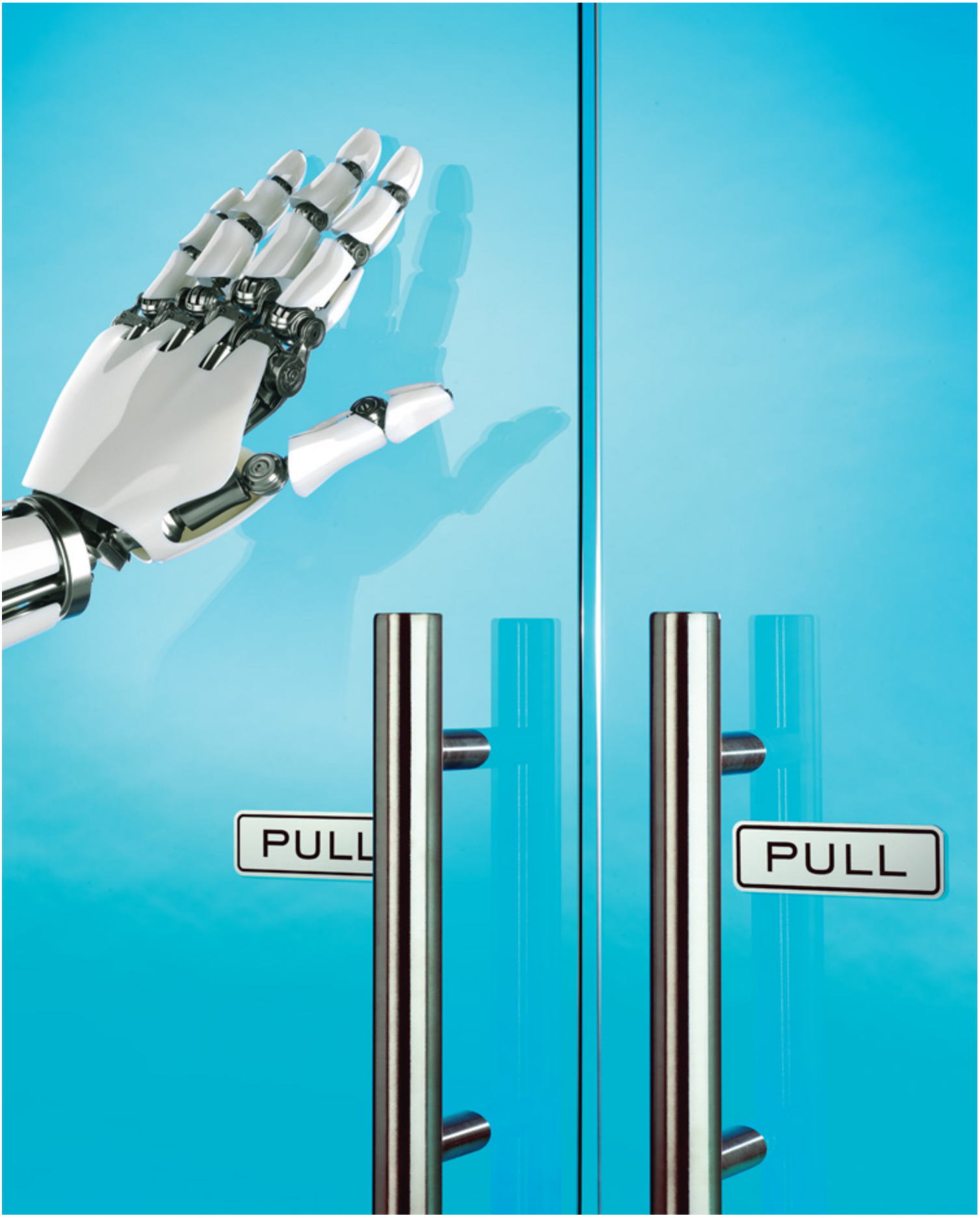
위기관리 & 운영관리

머신러닝이 선로를 벗어날 때

보리스 바빅(Boris Babic)),테오도로스 에브게뉴(Theodoros Evgeniou),I. 글렌 코언(I. Glenn Cohen),새라 거크(Sara Gerke)

매거진 | 1-2월호



머신러닝이 선로를 벗어날 때
리스크 관리 가이드

문제점
머신러닝 기술에 기반한 제품과 서비스가 급증하면서, 이를 개발하는 기업, 사용하는 기업, 관련 데이터를 공급하는 기업 등에 여러 가지 새로운 리스크가 대두되고 있다. 머신러닝 기반 시스템이 내리는 판단이 항상 윤리적이고 정확한 것은 아니기 때문이다.

