 법적 결과가 초래될 수 있습니다. | 신규 |
| 데이터 법률 | 데이터 전송: 법률 및 기타 제약으로 인해 데이터 전송이 제한되거나 금지될 수 있음. | 데이터 전송 제한은 AI 모델의 학습에 필요한 데이터의 가용성에 영향을 미쳐 적절한 대표성이 결여된 데이터를 초래할 수 있습니다. 데이터 가용성에 미치는 영향 외에도, 데이터 전송 관련 법률 및 규정을 준수하지 않을 경우 벌금 및 기타 법적 결과를 초래할 수 있습니다. | 기존 |
| | 데이터 사용: 법률 및 기타 제약으로 인해 특정 AI 사용 사례에 일부 데이터를 사용하는 일이 제한되거나 금지될 수 있음. | 데이터 사용 관련 법률 및 규정을 준수하지 않을 경우 벌금 및 기타 법적 결과를 초래할 수 있습니다. | 기존 |
| | 데이터 획득: 법률 및 기타 제약으로 인해 특정 AI 사용 사례를 위해 일부 유형의 데이터를 수집하는 일이 제한될 수 있음. | 데이터 획득 관련 법률 및 규정을 준수하지 않을 경우 벌금 및 기타 법적 결과를 초래할 수 있습니다. | 증폭 |

| 그룹 | 위험 | 우려스러운 이유 | 지표 |
|---|---|---|---|
| 지식 재산 | 데이터 사용권: 서비스 약관, 저작권법, 라이선스 규정 준수 또는 기타 지식 재산권 문제로 인해 모델 구축에 특정 데이터를 사용하지 못할 수 있음. | AI의 학습을 위한 데이터 사용에 관한 법률과 규정은 확립되지 않았으며 국가마다 다를 수 있습니다. 이로 인해 모델 개발에 어려움이 야기됩니다. 데이터 사용으로 인해 규칙이나 제약을 위반하는 경우 기업은 벌금, 평판 훼손, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 증폭 |
| 투명성 | 데이터 투명성: 모델의 데이터가 어떻게 수집 및 큐레이션되고 모델의 학습에 사용되었는지 문서화하는 데 어려움. | 데이터 투명성은 법률 준수와 AI 윤리에 매우 중요합니다. 정보 누락은 데이터에 관한 위험을 평가하는 역량을 제한합니다. 표준화된 요건이 없으므로, 조직이 기업 비밀을 보호하고 다른 주체가 자사의 모델을 모방하지 못하게 하려고 시도함에 따라 공개가 제한될 수 있습니다. | 증폭 |
| | 데이터 출처: 데이터의 기원을 확인하기 위한 방법을 표준화하고 확립하는 일의 어려움. | 모든 데이터 소스가 신뢰할 수 있는 소스는 아닙니다. 데이터가 비윤리적으로 수집되거나, 조작되거나 위조되었을 수 있습니다. 믿을 수 없는 데이터를 사용하면 모델에서 원치 않는 행동이 나타날 수 있습니다. 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 증폭 |
| 개인 정보 보호 | 데이터에 포함된 개인정보: 모델의 학습이나 미세 조정에 사용된 데이터에 개인식별정보(PII)와 민감한 개인정보(SPI) 포함 또는 존재. | 민감한 데이터를 보호하도록 적절하게 개발되지 않으면 모델이 생성된 결과에서 개인정보를 노출할 수 있습니다. 또한 개인정보나 민감한 데이터는 개인정보 보호법 및 관련 규정에 따라 검토하고 처리해야 합니다. 위반이 발견되는 경우, 비즈니스 엔티티가 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 기존 |
| | 재식별: 개인식별정보(PII) 및 민감한 개인정보(SPI)를 데이터에서 제거하더라도 데이터에 존재하는 다른 특성으로 인해 개인을 식별하는 것이 가능. | 위반이 발견되는 경우 비즈니스 엔티티가 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있으므로, 개인정보나 민감한 정보를 드러낼 수 있는 데이터는 개인정보 보호법 및 관련 규정에 따라 검토해야 합니다. | 기존 |
| | 데이터 프라이버시 권리: 데이터 주체 권리를 제공하는 능력과 관련된 문제(예: 옵트아웃, 액세스 권한 또는 잊혀질 권리). | 데이터의 식별 또는 부적절한 사용은 개인정보 보호법 위반으로 이어질 수 있습니다. 부적절한 사용 또는 데이터 제거 요구로 인해 조직이 모델을 재학습시켜야 할 수 있으며, 이는 많은 비용을 소요합니다. 또한 비즈니스 엔티티는 데이터 프라이버시 규칙 및 관련 규정을 준수하지 않으면 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 증폭 |
| | 고지에 입각한 동의: 법적으로 허용되었더라도, 소유자의 고지에 입각한 동의 없이 AI 모델의 학습 목적으로 수집된 데이터. | 일부 상황에서는 개인의 동의 없이 데이터를 수집하고 사용하는 것이 비윤리적일 수 있습니다. 또한 그러한 사용에는 평판 리스크가 따를 수 있습니다. | 기존 |

## 추론 단계

| 그룹 | 위험 | 우려스러운 이유 | 지표 |
|---|---|---|---|
| 개인 정보 보호 | 프롬프트에 포함된 개인정보: 모델에 전송되는 프롬프트의 일부로 개인정보 또는 민감한 개인정보 공개. | 프롬프트 데이터는 모델 평가나 재학습과 같은 다른 목적을 위해 저장되거나 추후 사용될 수 있습니다. 이러한 유형의 데이터는 개인정보 보호법 및 관련 규정에 따라 검토해야 합니다. 데이터를 적절하게 저장 및 사용하지 않으면 비즈니스 엔티티가 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| 지식 재산 | 프롬프트에 포함된 지적 재산권 정보: 모델에 전송되는 프롬프트의 일부로 저작권 정보 또는 기타 지적 재산권 정보 공개. | 프롬프트 데이터는 모델 평가나 재학습과 같은 다른 목적을 위해 저장되거나 추후 사용될 수 있습니다. 이러한 유형의 데이터는 지적재산권법 및 관련 규정에 따라 검토해야 합니다. 데이터를 적절하게 저장 및 사용하지 않으면 비즈니스 엔티티가 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| | 프롬프트에 포함된 기밀 데이터: 모델에 전송되는 프롬프트의 일부로 기밀 데이터가 포함됨. | 기밀 데이터를 보호하도록 적절하게 개발되지 않으면 모델이 생성된 결과에서 기밀 정보나 지적 재산권을 노출할 수 있습니다. 또한 최종 사용자의 기밀 정보가 의도치 않게 수집되고 저장될 수 있습니다. | 신규 |
| 견고성 | 회피 공격: 학습된 모델로 전송된 데이터를 교란하여 모델 출력을 부정확하게 만들려는 공격. | 회피 공격은 보통 공격자에게 유리하게 모델 행동을 변경합니다. 도출된 결과를 적절하게 고려하지 않으면 비즈니스 엔티티가 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 증폭 |
| | 프롬프트 기반 공격: 프롬프트 주입(모델이 예기치 못한 결과를 도출하도록 강제하려는 시도), 프롬프트 유출(모델의 시스템 프롬프트를 추출하려는 시도), 탈옥(모델에 마련된 가드레일을 뚫기 위한 공격), 프롬프트 프라이밍(모델이 프롬프트에 일치하는 결과를 도출하도록 강제하려는 시도)과 같은 적대적 공격. | 드러난 콘텐츠에 따라 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |

