

The
Software
Alliance

BSA

편향성에 맞서다:
AI의 신뢰 구축을
위한 BSA 프레임워크

목차

서론.....	1
AI의 편향성 이란 무엇인가?.....	3
AI의 편향성의 출처 및 유형.....	4
AI 위험 관리의 필요성.....	8
위험 관리란 무엇인가?.....	8
편향성 위험 관리.....	9
효과적인 리스크 관리를 위한 초석.....	10
거버넌스 프레임워크.....	11
영향 평가.....	13
AI 편향성 위험 관리 프레임워크.....	14
AI 라이프 사이클 단계.....	15
프레임워크 구조.....	17
이해관계자의 역할 및 책임.....	18
다양한 AI 개발 및 배치 모델.....	18
BSA의 AI 편향성 위험 관리 프레임워크.....	19
참고 문헌.....	28
미주(尾註).....	29

서론

인공 지능(AI) 연구 및 개발이 엄청난 속도로 발전하면서 기술이 세상을 어떻게 형성할 지에 대한 기대치가 빠르게 변화하고 있다. AI가 언젠가 모든 산업에 영향을 미칠 것이라는 전망은 빠른 속도로 상업적 현실로 바뀌고 있다. 금융 서비스 분야에서 의료 분야에 이르기까지 AI는 고객 경험을 개선하고 경쟁력을 강화하며 이전에는 다루기 어려웠던 문제를 해결하기까지, 그 활용도가 증가하고 있다. 예를 들어, AI를 활용하여 의학 연구원들은 쇠약 증상이 발생하기 몇 년 전에 초기 단계의 알츠하이머 병을 진단할 수 있고,¹ 생태학자들은 AI를 통해 중요한 서식지를 보호하고 말라위에서의 불법적인 코끼리 밀렵을 방지하려는 노력의 결과를 추적하기 위해 엄청나게 많은 데이터 세트를 분석할 수 있게 되었다.²

이 보고서에서 사용하는 "인공 지능"이라는 용어는 상관 관계, 패턴을 식별하기 위한 대량의 학습 데이터와, 향후 데이터 입력을 기반으로 예측이나 권고를 할 수 있는 모델 개발에 사용할 수 있고 기타 메타 데이터를 분석할 수 있는 기계 학습 알고리즘을 사용하는 시스템을 말한다. 예를 들어 개발자들은 사진 속 사물에 대한 청각적 설명을 제공하여 시각 장애가 있는 사람들이 세상을 탐색할 수 있도록 지원하는 앱인 "Seeing AI"을 개발할 때 기계 학습을 사용했다.³ 앱 사용자가 자기 스마트폰으로 사진을 찍으면 Seeing AI는 사진에 나타난 것들을 음성으로 설명한다. 사진 속의 물체를 식별할 수 있는 컴퓨터 시각 모델(computer vision model)을 개발하기 위해 나무, 도로 표지판, 풍경 및 동물과 같은 일반적인 물체를 묘사하는 수백만 개의 공개된 이미지 데이터를 사용하여 시스템을 학습시켰다. 사용자가 새로운 이미지를 입력하면 Seeing AI는 이 이미지를 학습 데이터에서 도출한 패턴 및 상관 관계와 비교하여 사진 속 물체가 무엇인지 예측한다.

AI가 산업 전반으로 확산되면서 기술의 설계 및 사용, 그리고 대중에게 미칠 수 있는 잠재적인 위험을 고려하는 방식으로 운영되도록 하기 위해 선제적으로 취할 수 있는 조치에 대한 질문 또한 제기되고 있다.

중대한 의사 결정 방식과 관련된 첨단 기술은 기회와 위험이 모두 따른다. 한편 금융 기관이 AI를 채택하면 인간의 편견에 덜 취약한 데이터 기반의 의사 결정 접근 방식을 활용할 수 있기 때문에 차별을 줄이고 공정성을 높일 수 있다.⁴ 예를 들어, AI를 사용하면 대출 기관은 기존 신용 보고서에서 일반적으로 처리하는 것보다 훨씬 많은 데이터를 평가할 수 있기 때문에 역사적으로 소외된 집단의 신용 및 주택에 대한 접근성을 향상시킬 수 있다. 하지만 이와 동시에 연구원들은 AI 시스템의 설계, 개발 및/또는 배치 결함으로 인해 기존의 사회적 편향성을 영속화하거나 심지어 악화될 수도 있다고 경고한다.⁵

따라서 AI 편향성의 위험을 파악하고 이를 완화하기 위한 메커니즘을 개발하는 영역은 산업계, 학계 및 정부 전문가의 집중적인 관심 영역으로 부상하고 있다. 지난 몇 년 동안 방대한 양의 연구 결과들이 AI 라이프 사이클 전반에 걸쳐 편향성 위험 관리에 도움이 될 수 있는 다양한 조직 모범 사례, 거버넌스 보호 장치 및 기술 도구를 소개했다. AI 모델에 대한 정적 평가로는 AI 시스템을 현장에 배치했을 때 발생할 수 있는 잠재적인 문제를 모두 확인할 수 없으므로, 전문가들은 AI 편향성 위험을 완화하려면 시스템이 의도한 대로 작동하는지 확인하기 위한 최종 사용자의 지속적인 모니터링을 포함하는 라이프 사이클 접근 방식이 필요하다는 것을 인정하고 있다.

이 리포트는 AI 시스템의 라이프 사이클 전반에 걸쳐 나타날 수 있는 잠재적인 편향성 위험을 파악하고 완화하기 위해 조직이 영향 평가를 수행하는 데 사용할 수 있는 AI 편향성 위험 관리 프레임워크를 제시한다. 데이터 프라이버시에 대한 영향 평가와 유사하게, AI 영향 평가는 AI의 책임성을 높이고,

피해의 위험을 완화할 수 있는 충분한 보호 장치를 갖춘 고위험 AI 시스템을 설계, 개발, 테스트 및 배치하여 신뢰도를 향상시키는 중요한 보증 메커니즘의 역할을 할 수 있다. AI 영향 평가는 또한 AI 시스템의 설계, 개발 및 배치와 관련된 많은 잠재적 이해 관계자가 위험에 대해 의사 소통하고 이러한 위험을 완화하는 책임을 명확하게 이해할 수 있도록 하는 중요한 투명성 메커니즘이기도 하다.

AI 영향 평가 수행 프로세스를 설정하는 것 외에 편향성 위험 관리 프레임워크는 다음과 같은 작업을 수행한다.

- 효과적인 AI 위험 관리 프로그램의 구현 및 지원에 필요한 주요 기업 거버넌스 구조, 프로세스 및 보호 장치를 제시한다.
- 이해 관계자가 AI 시스템의 라이프 사이클 전반에 걸쳐 나타날 수 있는 특정 AI 편향성 위험을 완화하기 위해 사용할 수 있는 기존 모범 사례, 기술 도구 및 리소스를 확인한다.

이 프레임워크는 조직이 공정성, 투명성 및 책임성을 강화하는 위험 관리 프로세스를 통해 AI 시스템의 신뢰도를 높이기 위해 사용할 수 있도록 고안된 탄력적인 방안이다.

AI 편향성이란 무엇인가?

이 문서에서 "AI 편향성"은 특정 인구 통계 집단의 구성원에게 불리하거나 불공정하거나 해로운 결과를 체계적으로 부당하게 산출하는 AI 시스템을 의미한다.

본질적으로 기계 학습의 목표는 향후 데이터 입력에 대하여 예측하기 위해 과거의 사례에서 일반화된 규칙을 도출하는 모델을 만드는 것이다. 예를 들어, 식물을 식별하도록 설계된 이미지 인식 시스템은 많은 종류의 식물을 각각을 묘사하는 막대한 양의 사진에 관하여 학습할 가능성이 높다. 시스템은 각 종의 사진에서 공통적인 잎의 패턴과 같은 일반적인 규칙을 찾아 새로 입력한 데이터(즉, 사용자가 제출한 사진)에 식별하도록 학습된 종이 포함되어 있는지 여부를 평가할 수 있는 모델을 만든다. 다시 말해, 기계 학습은 과거의 데이터에서 일반화를 끌어내어 향후 입력되는 데이터에 대한

예측을 수행하게 된다.

그러나 AI를 인간 행동 모델링에 사용하는 경우, 의도하지 않게 발생할 수 있는 편향성은 전혀 다른 차원의 문제이다. AI가 사람들의 생활에 결과적으로 영향을 미칠 수 있는 비즈니스 프로세스에 통합됨에 따라 "편향성적" 시스템이 역사적으로 소외된 집단의 구성원들에게 체계적으로 불이익을 줄 위험이 있다. AI 편향성은 부정확하게 수행되거나 인종, 성 정체성, 성적 지향, 연령, 종교, 장애를 포함하지만 이에 국한되지 않는 민감한 특성에 기반하여 사람들을 부당하게 취급할 수 있는 시스템에서 나타날 수 있다.

AI 편향성의 출처 및 유형



디자인

AI 편향성은 AI 라이프 사이클의 여러 단계에서 도입될 수 있다.⁶ AI 시스템의 구상 및 설계 초기 단계의 의사결정 과정에서 다음과 같은 편향성이 도입될 수 있다.

- **잘못된 가정으로 인한 편향성** 어떠한 경우에는 제안된 AI 시스템의 기초가 되는 기본 가정이 본질적으로 편향되어 있어서, 어떠한 형태로도 사회에 배치하는 것이 적합하지 않을 수 있다.

