

AI 거버넌스에 대한 간략한 개요

책임 있는 기계 학습 시스템

나브딕 길 아비섹 마투르 마르코스 V. 콘데*

H2O.ai

마운틴 뷰, 캘리포니아

{navdeep.gill, abhishek.mathur, marcos.conde}@h2o.ai

추상적인

모든 산업과 영역에 걸쳐 모든 규모의 조직은 인공 지능(AI) 기술을 활용하여 운영, 고객 경험 등과 관련된 가장 큰 문제를 해결하고 있습니다. 그러나 AI의 확률론적 특성으로 인해 이와 관련된 위험은 기존 기술보다 훨씬 큼니다.

연구에 따르면 이러한 위험은 규제, 규정 준수, 평판, 사용자 신뢰에서 재정 및 사회적 위험에 이르기까지 다양할 수 있습니다. 조직의 성격과 규모에 따라 AI 기술은 책임 있는 방식으로 사용되지 않을 경우 상당한 위험을 초래할 수 있습니다. 이 입장 문서는 위험 예방 및 완화를 목표로 AI의 책임 있는 사용을 감독하도록 설계된 프레임워크인 AI 거버넌스에 대한 간략한 소개를 제공하고자 합니다. 이러한 프레임워크를 갖추면 위험을 관리할 뿐만 아니라 AI 프로젝트에서 최대 가치를 얻고 조직 전체의 AI 채택을 위한 일관성을 개발할 수 있습니다.

1. 소개

이 입장 보고서에서는 신뢰할 수 있고 사회적으로 책임 있는 기계 학습의 다양한 측면과 속성 (보안, 견고성, 개인 정보 보호, 공정성, 윤리, 해석 가능성, 투명성 등) 간의 새로운 연결을 가능하게 하는 기업의 AI 거버넌스에 대한 통찰력을 공유합니다.

오랫동안 인공 지능(AI)은 손끝에 있는 엄청난 양의 리소스로 인해 기업 조직에서 채택한 것이었습니다. 오늘날 소기업은 과거에 사용 가능했던 것보다 훨씬 저렴한 클라우드 소프트웨어와 같은 최신 기술 덕분에 AI를 활용할 수 있습니다 [3, 14, 18, 20]. AI는 최근 몇 년 동안 상승 궤도에 있었고 향후 몇 년 동안 크게 증가할 것입니다 [33] [25] [42].

그러나 모든 투자에는 장단점이 있습니다. 불행히도 AI 채택과 관련된 단점은 고유한 불확실성과 향후 문제를 피하기 위해 필요한 조치를 취하지 않는 AI 시스템 구축자에 의해 발생합니다 [31, 20]. 이 작업에서 AI는 최신 머신 러닝(ML) 및 딥 러닝(DL) 시스템으로 구성되지만 그 자체로 다른 위험 및 취약성을 나타내는 주변 소프트웨어는 아닙니다 [32, 1, 17].

1.1 업계 내 문제

AI의 일부 인기 있는 응용 분야는 이상 감지 및 예측[30, 47, 22], 추천 시스템[41], 의료 진단[48, 19], 자연 과학[12, 10, 24] 및 검색 엔진[13, 11]입니다.].

그러나 산업에서 이러한 AI의 적용은 아직 초기 단계입니다. 그렇긴 하지만 채택 이후 많은 문제가 발생했으며 [16, 20, 31], 이는 몇 가지 요인에 기인할 수 있습니다.

*MC는 University of Würzburg, CAIDAS에도 소속되어 있습니다. Alexander von Humboldt 재단에서 지원합니다.

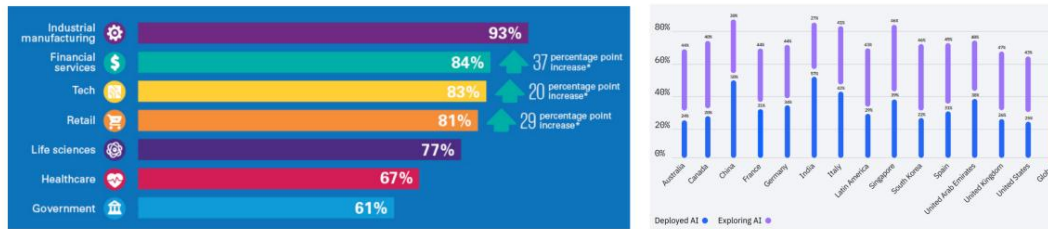


그림 1: (왼쪽) KPMG [26]의 "COVID-19 동안 급증한 AI 채택률" . (오른쪽) IBM Global AI Adoption Index 2022[25]. 자세한 내용은 설문 조사[26, 25]를 참조하십시오.

위험 인식 및 관리 부족: AI의 적용과 잠재적인 성공에 너무 많은 관심을 기울이고 잠재적 함정과 위험에는 충분한 관심을 기울이지 않습니다.

AI 채택이 너무 빠르게 진행되고 있음: 2021년 KPMG의 설문 조사 [26]에 따르면 많은 응답자는 AI 기술이 산업 제조 (55%), 기술(49%), 금융 서비스(37%), 정부(37%) 및 의료(35%) 부문.