원인
첫째, 머신러닝 기반 시스템이 판단을 내릴 때, 확률에 기반하는 경우가 많기 때문이다. 둘째, 시스템 작동 환경이 예상치 못한 방향으로 바뀔 수 있기 때문이다. 셋째, 복잡한 시스템 안에서 발생한 오류를 식별하고 원인을 분석하기가 어렵기 때문이다.

해결책
머신러닝 시스템이 지속적으로 진화할 수 있게 할 것인지, 아니면 고정된 버전을 주기적으로 출시할 것인지 경영진 차원의 결정이 필요하다. 또, 서비스 출시 전후로 적합한 검증 과정과 지속적인 모니터링 절차를 마련하고 운영해야 한다.

머신러닝이란 새로운 정보를 흡수해 이를 바탕으로 판단방식을 변화시켜 나가는 컴퓨터 프로그램을 의미한다. 그런데 만약 머신러닝 프로그램의 판단이 투자 손실이나 교통사고를 야기하거나 인재 채용, 대출 심사 등의 절차에서 불공정한 결정을 내릴 경우 어떻게 될까? 또, 머신러닝 기반 제품 및 서비스를 제공하는 기업은 자사의 프로그램이 스스로 진화하도록 내버려두는 것이 좋을까, 아니면 알고리즘을 '고정(lock)'한 이후 주기적으로 직접 업데이트하는 것이 좋을까? 알고리즘을 고정하기로 결정했다면, 업데이트 시점과 주기는 어떻게 정해야 할까? 여기서 내리는 결정에 따라 발생하는 리스크를 어떻게 분석하고 관리할 수 있을까?

어느 산업 분야에서든 머신러닝에 기반한 인공지능 기술이 점점 더 많은 서비스와 프로세스에 적용되고 있기 때문에, 기업 경영진과 이사회는 위에서 열거한 질문에 대답할 준비가 돼 있어야 한다. 본 아티클에서는 의료법, 윤리, 규제 및 머신러닝 분야에서 필자들이 진행한 연구를 바탕으로 머신러닝이 초래할 수 있는 리스크를 이해하고, 관리할 수 있는 주요 개념을 소개하고자 한다.

머신러닝에서 리스크가 발생하는 이유

머신러닝이 기존 디지털 기술과 다른 점은 바로 새로 공급되는 데이터에 적응하며, 독립적으로 점점 더 복잡해지는 문제에 대한 결정을 내릴 수 있다는 점이다. 이런 능력 덕분에, 머신러닝 기술은 어떤 금융상품을 사고팔지, 자율주행 자동차가 장애물을 어떻게 피해야 할지, 어떤 사람이 특정 질병에 걸렸는지 등 단순하지 않은 문제를 해결하는 데 사용되고 있다. 하지만 머신러닝의 알고리즘이 항상 아무 문제없이 작동하는 것은 아니다. 때때로 부정확한 판단이나 비윤리적인 결정을 내릴 수도 있다. 여기에는 다음과 같은 세 가지 근본적인 이유가 있다.

첫 번째는 바로 대부분의 알고리즘이 '확률'에 기반해 결정을 내린다는 점이다. 확률적 판단을 여러 번 중복해

서 내리기 때문에, '일부' 판단은 언제든지 틀릴 위험이 있다. 오류가 발생할 확률은 알고리즘 학습에 이용된 데이터의 질과 양, 사용된 머신러닝 기법 등 여러 요소에 따라 달라질 수 있다. 예를 들어, 복잡한 수리 모델을 활용하는 딥러닝deep learning 기법을 쓰느냐, 아니면 판단 규칙을 세워 결정을 내리는 분류 나무classification tree 기법을 사용하느냐에 따라 달라질 수 있는 것이다. 또, 사용된 알고리즘이 '설명 가능'한지도 중요하다. 설명 가능한 알고리즘이란 알고리즘이 판단을 내리는 과정을 기계가 아닌 사람이 이해할 수 있는지를 의미한다. 설명 가능한 알고리즘만을 사용해 머신러닝 시스템을 설계하면 정확도가 떨어질 수도 있다.

둘째로, 머신러닝이 작동하는 환경이 시간이 흐르며 달라지거나, 개발 당시 의도와는 다른 조건에서 적용될 수 있기 때문이다. 여러 이유가 있을 수 있지만, 가장 흔한 경우는 개념 변동concept drift과 공변량 변화covariate shift다.