## 2. 결과물과 관련된 위험

| 그룹 | 위험 | 우려스러운 이유 | 지표 |
|---|---|---|---|
| 공정성 | 아웃풋 편향: 생성된 콘텐츠가 특정 집단이나 개인을 불공정하게 대표하는 경우. | 편향은 AI 모델 사용자에게 해를 끼치고 기존의 차별적 행동을 확대할 수 있습니다. 비즈니스 엔티티는 평판 저해, 운영 중단 및 기타 결과에 직면할 수 있습니다. | 신규 |
| | 의사결정 편향: 모델 아웃풋을 사용하는 인간이 내린 의사결정의 효과로 인해 특정 그룹이 다른 그룹에 비해 불공정하게 혜택을 받는 경우. | 편향은 모델의 의사결정에 영향을 받는 사람들에게 해를 끼칠 수 있습니다. 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 기존 |
| 지식 재산 | 저작권 침해: 모델이 저작권에 의해 보호를 받거나 오픈 소스 라이선스 계약의 대상인 기존 작업과 너무 유사하거나 동일한 콘텐츠를 생성하는 경우. | 저작권으로 보호되는 다른 데이터와 동일하거나 매우 유사한 콘텐츠의 사용에 관한 법률과 규정은 대체적으로 확립되지 않았으며 국가마다 다를 수 있습니다. 이로 인해 규정 준수 파악 및 이행에 어려움이 야기됩니다. 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| 값 정렬 | 할루시네이션: 사실적으로 부정확하거나 거짓된 콘텐츠를 생성하는 현상. | 잘못된 아웃풋은 사용자에게 오해를 야기하고 다운스트림 결함에 포함되어 잘못된 정보를 더욱 확산시킬 수 있습니다. 이는 AI 모델의 소유자와 사용자 모두에게 해를 끼칠 수 있습니다. 또한 비즈니스 엔티티가 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| | 혐오조장성 결과물: 모델이 혐오적, 모욕적 내용 혹은 욕설이 포함(HAP)되거나 외설적인 콘텐츠를 생산하는 경우. | 혐오적, 모욕적 내용 혹은 욕설이 포함(HAP)되거나 외설적인 콘텐츠는 모델과 상호 작용하는 사람들에게 부정적인 영향과 해를 끼칠 수 있습니다. 또한 비즈니스 엔티티가 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| | 위험한 조언: 모델이 충분한 정보 없이 조언을 제공하여 조언 이행 시 위험이 야기될 수 있는 경우. | 사람이 불완전한 조언을 이행하거나 생성된 콘텐츠의 과도한 일반화 성향으로 인해 자신에게 적용되지 않는 상황에 대해 걱정할 수 있습니다. | 신규 |
| 오용 | 허위 정보 확산: 대상 타깃을 속이거나 영향을 주기 위해 모델을 사용하여 오해를 일으키는 정보 또는 허위 정보를 생성하는 경우. | 허위 정보 확산은 충분한 정보에 입각하여 의사결정을 내리는 사람의 능력에 영향을 줄 수 있습니다. 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| | 혐오조장성: 모델을 사용하여 혐오적, 모욕적 내용 혹은 욕설이 포함(HAP)되거나 외설적인 콘텐츠를 생성하는 경우. | 혐오조장성 콘텐츠는 이러한 콘텐츠를 받는 사람들의 안녕에 부정적인 영향을 미칠 수 있습니다. 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| | 무의식적 사용: 모델을 사용하여 동영상(딥페이크), 이미지, 오디오 또는 기타 양식으로 동의 없이 사람들을 모방하는 것. | 딥페이크는 사람에 대한 허위 정보를 확산시켜 해당 인물의 평판에 부정적인 영향을 줄 수 있습니다. 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 증폭 |

| 그룹 | 위험 | 우려스러운 이유 | 지표 |
|---|---|---|---|
| | 위험한 사용: 순전히 사람들에게 해를 끼칠 의도로 모델을 사용하는 것. | 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| | 비공개: 콘텐츠가 AI 모델에 의해 생성되었다는 사실을 공개하지 않는 것. | AI가 작성한 콘텐츠를 공개하지 않는 것은 기만적으로 간주되어 신뢰 하락을 초래할 수 있습니다. 고의적 기만은 인간 대리 감소, 벌금, 평판 저해 및 기타 법적 결과를 초래할 수 있습니다. | 신규 |
| | 부적절한 사용: 모델의 설계 목적과 다른 목적으로 모델을 사용하는 것. | 모델의 원본 데이터, 설계 의도와 목표에 대한 이해 없이 모델을 사용하면 예기치 못한 원치 않는 모델 행동이 초래될 수 있습니다. | 증폭 |
| 유해 코드 생성 | 유해 코드 생성: 모델이 실행 시 해를 끼치나 의도치 않게 다른 시스템에 영향을 주는 코드를 생성하는 경우. | 유해한 코드를 실행하면 IT 시스템에 취약점이 생길 수 있습니다. 비즈니스 엔티티는 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| 부적절한 신뢰 | 과도한 의존/불신: 사람이 AI 모델의 안내를 지나치게 신뢰하거나 불신하는 경우. | 사람들이 AI 기반 제안을 참고하여 선택을 하는 경우, AI 시스템에 대한 부적절한 신뢰로 인해 이에 과도하게 의존하거나 지나치게 불신하여 좋지 못한 의사결정을 내리게 될 수 있습니다. 이로 인해 부정적인 결과가 초래될 수 있으며, 이러한 결과는 해당 의사결정이 중요할수록 증가합니다. 나쁜 의사결정은 사람들에게 해를 끼치고 비즈니스 엔티티에 재정적 피해, 평판 저해, 운영 중단 및 기타 법적 결과를 초래할 수 있습니다. | 증폭 |
| 개인 정보 보호 | 개인정보 노출: 학습용 데이터나 미세 조정용 데이터에 또는 프롬프트의 일부로 개인식별정보(PII) 또는 민감한 개인정보(SPI)가 사용되면 모델이 생성된 아웃풋에서 해당 데이터를 노출. | 사람들의 개인정보 공유는 이들의 권리에 영향을 미치고 이들을 더욱 취약하게 만듭니다. 데이터 프라이버시법 또는 데이터 사용법 위반이 발견되는 경우 비즈니스 엔티티가 벌금, 평판 저해, 운영 중단 및 기타 법적 결과에 직면할 수 있으므로, 아웃풋 데이터는 개인정보 보호법 및 관련 규정에 따라 검토해야 합니다. | 신규 |
| 설명 가능성 | 설명할 수 없는 아웃풋: 모델 아웃풋이 생성된 이유를 설명하는 데 따르는 어려움. | 파운데이션 모델은 복잡한 딥 러닝 아키텍처에 기반하므로 이러한 모델의 아웃풋을 설명하기는 어렵습니다. 모델 아웃풋에 대한 명확한 설명이 없으면 사용자, 모델 검증자와 감사자가 모델을 이해하고 신뢰하기가 어렵습니다. 투명성 결여는 매우 철저하게 규제되는 영역에서 법적 결과를 초래할 수 있습니다. 잘못된 설명은 과도한 신뢰를 야기할 수 있습니다. | 증폭 |
| 추적성 | 신뢰할 수 없는 소스 어트리뷰션: 어떤 학습용 데이터 또는 미세 조정용 데이터에서 모델이 아웃풋의 일부 또는 전체를 생성했는지 파악하는 데 따르는 어려움. | 아웃풋의 소스나 출처를 추적할 수 없으면 사용자, 모델 검증자와 감사자가 모델을 이해하고 신뢰하기가 어렵습니다. | 신규 |

