AI 거버넌스 계층화 모델

인용	가세르, 우르스, 버질리오 A.F. 알메이다. 2017년. “AI 거버넌스 계층화 모델”. IEEE 인터넷 컴퓨팅 21(6): 58-62. doi:10.1109/mic.2017.4180835
출판 버전	doi:10.1109/MIC.2017.4180835
인용 링크	http://nrs.harvard.edu/urn-3:HUL.InstRepos:34390353
이용 약관	이 문서는 하버드 대학교의 DASH 저장소에서 다운로드 되었으며, 이하 명시된 바와 같이 개방형 엑세스 정책 문서에 적용되는 약관에 따라 제공됩니다. http:// nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-ofuse# OAP

AI 거버넌스 계층화 모델

Urs Gasser, Virgilio A.F. Almeida - 하버드 대학교

AI 기반 시스템은 블랙박스(black box)와 같아서, 개발자와 소비자, 정책입안자 사이에 대규모 정보 비대칭을 발생시킨다. 이러한 정보 격차를 해소하기 위해 이 논문에서는 AI 거버넌스에 대해 생각하기 위한 개념적 프레임워크를 제안한다.

사회의 많은 분야에서 디지털 기술과 빅데이터를 빠르게 채택하고 있다. 그 결과 AI, 자율 시스템, 알고리즘 의사 결정이 여러 인간 생활 영역과 자연스럽게 통합되는 경우가 많다. AI와 알고리즘 시스템은 이미 민간 부문과 공공 부문 모두에서 광범위하게 의사 결정을 하고 있다. 예를 들어 구글, 페이스북과 같은 개인 글로벌 플랫폼은 AI 기반 필터링 알고리즘을 사용하여 정보에 대한 접근을 제어한다. 자율주행차를 제어하는 AI 알고리즘은 승객과 보행자의 안전을 어떻게 보장할지 결정해야 한다.

보안 및 안전 의사결정 시스템을 포함한 다양한 애플리케이션은 AI 기반 얼굴 인식 알고리즘에 크게 의존한다. 그리고 스탠포드 대학의 최근 연구는 데이트 관련 사이트에서 사람들의 성별을 91%의 정확도로 추론할 수 있는 AI 알고리즘을 설명한다.

본 연구에서 증명된 AI의 능력에 경종을 울리며, 그리고 AI 기술이 광범위한 사용을 향해 나아가면서, 사회 일각에서는 이러한 기술의 광범위한 사용으로 인한 의도하지 않은 결과와 잠재적인 단점에 관해 우려를 표명하고 있다.

AI 생태계에 대한 투명성과 책임성, 설명가능성을 보장하기 위해서는 우리 정부와 시민사회, 민간, 학계가 토론 자리에서 AI와 자율시스템의 위험과 가능한 단점을 최소화하면서 이 기술의 잠재력을 최대한 활용할 수 있는 거버넌스 메커니즘을 논의해야 한다. 그러나 AI, 자율 시스템, 알고리즘에 대한 거버넌스 생태계를 설계하는 과정은 여러 가지 이유로 복잡하다. 옥스퍼드대 연구진이 지적하듯 의사결정 알고리즘, AI, 로봇공학을 위한 3개의 별도 규제 솔루션은 법적, 윤리적 과제와 관련이 없는 것으로 잘못 해석할 수 있어 오늘날의 시스템에서는 더 이상 정확하지 않다.

알고리즘, 하드웨어, 소프트웨어 및 데이터는 항상 AI 및 자율 시스템에 포함되어 있다. 미리 규제하는 것은 어떤 종류의 산업에서도 쉬운 일이 아니다. AI 기술이 빠르게 진보하고 있지만 여전히 개발 단계에 있을 뿐이다. 글로벌 AI 거버넌스 시스템은 문화적 차이를 수용할 수 있을 만큼 유연해야 하며, 서로 다른 국가 법률 시스템 간의 격차를 해소해야 한다. AI에 대한 거버넌스 구조를 설계하기 위해 우리가 취할 수 있는 많은 접근법이 있지만, 한 가지 방법은 인터넷 환경에서 작용하는 거버넌스 구조의 발전과 진화에서 영감을 얻는 것이다. 따라서, AI 시스템의 거버넌스와 관련된 다양한 문제에 대해 논의하고, AI에 대한 거버넌스와 자율 시스템, 알고리즘 의사 결정 프로세스에 대해 고찰하기 위한 개념적 프레임워크를 소개한다.

AI의 본질

병원, 법원, 학교, 가정, 그리고 인간의 의사결정을 지원하는 방법으로 AI 기반 애플리케이션이 점점 더 많이 채택되고 있지만, 현재 미국 연구진이 1950년대 중반에 만든 용어인 'AI'의 정의는 보편적으로 받아들여지지 않고 있다. 그 정의가 받아들여지지 않는 이유 중 하나는 기술적 관점에서 AI는 단일 기술이 아니라 음성 인식과 컴퓨터 비전, 어텐션과 메모리에 이르기까지 다양한 기술들의 하위 학문이라는 점이다.