개념 변동이란 투입된 변수와 결과값 사이의 관계가 시간이 흐르며 달라지거나 애초에 잘못 정의된 경우를 의미한다. 주식 트레이딩 알고리즘의 경우를 예로 들어보자. 만일 이 알고리즘을 훈련하는 데 경제성장률이 높고 시장변동성이 낮은 시기의 데이터만 사용했다면, 최근 코로나19 사태와 같이 경기가 침체되거나 시장에 혼란이 발생했을 때 모델의 성능이 저하될 수 있다. 시장 환경의 변화에 따라 기업의 레버리지 규모와 주식 수익률 등 변수 사이의 관계가 달라질 수 있기 때문이다. 비슷한 이유로, 경기 변동에 따라 신용평가 모델의 성능이 달라지기도 한다.

의료 분야에서도 개념 변동이 발생하는 경우를 찾아볼 수 있다. 피부 영상을 바탕으로 피부암을 진단하는 머신러닝 프로그램을 개발할 때, 피부색과 진단 결과 간 연관성을 제대로 파악하지 못한다면 알고리즘이 오진을 내리게 될 수도 있다. 개인의 피부색은 인종이나 직사광선 노출 정도와 같은 다양한 원인으로 달라질 수 있는데, 모델 개발에 사용되는 의료기록에는 이런 정보가 반영되지 않는 경우가 자주 있기 때문이다.

공변량 변화는 알고리즘이 학습하는 데 쓰인 데이터와 실제 사용 시점에 공급된 데이터가 다른 경우 발생한다. 알고리즘이 학습한 패턴이 안정적이고 개념 변동이 일어나지 않았더라도, 공변량 변화는 발생할 수 있다. 예를 들어, 의료용품 제조사가 도시지역 대형 병원의 데이터를 가지고 의료용 머신러닝 알고리즘을 개발했다고 해보자. 알고리즘을 시장에 출시한 후, 이를 실제로 사용하는 지방 의료시설의 데이터는 학습에 사용된 도시 병원 데이터와 상이할 수 있다. 도시지역 병원에는 지방에서는 보기 힘든 기저 질환을 앓고 있는 특정 사회계층의 환자 비율이 더 높을 수 있기 때문이다. 이런 데이터 괴리는 제품이 시장에 출시된 후에, 사전검증 단계에서보다 더 높은 오류율을 보이고 나서야 발견될 수도 있다. 빠른 속도로 다각화되는 시장에서 시스템의 작동 환경이 어떻게 바뀔지 예측하는 것은 점점 더 어려워지고 있다. 또 아무리 많은 데이터를 활용하더라도 실제 세상에서 발생하는 미묘한 차이를 모두 반영하는 것은 현실적으로 불가능하다.

머신러닝의 판단이 틀릴 수 있는 세 번째 이유는 머신러닝을 포함하는 전체 시스템이 복잡하게 구성돼 있다는 것이다. 의료진이 찍은 검사 영상을 분석해 당뇨망막병증과 황반부종을 진단하는 장비인 IDx-DR의 경우를 생각해보자. 미국의 FDA가 승인한 최초의 머신러닝 기반 자율 의료장비이기도 한데, 시스템에 입력된 영상의 선명도, 사용된 알고리즘의 종류, 알고리즘 학습에 사용된 데이터, 영상을 입력하는 의료진이 적절한 교육을 받았는지 등 다양한 요소에 따라 정확도가 달라질 수 있다. 이토록 많은 변수가 복잡하게 얽혀 있기 때문에, 아

무런 오류도 발생하지 않을 거라 기대하는 것은 고사하고, 오류가 발생했는지, 발생했다면 왜 발생했는지를 파악하는 것도 어려운 것이다.

하지만, 꼭 알고리즘이 잘못된 판단을 내리지 않더라도 머신러닝으로 인해 발생하는 리스크가 더 있는데, 바로 제3자 리스크와 도덕적 리스크다.

제3자 리스크

이것은 머신러닝 기술이 완전하지 않기 때문에 개별 기업이나 사용자가 통제할 수 없는 영역에서 발생하는 리스크를 의미한다.