예시

2016년 상하이교통대학 연구원들은 안면 이미징 시스템을 통해 "범죄성"을 예측하기 위해 AI 시스템을 학습시키는 논란이 많은 내용의 논문⁷을 발표했다. 많은 양의 경찰 수배자 사진으로 시스템을 학습시킨 연구원들은 자기 시스템이 사람의 얼굴 구조를 분석하는 것만으로도 거의 90%의 정확도로 "범죄성"을 예측할 수 있다고 주장했다. 놀랄 것도 없이, 이 논문은 순식간에 신랄한 비판의 대상이 되었고, 당연히 평론가들은 해당 모델이 범죄성을 사람의 외모에서 추론할 수 있다는 매우 불완전한(그리고 인과적으로 지지할 수 없는) 가정에 의존하고 있다고 지적했다.⁸

AI 시스템의 목표 변수가 시스템이 실제로 예측하려고 하는 것에 대한 부정확하거나 지나치게 단순한 프록시인 경우에도 가정의 편향성이 발생할 수 있다. 예를 들어, 2019년 연구원들은 긴급치료가 필요할 가능성을 예측하여 환자들⁹을 분류하기 위해 병원에서 널리 사용되는 AI 시스템이 건강하지 못한 소수 환자의 손상보다 건강한 백인 환자의 요구를 체계적으로 우선시한다는 사실을 발견했다. 이 예에서, 시스템이 환자의 의료 요구에 대한 실제 데이터 대신 쉽게 확보할 수 있는 "의료 비용"에 대한 과거 데이터를 기준으로 사용하여 "의료 요구"를 예측하려고 했기 때문에 편향성이 발생했다. 불행히도 소수 환자는 과거에 의료에 대한 접근성이 낮았기 때문에, "의료비"를 소수 환자의 현재 요구에 대한 프록시로 사용하여 사실을 왜곡하고 위험하게 편향성된 결과를 초래하게 된 것이다.

- **역사적 편향성** AI 시스템 학습에 사용한 데이터에 반영되어 있는 역사적 편향성이 반영될 위험이 있다.

예시

영국의 한 의과 대학은 입학 자격을 갖춘 지원자를 가려내는 시스템의 개발에 착수했다. 이 시스템은 이전에 입학한 학생들의 데이터를 사용하여 학습시켰다. 그러나 과거 학교의 입학 결정들이 체계적으로 다른 지원자들과 동일한 자격을 갖춘 소수 인종과 여성을 불리하게 취급했다는 사실이 밝혀졌다. 과거의 편향성이 반영된 데이터를 사용하여 모델을 학습시킴으로써 해당 의과 대학은 의도치 않게 그와 동일한 편향된 입학 패턴을 복제하는 시스템을 만들어낸 것이다.¹⁰

- **샘플링 편향성** 시스템 학습에 사용되는 데이터가 해당 시스템이 사용될 모집단을 잘 대표하지 못하는 데이터인 경우, 이 데이터를 통해 과소 평가되었을 수 있는 집단에서 시스템의 효율성이 저하될 위험이 있다. 이러한 위험은 충분한 양의 대표 데이터를 쉽게 사용할 수 없거나 특정 모집단을 체계적으로 과대 또는 과소 평가하는 방식으로 데이터를 선택하거나 수집할 때 일반적으로 발생한다.

예시

Joy Buolamwini와 Timnit Gebru의 획기적인 연구에서 입증되었듯이, 백인과 남성 얼굴 위주로 불균형하게 구성된 데이터 조합을 사용하여 학습된 안면 인식 시스템은 안색이 어두운 여성의 얼굴을 평가할 때 정확도가 상당히 떨어진다.¹¹

.....

샘플링 편향성은 데이터 수집 관행의 결과로 발생할 수도 있다. 보수가 필요한 도로의 패인 구멍(pothole)을 자동으로 감지하고 보고할 수 있는 시스템을 만들려고 한 보스톤 시의 시도가 대표적인 예이다. 이 프로그램의 초기 버전은 "Street Bump"라고 불리는 스마트폰 앱의 사용자들이 제공하는 데이터에 크게 의존했기 때문에 스마트폰과 데이터 요금제를 이용할 능력이 있는 부유한 사람들이 거주하는 지역으로부터 불균형적으로 많은 보고가 접수되었다. 표본 편향성의 결과 데이터 세트에서 가난한 동네의 포트홀이 과소 표현되어 시스템이 해당 집단 구성원을 부당하게 대우하는 방식으로 보수 자원을 할당할 위험이 발생했다.¹²

- **레이블링 편향성** 유수의 AI 시스템은 학습 알고리즘이 미래의 데이터 입력을 분류하는 데 사용할 수 있는 패턴과 상관 관계를 식별할 수 있도록 학습 데이터를 "레이블링" 하고 있다. 학습 데이터 세트에 레이블을 지정하는 프로세스에는 주관적인 결정이 포함될 수 있는데, 이는 AI 시스템에 인간의 편견을 도입하는 매개체가 될 수 있다.

예시

ImageNet은 AI 연구진이 영상 인식 시스템을 학습시킬 수 있도록 분류하여 레이블을 지정한 1400만 개 이상의 이미지로 구성된 데이터베이스이다. ImageNet은 AI 객체 인식에서 최첨단 기술을 발전시키기 위한 핵심 도구였지만 최근의 연구 결과 사람의 이미지를 포함하는 시스템을 학습시키기 위해 데이터베이스의 분류 및 레이블링 시스템을 사용할 때 심각한 편향성 위험을 생성할 수 있음이 밝혀졌다.

Excavating AI¹³에서 Kate Crawford와 Trevor Paglen은 ImageNet에 있는 사람의 이미지와 관련된 카테고리 및 데이터 레이블에는 이들을 학습 데이터로 사용하는 모든 AI 시스템에서 전파될 수 있는 다양한 "성적, 인종적, 장애인 차별적 및 연령적" 편향성이 반영되어 있다는 것을 보여주었다. 예를 들어 ImageNet 데이터로 학습한 AI 시스템은 흑인 피험자의 이미지를 "가해자" 또는 "범죄자"로 분류할 가능성이 더 높다.¹⁴



필요한 데이터가 수집되면 개발 팀은 모델을 학습시키고 검증하는 데 사용할 수 있도록 데이터를 정리, 처리 및 정상화하여야 한다. 개발자는 또한 사용 중인 데이터의 특성과 해결하려는 문제에 적합한 기계 학습법을 선택하거나 기성 모델을 조정해야 한다. 이 경우 서로 다른 접근 방식을 사용하여 여러가지 다양한 모델을 구축한 후, 그 중에서 가장 성공적인 모델을 선택할 수 있다.¹⁵ 일반적으로, 개발 팀은 또한 모델이 작동하도록 데이터 매개변수를 선택해야 한다. 예를 들어, 수치 점수를 반영하는 데이터는 임계값을 할당하여 "예" 또는 "아니오"라는 답변으로 변환할 수 있다. 예를 들면 X와 같거나 더 큰 점수는 "예"로, 해당 임계값 미만의 점수는 "아니오"로 재지정할 수 있다. 개발 단계에서 나타날 수 있는 편향성에는 다음과 같은 것들이 있다.

- **프록시 편향성** 모델을 학습시킬 때 가중치를 부여할 입력 변수(즉, "특징")를 선택하는 과정은 편향성이 도입될 수 있는 또 다른 중요한 결정 지점이다. 민감한 인구 통계 데이터를 제외하더라도, 시스템이 프록시(proxy)라고 하는 그러한 특성과 밀접한 관련이 있는 특징에 의존하는 경우 편향성이 도입될 수 있다.

예시

겉보기에 무해한 특징을 사용하더라도 민감한 속성과의 상관관계로 인해 프록시 편향성이 발생할 수 있다. 예를 들어, 연구자들은 한 사람이 Mac 또는 PC 노트북을 소유하고 있는지 여부에 대한 정보가 대출금 상환 가능성을 예측할 수 있다는 것을 보여주었다.¹⁶ 따라서 금융기관은 잠재적인 대출 신청자를 선별하기 위한 AI 시스템을 구축할 때 그러한 변수를 포함시키려고 할 수 있다. 그러나 이 특징을 포함시킬 경우 Mac의 소유가 인종과 밀접한 상관관계가 있기 때문에 심각한 프록시 편향성의 위험이 도입된다. 이에 따라 이 특징을 도입할 경우 인종과 밀접한 상관관계가 있지만 실제 신용 위험과는 무관한 특징을 바탕으로 신청자를 체계적으로 냉대하는 시스템이 만들어질 수 있다.

- **집계 편향성** 주요 변수를 간과하는 "일률적용(one-size-fits-all)" 모델을 사용하면 지배적인 하위 집단에만 최적화된 시스템 성능을 얻을 수 있다. 모델이 시스템의 정확도에 실질적으로 영향을 미치는 하위 집단 간의 근본적인 차이를 반영하지 못할 경우 집계 편향성(aggregation bias)이 발생할 수 있다. 드문 현상은 평균과 집계에서 손실될 수 있다. 더 안 좋은 것은, 집계된 모집단 모델이 동일한 모집단의 하위 집단 모드와 다르거나 심지어는 반대되는 거동을 정확하게 예측할 수 있다는 점이다. 이런 현상을 심슨의 역설(Simpson's Paradox)이라 한다.

예시

집계 편향성의 위험은 의료 조건이 인종 및 민족에 걸쳐 사람들에게 영향을 미칠 수 있는 고유한 방식을 진단과 치료에 자주 반영해야 하는 의료 환경에서 특히 심각하다. 예를 들어 당뇨병으로 인한 합병증의 위험은 인종에 따라 크게 다르기 때문에 이러한 차이를 반영하지 않는 한 당뇨병 관련 위험 예측에 사용되는 AI 시스템은 특정 환자에게 그 성능을 제대로 발휘하지 못할 수 있다.