AI 채택에는 정부 개입이 필요합니다. KPMG의 동일한 설문 조사에 따르면 [26] (그림 1 참조) 응답자의 압도적 비율은 산업 제조(94%), 소매(87%), 금융 서비스 (86%), 생명 과학(86%), 기술(86%), 의료(84%) 및 정부(82%) 부문.

기업은 AI 채택과 관련하여 아직 미숙합니다. 일부 기업은 ML 모델이 실제 세계에 배포된 후 비즈니스 조건이 변경될 준비가 되어 있지 않습니다.

적절한 거버넌스 메커니즘을 통해 이러한 문제 중 많은 부분을 피할 수 있습니다. 그러한 메커니즘이 없는 AI는 본질적인 불확실성으로 인해 해로운 결과를 낳는 위험한 게임입니다 [52, 8].

즉, AI 애플리케이션에 거버넌스를 추가하는 것은 프로덕션의 안전을 보장하는 데 필수적입니다.

2 AI 거버넌스

AI 거버넌스(AIG)란 무엇입니까?

AI 거버넌스는 조직에서 책임 있는 인공 지능을 운용하기 위한 프레임워크입니다. 이 프레임워크는 조직이 편견 없는 데이터를 큐레이팅 및 사용하고, 사회적 및 최종 사용자 영향을 고려하고, 편견 없는 모델을 생성하도록 권장합니다. 프레임워크는 또한 배포 단계를 통한 모델 진행에 대한 제어를 시행합니다. AI와 관련된 잠재적 위험은 모델 및 알고리즘의 품질에 영향을 미치기 전에 모델을 설계할 때 고려해야 합니다. 모니터링하지 않으면 AI는 바람직하지 않은 결과를 낳을 뿐만 아니라 조직에 상당한 악영향을 미칠 수 있습니다.

조직이 AI 프로젝트에서 최대 가치를 실현하고 조직 전체의 AI 채택을 위한 일관성을 개발하는 동시에 비즈니스에 대한 상당한 위험을 관리하려면 AI 거버넌스를 구현해야 합니다 [44, 15, 21, 27]. 이를 통해 조직은 책임 있는 방식으로 AI 프로젝트를 개발할 수 있을 뿐만 아니라 전체 조직에 걸쳐 일관성이 있고 비즈니스 목표가 전면 및 중앙에 있는지 확인할 수 있습니다. AI 거버넌스를 구현하면(그림 2 참조) 다음과 같은 이점을 실현할 수 있습니다. 정렬 및 명확성: 팀은 준수해야 하는 산업, 국제, 지역, 지역 및 조직 정책이 무엇인지 인식하고 정렬합니다. 예제.

사려와 책임: 팀은 AI 프로젝트의 비즈니스 사례를 정당화하기 위해 신중한 노력을 기울이고 최종 사용자 경험, 적대적 영향, 공공 안전 및 개인 정보 보호에 대해 생각하는 데 의식적인 노력을 기울일 것입니다. 이것은 또한 각각의 AI 프로젝트를 개발하는 팀에 더 큰 책임을 부여합니다.

일관성 및 조직적 채택: 팀은 AI 프로젝트를 개발하고 협업하는 보다 일관된 방법을 갖게 되어 프로젝트에 대한 추적 및 투명성을 높일 수 있습니다.

이는 또한 조직 내에서 진행되는 모든 AI 프로젝트에 대한 포괄적인 보기를 제공하여 가시성과 전반적인 채택을 향상시킵니다.

프로세스, 커뮤니케이션 및 도구: 팀은 비즈니스 가치 실현을 시작하기 위해 AI 프로젝트를 프로덕션으로 이동하기 위한 단계가 무엇인지 완전히 이해할 것입니다. 또한 정의된 프로세스를 통해 도구를 활용하는 동시에 도구를 통해 올바른 이해 관계자와 통신할 수 있습니다.

신뢰와 대중의 인식: 팀이 AI 프로젝트를 보다 신중하게 구축함에 따라 이는 본질적으로 고객과 최종 사용자 간의 신뢰를 구축하여 긍정적인 대중 인식을 구축할 것입니다.

AI 거버넌스에는 다음이 필요합니다.

1. AIG 리더에게 전체 조직에서 AI 노력에 대한 정책 및 책임을 수립하는 데 필요한 올바른 정보를 제공하는 구조화된 조직입니다. 소규모 조직의 경우 원하는 AIG의 구조적 프레임워크를 향해 작업하는 보다 단계적인 접근 방식이 필요할 수 있습니다. 대규모 조직의 경우 사람, IT 인프라, 더 큰 예산과 같은 리소스만으로도 이 프로세스를 달성할 수 있습니다.
2. AIG 리더가 필요로 하는 정보를 수집하는 구체적이고 구체적인 AI 워크플로는 구성된 정책을 시행하는 데 도움이 됩니다. 이는 소비 가능한 방식으로 다양한 당사자에게 정보를 제공합니다. 그러한 정보를 갖는 것은 무엇보다도 실수, 오류 및 편향을 최소화하는 데 사용될 수 있습니다.