머신러닝을 사용하지 않는 일반적인 경우에는 어떤 문제가 발생했을 때 믿을 만한 주변 정황에 근거해 문제의 원인을 재구성할 수 있다. 따라서, 최소한 기업 경영진은 자사가 얼마나 큰 책임을 지게 될지라도 가늠해볼 수 있다. 하지만 머신러닝 프로그램은 보통 더 복잡한 시스템의 일부로 얽혀있기 때문에, 어디서 문제가 발생했는지 파악하기 어려울 수 있다. 알고리즘 개발자, 시스템 구현 엔지니어, 기타 관련 업체 등 어떤 '제3자'의 책임인지 파악하거나, 알고리즘의 문제인지 데이터 문제인지, 데이터 때문이라면 사용자가 투입한 인풋의 문제인지 모델 개발에 사용된 훈련 데이터의 문제인지 알기 힘들다는 것이다. 설사 훈련 데이터에 문제가 있었다는 점을 파악했다 하더라도, 데이터의 출처가 한 곳이 아니라 다수의 업체이기 때문에 어떤 업체의 데이터 때문이었는지 찾기 힘들다. 또, 머신러닝 사용 환경의 변화와 머신러닝 기술에 내재된 확률적 특성으로 인해 책임 소재를 파악하는 것이 더욱 어려워진다. 꼭 누군가 관계자의 잘못이 아니더라도, 알고리즘 자체에 부정확한 판단을 내릴 가능성이 항상 내재하기에 의도치 않은 문제가 발생할 수 있다는 것이다.

경영진은 기업이 언제 법적 책임을 지게 될 수 있는지도 파악할 필요가 있는데, 이를 결정하는 법과 규정 또한 지속적으로 변화한다. 의료 분야를 예로 들자면, 현행 의료법 체계에서는 의료 행위에 대한 최종 결정 책임이 의사에게 있다고 보기 때문에 관련 소프트웨어 제조사에는 제조물 책임product liability을 적용하지 않는 경향이 있다. 하지만, 알고리즘 작동방식을 알 수 없는 블랙박스 프로그램이나 자율 시스템이 의료 분야에 적용돼 의료진의 개입이 최소화되거나 전혀 없이 진단 및 처방을 내리게 된다면 이야기가 달라질 수 있다. 만약, 머신러닝 알고리즘이 보통 경우보다 훨씬 높은 투약량을 처방하는 등 일반적이지 않은 결정을 내렸을 때, 이를 따르지 않아 발생한 사고에 대해서만 의료진이 책임을 지는 방향으로 규제가 바뀐다면 어떻게 되겠는가? 이런 식으로 규제가 바뀌게 되면 지금은 의료진이 지고 있는 법적 책임 리스크를 머신러닝 기반 의료장비의 제조사나 알고리즘 학습에 사용된 데이터 공급사, 알고리즘을 구현하고 공급한 회사 등의 기업이 대신 부담하게 될 수도 있다.

도덕적 리스크

머신러닝 기반 제품 및 서비스가 자율적으로 윤리적 딜레마를 해결해야 하는 상황이 생길 수도 있다. 이런 경우 규제나 상품 개발 측면에서 기업의 어려움과 리스크가 가중된다. 학계에서도 이에 대한 논의가 시작되고 있는데, 이런 문제를 '책임 있는 알고리즘 디자인responsible algorithm design'이라고 부른다. 여기서 발생하

는 난제 중 하나는 어떤 방식으로 프로그램이 도덕적 판단을 내리게 할지 결정하는 것이다. 예를 들어, 테슬라가 자율주행자동차를 프로그래밍할 때 비용편익에 기반한 공리주의적 관점을 택해야 할까 아니면 어떤 가치는 그 어떤 이익과도 바꿀 수 없다는 칸트적 관점을 택해야 할까? 공리주의적 관점을 택하기로 결정했더라도, 추상적 가치의 비용과 편익을 측정하는 것은 아주 어려운 일이다. 예를 들어, 노인 세 명과 중년 한 명 중 어느 쪽의 생명이 더 귀중한지 판단해야 하는 상황이라면, 이를 어떻게 프로그래밍할 수 있겠는가? 또, 프라이버시, 공정성, 정확성, 보안성 등의 가치 사이에서 기업이 균형을 맞춰야 할 때, 어디에 얼마만큼 비중을 둬야 하는 것일까? 이런 판단으로 인해 발생하는 리스크를 피할 방법은 있는가?