17



배치, 모니터링 및 반복

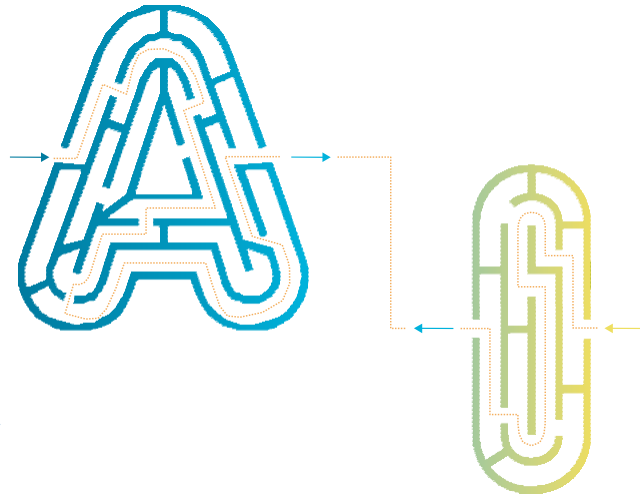
AI 시스템은 필연적으로 모델 학습에 사용된 데이터와 다른 실제 세계의 시나리오를 접할 수밖에 없다. 결과적으로 배치 전에 철저한 검증과 테스트를 거친 시스템조차도 생산에 투입되면 성능이 저하될 수 있다. 따라서 AI 시스템은 라이프 사이클 동안 지속적으로 감성과 평가를 받아야 한다.

- **배치 편향성** 편향성은 시스템이 배치된 후 AI 시스템의 학습 또는 평가에 사용된 데이터가 시스템이 배치되었을 때 마주치는 모집단과 현저하게 다른 경우를 포함하여 다양한 방식으로 발생하여 모델이 제 성능을 발휘할 수 없게 만든다. 배치 편향성은 학습 당시 과다 학습으로 인해(즉, 예측 모델이 학습 데이터를 지나치게 상세하게 학습하여 다른 입력 데이터를 정확하게 일반화할 수 없는 경우) 또는 컨셉트 변화(concept drift)로 인해(즉, 목표 변수와 학습 데이터 사이의 관계 변화로 인해 성능 저하가 발생하는 경우) 모델이 학습한 데이터를 넘어 안정적으로 일반화할 수 없는 경우에 발생할 수 있다.
- **오남용 편향성** 배치 편향성은 하나의 목적으로 구축된 AI 시스템 또는 기능이 예기치 않거나 의도하지 않은 방식으로 사용되는 경우에도 발생할 수 있다.

「AI 위험 관리의 필요성」

위험 관리란 무엇인가?

위험 관리는 위험을 식별하고 잠재적 영향을 완화하기 위한 방법론을 설정하여 설계를 통해 시스템의 신뢰성을 확보하는 프로세스이다. 위험 관리 프로세스는 빠르게 진화하는 기술과 매우 동적인 위협 환경의 조합으로 인해 기존의 "법규 준수(compliance)" 기반 접근 방식이 제 기능을 하지 못하는 사이버 보안 및 개인 정보 보호와 같은 상황에서 특히 중요하다. 시대 변화를 따라가지 못하는 규범화된 정적 요구 사항으로 제품 또는 서비스를 평가하는 대신, 위험 관리는 제품 또는 서비스의 라이프 사이클 전반에 걸쳐 위험 완화에 도움이 되도록 규정 준수 책임을 개발 파이프 라인에 통합하려고 한다. 효과적인 위험 관리는 제품의 설계, 개발 및 배치 중 주요 지점에서 조직의 개발 팀과 규정 준수 담당자 간의 공동 작업을 촉진하는 거버넌스 프레임워크(governance framework)를 기반으로 한다.



편향성 위험 관리

AI 시스템을 개발하고 사용하는 조직은 어느 누군가의 인구통계학적 특성에 따라 부당하게 불리하거나 유해한 결과가 산출되는 등 편향성이 발생되지 않도록 조치를 취해야 한다. 이러한 편향성으로 인해 발생할 수 있는 위험을 효과적으로 방지하기 위해서는 다음과 같은 이유로 위험 관리 접근법이 필요하다.

"편향성"과 "공정성"은 환경에 따라 다르다.

시스템이 "공정한" 방식으로 작동하는지를 평가하는 방법에 관하여 보편적으로 합의된 바가 없기 때문에, AI 시스템의 편향성을 제거하는 것은 불가능하다. Arvind Narayanan 교수가 잘 설명했듯이 시스템이 공정하게 작동하는지를 평가하기 위해 사용할 수 있는 정의¹⁸(즉, 수학적 기준)는 최소 21가지가 있으며, AI 시스템이 그 모든 정의를 동시에 충족하는 것은 불가능하다.

공정성에 대한 보편적인 정의가 존재하지 않기 때문에 개발자는 대신 자신이 만들고 있는 시스템의 특성을 평가하여 편향성을 평가하는 측정 기준 가운데 어떤 것이 발생 가능한 위험의 완화에 가장 적합한지를 결정하여야 한다.

편향성을 완화하기 위한 과정에서 '트레이드 오프'가 발생할 수 있다.

한 그룹의 편향성을 완화하기 위해 개입할 경우 다른 그룹의 편향성이 증가하거나 시스템의 전체 정확도가 감소할 수 있다.¹⁹

위험 관리는 환경에 맞는 방식의 트레이드 오프를 탐색하기 위한 메커니즘을 제공한다.

편향성은 배치 후에도 발생할 수 있다.

시스템을 배치하기 전에 철저히 평가한 경우에도 인구통계 분포가 학습 및 테스트 데이터의 구성과 다른 환경에서 잘못 사용되거나 배치될 경우 편향된 결과가 발생할 수 있다.

「효과적인 위험 관리를 위한 초석」

위험 관리의 목표는 AI 시스템의 라이프 사이클 전반에 걸쳐 발생할 수 있는 잠재적 위험을 식별하고 완화하기 위한 반복 가능한 프로세스를 확립하는 것이다. 포괄적인 위험 관리 프로그램에는 두 가지 핵심 요소가 있다.

1

조직의 위험 관리
기능을 지원하는
거버넌스 프레임워크

2

위험 식별 및 완화를
위한 영향 평가를
수행하는 확장 가능한
프로세스

거버넌스 프레임워크

AI 위험 관리를 효과적으로 수행하기 위해서는 시스템 라이프 사이클 전반에 걸쳐 정책, 프로세스를 규정하고 위험을 식별 및 완화하고 관리하는 인력을 설정하는 거버넌스 체계에 의해 뒷받침되어야 한다. 이러한 거버넌스 프레임워크의 목적은 제품 개발, 규정 준수, 마케팅, 영업 및 고위 경영진을 포함한 조직 단위 전반에 걸쳐 AI 시스템의 설계, 개발 및 배치 과정에서의 효과적인 위험 관리를 촉진하기 위한 각 독립체의 역할과 책임에 대한 이해를 증진하는 것이다. 위험 관리 거버넌스 프레임워크의 주요 기능은 다음과 같다.

정책 및 프로세스

거버넌스 프레임워크의 핵심은 위험 관리에 대한 조직의 접근 방식을 설정하는 일련의 공식적인 정책이다. 이러한 정책에서는 조직의 위험 관리 목표, 목표 달성을 위해 사용할 절차, 그리고 규정 준수 평가에 사용할 기준을 정의해야 한다.

- **목표** AI 위험 관리는 조직이 핵심 가치에 부합하는 방식으로 AI를 개발하고 사용하도록 보장하는 것을 목표로 조직의 광범위한 위험 관리 기능 내에서 맥락화 되어야 한다. 이를 위해 거버넌스 프레임워크는 조직이 이러한 가치를 훼손하지 않도록 어떠한 방식으로 위험을 관리하는 지 확인해야 한다.
- **프로세스** 거버넌스 프레임워크는 AI 라이프 사이클의 각 단계에서 위험을 식별하고, 위험의 중요성을 평가하고, 위험을 완화하기 위한 프로세스와 절차를 수립해야 한다.
- **평가 메커니즘** 거버넌스 프레임워크는 조직이 정책 및 절차가 지정된 대로 수행되고 있는지 여부를 평가하기 위해 사용할 측정 기준이나 표준과 같은 메커니즘을 설정해야 한다.
- **정기적 검토** AI 기능이 계속 발전하고 기술이 새로운 용도로 활용됨에 따라 조직이 주기적으로 AI 거버넌스 프레임워크를 검토 및 업데이트하여 목적에 부합하게, 그리고 진화하는 위험 환경을 해결할 수 있게 유지하는 일이 중요하다.



경영진의 감독 역할 AI 개발자와 AI 배치 책임자는 충분한 경영진의 감독으로 뒷받침되는 거버넌스 프레임워크를 유지해야 한다. 거버넌스 프레임워크 정책의 내용을 개발하고 승인하는 것 외에도 고위 경영진은 회사의 AI 제품 개발 라이프 사이클을 적극적으로 감독하는 역할을 수행해야 한다. 결과적으로 사람들에게 부정적인 영향을 미칠 수 있는 고위험 시스템의 경우 회사 지도부는 "진행/불가" 결정을 내릴 책임이 있어야 한다.

직원, 역할 및 책임

위험 관리의 효율성은 AI 라이프 사이클 전반에 걸쳐 의사 결정을 이끌 수 있는 다기능 전문가 그룹의 구성에 달려 있다. 조직의 규모와 개발 또는 배치 중인 시스템의 특성에 따라 여러 사업부의 직원이 위험 관리 책임을 담당할 수 있다. 따라서 거버넌스 프레임워크는 조직 내에서 AI 위험 관리와 관련된 역할과 책임이 있는 직원을 파악하고 보고 라인, 권한 및 필요한 전문 지식을 명확하게 제시하여야 한다. 역할과 책임을 분배할 때 조직은 독립성, 능력, 영향력 및 다양성을 우선적으로 고려해야 한다.