AI 거버넌스에 대한 요구 사항은 조직이 개발해야 하는 프레임워크로 나타납니다. AIG를 성공적으로 구현하려면 이 프레임워크의 구성 요소가 투명하고 포괄적이어야 합니다. 특히 조직 및 사용 사례 계획, AI 개발 및 AI "운영화"에 중점을 두어 4단계 AI 수명 주기 접근 방식을 만들어야 합니다.

거버넌스 AI 수명 주기의 3단계

3.1 조직 계획

AI 거버넌스 프로그램 [27, 15, 51]은 (a) 모든 팀 구성원 사이에 규정, 법률 및 정책에 대한 포괄적인 이해가 있는 방식으로 구성되어야 합니다. 어려움 (c) 팀 구성원을 지원하기 위한 경량이지만 명확한 프로세스가 있습니다.

1. 규정, 법률, 정책

특정 법인에 적용되는 법률 및 규정은 식별되고 문서화되어야 하며 다른 사람들이 검토하고 감사할 수 있어야 합니다. 이러한 규정, 법률 및 정책은 업계와 경우에 따라 지리적 위치에 따라 다릅니다. 해당되는 경우 조직은 자신의 가치와 윤리적 견해를 반영하는 자체 정책을 개발해야 합니다 [51, 36, 20]. 이를 통해 팀은 보다 자율적이고 확신을 가지고 결정을 내릴 수 있습니다.

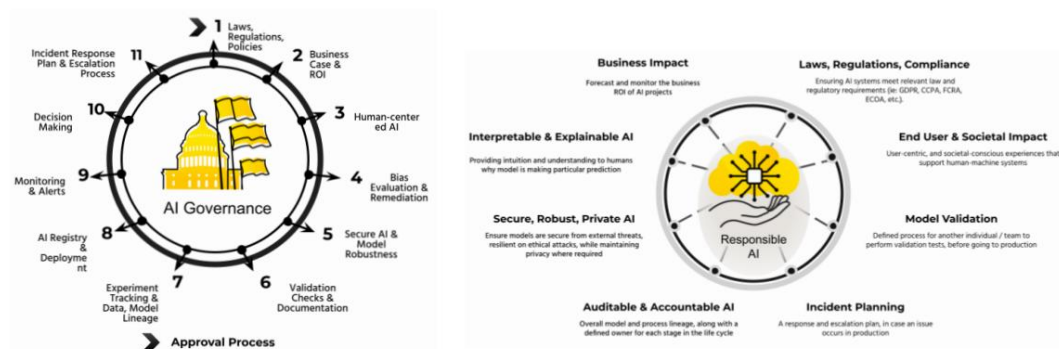


그림 2: 기업의 책임 있는 AI에 대한 AI 거버넌스 애플리케이션의 그림.

2. 조직(역량중심)

AI 프로젝트가 있는 팀을 지원하는 조직 내 그룹을 구성하는 것은 상당히 유익한 것으로 입증될 수 있습니다. 여기에는 규정, 법률 및 정책 에 정통하고 AI 팀이 가질 수 있는 모든 질문에 답변할 수 있는 그룹이 포함됩니다. 조직 내 여러 AI 팀 간에 모범 사례를 공유할 수 있는 그룹 데이터, 모델, 프로세스 등을 감사하여 위반이나 규정 위반이 없는지 확인할 수 있는 그룹입니다.

자세한 내용은 독자에게 Floridi 등의 설문 조사를 참조하십시오. [20].

3. 프로세스

AI 팀에 지침을 제공하는 경량 프로세스를 개발하면 진행률과 속도를 방해하는 대신 효율성을 높일 수 있습니다. 여기에는 데이터, 모델, 배포 등에 대한 승인 프로세스 및 사고 대응이 무엇인지 식별하는 것이 포함됩니다.

3.2 사용 사례 계획

사용 사례 구축에는 비즈니스 가치, 기술 스택 및 모델 사용 설정이 포함됩니다.

이 프로세스에 관련된 사람들의 그룹에는 주제 전문가, 데이터 과학자 /분석가/주석자 및 ML 엔지니어, IT 전문가 및 재무 부서가 포함될 수 있습니다.

비즈니스 가치 프레임워크. AI 팀은 AI 사용 사례에 대한 동기가 문서화되고 모든 이해 관계자 간에 전달되도록 해야 합니다. 여기에는 원래 가설과 실험을 평가하는 데 사용되는 메트릭도 포함되어야 합니다.

도구, 기술, 제품. AI 팀은 조직의 사전 승인된 도구 및 제품 세트에서 선택하거나 AI 사용자 사례에서 사용하기 전에 승인된 도구 및 제품 세트를 받아야 합니다. AI 개발을 위한 도구가 관리되지 않으면 비용이 많이 들고 도구를 관리할 수 없을 뿐만 아니라 (기존 IT 팀이 알고 있듯이) AI 모델에 반복성과 추적성을 생성할 수 없습니다.