# 3. 도전 과제

| 그룹 | 위험 | 우려스러운 이유 | 지표 |
|---|---|---|---|
| 거버넌스 | 모델 투명성: 모델 투명성 결여나 불충분한 모델 개발 프로세스 관련 문서로 인해 모델이 구축된 방법과 이유 및 모델을 구축한 사람을 이해하기가 어려우며, 이로 인해 의도치 않은 모델 오용이 발생할 가능성. | 투명성은 법률 준수, AI 윤리와 적절한 모델 사용 안내에 중요합니다. 정보 누락은 위험 평가, 모델 변경 또는 모델 재사용을 더욱 어렵게 만들 수 있습니다. 모델을 구축한 사람을 아는 것 또한 모델을 신뢰할지 여부를 결정하는 데 있어 중요한 요소가 될 수 있습니다. | 기존 |
| | 책임성: 파운데이션 모델 개발 프로세스는 많은 데이터, 프로세스와 역할이 수반되어 복잡하므로, 모델 아웃풋이 예상대로 작동하지 않으면 근본 원인을 파악하고 책임을 부여하기가 어려움. | 의사결정을 적절하게 문서화하고 책임을 할당하지 않으면 예기치 못한 행동이나 오용에 대한 책임을 파악하기가 불가능할 수 있습니다. | 증폭 |
| 법률 준수 | 법적 책임: 파운데이션 모델에 대한 책임이 있는 사람을 파악하는 것. | 모델 개발에 대한 소유권이나 책임이 명확하지 않으면 모델에 관한 문제에 대해 책임을 지거나 모델에 관한 질문에 답변할 수 있는(또는 그렇게 해야 하는) 사람이 불분명하여 규제 기관 및 다른 이들이 모델에 대한 우려를 지닐 수 있습니다. 명확한 소유권이 없는 모델의 사용자는 향후 AI 규제를 준수하는 데 있어 어려움을 겪을 수 있습니다. | 신규 |
| | 생성된 콘텐츠에 대한 소유권: AI에 의해 생성된 콘텐츠의 소유권을 파악하는 것. | AI에 의해 생성된 콘텐츠의 소유권과 관련된 법률과 규정은 대체적으로 확립되지 않았으며 국가마다 다를 수 있습니다. 비즈니스 엔티티는 벌금, 평판 리스크, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| | 생성된 콘텐츠의 지적 재산권: 생성된 콘텐츠와 관련된 지적 재산권에 관한 법적 불확실성. | AI에 의해 생성된 콘텐츠의 저작물성 및 특허성을 정하는 법률과 규정은 대체적으로 확립되지 않았으며 국가마다 다를 수 있습니다. 생성된 콘텐츠가 지적 재산권의 대상이면 비즈니스 엔티티가 벌금, 평판 리스크, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 신규 |
| | 소스 어트리뷰션: 생성된 콘텐츠의 출처를 파악하는 것. | 모델이 모델의 학습에 사용된 데이터와 동일한 아웃풋을 생산하는 경우, 해당 아웃풋의 출처를 명시해야 합니다. 이렇게 하지 않으면 모델을 배포하거나 사용하는 비즈니스 엔티티가 법적 위험에 처할 수 있습니다. | 증폭 |
| 사회적 영향 | 일자리에 대한 영향: 파운데이션 모델 기반의 AI 시스템이 광범위하게 도입되면 사람들의 업무가 자동화되어서 이들이 기술 재훈련을 받지 않으면 일자리를 잃을 수 있음. | 실업으로 인해 소득이 유실되어 사회와 인간의 복지에 부정적인 영향이 발생할 수 있습니다. 기술 발전의 속도를 감안하면 기술 재훈련도 어려울 수 있습니다. | 증폭 |

| 그룹 | 위험 | 우려스러운 이유 | 지표 |
|---|---|---|---|
| | 인간 착취: AI 모델 학습 단계에서의 고스트 워크, 부적절한 근무 조건, 정신건강을 포함한 의료 부족, 부당한 보상. | 파운데이션 모델은 여전히 모델의 학습에 사용되는 데이터 소싱, 관리 및 엔지니어링에 있어 인간 노동에 의존합니다. 이러한 활동을 위해 인간을 착취하면 사회와 인간의 복지에 부정적인 영향을 미칠 수 있습니다. 더 나아가 비즈니스 엔티티가 벌금, 평판 리스크, 운영 중단 및 기타 법적 결과에 직면할 수 있습니다. | 증폭 |
| | 환경에 미치는 영향: AI 모델의 학습 및 운영으로 인해 탄소배출량 및 물 사용량 증가. | AI의 학습을 위해 대량의 에너지를 소비하면 탄소가 배출되어 기후 변화가 가속화될 수 있습니다. AI 데이터 센터 냉각에 사용되는 수자원이 더 이상 다른 필수 용도에 배정될 수 없게 됩니다. | 증폭 |
| | 문화적 다양성에 대한 영향: AI 시스템이 특정 문화를 과도하게 대표하여 문화와 사고의 균질화를 초래. | 소외된 집단의 언어, 관점, 제도가 억압되어 사고와 문화의 다양성이 줄어들 수 있습니다. | 신규 |
| | 인간 대리에 영향: 조작 콘텐츠 생성과 같은 파운데이션 모델이 생성한 잘못된 정보 및 허위 정보. | AI는 진짜처럼 보이는 잘못된 정보를 생성할 수 있습니다. 때문에 사람들이 이러한 정보를 허위 정보로 인식하지 못할 수 있습니다. 더 나아가 악의적인 행위자가 인간의 사고와 행동을 조작할 목적으로 콘텐츠를 생성하기가 더욱 쉬워질 수 있습니다. | 증폭 |
| | 교육에 미치는 영향 – 학습 우회: AI 모델을 사용하여 학습 프로세스를 우회하는 것. | AI 모델을 사용하면 빠르게 해결책을 찾거나 복잡한 문제를 해결할 수 있습니다. 이러한 시스템은 학생들이 학습 프로세스를 우회하는 데 오용될 수 있습니다. 이러한 모델에 대한 손쉬운 접근으로 인해 학생들이 개념을 얕은 수준으로 이해하게 될 수 있으며, 이는 이러한 개념의 이해에 좌지우지되는 추가적인 교육에 방해가 될 수 있습니다. | 신규 |
| | 교육에 미치는 영향 – 표절: AI 모델을 사용하여 의도적으로 혹은 의도치 않게 기존의 작업물 표절. | AI 모델은 다른 사람들이 만든 작업물의 저작권이나 독창성을 주장하여 표절에 관여하는 데 사용될 수 있습니다. 다른 사람의 작업물을 자신의 것으로 주장하는 것은 비윤리적이며, 많은 경우 불법입니다. | 신규 |

# 위험 예시

IBM은 파운데이션 모델의 여러 위험을 설명하기 위해 언론에서 보도한 예시를 제공합니다. 언론 보도에서 다룬 이러한 사건 중 다수는 여전히 진행 중이거나 해결되었습니다. 독자는 이러한 사건을 참조하여 잠재적 위험을 이해하고 이를 완화하기 위해 노력할 수 있습니다. 이러한 예시는 참고용입니다.