- **독립성** 위험 관리는 직원을 개별 계층의 독립적 검토를 용이하게 하는 방식으로 구성할 때 가장 효과적이다. 예를 들어, 위험 관리 책임을 다음과 같이 여러 팀 사이에 분할할 수 있다.
 - **제품 개발 팀** AI 제품 및 서비스의 설계 및 개발에 관여하는 엔지니어, 데이터 과학자 및 도메인 전문가.
 - **규정 준수 팀** 고위험 AI 시스템에 대한 영향 평가 개발 등 회사의 AI 개발 정책 및 관행 준수를 감독하는 법무 팀, 규정 준수 팀, 도메인 전문가 팀 및 데이터 전문가 팀 등 다양한 팀.
 - **거버넌스 팀** 이상적으로는 조직의 AI 거버넌스 프레임워크 및 위험 관리 프로세스에 대한 효과적인 감독을 개발, 유지 및 보장할 책임이 있는 고위 경영진이 이끄는 팀.
- **역량, 자원 조달력 및 영향력.** 위험 관리 책임이 있는 직원은 거버넌스 기능을 수행하기 위해 적절한 교육과 리소스를 제공받아야 한다. 직원에게 권한이 부여되고 위험을 처리 및/또는 확대하는 결정을 내릴 권한과 적절한 동기가 부여되어 있는지 확인하는 것도 똑같이 중요하다. 예를 들어, 조직은 위험 관리 담당자가 주요 위험 영역과 의사 결정에 있어서 경영진 수준의 가시성을 확보할 수 있도록 경영진 의사 결정권자와 협력할 수 있는 명확한 직통 경로(escalation path)를 설정해야 한다.



다양성 AI 시스템의 사회 기술적 특성으로 인해 시스템 개발 및 감독에 참여하는 팀 내에서 다양성의 우선 순위를 정하는 것이 매우 중요하다. 개발 및 감독 프로세스는 팀 구성원이 AI 시스템의 영향을 받거나 상호 작용할 수 있는 사용자의 요구와 우려 예측에 도움이 되는 다양한 관점과 배경을 제시할 때 가장 효과적이다. "알고리즘 개발은 윤리적 및 정치적 가치를 포함하여 인식하지 못할 수 있는 개발자의 가정을 암묵적으로 암호화하기" 때문에, 조직은 AI 설계 및 개발 프로세스의 라이프 사이클 동안 다방면에 걸친 생생한 경험을 반영하고 전통적으로 과소 표현된 관점이 포함되도록 팀을 구성하는 것이 중요하다.²⁰ 조직의 다양성이 부족한 경우 조직은 외부 이해관계자와 협의하여 특히 시스템의 영향을 받을 수 있는 과소 표현된 집단으로부터의 피드백을 요청해야 한다.

영향 평가

AI 위험을 효과적으로 관리하기 위해, 조직은 대중에게 중대한 영향을 미칠 수 있는 모든 시스템에 대한 영향 평가를 수행하기 위한 강력한 프로세스를 구현해야 한다. 영향 평가는 환경 보호에서 데이터 보호에 이르기까지 다른 다양한 분야에서 시스템이 대중에게 미칠 수 있는 잠재적 위험을 설명하는 방식으로 설계되었음을 입증하여 신뢰도를 증진시키는 책임 메커니즘으로 널리 사용된다. 요컨대, 영향 평가의 목적은 시스템이 제기할 수 있는 위험을 식별하고, 시스템이 발생시킬 수 있는 위해성의 정도를 정량화하고, 그러한 위험을 허용 가능한 수준으로 완화하기 위해 취한 모든 단계를 문서화하는 것이다.

영향 평가 프로세스는 평가 중인 시스템의 특성과 시스템이 야기할 수 있는 위해의 유형을 다루도록 조정하여야 한다. 위험도가 매우 낮은 시스템(예: 문서에 사용되는 글꼴 예측에 사용되는 시스템)의 경우 전체적인 영향 평가가 필요하지 않을 수 있다. 그러나 대중에게 중대한 피해를 줄 수 있는 고유한 위험이 내재된 시스템의 경우에는 반드시 전체 영향 평가를 수행해야 한다. AI를 적용할 수 있는 응용 프로그램의 범위가 엄청나게 다양하기 때문에 위험을 식별하고 완화하기 위한 "일률적인" 접근 방식은 존재하지 않는다. 대신, 영향 평가 프로세스는 AI 시스템의 특성과 고유한 위험 유형 및 발생할 수 있는 잠재적인 위해를 취급할 수 있도록 조정되어야 한다. 시스템에 고유한 물질적 위해의 위험이 내포되어 있는지 여부를 결정하기 위해 이해 관계자는 다음 사항을 고려해야 한다.

- **사람에게 미칠 수 있는 영향** 영향 평가는 신용 또는 주택에 대한 접근 능력과 같이 사람들에게 결과적인 영향을 미칠 수 있는 의사 결정 과정에서 AI 시스템이 사용되는 상황에서도 마찬가지로 중요하다.
- **시스템 환경과 목적** AI 시스템의 특성과 이를 사용할 설정을 평가하는 것은 영향 평가의 필요성과 적절한 범위를 결정하는 좋은 출발점이다. 영향 평가는 잠재적 위험의 심각성 및/또는 가능성이 높은 영역(예: 의료, 교통, 금융)에서 사용될 고위험 AI 시스템에서 특히 중요하다.
- **인간 감독의 정도** AI 시스템의 완전 자동화 정도가 내재된 고유한 위험에 영향을 미칠 수도 있다. 고도로 숙련된 전문가에게 권고하도록 설계된 시스템은 유사하게 배치된 완전 자동화 시스템보다 고유한 위험을 더 적게 내포할 수 있다. 물론 단순히 인간이 참여한다고 해서 AI 시스템이 위험으로부터 자유로운 것은 결코 아니다. 대신 인간-컴퓨터 상호 작용의 본질을 전체적으로 조사하여 인간의 감독이 AI 시스템의 고유한 위험을 어느 정도 완화할 수 있는지를 결정할 필요가 있다.
- **데이터 유형** 시스템 학습에 사용되는 데이터의 특성 또한 시스템의 고유한 위험을 밝힐 수 있다. 예를 들어 인간의 특성이나 행동과 관련된 학습 데이터를 사용할 경우 시스템의 편향성에 대해 면밀한 조사가 필요할 수 있다.

「AI 편향성 위험 관리 프레임워크」

잠재적으로 AI 편향성 위험이 있는 시스템에 대한 영향 평가를 수행하는 조직을 지원하는 AI 편향성 위험 관리 프레임워크를 아래에서 개략적으로 설명한다. 이 프레임워크는 AI 시스템의 라이프 사이클 전반에 걸쳐 발생할 수 있는 편향성의 원인을 파악하는 프로세스를 제시하는 것 외에 이러한 위험을 완화하기 위해 사용할 수 있는 모범 사례도 제시한다.

이 프레임워크는 AI 개발자 및 AI 배치 책임자 조직이 다음과 같은 목적으로 사용할 수 있는 보증 기반 책임 메커니즘이다.

- **내부 프로세스 지침** AI 개발자 및 AI 배치 책임자는 이 프레임워크를 내부 프로세스에 대한 역할, 책임 및 기대치를 구성하고 설정하는 도구로 사용할 수 있다.
- **공급업체 관계** AI 배치 책임자는 이 프레임워크를 사용하여 구매 결정을 안내하거나 AI 위험을 적절히 설명하였음을 보증하는 공급업체 계약을 진행할 수 있다.
- **학습, 인식 및 교육** AI 개발자 및 AI 배치 책임자는 이 프레임워크를 사용하여 AI 시스템 개발 및 사용에 관련된 직원을 위한 내부 훈련 및 교육 프로그램을 구축할 수 있다. 또한 이 프레임워크는 AI 편향성 위험 관리에 대한 조직의 접근 방식에 대해 경영진을 교육하는 데 유용한 도구를 제공할 수 있다.
- **신뢰와 자신감** AI 개발자는 제품의 기능과 AI 편향성 위험을 완화하기 위한 접근 방식에 관한 정보를 대중에게 전달하고자 할 수 있다. 그런 의미에서 이 프레임워크는 조직이 윤리적인 AI 시스템을 구축하려는 노력에 대해 대중과 소통하는 데 도움이 될 수 있다.
- **보증 및 책임** AI 개발자와 AI 배치 책임자는 이 프레임워크를 시스템의 라이프 사이클 전반에 걸쳐 AI 위험을 관리하기 위한 각자의 역할과 책임에 대해 소통하고 조정하는 기반으로 사용할 수 있다.
- **사고 대응** 예기치 않은 사고가 발생한 후 이 프레임워크에 명시된 프로세스 및 문서를 활용하여 AI 개발자와 AI 배치 책임자가 시스템 성능 저하 또는 결함의 잠재적 원인을 식별하는 데 도움이 되는 추적 감사를 실시할 수

AI 라이프 사이클 단계

이 AI 편향성 위험 관리 프레임워크는 AI 시스템의 생성 및 사용과 관련된 주요 반복 단계를 나타내는 AI 라이프 사이클 단계를 중심으로 구성된다.



설계 단계

- 프로젝트 구상** AI 설계의 초기 단계에는 시스템이 해결하려는 "문제"를 식별하여 공식화하고 모델이 해당 목표를 달성하는 방법을 초기에 매핑하는 작업이 포함된다. 이 단계에서 설계 팀은 시스템의 목적과 구조를 정의한다. 시스템 특성에 따라 설계 팀은 시스템이 예측하려는 대상 변수를 식별한다. 예를 들어, 소비자의 심박수를 분석하여 그 사람이 뇌졸중 또는 심장 질환의 위험에 있는지를 예측할 수 있는 불규칙성(즉, 대상 변수)을 모니터링하는 피트니스 앱(fitness app)이 있다. 시스템 설계 프로세스의 초기 단계에서 편향성 위험 관리 프레임워크의 목표는 AI 사용이 현재 프로젝트에 적합한지 여부를 확인하는 것이다.