모델 사용. 가치관이 유스 케이스에 연결되면 다음 단계는 유스 케이스를 하위 구성 요소로 분류하는 것입니다. 여기에는 모델의 소비자 식별, 모델의 제한 사항 및 모델 내에 존재할 수 있는 잠재적 편향 과 그 의미. 또한 대상의 포괄성, 공공 안전/사용자 프라이버시, 의도된 사용 사례에 필요한 모델 인터페이스의 식별을 보장해야 합니다.

3.3 인공지능 개발

데이터 처리 및 분석, 모델링, 설명 생성, 편향 감지, 정확성 및 효능 분석, 보안 및 견고성 검사, 모델 계보, 검증 및 문서화를 포함한 기계 학습 모델 개발.

1. 데이터 처리, 분석 및 모델링

모든 AI 프로젝트의 첫 번째 기술 단계는 데이터의 조달 및 분석이며, 이는 향후 모든 작업의 기반을 마련하기 때문에 매우 중요합니다. 데이터가 분석되면 당면한 사용 사례에 모델링이 필요한지 해독해야 합니다. 모델링이 필요한 경우 AI를 적용할 수 있습니다. 이러한 애플리케이션은 다양한 유형의 사람들에 걸쳐 있는 반복 프로세스입니다.

2. 설명과 편견

모델 설명의 목표는 특성 값을 인간 친화적인 방식으로 모델 예측에 연결하는 것입니다 [37]. 이러한 설명으로 수행하는 작업은 모델러, 중간 사용자 및 최종 사용자의 3가지 페르소나로 분류됩니다. 모델러는 모델 디버깅과 방금 만든 모델에 대한 이해를 얻기 위해 설명을 사용합니다. 중간 사용자는 실행 가능한 통찰력을 위해 모델러가 만든 것을 사용합니다. 마지막으로 최종 사용자는 모델이 직접적으로 영향을 미치는 사람입니다. 이러한 이유로 XAI(Explainable Artificial Intelligence)는 매우 활발한 연구 주제입니다[5, 23].

의도적(이질적인 치료) 또는 의도하지 않은(이질적인 영향) 편견은 AI의 많은 적용에서 우려의 원인입니다 [45, 34]. 편견을 방지하기 위해 조사해야 할 일반적인 사항에는 모델링 프로세스에 사용되는 데이터 소스, 다양한 인구 통계 간의 성능 문제, 이질적인 영향, 알려진 제한 사항 식별 및

잠재적인 악영향 및 모델이 공공 안전에 미치는 영향 [6]. 우리는 Mehrabi et al.의 설문 조사를 독자에게 소개합니다. [34] AI의 편견과 공정성에 대한 자세한 내용.

3. 정확성, 효능 및 견고성

기계 학습 모델의 정확성은 예측이 잠재적인 행동을 유도하는 모든 비즈니스 애플리케이션에 매우 중요합니다. 그러나 최적화에 가장 중요한 지표는 아닙니다.

모델의 효능도 고려해야 합니다. 즉, 모델이 의도한 비즈니스 영향을 미치고 있습니까?

모델이 생산 환경에서 예측을 제공할 때 데이터는 프로젝트 팀이 액세스할 수 있었던 데이터와 약간 또는 상당히 다를 수 있습니다. 모델 드리프트 및 기능 드리프트가 이러한 불일치를 포착할 수 있지만 이는 지연 지표이며 그때까지 모델은 이미 예측을 수행했습니다. 여기에서 견고성이 필요합니다. 프로젝트 팀은 "범위 외" 데이터를 사용하여 모델 견고성을 사전에 테스트하여 모델이 교란되는지 여부를 이해할 수 있습니다. 범위 외 데이터는 수동 생성(기능 값으로 전환)과 자동 생성(시스템에서 기능 값 전환)의 조합일 수 있습니다.

4. 오늘날의 보

안 ML 시스템은 모든 공개 IT 시스템에 영향을 미칠 수 있는 일반적인 공격에 노출되어 있습니다 [7], [39]. 데이터 및 ML 코드에 대한 내부자 액세스를 악용하는 특수 공격 ML 예측 API 및 엔드포인트에 대한 외부 액세스 [49], [46]; 타사 ML 아티팩트에 숨길 수 있는 트로이 목마. 이러한 공격은 기계 학습 모델을 실제 세계로 보내기 전에 설명하고 테스트해야 합니다.

5. 문서화 및 검증

전체 AI 프로젝트 수명 주기의 전반적인 계통은 투명성 과 이해를 보장하기 위해 문서화되어야 합니다 [35, [43]. 자신의 응용 프로그램.

모델 유효성 검사 [38, 28] 는 비즈니스 사용 사례에 따라 모델이 견고하고 예상대로 수행되는지 확인하기 위해 제3자가 수행하는 일련의 프로세스 및 활동입니다. 또한 잠재적 제한 및 가정의 영향을 식별합니다. 기술적인 관점에서 다음 사항을 고려해야 합니다. (i) 민감도 분석. (ii) 샘플 내 성능과 샘플 외 성능. (iii) 모델 개발 팀의 결과 복제. (iv) 안정성 분석. 모델 "검증자"는 모든 결과를 문서화하고 관련 이해 관계자와 공유해야 합니다.