도덕적 리스크는 특정 집단에 대한 편향성의 형태로도 나타날 수 있다. 일례로, 안면인식 알고리즘은 유색 인종의 얼굴을 잘 인식하지 못하고, 피부 병변 진단 알고리즘의 정확도는 인종에 따라 달라질 수 있다. 신용평가 프로그램이나 범죄자의 재범률을 예측하는 알고리즘이 히스패닉이나 흑인에게 부당하게 불이익을 주는 경우도 종종 있다. 다양한 산업에서 사용되는 머신러닝 시스템이, 일부 영역에서는 특정 집단에 불공정하게 보일 수도 있다는 것이다.

이런 편향성 문제는 무엇이 공정한 것인지를 정의하고 알고리즘에 구현하는 방식이 한둘이 아니고 때로는 상충되기도 하기 때문에 더욱 심화된다. 예를 들어, 대출심사 알고리즘을 설계할 때 두 사람의 신용등급이 같다면 인종이나 성별 등 집단 정체성group identity을 이유로 차별하지 못 하도록 하더라도, 결과적으로는 신용도가 높은데도 소수집단에 속한 사람들의 대출 신청을 거절하게 되기도 한다. 따라서, 기업은 이러지도 저러지도 못하는 딜레마에 빠지게 된다. 누구에게 대출을 해줄지 결정하는 알고리즘을 사용할 때, 공정성을 정의하는 여러 방식 중 하나에 부합하지 않아 특정 집단을 차별한다는 비판을 피하기 어렵기 때문이다. 게다가 문화에 따라 윤리적 가치관이 달라지기도 한다. 글로벌 시장에서 활동하는 기업에는 또 다른 딜레마가 발생하는 것이다. 일례로, 유럽연합 집행위원회는 2020년 2월 발행한 인공지능백서에서 '유럽의 가치체계'에 맞춰 AI를 개발해야 한다고 요구한 바 있다. 하지만, 이런 유럽식 AI를 다른 가치 체계를 가진 타 지역에 아무런 잡음 없이 수출할 수 있겠는가?

마지막으로, 통계 모델의 불안정성 때문에 도덕적 리스크가 발생할 수도 있다. 모델의 불안정성이란 투입된 변수가 비슷함에도 전혀 다른 결과가 나오는 것을 의미한다. 알고리즘이 안정적이지 않으면 여러 조건이 유사한 두 사람을 불공평하게 차별하게 될 수도 있다는 것이다.

물론, 앞에 말한 여러 문제점 때문에 머신러닝 알고리즘을 아예 사용하지 말아야 한다는 이야기는 아니다. 머신러닝 기술을 통해 만들어낼 수 있는 다양한 기회를 십분 활용함과 동시에 관련 리스크를 적절히 관리하는 데에 경영진의 각별한 관심이 필요하다는 것이다.

고정하느냐 풀어주느냐

머신러닝을 사용하기로 했다면, 그 다음에는 알고리즘이 지속적이고 유동적으로 진화할 수 있도록 둬야 할지 아니면 검증을 거친 고정 버전만 단계적으로 출시할지 결정해야 한다. 하지만, 검증된 고정 버전만 출시한다고 해서 상기한 리스크가 해결되느냐는 또다른 문제다.

이에 대한 논의는 의료 분야에서 활발히 이뤄져왔다. 지금까지는 '의료기기로서의 소프트웨어Software as a Medical Device'(하드웨어 없이도 의료 목적의 기능을 수행하는 소프트웨어를 의미)가 미 FDA의 허가를 받기 위해서는 알고리즘이 고정돼 있어야 했다. 본인들이 파악할 수 없는 방식으로 진단 및 치료 절차가 계속 바뀌는 의료기기는 허가해주지 않겠다는 것이다. 하지만 요즘 들어서는 FDA를 비롯한 여러 규제기관 사이에도 알고리즘이 고정돼 있다고 해서 다음과 같은 위험요소가 사라지는 것은 아니기 때문에, 알고리즘 고정 여부가 리스크의 정도를 결정하지는 않는다는 인식이 자리잡고 있다.

판단의 부정확성. 머신러닝 알고리즘을 고정한다고 해도 알고리즘이 대부분 확률 추정치에 근거해 판단을 내린다는 사실이 바뀌지는 않는다. 또, 일반적으로 데이터투입량이 늘어나면 알고리즘 성능이 개선되기는 하지만, 항상 그런 것도 아닐 뿐더러 성능 개선폭도 일정하지 않다. 머신러닝 시스템의 종류나 데이터의 양에 따라 개선 폭이 달라질 수 있기 때문이다. 알고리즘이 고정되지 않았을 경우에도 판단의 정확도가 어떻게 달라질지 파악하는 일이 어렵기는 하지만, 이를 위해 노력하는 것은 중요하다.