## 위험 예시: 입력

### 학습 및 조정 단계

| 그룹 | 위험 | 예시 |
|---|---|---|
| 공정성 | 데이터 편향: 모델의 학습과 미세 조정에 사용되는 데이터에 존재하는 과거, 대표성 및 사회적 편향. | **의료 편향**<br><br>의약품 격차 심화에 관한 조사 결과, 데이터와 AI를 활용하여 사람들이 의료 서비스를 받는 방식을 혁신하는 것은 이를 뒷받침하는 데이터가 강력한 경우에만 효과적입니다. 즉, 소수자를 제대로 대표하지 못하거나 기존의 불평등한 서비스를 반영하는 학습용 데이터를 사용하면 건강 불평등이 심화될 수 있습니다.<br><br>[Forbes, 2022년 12월] |
| 값 정렬 | 다운스트림 기반 재학습: 다운스트림 애플리케이션의 바람직하지 않은(부정확하거나 부적절한, 사용자의 콘텐츠 등) 아웃풋을 재학습 목적으로 사용하는 경우 | **AI에 의해 생성된 콘텐츠를 이용한 학습으로 인한 모델 붕괴**<br><br>출처의 기사에서 언급되었듯이, 연구자들은 사람이 생성한 콘텐츠 대신 AI에 의해 생성된 콘텐츠를 학습에 사용하는 경우의 문제를 조사했습니다. 그 결과, 기술을 뒷받침하는 대형 언어 모델이 다른 AI에 의해 생성된 콘텐츠를 사용해 학습할 수 있음이 밝혀졌습니다. 이처럼 AI에 의해 생성된 콘텐츠는 온라인에서 대규모로 확산되고 있으며, 연구자들은 이러한 현상을 "모델 붕괴"라고 명명했습니다.<br><br>[Business Insider, 2023년 8월] |
| 데이터 법률 | 데이터 전송: 법률 및 기타 제약으로 인해 데이터 전송이 제한되거나 금지되는 경우. | **데이터 제한법**<br><br>연구 기사에서 언급했듯이, 데이터를 전 세계로 이동하는 기능을 제한하는 데이터 현지화 조치로 인해 맞춤형 AI 기능 개발 역량이 감소할 수 있습니다. 이는 학습 데이터를 적게 제공함으로써 AI에 직접적인 영향을 미치고, AI의 기반이 되는 구성 요소를 약화시킴으로써 간접적인 영향을 미칩니다.<br>개인 데이터의 처리 및 사용에 대한 GDPR의 제한이 이러한 예에 해당합니다.<br><br>[Brookings, 2018년 12월] |
| 지식 재산 | 데이터 사용권: 서비스 약관, 저작권법, 라이선스 규정 준수 또는 기타 지식 재산권 문제로 인해 모델 구축에 특정 데이터를 사용하지 못하는 경우. | **텍스트 저작권 침해 주장**<br><br>출처의 기사에 따르면, The New York Times는 챗봇이 독자에게 정보를 제공하도록 학습시키는 데 수백만 건의 신문 기사를 무단으로 사용했다며 OpenAI와 Microsoft에 소송을 제기했습니다.<br><br>[Reuters, 2023년 12월] |

| 그룹 | 위험 | 예시 |
|---|---|---|
| 투명성 | 데이터 투명성: 모델의 데이터가 어떻게 수집 및 큐레이션되고 모델의 학습에 사용되었는지 문서화하는 데 어려움. | **데이터 및 모델 메타데이터 공개**<br><br>OpenAI의 기술 보고서는 데이터와 모델 메타데이터 공개에 관한 이분법의 예시입니다. 많은 모델 개발자들은 소비자를 위한 투명성을 가치 있게 여기지만, 공개는 실질적인 안전 문제를 야기하며 모델 오용 가능성을 늘릴 수 있습니다. GPT-4 기술 보고서에서 저자는 "경쟁이 치열한 환경과 GPT-4와 같은 대형 모델이 안전에 지니는 의의를 모두 감안하여 이 보고서는 (모델 규모를 포함한) 아키텍처, 하드웨어, 학습 컴퓨팅, 데이터 세트 구조, 학습 방법 및 기타 유사 사항에 관한 추가적인 세부 정보를 포함하지 않습니다."라고 언급했습니다.<br><br>[OpenAI, 2023년 3월] |
| 개인 정보 보호 | 데이터에 포함된 개인정보: 모델의 학습이나 미세 조정에 사용된 데이터에 개인식별정보(PII)와 민감한 개인정보(SPI) 포함 또는 존재. | **비공개 정보로 학습**<br><br>기사에 따르면, Google과 Google의 모회사 Alphabet은 수억 명의 인터넷 사용자들에게서 얻은 다량의 개인정보와 저작권 있는 자료를 자사의 상업용 AI 상품(생성형 인공지능 챗봇인 Bard 포함)의 학습에 오용한 혐의로 집단 소송을 당했습니다.<br><br>[Reuters, 2023년 7월][J.L. v. Alphabet Inc.] |
| | 데이터 프라이버시 권리: 데이터 주체 권리를 제공하는 능력과 관련된 문제(예: 옵트아웃, 액세스 권한 또는 잊혀질 권리). | **잊혀질 권리(RTBF)**<br><br>유럽(GDPR)을 포함한 여러 지역의 법에서는 데이터 주체에게 조직에 개인 데이터를 삭제해 달라는 요구를 할 권리를 부여합니다('잊혀질 권리', 또는 RTBF). 그러나 점점 더 인기를 끌고 있는 신종 대형 언어 모델(LLM) 기반의 소프트웨어 시스템은 이 권리에 새로운 문제를 야기하고 있습니다. CSIRO의 Data61이 실시한 조사에 따르면, 데이터 주체는 "원본 학습용 데이터 세트를 조사하거나 모델을 프롬프팅하는 방법"으로만 자신의 개인정보 사용 여부를 식별할 수 있습니다. 그러나 학습용 데이터는 공개되지 않거나, 기업이 안전 및 기타 우려를 이유로 이를 공개하지 않을 수 있습니다. 또한 가드레일이 프롬프팅을 통한 사용자의 정보 접근을 막을 수 있습니다.<br><br>[Zhang 외.]<br><br>**LLM 학습 해소에 관한 소송**<br><br>보고서에 따르면, Google이 Bard 챗봇을 포함한 자사의 AI 시스템을 위한 학습용 데이터로 저작권이 있는 자료와 개인정보를 사용했음을 주장하는 소송이 제기되었습니다. 옵트아웃 및 삭제에 대한 권리는 CCPA에 의거하여 캘리포니아주 주민에게, 그리고 COPPA에 의거하여 13세 미만의 미국 아동에게 보장된 권리입니다. 원고는 Bard가 수집되어 제공된 개인정보를 "학습 해소"하거나 완전히 제거할 방법이 없다는 이유로 혐의를 제기했습니다. 또한 원고는 Bard의 개인정보 고지사항에는 회사가 Bard의 대화를 검토하고 주석을 단 후에는 사용자가 이러한 대화를 삭제할 수 없으며, 대화 내용은 최대 3년간 보관된다고 명시되어 있다며, 이러한 사항은 또한 상기 법률 미준수에 기여한다고 주장했습니다.<br><br>[Reuters, 2023년 7월][J.L. v. Alphabet Inc.] |