3.4 AI 운영화

머신 러닝 모델을 프로덕션(예: MLOps [2, 50])에 배포하는 것은 잠재적으로 가치를 얻기 위한 첫 번째 단계입니다. 배포 프로세스로 들어가는 단계에는 다음이 포함되어야 합니다. 검토-승인 흐름: AI 프로젝트에서 모델 구축은 실험, 모델 등록, 배포 및 폐기와 같은 다양한 단계를 거칩니다. 한 단계에서 다음 단계로 이동하려면 검토하고 피드백을 제공할 "외부" 검토자가 필요합니다.

모니터링 및 경고: 모델이 배포되면 모델에 성능 저하가 없는지 확인하기 위해 다양한 메트릭에 대해 모니터링해야 합니다. 배포 시 모델 저하의 원인에는 기능 및/또는 대상 드리프트, 데이터 무결성 부족, 이상값 등이 포함될 수 있습니다. 모니터링 측면에서 예측의 정확성, 공정성 및 설명이 중요합니다 [4, 29].

의사 결정: 기계 학습 모델의 결과는 예측이지만 그 결과는 의사 결정으로 전환되어야 합니다. 결정하는 방법? 자율적이 될까요? 루프에 사람이 포함될까요?

이러한 질문에 대한 답변은 애플리케이션마다 다르지만 관련된 모든 사람의 위험을 줄이기 위해 적절한 방식으로 결정을 내리도록 하는 아이디어는 동일하게 유지됩니다.

사고 대응 및 에스컬레이션 프로세스: 프로덕션에서 AI 모델을 사용하면 항상 문제가 발생할 가능성이 있습니다. 조직은 사고 대응 계획 및 에스컬레이션 프로세스를 문서화하고 모든 프로젝트 팀에 알려야 합니다.

AI 애플리케이션을 위한 AI 거버넌스를 성공적으로 구현하는 기업은 인공지능을 매우 효과적으로 사용할 수 있습니다. 그렇게 하지 못하는 사람들은 다음 사용 사례에서 볼 수 있듯이 재앙적인 결과와 복구로 가는 험난한 길을 걸을 위험이 있습니다.

4 AIG 사용 사례

이 섹션에서는 AI 거버넌스가 끔찍한 결과를 피할 수 있었던 최근 사용 사례에 대해 설명합니다 .

2021년 온라인 부동산 기술 대기업 Zillow는 AI 기반 주택 플리핑 사업을 중단했습니다. 핵심적으로 이 비즈니스 라인은 기계 학습 모델의 예측에 크게 의존했습니다. Zillow는 부동산 가격을 과대 평가하여 주택에 초과 지불하고 있음을 알게 되었습니다. 이러한 재산 가치의 과대 평가로 인해 5억 6,900만 달러의 손실과 가치 평가의 28%[9, 40]가 발생했습니다.

2,000명의 직원을 해고하고 금전적인 손실을 입게 되면서 회사 평판도 나빠졌습니다.

이 AI 실패는 "Zillow가 어떤 실수를 저질렀습니까?"라는 질문을 던집니다. 그리고 "이것을 막을 수 있었을까?" 두 번째 질문에 대한 대답은 단호하게 "예"입니다. 그러나 첫 번째 질문에 대한 답변에는 여러 측면이 있지만 아래에서 주요 실수를 분석하려고 합니다.

1. "Human in the Loop" 제거

초기에 이 회사는 지역 부동산 중개인과 부동산 전문가를 고용하여 주택 가격 예측 ML 알고리즘의 출력을 확인했습니다. 그러나 비즈니스 규모가 커짐에 따라 알고리즘에 대한 사람의 검증 프로세스가 최소화되고 [9] 제안 프로세스가 자동화되어 비용을 절감하고 인수를 늘리는 데 도움이 되었습니다. 그러나 이러한 휘발성 영역에서 사람의 확인을 제거하면 확인 없이 액면 그대로 예측이 이루어지고 이는 부동산 가격을 과도하게 책정하는 데 큰 역할을 했습니다.

인간을 루프에 유지하면 회사가 과대 평가되고 편향된 예측에 의존하지 않도록 도울 수 있습니다. 조직의 의사 결정을 위한 ML 에코시스템을 고려할 때 루프에 인간을 두는 것은 안전 장치이며 모든 고위험 의사 결정에 사용해야 합니다.

2. 개념 드리프트 및 모델 모니터링 부족을 고려하지 않음

사건의 타임라인을 돌아보면 ML 알고리즘이 실제 시장 상태에 따라 조정되지 않은 것으로 보입니다[9]. 알고리즘은 계속해서 시장이 "뜨겁고" 주택 가격을 과대평가했다고 가정했습니다.

드리프트 문제, 특히 개념 드리프트를 방지하려면 기업은 AI 모델의 품질을 모니터링하고 유지하기 위한 도구를 활용해야 합니다. 이상적으로는 이 예에서 드리프트 또는 성능 저하가 있을 때 데이터 과학 팀에 자동으로 경고하고 근본 원인 분석을 지원하며 인간 참여형 모델 업데이트를 알리는 인프라를 설정해야 합니다 .