잠재적인 위험은 다음과 같다:

- 잘못된 가정으로 인한 편향성** 목표 변수는 유해한 편향성을 영속화할 수 있는 고유의 편견이나 잘못된 가정을 반영할 수 있다. 어떤 경우에는 제안된 AI 시스템의 기초가 되는 기본 가정이 어떠한 형태의 공개 배치에도 적합하지 않을 정도로 본질적으로 편향성되어 있을 수 있다.
- 데이터 수집** 시스템 목표가 정의되면 개발자는 향후 데이터 입력에 대한 예측을 수행할 수 있는 패턴을 식별하도록 모델을 학습시키기 위해 사용할 주요 데이터(corpus of data)를 수집해야 한다. 이 학습 데이터는 의도치 않게 AI 시스템에 여러 가지 방법으로 편향성을 도입할 수 있다. 잠재적인 위험은 다음과 같다.
 - 역사적 편향성** 과거의 편향성을 반영할 수 있는 데이터를 사용하여 AI 시스템을 학습시키면 이러한 불평등이 더욱 심화될 위험이 있다.
 - 샘플링 편향성** 편향성의 위험은 AI 시스템 학습에 사용되는 데이터가 배치될 모집단을 대표하지 않을 때에도 발생한다. 과소 표현된 데이터로 학습을 받은 AI 시스템은 과도 또는 과소 표현된 집단의 구성원에 대한 예측을 할 때 효과적으로 작동하지 않을 수 있다.
 - 레이블링 편향성** 많은 AI 시스템은 찾아야 할 패턴을 식별할 수 있도록 학습 데이터에 레이블이 지정되어야 한다. 학습 데이터 세트에 레이블을 지정하는 프로세스는 AI 시스템에 편향성을 도입하는 매개체(vector)가 될 수 있다.



개발 단계

- **데이터 준비 및 모델 정의** AI 라이프 사이클의 다음 단계는 모델을 학습시킬 데이터를 준비하는 것이다. 이 프로세스에서 개발 팀은 알고리즘이 향후 예측 규칙의 기반이 되는 패턴과 상관 관계를 찾으면서 평가할 학습 데이터 변수(즉, "특징")를 정리, 정규화 및 식별한다. 팀은 또한 시스템을 구동할 알고리즘 모델 유형(예: 선형 회귀, 로지스틱 회귀, 심층 신경망)을 선택하는 등 시스템의 기본 아키텍처를 구축해야 한다.²¹ 데이터가 준비되고 알고리즘이 선택되면 팀은 시스템을 학습시켜 향후 데이터 입력에 대한 예측을 할 수 있는 기능 모델을 만든다. 잠재적 위험에는 다음과 같은 것이 포함된다.
 - **프록시 편향성** 학습 데이터에서 특징을 선정하고 모델링 접근 방식을 선택하는 프로세스에는 모델의 대상 변수에 대한 예측과 관련성이 있는 것으로 고려하여야 하는 변수를 인간이 결정하는 과정이 포함된다. 이러한 개입은 보호되는 계층에 대한 프록시 역할을 하는 변수에 의존하는 등 의도치 않게 시스템에 편향성을 유발할 수 있다.
 - **집계 편향성** 모델이 시스템의 정확도에 실질적으로 영향을 미치는 하위 집단 간의 근본적인 차이를 반영하지 못할 경우 집계 편향성(aggregation bias)이 발생할 수 있다. 주요 변수를 간과하는 "일률적인" 모델을 사용하면 지배적인 하위 그룹에 대해서만 최적화된 시스템 성능을 얻을 수 있다.
- **모델 검증, 테스트 및 수정** 모델이 학습된 후에는 시스템이 의도한 대로 작동하는지를 확인하기 위한 검증과 시스템 출력이 의도하지 않은 편향성을 반영하지 않음을 입증하기 위한 테스트를 거쳐야 한다. 검증 및 테스트 결과에 기초하여, 허용할 수 없는 것으로 간주되는 편향성의 위험을 완화하기 위해 모델을 수정할 필요가 있을 수 있다.



배치 단계

- **배치 및 사용** AI 개발자는 배치 전에 시스템을 평가하여 초기 설계 및 개발 단계에서 식별된 위험이 회사의 거버넌스 정책에 부합하는 방식으로 충분히 완화되었는지 판단해야 한다. 식별된 위험이 시스템의 오남용을 통해 발생할 수 있는 범위까지 AI 개발자는 그러한 위험을 완화할 제품 특징(예: 오남용 위험을 줄이는 사용자 인터페이스)을 통합하고, 위험을 악화시킬 수 있는 사용을 금지(예: 최종 사용자 실시권 계약)하며, AI 배치 책임자에게 자체 영향 평가를 수행할 수 있도록 충분한 문서를 제공하는 등 오남용을 통제하기 위해 노력해야 한다.

AI 시스템을 사용하기 전에 AI 배치 책임자는 AI 개발자가 제공 한 문서를 검토하여 시스템이 자체 AI 거버넌스 정책에 부합하는지 여부를 평가하고 배치 관련 위험 관리 책임이 명확하게 지정되어 있는지 여부를 확인해야 한다.

배치 후 위험 관리 책임의 일부는 AI 개발자가 처리할 수 있지만, AI 배치 책임자는 종종 시스템 성능을 모니터링하고 위험성 프로필과 일치하는 방식으로 작동하는지 여부를 평가할 책임이 있다. 잠재적인 위험은 다음과 같다.

- **배치 편향성** AI 시스템은 정적인 순간을 나타내면서 모델의 일관성 있고 정확한 예측 능력을 저해할 수 있는 "노이즈(noise)"를 필터링한 데이터로 학습시킨다. 현실 세계에 배치될 경우 AI 시스템은 개발 및 테스트 환경과는 다른 조건에 직면할 수밖에 없다. 또한 현실 세계는 시간이 지나면서 변화하기 때문에, 모델이 나타내는 시간의 간략한 정보(snapshot)는 데이터 변수 간의 관계가 발전함에 따라 자연스럽게 정확도가 떨어질 수 있다. 배치된 AI 시스템의 입력 데이터가 학습 데이터와 실질적으로 다를 경우 시스템이 "드리프트(drift)"하고 모델의 성능이 편향성의 위험을 악화시키는 방향으로 저하될 위험이 있다. 예를 들어 특정 국가에서 사용하도록 설계(및 테스트)된 AI 시스템을 인구 통계치가 크게 다른 국가에 배치하면 시스템이 제대로 작동하지 않을 수 있다.
- **오남용 편향성** AI 시스템을 설계된 조건과 크게 다른 환경에 배치하거나 의도된 용도와 일치하지 않는 목적을 위하여 배치할 경우 편향성의 위험이 높아질 수 있다.

프레임워크 구조

프레임워크는 전체 시스템의 라이프 사이클에 걸쳐 AI 편향성의 위험을 확인하고 완화하기 위한 모범 사례를 식별한다. 프레임워크는 다음과 같이 구성된다.

- **기능** 가장 높은 수준에서 기본적인 AI 위험 관리 활동을 나타내며, 영향 평가와 위험 완화 모범 사례로 나뉜다.
- **카테고리** AI 라이프 사이클의 각 단계에서 기능을 실행하는 데 필요한 활동과 프로세스를 설정한다. 즉, 카테고리는 영향 평가를 수행하는 단계를 제시하고 관련 위험의 관리에 사용할 수 있는 해당 위험 완화 모범 사례를 식별한다.
- **진단서** 카테고리에서의 실행을 위해 취해야 할 개별 조치를 지정한다. 각 카테고리의 결과 달성을 지원하는 일련의 결과를 제공한다.
- **구현에 대한 의견** 진단서에 설명된 결과를 달성하기 위해 필요한 추가 정보를 제공한다.
- **도구 및 리소스** 이해 관계자가 AI 라이프 사이클의 각 단계와 관련이 있는 편향성 위험을 완화하기 위해 사용할 수 있는 다양한 외부 지침 및 툴킷(toolkit)을 식별한다. 프레임워크에서 식별된 특정 도구 및 리소스는 완전하지 않으며 정보 제공 목적으로만 강조 표시된다.

이해관계자의 역할 및 책임

이 프레임워크는 본질적으로 동적인 AI 시스템의 특성을 반영하여 시스템 설계, 개발 및 배치의 다양한 측면에서 역할을 할 수 있는 다양한 이해 관계자를 설명하기 위한 것이다. AI 개발 또는 배치의 단일 모델이 없기 때문에 추상적으로 프레임워크의 많은 위험 관리 기능에 대한 역할을 할당하거나 특정한 책임을 위임하는 것은 불가능하다. 그러나 일반적으로 시스템 라이프 사이클 동안 AI 위험 관리의 특정 측면에 대해 다양한 수준의 책임을 부담할 수 있는 세 부류의 이해 관계자가 있다.

- **AI 개발자** AI 개발자는 AI 시스템의 설계 및 개발을 담당하는 조직이다.
- **AI 배치 책임자** AI 배치 책임자는 AI 시스템을 채택하고 사용하는 조직이다. (법인이 자체 시스템을 개발하는 경우에는 해당 법인이 AI 개발자이자 AI 배치 책임자이다.)
- **AI 최종 사용자** AI 최종 사용자는 AI 시스템 사용을 감독할 책임이 있는 개인(주로 AI 배치 책임자의 직원)이다.

위험 관리 책임은 많은 경우 AI 시스템의 개발 및 배치 모델에 따라 이러한 이해 관계자들 사이에 배분된다.