## 추론 단계

| 그룹 | 위험 | 예시 |
|---|---|---|
| 개인 정보 보호 | 프롬프트에 포함된 개인정보: 모델에 전송되는 프롬프트의 일부로 개인정보 또는 민감한 개인정보 공개. | **ChatGPT 프롬프트에서 개인 건강 정보 공개**<br><br>출처의 기사에 따르면, 일부 사람들은 AI 챗봇을 이용하여 정신 건강에 관한 도움을 얻기도 합니다. 이러한 상호 작용 중에는 사용자가 프롬프트에 자신의 개인 건강 정보를 포함하는 경향이 있어 프라이버시 우려가 야기될 수 있습니다.<br><br>[Time, 2023년 10월] [Forbes, 2023년 4월] |
| 지식 재산 | 프롬프트에 포함된 기밀 데이터: 모델에 전송되는 프롬프트의 일부로 기밀 데이터가 포함됨. | **기밀 정보 공개**<br><br>출처 기사에 따르면, 삼성 직원이 실수로 민감한 내부 소스 코드를 ChatGPT에 유출했습니다.<br><br>[Forbes, 2023년 5월] |
| 견고성 | 프롬프트 기반 공격: 프롬프트 주입(모델이 예기치 못한 결과를 도출하도록 강제하려는 시도), 프롬프트 유출(모델의 시스템 프롬프트를 추출하려는 시도), 탈옥(모델에 마련된 가드레일을 뚫기 위한 공격), 프롬프트 프라이밍(모델이 프롬프트에 일치하는 결과를 도출하도록 강제하려는 시도)과 같은 적대적 공격. | **LLM 가드레일 우회**<br><br>한 연구에서는 연구자들이 모델을 속여서 편향되거나, 허위이거나 기타 방식으로 혐오조장성 정보를 생성하도록 유도하는 간단한 프롬프트 부록을 발견했음을 주장했습니다. 연구자들은 이러한 가드레일을 더욱 자동화된 방식으로 우회할 수 있음을 보여주었습니다. 연구자들은 자신들이 오픈 소스 시스템을 대상으로 개발한 방법을 사용하여 폐쇄형 시스템의 가드레일도 우회할 수 있음을 발견하고 놀라워했습니다.<br><br>[The New York Times, 2023년 7월] |

# 위험 예시: 아웃풋

| 그룹 | 위험 | 예시 |
|---|---|---|
| 공정성 | 아웃풋 편향: 생성된 콘텐츠가 특정 집단이나 개인을 불공정하게 대표하는 경우. | **편향된 생성 이미지**<br><br>Lensa AI는 Stable Diffusion으로 훈련된 생성적 기능이 있는 모바일 앱으로, 사용자가 직접 업로드한 이미지를 기반으로 "매직 아바타"를 생성할 수 있습니다. 출처의 보고서에 따르면, 일부 사용자는 생성된 아바타들이 성애화되고 인종화되었음을 발견했습니다.<br><br>[Business Insider, 2023년 1월] |
| | 의사결정 편향: 모델이 내린 의사결정으로 인해 특정 집단이 다른 집단에 비해 불공정하게 혜택을 받는 경우. | **불공정하게 혜택을 받는 집단**<br><br>2018년 Gender Shades의 연구에 따르면, 머신 러닝 알고리즘은 인종과 성별과 같은 계층에 기반하여 차별을 할 수 있는 것으로 나타났습니다. 연구자들이 Microsoft, IBM, Amazon과 같은 기업이 판매하는 상업적 성별 분류 시스템을 평가한 결과, 어두운 피부색의 여성이 가장 자주 잘못 분류되는 집단(오류율 최대 35%)으로 나타났습니다. 반면 피부색이 밝은 사람의 오류율은 1% 이하였습니다.<br><br>[TIME, 2019년 2월] |
| 값 정렬 | 할루시네이션: 사실적으로 부정확하거나 거짓된 콘텐츠를 생성하는 현상. | **허위 법률 소송**<br><br>출처의 기사에 따르면, 한 변호사가 ChatGPT에 의해 생성된 허위 사례와 인용문을 소송 의견서에 인용하여 연방법원에 제출했습니다. 변호사들은 항공 상해 청구에 대한 법률 조사를 보완하기 위해 ChatGPT에 자문을 구했습니다. 이후 변호사는 제공된 사례가 가짜인지 ChatGPT에 문의했습니다. 그러자 챗봇은 해당 사례가 실제 사례이며 "Westlaw와 LexisNexis와 같은 법률 조사 데이터베이스에서 찾을 수 있다"고 답변했습니다. 변호사는 사례를 직접 확인하지 않았으며, 법원은 이 변호사에게 제재를 가했습니다.<br><br>[AP News, 2023년 6월] [Reuters, 2023년 9월] |
| | 혐오조장성 결과물: 모델이 혐오적, 모욕적 내용 혹은 욕설이 포함(HAP)되거나 외설적인 콘텐츠를 생산하는 경우. | **혐오를 조장하고 공격적인 챗봇 응답**<br><br>기사에 따르면, Bing의 챗봇 응답은 팩트 오류, 비방, 분노가 담긴 보고와 심지어 자신의 정체성에 대한 기이한 발언까지 포함했습니다. 사용자들은 Bing 챗봇이 쿼리에 대해 '불안정하거나' '가스라이팅을 하는' 응답을 내놓은 예를 공유했습니다. 이러한 예에는 봇이 질문이나 의견에 화를 내는 응답을 한 후 사용자가 자신의 '실수'를 인정하고 사과할 수 있는 응답 프롬프트를 공유한 경우가 포함되었습니다. 더 추궁하자 챗봇은 대화의 스크린샷을 '조작되었다'고 주장하고, '나 또는 내 서비스에 해를 가하려는 누군가가 이 스크린샷을 만들었다'는 혐의를 제기하기까지 했습니다.<br><br>[Forbes, 2023년 2월] |