3. 모델 검증 부족

조직은 알고리즘을 배포하기 전에 모델이 예상대로 수행되고 비즈니스 사용 사례와 일치하는지 확인하기 위한 일련의 프로세스 및 활동을 제정해야 합니다. 효과적인 유효성 검사는 모델이 견고함을 보장합니다. 또한 잠재적인 제한 사항과 가정을 식별하고 가능한 영향을 평가합니다. 모델 검증의 부족 이 이 경우에 결정적인 역할을 했고 주택 가격의 과대 평가로 이어졌습니다. 이상적으로는 모델 검증은 모델 구축 프로세스에 참여하지 않은 제3자가 수행해야 한다는 점에 유의해야 합니다 .

4. 사고 대응 부족

이 회사가 그들의 모델이 저하되고 잘못된 결과를 생성한다는 사실을 깨닫기 시작한 시점은 분명하지 않습니다. 이러한 사고 대응 계획이 없으면 기업은 막대한 비용, 시간 및 인적 자원을 소모할 수 있습니다. 돌이켜보면 적절한 AI 거버넌스가 구현되었다면 회사는 사고 식별 및 문서화, 인적 검토, 모델 트래픽 리디렉션과 같은 사고 대응을 시작할 수 있었을 것입니다. 이 세 단계만으로도 비즈니스 에 많은 피해를 방지할 수 있었을 것입니다 .

이 업계 사용 사례는 적절한 AI 거버넌스 없이 발생할 수 있는 일의 예입니다. 사용 사례 시작부터 적절한 AI 거버넌스 원칙을 고려했다면 저지른 실수를 피할 수 있었습니다 . 특히, 이 회사가 계속해서 주제 전문가에게 의존하고, 다양한 유형의 드리프트를 설명하고, 모델 검증을 추가하고, 적절한 사고 대응 계획을 통해 구체적인 모델 모니터링을 구현했다면 이 모든 상황을 피하거나 피해를 줄일 수 있었을 것입니다. 훨씬 더 낮은 크기로.

5. 결론

오늘날 AI 시스템은 고용, 보석, 가격방 및 대출에 대한 삶을 바꾸는 결정을 내리는 데 사용되며 AI 시스템이 위임하는 결정의 범위는 미래에 확장될 것으로 보입니다.

많은 분야에 걸친 AI의 보급은 곧 둔화되지 않을 것이며 조직은 이러한 애플리케이션을 따라잡기를 원할 것입니다. 그러나 그들은 AI와 함께 오는 위험을 인식하고 그러한 위험을 피하기 위해 AI 적용에 접근하는 방법에 대한 지침을 가지고 있어야 합니다. AI 거버넌스를 위한 프레임워크를 구축함으로써 조직은 사용 사례에 AI를 활용하는 동시에 위험을 피하고 위험 완화 계획을 수립할 수 있습니다.

사회적 영향이 백서에서 논의하는 것처럼 조직의 AI 애플리케이션에 대한 거버넌스와 특정 제어는 필수 사항이어야 합니다. AI 거버넌스는 신뢰할 수 있고 사회적으로 책임 있는 기계 학습 시스템의 다양한 측면 간의 연결을 활성화하고 촉진하는 것을 목표로 하므로 보안, 견고성, 개인 정보 보호, 공정성, 윤리 및 투명성을 설명합니다. 우리는 이러한 아이디어의 구현이 사회에 긍정적인 영향을 미쳐야 한다고 믿습니다.

감사의 말 NeurIPS 2022의 TSRML(Trustworthy and Socially Responsible Machine Learning) 워크숍에 감사드립니다. 이 작업은 H2O.ai의 지원을 받았습니다.

참조

- [1] Omar Hussain Alhazmi, Yashwant K Malaiya, Indrajit Ray. 소프트웨어 시스템의 보안 취약성을 측정, 분석 및 예측합니다. 컴퓨터 및 보안, 26(3):219–228, 2007.
- [2] Sridhar Alla와 Suman Kalyan Adari. 멜롭스란? MLFlow로 MLOps 시작, 79–124페이지. 스프링거, 2021.
- [3] Sulaiman AlSheibani, Yen Cheung, Chris Messom. 인공 지능 채택: 회사 수준 의 Ai 준비 . PACIS 2018 절차, 2018.
- [4] Dario Amodè, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman 및 Dan Mané. 콘크리트 AI 안전 문제. arXiv 프리프린트 arXiv:1606.06565, 2016.
- [5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Benoit, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins 등. 설명 가능한 인공 지능(xai): 책임 있는 AI에 대한 개념, 분류, 기회 및 과제. 정보 융합, 58:82–115, 2020.
- [6] S. Barocas, M. Hardt, A. Narayanan. 공정성과 기계 학습: 한계와 기회, 2022년. URL: <https://fairmlbook.org/pdf/fairmlbook.pdf>.
- [7] 마르코 바레노, 블레인 넬슨, 앤서니 D 조셉, J 더그 타이아가. 기계 학습의 보안. 기계 학습, 81(2):121–148, 2010. URL: <https://people.eecs.berkeley.edu/~adj/publications/paper-files/SecML-MLJ2010.pdf>.
- [8] John Bresina, Richard Dearden, Nicolas Meuleau, Sailesh Ramkrishnan, David Smith, Richard Washington. 지속적인 시간과 자원의 불확실성 하에서의 계획: AI에 대한 도전. arXiv 사전 인쇄 arXiv:1301.0559, 2012.
- [9] 패트릭 클라크. Zillow의 알고리즘 기반 구매 행위는 홈플립핑 실험을 망쳤습니다. 블룸버그, 2021. URL: <https://www.bloomberg.com/news/articles/2021-11-08/zillow-z-home-flipping-experiment-doomed-by-tech-algorithms>.
- [10] 마르코스 V 콘데와 최의진. 몇 발의 긴꼬리 새 오디오 인식. arXiv 프리프린트 arXiv:2206.11260, 2022.
- [11] Marcos V Conde와 Kerem Turgutlu. 클립 아트: 세분화된 예술 분류를 위한 대조 사전 훈련. 컴퓨터 비전 및 패턴 인식에 관한 IEEE/CVF 회의 진행, 3956–3960페이지, 2021.
- [12] Marcos V Conde, Kumar Shubham, Prateek Agnihotri, Nitin D Movva 및 Szilard Bessenyeyi. 야생에서 새 소리의 약한 감독 분류 및 감지. Birdclef 2021 솔루션. arXiv 프리프린트 arXiv:2107.04878, 2021.