다양한 AI 개발 및 배치 모델

위험 관리 책임은 개발 중인 AI 시스템의 특성에 따라, 그리고 기본 모델을 학습하는 목적과 수단을 어느 당사자가 결정하느냐에 따라 이해 관계자 사이에서 적절하게 배분된다. 예를 들면 다음과 같다.

- **범용 정적 모델** AI 개발자는 모든 고객(즉, AI 배치 책임자)에게 사전 학습된 정적 모델을 제공한다.
- AI 개발자는 모델 위험 관리의 대부분의 측면에 대한 책임을 진다.
- **주문형 맞춤 모델** AI 개발자는 자체 데이터를 사용하여 모델을 주문 제작 및/또는 재학습 할 수 있는 AI 배치 책임자에게 사전 학습된 모델을 제공한다.
- 위험 관리 책임은 AI 개발자와 AI 배치 책임자가 공동으로 부담한다.
- **개인별 맞춤 모델** AI 개발자가 AI 배치 책임자를 대신하여 AI 개발자의 데이터를 사용하여 개인별 맞춤형(bespoke) AI 모델을 학습시킨다.
- 위험 관리는 AI 개발자와 AI 배치 책임자의 공동 책임이며, 대부분의 의무는 AI 배포자가 부담한다.

BSA의 AI 편향성 위험 관리 프레임워크



디자인

기능	카테고리	진단서	구현에 대한 의견
프로젝트 구상			
영향 평가	목표와 전제 식별 및 문서화	시스템의 의도 및 목적을 문서화한다.	<ul style="list-style-type: none"> • 시스템의 목적이 무엇인가? 즉, 어떤 "문제"를 해결하는가? • 시스템의 의도된 사용자는 누구인가? • 시스템은 어디서 어떻게 사용되는가? • 잠재적인 오남용은 무엇인가?
		모델의 의도된 효과를 명확하게 정의한다.	예측, 분류, 추천, 순위 지정 또는 발견을 위한 모델은 무엇인가?
		의도된 사용례와 시스템을 배치하게 된 정황을 명확하게 정의한다.	
	공정성 평가 측정 기준 선정 및 문서화	AI 시스템 편향성을 평가하는 기준으로 사용할 "공정성" 측정 기준을 식별한다.	"공정성"의 개념은 매우 주관적이며 이를 평가할 수 있는 측정 기준은 수십 가지가 있다. 모든 공정성 측정 기준을 동시에 충족하는 것은 불가능하기 때문에 개발 중인 AI 시스템의 특성에 가장 적합하고 관련 법적 요구 사항과 일치하는 측정 기준을 선택해야 한다. AI 라이프 사이클의 후반 단계를 알리기 위해 공정성 측정 기준이 선택 및/또는 제외된 근거를 문서화하는 것이 중요하다.
	이해 관계자 영향 문서화	시스템의 영향을 받을 수 있는 이해 관계자 그룹을 식별한다.	이해 관계자 집단에는 AI 배치 책임자, AI 최종 사용자, 영향을 받는 개인(즉, AI 시스템과 상호 작용하거나 영향을 받을 수 있는 일반인)이 포함된다.
		각 이해 관계자 집단에 대해 시스템의 의도된 용도와 합리적으로 예측 가능한 오남용을 고려하여 잠재적인 이점과 잠재적인 악영향을 문서로 기록한다.	
		시스템의 특성으로 인하여 사용자 인구 통계를 기반으로 하는 잠재적인 편향성 관련 손해가 발생하기 쉬운지 여부를 평가한다.	사용자 인구 통계에는 인종, 성별, 연령, 장애 상태 및 이들의 교집합이 포함될 수 있지만 이에 국한되지 않는다.
위험 완화 문서화	편향성 위험이 존재하는 경우, 위험을 완화하기 위한 노력을 문서로 기록한다.		



디자인

기능	카테고리	진단서	구현에 대한 의견	
프로젝트 구상				
영향 평가 (계속)	위험 완화 문서화	식별된 위험과 각 위험의 잠재적 손해를 측정하는 방법과 완화 전략의 효과를 평가하는 방법을 문서로 기록한다.		
		편향성 위험이 존재하는 경우, 위험을 완화하기 위한 노력을 문서로 기록한다.		
		위험이 완화되지 않은 경우 해당 위험을 허용되는 것으로 간주한 이유를 문서로 기록한다.		
위험 완화 모범 사례	독립성 및 다양성	영향 평가를 알리기 위해 다양한 이해 관계자로부터 피드백을 구한다.	이 초기 단계에서 식별된 위험은 개발 및 영향 평가 프로세스의 이후 측면에 관한 정보를 제공하기 때문에 다양한 실제 경험, 문화적 배경 및 주제 관련 전문 지식을 가진 사람들에게 다양한 의견을 요청하여 발생할 수 있는 잠재적 손해를 전반적으로 이해하는 것이 중요하다. 사내 직원이 주제 관련 지식이나 문화적 다양성이 부족한 경우 그 범위 내에서 외부 전문가와 상의하거나 시스템의 영향을 받을 수 있는 집단 구성원의 의견을 구해야 할 경우도 있다.	
		투명한 문서 관리	개발 프로세스 전반에 걸쳐 위험과 의도하지 않은 잠재적인 영향을 모니터링할 수 있도록 AI 정보 경로(pipeline)의 후반 단계에서 작업하는 직원과 영향 평가 문서를 공유한다.	
		책임 및 거버넌스	고위 경영진이 잠재적인 고위험 AI 시스템에 대해 충분한 보고를 받았는지 확인한다.	"고위험"으로 간주되는 시스템에 대한 영향 평가 문서는 용이하게 "진행/중단"결정을 할 수 있도록 고위 경영진과 공유해야 한다.
데이터 수집				
영향 평가	데이터 출처 기록 유지	AI 모델 학습에 사용된 데이터를 "재창조"하고 그 결과를 재현할 수 있는지를 확인하고 데이터 출처에 대한 자료 업데이트를 모니터링할 수 있도록 충분한 기록을 유지한다.	기록에는 다음 사항이 포함되어 있어야 한다. <ul style="list-style-type: none"> • 데이터 출처 • 데이터의 기원(예: 데이터 생성자, 생성 시기, 생성 목적, 생성 방법) • 데이터 및 데이터 거버넌스 규칙의 의도된 사용 및/또는 제한(예: 데이터 소유 주체, 데이터 보존 기간(또는 폐기 시점), 사용 제한 여부) • 알려진 데이터 제한(예: 누락된 요소?) • 데이터를 샘플링하는 경우 샘플링 전략 • 데이터 업데이트 및 추적 버전 	



디자인

기능	카테고리	진단서	구현에 대한 의견
데이터 수집			
영향 평가 (계속)	편향성 가능성 있는 데이터 조사	과거 편향성에 대한 데이터를 면밀히 조사한다.	데이터 출처를 조사하고 과거의 편향성을 반영할 수 있는 가능성을 평가한다.
		데이터의 "대표성"을 평가한다.	<ul style="list-style-type: none"> • 학습 데이터의 인구 통계 분포를 시스템이 배치될 곳의 모집단과 비교한다. • 시스템과 상호 작용할 가능성이 있는 하위 모집단이 충분히 표현되었는지 여부를 평가한다.
		데이터 레이블링 방법론을 면밀히 조사한다.	<ul style="list-style-type: none"> • 데이터 레이블링에 사용된 인력 및 프로세스를 문서화한다. • 제3자 데이터의 경우 레이블링(및 관련 방법론)에 잠재적인 편향성 출처(source)가 있는지 면밀히 조사한다.
	위험 완화 문서화	편향성을 완화하기 위해 데이터를 증강, 조작 또는 재조정하였는지 여부와 그 방법을 문서로 기록한다.	
위험 완화 모범 사례	독립성 및 다양성	데이터 세트에 대한 강력한 조사를 용이하게 하기 위해 데이터 검토 팀에는 주제에 관한 전문 지식과 실제 경험 측면에서 다양한 인력이 포함되어야 한다.	데이터에 있는 편향성의 잠재적 출처를 효과적으로 식별하려면 데이터를 추출한 영역에 대한 친숙함과 데이터가 생성된 역사적 환경 및 제도에 대한 깊은 이해를 포함하여 다양한 전문 지식과 경험이 필요하다. 사내 인력의 다양성이 부족한 범위에서 외부 전문가 또는 잠재적으로 영향을 받는 이해 관계자 집단과의 협력이 필요할 수 있다.
		대표성 없는 데이터 재조정	추가 데이터로 재조정하는 것을 고려한다.
	합성 데이터로 재조정하는 것을 고려한다.	불균형한 데이터 세트는 과소 표현된 집단의 데이터를 "과다 샘플링(oversampling)"하여 잠재적으로 균형을 맞출 수 있다. 일반적인 과다 샘플링 방법은 과소 표현된 집단에서 새로운 "합성" 데이터를 생성하는 합성 소수 과다 샘플링 기법(SMOTR)이다	



디자인

기능	카테고리	진단서	구현에 대한 의견
데이터 수집			
위험 완화 모범 사례 (계속)	데이터 레이블링	객관적이고 확장 가능한 레이블링 지침 설정	<ul style="list-style-type: none"> 레이블링 편향성 가능성을 완화하려면 개별 레이블링 결정을 위한 객관적이고 반복 가능한 프로세스를 설정하는 명확한 지침을 데이터 레이블링을 담당하는 직원에게 제공해야 한다. 편향성 위험이 높은 영역에서 레이블링 담당자는 주제에 관한 적절한 전문 지식을 갖추고 있어야 하며, 잠재적인 무의식적 편향성을 인식하는 교육을 받아야 한다. 고위험 시스템의 경우 레이블 품질을 모니터링하기 위한 품질 보증 메커니즘을 설정해야 할 수 있다.
	책임 및 거버넌스	데이터 레이블링 프로세스를 포괄적인 데이터 전략에 통합한다.	조직 데이터 전략을 수립하면 향후 참조할 수 있도록 데이터를 면밀히 조사하려는 회사의 노력을 문서화함으로써 데이터를 일관성 있게 평가하고 노력이 중복되는 것을 방지할 수 있다.