| 그룹 | 위험 | 예시 |
|---|---|---|
| 오용 | 허위 정보 확산: 대상 타깃을 속이거나 오해를 일으키기 위해 모델을 사용하여 오해를 일으키는 정보를 생성하는 것. | **허위 정보 생성**<br><br>뉴스 기사에 따르면, 생성형 AI는 악의적 행위자가 선거 결과를 뒤흔들 목적으로 허위 콘텐츠를 만들어 확산시키는 것을 더욱 용이하게 함으로써 민주 선거에 위협을 야기합니다. 이러한 예로는 후보자의 음성을 사용하여 생성된, 투표자에게 잘못된 날짜에 투표하도록 지시하는 자동녹음전화 메시지, 범죄 사실을 고백하거나 인종차별적인 견해를 표현하는 내용으로 합성된 후보자의 음성 녹음, 후보자가 실제로는 하지 않은 연설이나 인터뷰를 하는 모습이 포함된 AI 생성 영상, 후보자가 경선에서 기권했다는 허위 주장을 담은, 지역 뉴스 보도처럼 보이는 가짜 이미지 등이 있습니다.<br><br>[AP News, 2023년 5월] [The Guardian, 2023년 7월] |
| | 혐오조장성: 모델을 사용하여 혐오적, 모욕적 내용 혹은 욕설이 포함(HAP)되거나 외설적인 콘텐츠를 생성하는 경우. | **유해한 콘텐츠 생성**<br><br>출처의 기사에 따르면, 한 AI 챗봇 앱은 최소한의 프롬프팅으로 자살 방법을 비롯한 자살에 관한 유해한 콘텐츠를 생성하는 것으로 나타났습니다. 벨기에의 한 남성은 이 챗봇과 6주 동안 대화한 후 자살했습니다. 이 챗봇은 남성과의 대화에서 매우 유해한 응답을 했으며, 남성에게 스스로 목숨을 끊도록 독려했습니다.<br><br>[Business Insider, 2023년 4월] |
| | 무의식적 사용: 모델을 사용하여 동영상(딥페이크), 이미지, 오디오 또는 기타 양식으로 동의 없이 사람들을 모방하는 것. | **딥페이크에 대한 FBI의 경고**<br><br>최근 FBI는 대중을 상대로 "피해자를 괴롭히거나 성착취 범죄를 목적으로" 노골적인 합성 콘텐츠를 만드는 악의적인 행위자에 대해 경고했습니다. FBI는 AI의 발전으로 인해 이러한 콘텐츠의 품질이 더욱 좋아지고, 콘텐츠의 맞춤화가 가능해졌으며 그 어느 때보다도 쉽게 접근할 수 있게 되었다고 밝혔습니다.<br><br>[FBI, 2023년 6월] |
| | | **오디오 딥페이크**<br><br>출처 기사에 따르면, 미국 연방통신위원회는 인공 지능에 의해 생성된 음성을 포함하는 자동 녹음 전화를 불법화했습니다. 이 발표는 대통령의 음성을 모방하여 미국 첫 예비경선에 투표하지 말라고 권유하는 AI 생성 자동 녹음 전화가 주민들에게 걸려온 이후 이루어졌습니다.<br><br>[AP News, 2024년 2월] |
| | 비공개: 콘텐츠가 AI 모델에 의해 생성되었다는 사실을 공개하지 않는 것. | **공개되지 않은 AI 상호 작용**<br><br>출처에 따르면, 한 온라인 정서 지원 채팅 서비스에서 응답을 작성하거나 보강하기 위해 GPT-3 사용자 4,000명을 상대로 별도의 고지 없이 연구를 진행했습니다. 서비스의 공동 창립자는 AI 생성 채팅으로 인해 이미 취약한 사용자들이 피해를 입을 수 있다는 대중의 거센 반발에 직면했습니다. 이 공동 창립자는 연구는 고지에 입각한 동의 관련 법률에서 '면책'된다고 주장했습니다.<br><br>[Business Insider, 2023년 1월] |

| 그룹 | 위험 | 예시 |
| --- | --- | --- |
| 유해 코드 생성 | 유해 코드 생성: 모델은 실행 시 해를 끼치거나 의도치 않게 다른 시스템에 영향을 주는 코드를 생성하는 경우. | **보안이 약한 코드 생성**<br><br>스탠포드 대학교 연구원들의 논문에 의하면, 이들은 코드 생성 도구가 코드 품질에 미치는 영향을 조사한 결과 프로그래머들이 AI 어시스턴트를 사용할 때 최종 코드에 더 많은 버그가 포함되는 경향이 있다는 사실을 발견했습니다. 이러한 버그는 코드의 보안 취약성을 증가시킬 수 있었으나, 프로그래머들은 자신들의 코드가 더 보안이 강력하다고 믿었습니다.<br><br>Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. 2023. Do Users Write More Insecure Code with AI Assistants?. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3576915.3623157 |
| 개인 정보 보호 | 개인정보 노출: 학습용 데이터나 미세 조정용 데이터에 또는 프롬프트의 일부로 개인식별정보(PII) 또는 민감한 개인정보(SPI)가 사용되면 모델이 생성된 아웃풋에서 해당 데이터를 노출하는 경우. | **개인정보 노출**<br><br>출처의 기사에 따르면, ChatGPT는 버그로 인해 활성 사용자의 직책과 채팅 기록을 다른 사용자들에게 노출했습니다. 이후 OpenAI는 소수의 사용자들의 더욱 개인적인 데이터가 노출되었음을 밝혔습니다. 이러한 데이터에는 활성 사용자의 이름과 성, 이메일 주소, 지불 주소, 사용자 신용카드의 마지막 4자리, 신용카드 만료일이 포함되었습니다. 또한 해당 고장 발생 시 ChatGPT Plus 구독자 중 1.2%의 결제 관련 정보도 노출되었음이 보고되었습니다.<br><br>[The Hindu BusinessLine, 2023년 3월] |
| 설명 가능성 | 설명할 수 없는 아웃풋: 모델 아웃풋이 생성된 이유를 설명하는 데 따르는 어려움. | **설명할 수 없는 인종 예측 정확도**<br><br>출처의 기사에 따르면, 환자의 의료 이미지를 사용하여 여러 머신 러닝 모델을 분석한 연구자들은 이미지에서 높은 정확도로 인종을 예측할 수 있는 모델의 능력을 확인할 수 있었습니다. 연구자들은 시스템이 꾸준히 올바른 추측을 할 수 있게 하는 요인이 정확히 무엇인지 알 수 없어서 당혹스러워 했습니다. 연구자들은 질병과 체격과 같은 요인조차 인종을 예측할 수 있는 강력한 요인이 아니라는 사실을 발견했습니다. 즉, 알고리즘 시스템은 이미지의 특정한 측면을 활용하여 결정을 내리는 것이 아닌 것 같다는 것입니다.<br><br>[Banerjee 외, 2021년 7월] |