환경 변화. 머신러닝 시스템이 판단을 내리는 환경이 지속적으로 변화하는지, 변화한다면 어떻게 변화하는지도 중요한 요소다. 일례로, 자동차의 자율주행 알고리즘은 주변 운전자의 행동에 따라 초단위로 바뀌는 환경 속에서 작동해야 한다. 자산 가격 책정, 신용평가, 트레이딩 등 금융 분야에서 사용되는 알고리즘은 경기 변동에 따라 달라지는 시장 환경에 적응해야 한다. 어떤 환경에서든 적절한 판단을 내릴 수 있도록 머신러닝 알고리즘이 환경에 발맞춰 진화할 수 있게 하는 것이 관건이다.

제3자 리스크. 알고리즘을 고정한다고 해도 그 알고리즘을 포함하는 전체 시스템까지 단순해지는 것은 아니다. 외부 업체에서 제공한 알고리즘 개발용 데이터에 하자가 있거나 사용자의 숙련도에 차이가 있어 발생하는 오류는 막을 수 없다는 것이다. 앞서 설명했듯이 다양한 데이터 공급업체, 알고리즘 개발자, 구현 엔지니어, 사용자 사이에서 책임 소재를 분명히 하기 어려울 수 있다.

도덕적 리스크. 머신러닝 시스템에 내재된 불완전성이나 편향성을 인지하지 못하고 알고리즘을 고정할 경우, 이런 결함까지 같이 고정돼 버릴 위험이 있다. 예를 들어, 유방 조영검사 영상을 분석해 유방암을 진단하는 알고리즘을 개발 단계에 고정해 놓을 경우, 개발 당시의 데이터에 포함되지 않은 집단에 대해서는 알고리즘이 학습할 수 없게 된다. 인종에 따라 평균 유방밀도가 다르기 때문에, 학습 데이터에 제대로 반영되지 않은 인종의 여성에 대해서는 진단 정확도가 떨어지게 되는 것이다. 마찬가지로, 특정 사회경제적 계층에 국한된 데이터만 가지고 개발한 신용평가 알고리즘을 고정해 놓으면 다른 계층을 차별하게 될 수도 있다. 즉, '레드라이닝redlining'(특정 지역 주민에게 대출과 같은 금융 서비스를 제공하는 것을 거부하는 행위)과 다를 바 없는 현상

이 발생할 수도 있는 것이다. 이런 문제를 해결하기 위해서는 기존에 학습하지 못한 집단의 데이터가 투입되었을 때 즉각적으로 알고리즘이 스스로 업데이트할 수 있어야 한다. 하지만 고정되지 않은 머신러닝 알고리즘을 사용하더라도 특정 집단의 데이터만 계속해서 학습해서 다른 집단을 차별할 위험이 있기도 하다. 또, 언제 이런 차별이 발생하는지를 식별하는 것도 쉬운 일이 아니다.

경영진이 할 수 있는 일

그렇다면, 머신러닝 기술을 사용할 때 발생하는 리스크를 해결하기 위해 경영진이 해야 할 일은 무엇일까? 적절한 관리 절차를 개발하고 경영진 및 이사회의 기술 이해도를 제고하는 것은 물론이고 머신러닝에 대해 합리적인 의문과 적절한 사고방식을 갖추는 것도 중요한 일이다.

1.머신러닝을 사람처럼 대하기. 머신러닝 프로그램을 죽어있는 기술이 아니라 살아있는 생명체처럼 여기는 사고방식이 필요하다. 채용과정에서 지능검사를 거쳤다고 해서 새로 뽑은 직원이 기존 직원들과 함께 얼마나 잘 일할 수 있을지 알 수 없는 것처럼, 알고리즘 개발자와 연구진의 테스트만으로 실제 상황에서 머신러닝이 얼마나 잘 작동할지를 예측할 수 없다. 직원, 소비자 등 다양한 이용자가 머신러닝을 어떻게 활용하고 머신러닝이 내린 결정에 어떻게 반응할지 면밀히 점검할 수 있도록 경영진 차원의 노력이 필요하다. 이를 위해서는 여러 방법을 생각해볼 수 있다. 꼭 규제상의 의무가 없더라도, 무작위 대조 시험randomized controlled trial 등의 기법을 활용해 제품 출시 전에 성능, 안정성, 공정성 등을 확인해 볼 수 있다. 출시 후에도 다양한 소비자의 반응을 분석해 머신러닝 프로그램의 판단이 소비자 유형에 따라 어떻게 달라지는지 점검해야 할 수도 있다. 또, 동일 상황에서 머신러닝이 내린 판단과 머신러닝 없이 내려진 판단을 비교할 필요도 있다. 무작위 대조 시험을 거치지 않은 경우, 대규모 출시 이전에 일부 시장에서만 일종의 베타 테스트를 거쳐 사용자 숙련도, 데이터 출처, 환경 변화 등 실제적인 변수에 프로그램이 어떻게 반응하는지를 파악하는 것도 한 가지 방법이다. 알고리즘이 실제 시장과 상황에서 제대로 작동하지 못한다면, 그것을 보완하거나 폐기할 필요가 있다는 뜻이다.