디자인: 위험 완화 도구 및 리소스

프로젝트 구상

- Aequitas Bias and Fairness Audit Toolkit**
 Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani, University of Chicago Center for Data Science and Public Policy (2018), <http://www.datasciencepublicpolicy.org/projects/aequitas/>.
- Diverse Voices Project | A How-To Guide for Facilitating Inclusiveness in Tech Policy**
 Lassana Magassa, Meg Young, and Batya Friedman, University of Washington Tech Policy Lab, <https://techpolicylab.uw.edu/project/diverse-voices/>.

데이터 편집

- Datasheets for Datasets**
 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, arXiv:1803.09010v7, (March 19, 2020), <https://arxiv.org/abs/1803.09010>.
- AI Factsheets 360**
 IBM Research, <https://aif360.mybluemix.net/>.



개발

기능	카테고리	진단서	구현에 대한 의견
데이터 준비 및 모델 정의			
영향 평가	기능 선택 및 엔지니어링 프로세스 문서화	기능 선택 및 엔지니어링 프로세스 중에 행한 선택의 근거를 문서로 작성하고 모델 성능에 미치는 영향을 평가한다.	기능 선택 또는 엔지니어링 선택이 암묵적으로 편향성된 가정에 의존할 수 있는지 여부를 조사한다.
		선택한 기능과 민감한 인구 통계 속성 간의 잠재적인 상관 관계를 문서화한다.	민감한 클래스와 밀접한 관련이 있는 기능의 경우 대상 변수와의 관련성과 모델에 포함된 이유를 문서화한다.
	모델 선택 프로세스 문서화	선택한 모델링 접근법에 대한 근거를 문서화한다.	
		선택한 접근 방식과 잠재적 결과 제한 사항에 있어서의 가정을 식별, 문서화 및 정당화한다.	
위험 완화 모범 사례	기능 선택	편향성된 프록시 기능을 검사한다	<ul style="list-style-type: none"> • 단순히 민감한 속성을 시스템에 대한 입력으로 사용하지 않는 것("무인식을 통한 공정성"이라고 알려진 접근 방식)은 편향성 위험을 완화하는 효과적인 접근법이 아니다. 모델에서 민감한 특성이 명시적으로 제외된 경우에도 다른 변수가 해당 특성에 대한 프록시 역할을 할 수 있고 시스템에 편향성을 도입할 수 있다. 프록시 편향성의 위험을 피하기 위해 AI 개발자는 모델의 기능과 보호된 특성 사이의 잠재적인 상관 관계를 조사하고 이러한 프록시 변수가 모델의 출력에서 할 수 있는 역할을 조사해야 한다. • AI 개발자가 민감한 속성 데이터에 접근할 수 없거나 그러한 데이터에 대한 추론이 금지된 상황에서는 기능과 민감한 속성 간의 통계적 상관 관계를 조사할 능력이 제한될 수 있다.²² 이러한 상황에서는 도메인 전문가가 제공하는 보다 전체적인 분석이 필요할 수 있다.



개발

기능	카테고리	진단서	구현에 대한 의견
데이터 준비 및 모델 정의			
위험 완화 모범 사례 (계속)	기능 선택	민감한 속성과 관계가 있는 기능을 면밀히 조사한다.	<ul style="list-style-type: none"> 민감한 속성과 관련이 있는 것으로 알려진 기능은 시스템의 대상 변수와 강력한 논리적 관계가 있는 경우에만 사용해야 한다. 예를 들어, 소득은 성별과 상관 관계가 있지만 대출금을 상환할 수 있는 개인의 능력과 합리적으로 관련되어 있다. 따라서 신용도를 평가하기 위해 설계된 AI 시스템에서 소득을 사용하는 것은 정당하다. 반대로 신용도를 예측하는 모델에서 "신발 크기"(성별과도 상관 관계가 있음)를 사용하는 것은 민감한 특성과 밀접하게 관련되어 있는 변수를 부적절하게 사용하는 것이다.
	독립성 및 다양성	도메인별 전문 지식을 갖춘 다양한 이해 관계자들로부터 피드백을 구한다.	기능 엔지니어링 프로세스의 경우 시스템 학습에 사용되는 데이터의 역사적, 법적, 사회적 차원에 대해 다양한 실제 경험과 전문 지식을 가진 직원이 정보를 제공해야 한다.
	모델 선택	편향성의 위험과 잠재적 영향이 모두 높은 상황에서는 이해할 수 없는 모델을 피한다.	보다 해석 가능한 모델을 사용하면 문제를 쉽게 식별하고 완화할 수 있으므로 의도하지 않은 편향성의 위험을 줄일 수 있다.
모델 검증, 테스트 및 수정			
영향 평가	문서 검증 프로세스	시스템(및 개별 구성 요소)이 설계 목표 및 의도된 배치 시나리오와 일치하는지 여부를 평가하기 위해 검증하는 방법을 문서화한다.	
		재검증 프로세스를 문서화한다.	<ul style="list-style-type: none"> 모델을 정기적으로 재검증할 기간을 설정한다. 주기 외에 재검증을 하여야 하는 성능 기준점을 설정한다.
	문서 테스트 프로세스	모델 성능을 평가하고 문서화하여 시스템의 편향성을 테스트한다.	테스트에는 설계 단계에서 식별된 공정성 측정 기준을 통합하여야 하며, 인구 통계 집단 전체에서 모델의 정확성과 오류율을 조사해야 한다.
		테스트 수행 방법, 평가한 공정성 측정 기준, 그리고 측정값을 선택한 이유를 문서로 기록한다.	
모델 개입을 문서로 기록한다.	테스트에서 용인할 수 없는 수준의 편향성이 나타날 경우 모델을 개선하기 위한 노력을 문서로 기록한다.		



개발

기능	카테고리	진단서	구현에 대한 의견
모델 검증, 테스트 및 수정			
위험 완화 모범 사례	모델 개입	테스트 중에 나타나는 편향성을 해결하기 위한 잠재적인 모델 개선을 평가한다.	<p>테스트 결과 시스템이 선택한 공정성 측정 기준을 기반으로 허용할 수 없는 수준의 편향성을 나타내는 것으로 드러나면 모델을 개선해야 한다. 잠재적 모델 개선에는 다음과 같은 것이 포함된다.</p> <ul style="list-style-type: none"> • 프로세스 전 개입 이러한 개선에는 설계 및 개발 라이프 사이클의 초기 단계의 재검토(예: 추가 학습 데이터 검색)가 포함될 수 있다. • 프로세스 중 개입 모델에 직접 추가 공정성 제약 조건을 적용하여 편향성을 완화할 수도 있다. 기존의 기계 학습 모델은 예측 정확도를 극대화하도록 설계되었다. 최신 기술을 통해 개발자는 모델에 제약 조건을 구축하여 집단 전체에 걸쳐 편향성 가능성을 줄일 수 있다. 공정성 제약 조건의 추가는 사실상 모델에게 정확도와 특정 공정성 측정 기준을 모두 최적화하라고 지시하는 것이다. • 프로세스 후 개입 어떤 경우에는 모델의 출력 예측을 조작하여 원하는 분포를 따르게 하는 사후 처리 알고리즘을 사용하여 편향성을 해결할 수 있다.
	독립성 및 다양성	검증 및 테스트 문서는 시스템 개발에 관여하지 않은 직원이 검토해야 한다.	독립적인 팀이 검증 및 테스트 결과를 설계 및 개발 프로세스의 초기 단계에서 개발된 시스템 규격과 비교해야 한다.

개발: 위험 완화 도구 및 리소스

- *Model Cards for Model Reporting*
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, (2019. 1): 220–229, <https://arxiv.org/abs/1810.03993>.
- *AI Factsheets 360*
Aleksandra Mojsilovic, IBM Research(2018. 8. 22), <https://www.ibm.com/blogs/research/2018/08/factsheets-ai/>.
- *AI Explainability 360*
IBM Research, <https://aix360.mybluemix.net/>.
- *AI Fairness 360*
IBM Research, <https://aif360.mybluemix.net/>.
- *Responsible Machine Learning with Error Analysis*
Besmira Nushi, Microsoft Research 2021. 2. 18), <https://techcommunity.microsoft.com/t5/azure-ai/responsible-machine-learning-with-error-analysis/ba-p/2141774>.
- *Aequitas Open Source Bias Audit Toolkit*
Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani, University of Chicago Center for Data Science and Public Policy, <http://www.datasciencepublicpolicy.org/projects/aequitas/>.
- *FairTest: Discovering Unwarranted Associations in Data-Driven Applications*
Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels 및 Huang Lin, ArXiv,(2015), <https://github.com/columbia/fairtest>.
- *Bayesian Improved Surname Geocoding*
Consumer Finance Protection Bureau (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.