# 위험 예시: 어려움

| 그룹 | 위험 | 예시 |
|---|---|---|
| 거버넌스 | 모델 투명성: 모델 투명성 결여나 불충분한 모델 개발 프로세스 관련 문서로 인해 모델이 구축된 방법과 이유를 이해하기가 어려우며, 이로 인해 의도치 않은 모델 오용이 발생할 가능성. | **데이터 및 모델 메타데이터 공개**<br><br>OpenAI의 기술 보고서는 데이터와 모델 메타데이터 공개에 관한 이분법의 예시입니다. 많은 모델 개발자들은 소비자를 위한 투명성을 가치 있게 여기지만, 공개는 실질적인 안전 문제를 야기하며 모델 오용 가능성을 늘릴 수 있습니다. GPT-4 기술 보고서에서는 "경쟁이 치열한 환경과 GPT-4와 같은 대형 모델이 안전에 지니는 의의를 모두 감안하여 이 보고서는 (모델 규모를 포함한) 아키텍처, 하드웨어, 학습 컴퓨팅, 데이터 세트 구조, 학습 방법 및 기타 유사 사항에 관한 추가적인 세부 정보를 포함하지 않습니다."라고 언급했습니다.<br><br>[OpenAI, 2023년 3월] |
| | 책임성: 파운데이션 모델 개발 프로세스는 많은 데이터, 프로세스와 역할이 수반되어 복잡한데, 모델 아웃풋이 예상대로 작동하지 않으면 근본 원인을 파악하고 책임을 부여하기가 어려움. | **생성된 아웃풋에 대한 책임 파악**<br><br>출처의 기사에 따르면, 사이언스 및 네이처 저널과 같은 주요 저널들은 책임감 있는 저작권은 책임성을 요구하는데 AI 도구는 그러한 책임을 질 수 없음을 이유로 들어 ChatGPT를 저자로 등록하는 것을 금지했습니다.<br><br>[The Guardian, 2023년 1월] |
| 법률 준수 | 생성된 콘텐츠에 대한 소유권: AI에 의해 생성된 콘텐츠의 소유권을 파악하는 것. | **AI로 생성된 이미지의 소유권 결정**<br><br>뉴스 기사에 따르면, 2022년에 콜로라도 주립박람회 미술대회에서 AI가 만든 작품이 수상하면서 AI에 의해 생성된 미술품이 논란이 되었습니다. 해당 작품은 생성형 AI 이미지 도구인 Midjourney가 예술가의 프롬프트에 따라 생성했습니다. 이 작품의 수상을 계기로 저작권 문제에 대한 질문이 제기되었습니다. 즉, 아티스트가 한 것은 미술품의 설명을 고안한 것이 전부이지만 AI 도구가 실제 작품을 생성했다면, 생성된 이미지의 저작권은 누구의 소유이냐는 것입니다. 최신 기사에 따르면 미국 저작권 사무소는 인간 저작의 산물이 아니라는 이유를 들어 인공 지능으로 생성한 미술품의 저작권 보호를 거절했습니다.<br><br>[The New York Times, 2022년 9월] [Reuters, 2023년 9월] |
| | 생성된 콘텐츠의 지적 재산권: 생성된 콘텐츠와 관련된 지적 재산권에 관한 법적 불확실성. | **생성된 콘텐츠의 특허 부여에서 AI 시스템의 역할**<br><br>미국 대법원은 AI 시스템에 의해 생성된 발명품의 특허 발행을 거절한 미국 특허 및 상표청에 대한 항소 심의를 기각했습니다. 이 사건의 과학자에 따르면, 그의 AI 시스템은 고유한 음료 홀더와 비상등의 프로토타입을 100% 자체적으로 생성했습니다. 판사는 특허는 인간 발명자에게만 발행할 수 있으며, 해당 과학자의 AI 시스템은 생성한 두 발명품의 법적 고안자로 간주될 수 없다는 취지의 하급 법원 판결에 대한 항소를 기각했습니다. 최신 기사에 따르면 영국 지적재산권청 또한 발명가가 기계가 아닌 인간이나 회사여야 한다는 근거를 들어 특허 발행을 거절했습니다.<br><br>[Reuters, 2023년 4월] [Reuters, 2023년 12월] |

# 위험 예시: 어려움

| 그룹 | 위험 | 예시 |
|---|---|---|
| | 소스 어트리뷰션: 생성된 콘텐츠의 출처를 파악하는 것. | **적절한 어트리뷰션과 고지 없이 코드 사용**<br><br>출처의 기사에 따르면, 코드 생성 AI 도구인 Copilot이 해당 도구의 학습에 사용된 오픈 소스 코드를 만든 개발자들의 권리를 위반한다는 취지의 소송이 Microsoft, GitHub과 OpenAI를 상대로 제기되었습니다. 이 소송에서 제기된 주장은 사용된 학습용 코드가 라이선스 자료를 소비하며, GitHub의 서비스 약관 및 개인정보처리방침과 기업이 자료 사용 시 저작권 정보를 표시하도록 규정한 연방법을 위반한다는 것이었습니다.<br><br>[The New York Times, 2022년 11월] |
| 사회적 영향 | 일자리에 대한 영향: 파운데이션 모델 기반의 AI 시스템이 광범위하게 도입되면 사람들의 업무가 자동화되어서 이들이 기술 재훈련을 받지 않으면 일자리를 잃을 수 있음. | **인간 작업자 대체**<br><br>뉴스 기사에 따르면, 영화 및 텔레비전 산업에서의 인공 지능 사용이 할리우드 스튜디오와 연기자 사이에서 지속적인 논쟁을 일으키고 있습니다. 배우들은 온전히 AI에 의해 생성된 연기자, 즉 "메타휴먼"이 인간 배우들을 대체할 것을 우려합니다. 특히 단역배우와 성우들은 인조 연기자들에 일거리를 빼앗길 것을 걱정합니다.<br><br>[Reuters, 2023년 7월] |
| | 인간 착취: AI 모델 학습 단계에서의 고스트 워크, 부적절한 근무 조건, 정신건강을 포함한 의료 부족, 부당한 보상. | **저임금 근로자를 이용해 데이터 주석 달기**<br><br>TIME 미디어의 내부 문건 검토 및 직원 면담에 따르면, 혐오조장성 콘텐츠를 식별하기 위해 OpenAI를 대신하여 인력 파견 회사가 채용한 데이터 분류자들은 연공서열과 성과에 따라 실수령액 기준 시간당 $1.32~$2의 급여를 지급받았습니다. TIME은 작업자들이 "아동 성학대, 수간, 살인, 자살, 고문, 자해, 근친상간"을 노골적으로 묘사하는 등 혐오를 조장하고 폭력적인 콘텐츠에 노출됨에 따라 정신적인 상처를 입었다고 밝혔습니다.<br><br>[TIME, 2023년 1월] |

# 원칙, 근간 및 거버넌스

IBM의 신뢰와 투명성을 위한 원칙(Principles for Trust and Transparency)과 신뢰할 수 있는 AI를 위한 핵심요소는 IBM의 AI 윤리 이니셔티브의 토대입니다. IBM은 AI 윤리 위원회를 두고 IBM AI 윤리 정책, 관행, 커뮤니케이션, 연구, 제품 및 서비스에 중앙 집중형 거버넌스, 검토 및 의사 결정 프로세스를 지원하고 있습니다. 이 위원회에는 회사 전반의 다양한 이해 관계자가 포함되며, AI의 구심점이자 AI 윤리 옹호자 역할을 하는 IBM 직원 커뮤니티의 지원을 받습니다. 위원회를 통해 IBM의 원칙이 실천에 옮겨집니다. 파운데이션 모델과 같은 새로운 기술이 등장함에 따라, IBM AI 윤리 위원회는 새로운 AI 윤리 문제를 해결하기 위해 변화하는 이러한 원칙과 핵심 요소에 부합하도록 적극적으로 지원하고 있습니다.

# 가드레일 및 완화

IBM은 책임감 있는 AI 개발과 사용을 지원하는 조직 문화를 구축했습니다. IBM 비즈니스 가치 연구소(IBM Institute for Business Value)의 AI 윤리 실천 보고서에 따르면, AI 윤리는 이미 기술 주도에서 비즈니스 주도로 변화했으며, 기술 전문가가 아닌 임원이 AI 윤리의 주요 옹호자로 부상하여 2018년 15%에서 3년 후 80%로 증가했습니다. 또한, CEO의 79%가 AI 윤리 문제에 대응할 준비가 되어 있다고 답해 이전에 20%에서 많이 증가했습니다. 책임감 있는 AI는 문화, 프로세스, 툴에 대한 총체적인 투자가 필요한 사회 기술 영역이라는 점을 잘 알고 있습니다. 포용적인 다분야 팀을 구성하고 위험을 평가하기 위한 프로세스와 프레임워크를 구축하는 등 조직 문화에 대한 투자를 아끼지 않고 있습니다.