2. 규제 담당자처럼 생각하고 검증부터 하기. 기업은 상품을 시장에 출시하기 전에 적절한 인증 절차를 거칠 수 있도록 계획을 세워야 한다. 실제로 규제를 담당하는 공공 기관에서 어떤 요소를 눈여겨보는지 참고하면 좋은 로드맵을 세울 수 있다. 일례로, 2019년 FDA는 머신러닝 기반 소프트웨어를 의료기기로 간주해 의료 소프트웨어를 지속적으로 발전시키면서도 환자의 안전을 보장하기 위한 새로운 규제 프레임을 제안한 바 있다. 눈여겨볼 내용은 소프트웨어 개발 기업의 내부 점검 절차뿐만 아니라 조직문화 등까지 평가 요소에 포함했고, 인증 절차를 거치지 않은 기업의 경우, 문제 발생시 책임 소재를 물을 수도 있다는 것이다.

또 다양한 스타트업들이 머신러닝 제품이나 프로세스에 편향성, 불공정성 등 내재된 결점이 있는지 평가하고 인증하는 서비스를 제공하고 있다. 국제전기전자공학회, 국제표준화기구 같은 전문 조직도 관련 인증 절차 및 표준을 개발하고 있고, 구글 등의 기업도 AI 학습에 사용된 데이터, AI가 내리는 판단, 사용자에게 미치는 영향 등을 종합적으로 분석하는 AI 윤리 분야의 서비스를 제공한다. 마지막으로, 개별 기업이 자사의 제품과 서비스에 특화된 점검 및 인증 프레임을 개발해야 할 경우도 있다.

3. 끊임없이 점검하기. 머신러닝 기술 자체와 기술 사용 환경이 계속해서 변화하기 때문에 개발 의도와는 다른 결과가 발생할 위험이 있다. 그렇기 때문에 적절한 제한선을 설정해 놓고 그 범위 안에서만 프로그램이 작동하는지 확인해야 한다. 꼭 머신러닝 영역이 아니더라도 다른 분야에서도 좋은 예시를 찾아볼 수 있다. 일례로, FDA의 의약품 감시체계인 센티넬 이니셔티브Sentinel Initiative는 전자의료기록 등 다양한 데이터를 활용해 의약품의 안정성을 점검하고 기준을 충족하지 못할 경우 판매를 중단시킬 수 있다. 또, 제조업이나 에너지산업, 사이버보안 업체에서 현재 운용하는 예방 정비preventative maintenance 기법이나 절차도 여러모로 비슷한 부분이 있다. 예를 들어, IT 시스템의 보안성을 점검하기 위해 취약점을 공격하는 일명 '적대적 공격adversarial attack' 기법을 AI에 적용해 모델의 취약점을 점검해볼 수 있다.

적절한 질문 던지기. 머신러닝 기술을 사용하는 기업 경영진과 유관 기관의 담당자들은 다음 네 가지 요소를 면밀하게 점검하고 끊임없이 질문해야 한다.

→ 정확성 및 성능. 머신러닝 기반 시스템의 알고리즘을 고정시키지 않을 경우, 앞으로 추가적인 데이터를 통해 학습하면서 얼마나 성능이 개선될 것인가? 성능이 개선되면 기업의 비즈니스 모델에는 어떤 영향이 있는가? 알고리즘 고정 여부에 따라 어떤 장단점이 있는지 소비자가 얼마나 잘 이해할 수 있는가?