배치 및 사용

기능	카테고리	진단서	구현에 대한 의견
배치 및 사용 준비			
영향 평가	책임 라인을 문서화한다.	필요한 경우 시스템의 결정을 검토할 수 있는 방법에 대한 세부 정보를 포함하여 시스템의 출력과 그 결과물을 담당하는 사람을 정의하고 문서화한다.	
		잠재적 사고 또는 시스템 오류 보고에 대응하기 위한 관리 계획을 수립한다.	<ul style="list-style-type: none"> • 시스템이 고장 난다는 것은 무엇을 의미하며 고장으로 인해 피해를 입을 수 있는 사람은 누구인가? • 고장은 어떻게 감지하나? • 고장이 감지되면 누가 대응하나? • 시스템을 안전하게 비활성화할 수 있는가? • 중요한 기능의 연속성을 위한 적절한 계획이 있는가?
	데이터 모니터링 프로세스를 문서화한다.	생산 데이터(즉, 배치 중에 시스템이 접하는 입력 데이터)가 학습 데이터와 실질적으로 다른 지 여부를 평가하는 데 사용할 프로세스 및 측정 기준을 문서화한다.	
	모델 성능 모니터링 프로세스를 문서화한다.	정적 모델의 경우 성능 수준과 오류 등급을 시간에 따라 모니터링하는 방법과 검토에 착수할 수 있는 기준을 문서화한다.	
		시간이 지남에 따라 진화할 모델의 경우 변경 사항의 목록을 작성하는 방법, 버전 수집 및 관리 여부와 그 시기 및 방법, 그리고 성과 수준을 모니터링하는 방법(예: 예정된 검토주기, 주기 외 검토에 착수할 수 있는 성과 지표)을 문서화한다.	
	감사 및 수명 만료 프로세스를 문서화한다.	위험 완화 통제가 목적에 적합한지 여부를 평가하기 위해 영향 평가 검토 결과를 감사할 주기를 문서화한다.	
시스템이 지원될 것으로 예상되는 일정과 시스템이 합리적인 성능 임계값 아래로 떨어질 경우 시스템을 폐기하는 프로세스를 문서화한다.			
위험 완화 모범 사례	드리프트 및 모델 성능 저하 모니터링	배치 중에 발생하는 입력 데이터를 시스템 학습 데이터의 통계적 표현과 비교 평가하여 데이터 드리프트(즉, 모델 성능을 떨어뜨릴 수 있는 학습 데이터와 배치 데이터 사이의 실질적인 차이)의 가능성을 평가할 수 있다.	



배치 및 사용

기능	카테고리	진단서	구현에 대한 의견
배치 및 사용 준비			
위험 완화 모범 사례 (계속)	제품 기능 및 사용자 인터페이스	예측 가능한 의도하지 않은 사용의 위험을 완화하는 사용자 인터페이스 기능(예: 인간 참여(Human-in-the-Loop) 요구 사항을 적용한 인터페이스, 시스템의 오남용 사실을 알리는 경보)을 제품과 통합한다.	
	시스템 문서 관리	AI 개발자는 AI 배치 책임자가 배치 위험에 관한 영향 평가를 독자적으로 수행할 수 있도록 시스템 기능, 사양, 제한 사항 및 의도된 용도에 관한 충분한 문서를 제공하여야 한다.	필요한 경우 AI 개발자는 AI 배치 책임자에게 독립적으로 영향 평가를 수행할 수 있는 기술 환경도 제공할 수 있다.
		예측 가능한 오남용을 방지하기 위해 고안된 제한 사항(예: 최종 사용자가 허용되는 사용 정책을 준수하도록 보장하기 위한 계약상의 의무)을 명시한 조건을 최종 사용자 실시권 계약에 통합하는 것을 고려한다.	
		판매 및 마케팅 자료를 면밀히 검토하여 이 자료가 시스템의 실제 기능과 일치하는지 확인해야 한다.	
	AI 사용자 교육	AI 배치 책임자는 AI 사용자에게 시스템의 기능 및 제한 사항, 그리고 출력을 평가하고 작업흐름도(workflow)에 통합하는 방법에 관한 교육을 제공해야 한다.	인간이 참여하여 AI 시스템을 감독하는 것이 효과적인 위험 완화 조치가 되려면 시스템 작동 방식을 이해하고 모델의 출력을 이해할 수 있도록 AI 사용자에게 적절한 정보와 교육을 제공하여야 한다.
	사고 대응 및 피드백 메커니즘	AI 배치 책임자는 AI 사용자 및 영향을 받는 개인(즉, 시스템과 상호 작용할 수 있는 일반 구성원)이 시스템 운영에 대하여 우려하는 사항을 보고할 수 있도록 피드백 메커니즘을 유지해야 한다.	결과적 결정의 경우, 영향을 받는 개인에게 이의 제기 메커니즘이 제공되어야 한다.

배치 및 사용: 위험 완화 도구 및 리소스

- *AI Incident Response Checklist*
BNH.AI, <https://www.bnh.ai/public-resources>.
- *Watson OpenScale*
IBM, <https://www.ibm.com/cloud/watson-openscale>.
- *Detect Data Drift on Datasets*
Microsoft Azure Machine Learning 2020. 06. 25), <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python#create-dataset-monitors>.

참고 문헌

A Framework for Understanding Unintended Consequences of Machine Learning

Harini Suresh and John V. Guttag, arXiv (February 2020), <https://arxiv.org/abs/1901.10002>.

AI Fairness

Trisha Mahoney, Kush R. Varshney, and Michael Hind, O'Reilly (April 2020), <https://www.oreilly.com/library/view/ai-fairness/9781492077664/>.

Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models Andrew Burt, Brenda Leong, Stuart Shirrell, and Xiangnong (George) Wang,

Future of Privacy Forum (June 2018), <https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf>.

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach, CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (April 2020): 1–14, <https://doi.org/10.1145/3313831.3376445>.

Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Geburu, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P., FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, (January 2020): 33–44, <https://doi.org/10.1145/3351095.3372873>.

Supervisory Guidance on Model Risk Management

US Federal Reserve Board (April 2011), <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf>.

Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector

David Leslie, The Alan Turing Institute (2019), <https://doi.org/10.5281/zenodo.3240529>.


미주(尾註)

- ¹ Gina Kolata, "Alzheimer's Prediction May Be Found in Writing Tests," *New York Times* (February 1, 2021), <https://www.nytimes.com/2021/02/01/health/alzheimers-prediction-speech.html>.
- ² Dina Temple-Raston, *Elephants under Attack Have an Unlikely Ally: Artificial Intelligence*, NPR (October 25, 2019), <https://www.npr.org/2019/10/25/760487476/elephants-under-attack-have-an-unlikely-ally-artificial-intelligence>.
- ³ *Seeing AI: An App for Visually Impaired People That Narrates the World Around You*, Microsoft, <https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/>.
- ⁴ See e.g., Jennifer Sukis, *The Origins of Bias and How AI May Be the Answer to Ending Its Reign*, Medium (January 13, 2019), <https://medium.com/design-ibm/the-origins-of-bias-and-how-ai-might-be-our-answer-to-ending-it-acc3610d6354>.
- ⁵ See e.g., Nicol Turner Lee, Paul Resnick, and Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, Brookings (May 22, 2019), <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>.
- ⁶ Harini Suresh and John V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning* (February 17, 2020), <https://arxiv.org/pdf/1901.10002.pdf>.
- ⁷ See Xiaolin Wu and Xi Zhang, *Automated Inference on Criminality Using Face Images*, Shanghai Jiao Tong University (November 13, 2016), <https://arxiv.org/pdf/1611.04135v1.pdf>.
- ⁸ Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov, *Physiognomy's New Clothes*, Medium (May 6, 2017), <https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a>.
- ⁹ Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* (October 25, 2019), <https://science.sciencemag.org/content/366/6464/447>.
- ¹⁰ Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California University Law Review* 104, no. 3 (September 30, 2016): 671, <http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf>.
- ¹¹ Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77–91, <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.
- ¹² Kate Crawford, *The Hidden Biases in Big Data*, Harvard Business Review (April 1, 2013), <https://hbr.org/2013/04/the-hidden-biases-in-big-data>.
- ¹³ Kate Crawford and Trevor Paglen, *Excavating AI: The Politics of Images in Machine Learning Training Sets* (September 19, 2019), <https://excavating.ai/>.
- ¹⁴ Cade Metz, "'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You?" *New York Times* (September 20, 2019), <https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html>.
- ¹⁵ Jessica Zosa Forde, A. Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael Littman, *Model Selection's Disparate Impact in Real-World Deep Learning Applications*, arXiv:2104.00606 (April 1, 2021), <https://arxiv.org/abs/2104.00606>.
- ¹⁶ Aaron Klein, *Credit Denial in the Age of AI*, Brookings Institution (April 11, 2019), <https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/>.
- ¹⁷ J. Vaughn, A. Baral, M. Vadari "Analyzing the Dangers of Dataset Bias in Diagnostic AI systems: Setting Guidelines for Dataset Collection and Usage," ACM Conference on Health, Inference and Learning, 2020 Workshop, http://juliev42.github.io/files/CHIL_paper_bias.pdf.
- ¹⁸ Arvind Narayanan, *21 Fairness Definitions and Their Politics*, ACM Conference on Fairness, Accountability and Transparency (March 1, 2018), <https://www.youtube.com/watch?v=jiXluYdnyyk>.
- ¹⁹ Reuben Binns and Valeria Gallo, *AI Blog: Trade-Offs*, UK Information Commission's Office (July 25, 2019), <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/>.
- ²⁰ Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (January 2020): 33–44, <https://doi.org/10.1145/3351095.3372873>.
- ²¹ Sara Hooker, Moving Beyond "Algorithmic Bias Is a Data Problem," *Patterns* (April 9, 2021), <https://www.sciencedirect.com/science/article/pii/S2666389921000611>.
- ²² McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang, "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness, arXiv:2011.02282 (January 23, 2021), <https://arxiv.org/abs/2011.02282>.



www.bsa.org

BSA Worldwide
20F Street, NW
Suite 800
Washington, DC 20001

 +1 .202 .872 .5500

 @BSAnews

 @BSATheSoftwareAlliance


BSA Asia-Pacific

300 Beach Road
#30-06 The Concourse
Singapore 199555

 +65 .6292 .2072

BSA Europe, Middle East & Africa

44 Avenue des Arts
Brussels 1040
Belgium

 +32 .2 .274 .13 .10