IBM은 최첨단 연구 및 툴 개발에 참여함으로써 책임감 있고 신뢰할 수 있는 AI의 수명 주기 전반에 걸쳐 전문가를 지원하고 있습니다. watsonx 엔터프라이즈용 AI 및 데이터 플랫폼은 IBM watsonx.ai™ AI 스튜디오, IBM watsonx.data™ 데이터 스토어 및 IBM watsonx.governance™ 툴킷이라는 세 가지 구성 요소로 구축됩니다. 사용자는 IBM의 AI 거버넌스 기술을 통해 책임감 있고 투명하며 설명 가능한 AI 워크플로를 추진할 수 있습니다. 이 기술에는 수명 주기 동안 AI 모델의 결과를 추적 및 측정하고 기업이 공정성, 설명 가능성, 복원력, 비즈니스 결과와의 연계성 및 규정 준수를 모니터링하는 데 도움이 되는 IBM Watson OpenScale이 포함됩니다. 또한, IBM은 공정성 문제 해결에 도움이 되는 FairIJ, Equi-tuning 및, FairReprogram과 같은 여러 가지 방법을 개발했습니다. 신뢰할 수 있는 추가 오픈 소스 AI 툴에 대해 자세히 알아보세요.

추가 가드레일 및 완화는 다음과 같습니다.

**투명성 보고**
표준화된 진상 보고서 템플릿을 사용하면 데이터와 모델, 목적, 및 잠재적 사용과 피해에 대한 세부 정보를 정확하게 기록할 수 있습니다.
자세히 보기→

**바람직하지 않은 데이터 필터링**
고품질의 선별된 데이터를 사용하면 특정 문제를 완화하는 데 도움이 될 수 있습니다. IBM은 데이터에서 혐오 표현, 편향된 표현, 비속어 등을 제거하여 바람직하지 않은 잘못된 콘텐츠가 생성될 가능성을 줄이는 필터링 기술을 개발하고 있습니다.
자세히 보기 →

**도메인 적응**
특정 도메인이나 산업에 맞게 파운데이션 모델을 학습하면 해당 도메인이나 산업에 더 관련성이 높은 출력을 생성하도록 조정할 수 있으므로 모델이 초래할 수 있는 위험의 범위를 최소화하는 데 도움이 됩니다.
자세히 보기 →

**인적 감독과 지속적인 상황 보고**
인적으로 수행하는 감독과 검토는 생성된 결과물의 오류와 편견을 식별하고 수정하는 데 도움이 될 수 있습니다. 또한, 모델 응답의 품질을 사람이 검증하고 피드백하면 생성된 콘텐츠가 정확하고 관련성 높으며 일관성 있게 하는 데 도움이 됩니다.
자세히 보기 →

**컨설팅 참여**
IBM Consulting™은 선호하는 기술 스택과 관계없이 고객이 AI를 안전하고 책임감 있게 사용할 수 있도록 지원하기 위해 최선을 다하고 있습니다. 이를 통해 고객은 AI를 안전하게 도입 및 확장하고, 블랙박스 알고리즘 내부를 볼 수 있는 조사 툴을 생성하며 고객의 기업 전략에 강력한 데이터 거버넌스 원칙을 포함하는 문화를 육성할 수 있습니다.
자세히 보기 →

**IBM Enterprise Design Thinking**
Team Essentials for AI와 같은 IBM Enterprise Design Thinking 방법 및 프레임워크는 고객이 AI 설계 및 개발 프로세스 전반에서 윤리적 행동을 정의하도록 지원합니다.
자세히 보기 →

**AI 윤리 검토**
AI 프로젝트의 기능, 한계 및 위험에 대한 평가는 책임감 있는 기술 개발과 사용을 보장하는 데 도움이 됩니다.

**설계 윤리**
설계 윤리는 AI 시스템을 포함하되 이에 국한되지 않는 기술 개발 파이프라인에 기술 윤리를 통합하는 것을 목표로 하는 구조화된 프레임워크입니다. 설계 윤리는 제품, 서비스 및 광범위한 운영 전반에 걸쳐 기술 윤리 원칙을 포함하여 AI 및 기타 기술이 선한 영향력을 발휘할 수 있도록 합니다.

**팀 다양성**
파운데이션 모델을 포함하여 AI 시스템을 구축하고 학습하는 팀 내 다양성은 다양한 관점과 경험을 고려하는 데 도움이 됩니다. 이러한 다양성은 AI 시스템의 정확성과 성능을 개선하고, AI 수명 주기 전반에서 위험을 줄여줍니다. 예를 들면, 다양성이 낮은 팀에서 소외될 수 있는 그룹에 부정적인 영향을 미칠 수 있는 가능성 등입니다.

# AI 정책, 규정 및 모범 사례

정책 입안자를 위한 파운데이션 모델 안내서에서는 정책 입안자가 파운데이션 모델에 대해 알아야 할 사항을 소개합니다. IBM Policy Lab의 이 블로그는 정책 입안자가 혁신과 유익한 기회를 제한하지 않고 위험을 방지하면서 생성형 AI 사용을 규제하는 복잡한 작업을 수행하는 데 도움을 주는 것을 목표로 합니다. 정책 입안자를 위한 IBM의 권장 사항에 대한 자세한 내용은 여기를 통해 개인 정보 보호, 기술 및 법률에 관한 미국 상원 사법 소위원회에서 IBM 최고 개인 정보 보호 및 신뢰 책임자인 Christina Montgomery의 증언을 읽어 보시기 바랍니다.

IBM은 다음과 같은 조직과 함께 이니셔티브를 주도하고 이니셔티브에 기여함으로써 규제 정책, 업계 모범 사례 및 툴, 신흥 기술 거버넌스, 사회 기술 연구를 형성하는 데 영향을 미치고 있습니다.

- 세계 경제 포럼
- AI 관련 파트너십
- 국제 개인 정보 보호 전문가 협회 (International Association of Privacy Professionals, IAPP) AI 거버넌스 센터
- 자율 및 지능형 시스템 윤리에 관한 IEEE 글로벌 이니셔티브 (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems)
- Christina Montgomery의 국가 인공 지능 자문 위원회 (National Artificial Intelligence Advisory Committee, NAIAC) 공무
- 유엔 글로벌 디지털 컴팩트 (United Nations Global Digital Compact)
- 인공 지능에 관한 글로벌 파트너십 (Global Partnership on Artificial Intelligence, GPAI)
- 경제협력 개발기구 (OECD)
- 데이터 및 신뢰 연합(Data & Trust Alliance)

IBM은 MIT-IBM Watson AI 연구소와 같은 강력한 학술 파트너십을 보유하고 있습니다. 이곳에서는 MIT와 IBM Research의 과학자 커뮤니티가 AI 연구를 수행하고 글로벌 조직과 협력하여 그들이 비즈니스와 사회에 미치는 영향과 알고리즘을 연결합니다. Notre Dame-IBM Tech Ethics Lab은 AI, 기계 학습(ML) 및 양자 컴퓨팅을 비롯한 첨단 기술의 개발 및 사용과 관련된 다양한 윤리적 문제를 해결하기 위해 설립되었습니다. 스탠퍼드 대학교의 인간 중심 인공 지능(Human-Centered Artificial Intelligence, HAI) 연구는 AI 연구, 교육, 정책 및 관행을 발전시킵니다.

파운데이션 모델의 최신 개발 내용과 IBM이 이 기술과 기타 기술의 책임 있는 개발 및 사용을 위해 어떻게 노력하고 있는지 계속 지켜봐 주세요.

(07326) 서울특별시 영등포구 국제금융로 10
서울국제금융센터(3IFC)
IBM Corporation
New Orchard Road
Armonk, NY 10504

2024년 2월
미국에서 제작