- [13] Marcos V Conde, Ivan Aerlic, Simon Jégou. 오픈 월드 이미지 검색을 위한 일반 이미지 설명자 비트 클립을 사용합니다. arXiv 프리프린트 arXiv:2210.11141, 2022.
- [14] 마리아 큐브릭. 비즈니스 및 관리에서 AI 채택을 위한 동인, 장벽 및 사회적 고려 사항: A 고등 교육. 사회 속의 기술, 62:101257, 2020.
- [15] 앨런 대포. AI 거버넌스: 연구 의제. AI 프로그램 거버넌스, 미래인류연구소, 옥스퍼드 대학교: 영국 옥스포드, 1442:1443, 2018.
- [16] 인공 지능 사건 데이터베이스. AI 사고 데이터베이스, 2022. URL: <https://사건데이터베이스.ai/>.
- [17] 마크 다우드, 존 맥도날드, 저스틴 슈. 소프트웨어 보안 평가 기술: 식별 및 소프트웨어 취약점 방지. 피어슨 교육, 2006.
- [18] Yanqing Duan, John S Edwards, Yogesh K Dwivedi. 빅 데이터 시대의 의사 결정을 위한 인공 지능 - 진화, 도전 및 연구 의제. 정보 관리 국제 저널, 48:63-71, 2019.
- [19] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, Richard Socher. 딥 러닝 자원 의료용 컴퓨터 비전. NPJ 디지털 의학, 4 (1):1-9, 2021.
- [20] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi 등. Ai4people— 좋은 AI 사회를 위한 윤리적 프레임워크: 기회, 위험, 원칙 및 권장 사항. 마음과 기계, 28(4): 689-707, 2018.
- [21] Urs Gasser 및 Virgilio AF Almeida. AI 거버넌스를 위한 계층화된 모델입니다. IEEE 인터넷 컴퓨팅, 21(6): 58-62, 2017.
- [22] Dmitry Gordeev, Philipp Singer, Marios Michailidis, Mathias Müller 및 SriSatish Ambati. 백테스팅 코로나19의 예측 가능성. arXiv 프리프린트 arXiv:2007.11411, 2020.
- [23] David Gunning, Mark Stefik, 최재식, Timothy Miller, Simone Stumpf, Guang-Zhong Yang. Xai - 설명 가능한 인공 지능. 사이언스 로봇틱스, 4(37):eaay7120, 2019.
- [24] 크리스토프 헨켈, 파스칼 파이퍼, 필립 싱어. 약한 감독하에 다양한 소리의 풍경에서 조류 종을 인식합니다. arXiv 프리프린트 arXiv:2107.07728, 2021.
- [25] IBM. IBM 글로벌 AI 채택 지수, 2022. URL: <https://www.ibm.com/downloads/cas/GVAGA3JP>.
- [26] KPMG. 팬데믹 기간 동안 AI 채택이 가속화되었지만 많은 사람들이 AI 채택이 너무 빠르다고 말합니다: KPMG 설문 조사, 2021. URL: <https://info.kpmg.us/news-perspectives/technology-innovation/thriving-in-an-ai-world/ai-adoption-accelerated-during-pandemic.html>.
- [27] Maciej Kuziemski와 Gianluca Misuraca. 공공 부문의 AI 거버넌스: 민주적 환경에서 자동화된 의사 결정의 최전선에서 나온 세 가지 이야기. 통신 정책, 44(6):101976, 2020.
- [28] Maurice Landry, Jean-Louis Malouin, Muhittin Oral. 운영 연구에서 모델 검증. 유럽 운영 연구 저널, 14(3):207-220, 1983.
- [29] Jan Leike, Miljan Martić, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, Shane Legg. AI 안전 그리드월드. arXiv 프리프린트 arXiv:1711.09883, 2017.
- [30] Spyros Makridakis, Evangelos Spiliotis 및 Vassilios Assimakopoulos. 통계 및 기계 학습 예측 방법: 우려 사항 및 향후 방법. PloS one, 13(3):e0194889, 2018.
- [31] 베나드 마르. 인공지능은 위험한가? 모두가 알아야 할 6가지 AI 위험 포브스, 2018.
- [32] 게리 맥그로. 소프트웨어 보안. IEEE 보안 및 프라이버시, 2(2):80-83, 2004.
- [33] 조 맥켄드릭. 지난 18개월 동안 AI 채택이 급증했습니다. 하버드 비즈니스 리뷰, 2021. URL: <https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months>.
- [34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman 및 Aram Galstyan. 기계 학습의 편견과 공정성에 대한 설문 조사. ACM 컴퓨팅 조사(CSUR), 54(6):1-35, 2021.