→ 편향성. 알고리즘 학습에 어떤 데이터가 사용됐는가? 실제 대상 집단과 비교해봤을 때 누락된 집단은 없는가? 알고리즘을 고정한다면 조금 더 공평한 판단을 내릴 것이라고 예상할 수 있는가? 알고리즘에서 오류가 발생할 경우 특정 집단이 특히 더 큰 피해를 입을 수 있는가? 알고리즘이 차별적인 판단을 하지 못하도록 하는 '가드레일'을 설치할 수 있는가?

→ 환경. 제품 및 서비스의 사용 환경은 시간이 지날수록 어떻게 변할 것인가? 알고리즘이 판단을 내려서는 안 되는 조건이 있는가? 있다면, 어떤 조건인가? 환경 변화에 따라 알고리즘이 적절한 방향으로 변하도록 할 수 있는가? 환경이 너무 많이 변해서 더 이상 서비스를 제공할 수 없어지는 시점은 언제인가? 제품과 서비스가 적응할 수 있는 환경의 범위는 어느 정도인가? 서비스 제공 기간에 걸쳐서 머신러닝 시스템의 안전성을 담보할 수 있는가?

→ 제3자 리스크. 알고리즘 성능에 영향을 끼칠 수 있는 데이터 공급업체 등의 제3자는 누구인가? 사용자에 따라서 알고리즘의 결과물이 달라지는가? 우리 회사의 데이터나 알고리즘을 사용하는 다른 업체는 어떤 곳이 있고, 이곳에서 발생한 문제 때문에 우리 회사가 법적인 책임을 져야 할 가능성이 있는가? 우리 회사가 개발한 알고리즘을 다른 업체가 사용할 수 있게 해야 하는가?

리스크 관련 원칙 세우기. 머신러닝과 관련해 발생하는 리스크를 관리할 수 있는 자체 가이드라인을 마련해야 한다. 구글이나 마이크로소프트가 좋은 예시다. 알고리즘의 공정성을 어떻게 정의할지와 같은 구체적인 내용을 결정할 수 있는, 관련 리스크에 특화된 가이드라인이 좋다. 예를 들어, 채용 과정에 사용하는 알고리즘의 경우 직관성, 공정성, 투명성이 중요한 기준이 되겠지만, 원자재 선물 가격을 예측하는 알고리즘이라면 이런 가치보다는 최대로 발생할 수 있는 손실 규모 같은 기준이 더 중요할 것이다.

다행스럽게도, 리스크 관리 가이드라인을 마련하고 적용하기 위해 무에서 유를 창조할 필요는 없다. 지난 몇 해 동안 여러 곳에서 만든 예시를 참고할 수 있기 때문이다. 예를 들어, 2019년 OECD는 국제적으로 통용되는 최초의 AI 관련 원칙을 담은 OECD 인공지능 권고안을 채택한 바 있다. 인권, 법치, 다양성, 민주주의, 포괄적 성장, 지속가능한 성장 등의 가치를 존중하고 혁신성, 신뢰성, 투명성을 갖춘 AI를 활성화하고, 안전성, 보안, 지속적인 리스크 관리에 집중하는 것을 골자로 한다. 또, 최근 출범한 OECD AI 정책 관측소AI Policy Observatory를 통해 각국 정부의 AI 관련 정책을 비롯한 다양한 자료를 참고할 수도 있다.

머신러닝에는 엄청난 잠재력이 있다. 하지만, 머신러닝을 비롯한 여러 AI 기술이 우리 사회와 경제 체계에 자

리잡게 되면서 관련 리스크도 같이 커질 것이다. 기업의 관점에서 머신러닝 기술로 인해 발생하는 리스크를 관리하는 것이 기술 자체를 도입하는 것만큼이나, 아니 어쩌면 그보다 더 중요할 수 있다. 새롭게 발생하는 리스크를 적절히 관리할 수 있는 방안을 마련하지 않는 기업은 시장의 호응을 얻기 어려울 것이다.

보리스 바빅(Boris Babic)은 인시아드 의사결정학(Decision Science) 조교수다.
I. 글렌 코언(I. Glenn Cohen)은 하버드법학대학원 부학장이자 교수이며, 보건법 정책, 바이오기술, 바이오윤리를 연구하는 페트리-프롬센터의 센터장이다.
테오도로스 에브게니우(Theodoros Evgeniou)는 인시아드 의사결정학 및 기술경영학 교수이다.
새라 거크(Sara Gerke)는 하버드법학대학원 페트리-프롬센터의 의학, AI, 법학 분야 연구원이다.

번역 윤성헌 **에디팅** 배미정