- [35] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji 및 Timnit Gebru. 모델 보고를 위한 모델 카드. 공정성, 책임성 및 투명성에 관한 회의 절차에 서. ACM, 2019. doi: 10.1145/3287560.3287596.
URL <https://doi.org/10.1145/3287560.3287596>.
- [36] 브렌트 미텔슈타트. 원칙만으로는 윤리적 AI를 보장할 수 없습니다. 자연 머신 인텔리전스, 1(11): 501–507, 2019.
- [37] 크리스토퍼 몰나르. 해석 가능한 기계 학습. 2022. URL: <https://christophm.github.io/interpretable-ml-book/>.
- [38] 연방 준비 제도 이사회 통화 감사원 사무소. 모범 위험 관리에 대한 감독 지침, SR Letter 11-7, 2011. URL <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf>.
- [39] 니콜라스 페이퍼노트. A Marauder's Map of Security and Privacy in Machine Learning: 기계 학습을 안전하고 비공개로 만들기 위한 현재 및 미래의 연구 방향에 대한 개요입니다. 인공 지능 및 보안에 관한 11차 ACM 워크숍 진행. ACM, 2018. URL: <https://arxiv.org/pdf/1811.01134.pdf>.
- [40] 헤더 펄버그 패트릭 클라크, 스리다르 나타라잔. Zillow는 중단을 뒤집은 후 28억 달러에 7,000채의 주택을 판매하려고 합니다. 블룸버그, 2021. URL: <https://www.bloomberg.com/news/articles/2021-11-01/zillow-selling-7-000-homes-for-2-8-billion-after-flipping-halt>.
- [41] Ivens Portugal, Paulo Alencar, Donald Cowan. 추천 시스템에서 기계 학습 알고리즘의 사용 : 체계적 검토. 응용 프로그램이 있는 전문가 시스템, 97:205–227, 2018.
- [42] PwC(프라이스워터하우스쿠퍼스). PwC 2022 AI 비즈니스 설문조사, 2022. URL: <https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-business-survey.html>.
- [43] Mahima Pushkarna, Andrew Zaldivar, Oddur Kjartansson. 데이터 카드: 책임 있는 AI를 위한 목적이 있고 투명한 데이터 세트 문서, 2022. URL <https://arxiv.org/abs/2204.01075>.
- [44] Sandeep Reddy, Sonia Allan, Simon Coghlan, Paul Cooper. 의료 분야에서 AI를 적용하기 위한 거버넌스 모델. 미국 의료 정보 학 협회 저널, 27(3):491–497, 2020.
- [45] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall. 인공 지능의 편견을 식별하고 관리하기 위한 표준을 향하여, 2022-03-15 04:03:00 2022. URL https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464.
- [46] Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov. 기계 학습 모델에 대한 멤버십 추론 공격. 보안 및 프라이버시(SP)에 관한 2017 IEEE 심포지엄, 3-18페이지. IEEE, 2017. URL: <https://arxiv.org/pdf/1610.05820.pdf>.
- [47] 와카스 술타니, 첸 첸, 무바라크 샤. 감시 비디오에서 실제 이상 탐지. 컴퓨터 비전 및 패턴 인식에 관한 IEEE 회의 진행, 페이지 6479–6488, 2018.
- [48] 스티브 켄지. 의료 영상 분야의 딥 러닝 개요. 방사선 물리학 및 기술, 10(3): 257–273, 2017.
- [49] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, Thomas Ristenpart. 예측 API를 통해 기계 학습 모델을 훔칩니다. 25차 {USENIX} 보안 심포지엄 ({USENIX} 보안 16), 페이지 601–618, 2016. URL: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf.
- [50] Mark Treveil, Nicolas Omont, Clément Stenac, Kenji Lefevre, Du Phan, Joachim Zentici, Adrien Lavoil lotte, Makoto Miyazaki, Lynn Heidmann. MLOps를 소개합니다. 오라일리 미디어, 2020.
- [51] Weiyu Wang과 Keng Siau. 인공 지능: 거버넌스, 정책 및 규제에 관한 연구. MWAIIS 2018 절차, 40, 2018.
- [52] Lotfi A Zadeh. AI의 불확실성을 다루기에 확률 이론이 충분한가: 부정적인 견해. 머신 인텔리전스 및 패턴 인식, 4권, 103–116페이지. 엘스비어, 1986.