그러나 현상학적 관점에서 AI라는 용어는 첨단 건강 진단 시스템, 차세대 디지털 튜터, 자율주행 자동차 및 기타 AI 기반 애플리케이션 공유에 나타나는 일정 수준의 자율성을 가리키는 포괄적 용어로 자주 사용된다. 종종 이러한 애플리케이션은 차례로 인간의 행동에 영향을 미치며 시스템의 설계자가 예측할 수 없는 방식으로 동적으로 진화한다. 이러한 맥락에서, 약한(또는 좁은) AI와 강한(또는 일반적인) AI를 자주 구별하게 되고, 이것이 AI의 본질을 논의할 때 도움이 된다. 약한(weak) AI는 게임 실행, 음성 인식, CT 스캔에서 특정 패턴 감지 등 비교적 좁은 작업에 집중하는 현재의 애플리케이션 세대를 말한다. 이와는 대조적으로, 강한(strong) AI는 기계가 어떤 문제에든 지능을 적용할 수 있는 능력을 가지고 있다는 점에서 진정한 지능과 자각을 가진 기계를 말한다. 현재 강력한 AI의 기술적 가능성과 잠재적 사회적 영향이 논란이 되고 있는 반면, 현재 약한 AI를 적용하는 것은 이미 주목받고 있는 일련의 실제 거버넌스 문제로 이어지고 있다.

AI 거버넌스 과제

신기술이 널리 보급되는 전형적인 패턴에 따라, 정책입안자와 기타 이해관계자들은 AI 기반 기술의 위험과 해악에 주로 관심을 기울이고 있다. 다시 말하면, 디지털 기술이 사회에 미치는 영향에 대한 이전의 논의들과 유사하게, AI, 자율 시스템, 알고리즘과 관련된 과제는 반드시 해결해야 할 실질적인 문제(정책, 법률, 거버넌스, 윤리적 고려사항 포함)의 형태로 제시되고 논의되는 경우가 많다.

예를 들어, 한 주요 전문가가 최근 제시한 AI 정책에 대한 로드맵은 AI 애플리케이션이 새로운 문제를 야기하거나 기존의 정책 우려와 압박 지점을 극대화하는 다음과 같은 핵심 이슈와 질문들의 목록을 제시한다.

• *정의와 평등.* 공정성, 책임성, 투명성과 같은 인간의 가치를 반영하고 불평등과 편견을 없애기 위해 AI 시스템을 어느 정도까지 설계하고 운영할 수 있는가?

• *무력 사용.* AI 기반 시스템이 현재 무력 사용에 대한 결정을 내리는 데 관여하고 있기 때문에(자율 무기의 경우) 인간의 통제가 얼마나 필요한가? AI 기반의 결정에 누가 책임을 져야 하는가?

• *안전 및 인증.* 특히 AI 기반 시스템에 물리적 표식이 있는 경우, 표준 설정 및 인증을 통해 안전 임계값을 정의하고 검증하는 방법은 무엇인가?

• *개인 정보 보호.* AI 시스템이 데이터에 의해 활성화되고 작동될 때, 정부 보안 감시 또는 기업이 고객에게 미치는 영향 측면에서 차세대 기술이 개인 정보 보호에 미치는 영향과 새로운 위협은 무엇입니까?

• *노동력 및 과세 전환.* AI 기반 로봇은 인간이 이전에 수행했던 일을 어느 정도까지 대체하거나 노동의 의미를 변화시킬 수 있을까? 로봇이 세금을 내지 않는다면 AI가 공공재정에 미치는 영향은 무엇일까?

이러한 실질적인 문제 목록(지적 재산 또는 책임 등)은 투명성, 책임성, 설명 가능성, 포괄성 및 공정성, 글로벌 거버넌스 및 AI 기반 시스템의 다양한 적용 영역에 걸쳐 있는 교차 검증 주제로 보완될 수 있다(참조).

AI 거버넌스 모델

앞에서 언급한 문제를 다루는 AI에 대한 미래 거버넌스 모델을 고려할 때, 그러한 목록을 넘어 AI 기반 기술의 '규제'(광범위하게 정의된)와 관련된 더 큰 구조적 과제 중 일부를 고려하는 것이 도움이 되거나 필요할 수 있다. 아래에서 미래의 AI 거버넌스 모델에 대한 설계 요건으로 변환되는 세 가지 과제를 강조하고 있다.

정보의 비대칭성. AI는 수십억 명의 삶을 구성할 수 있는 잠재력을 가지고 있지만, 소수의 전문가들만이 그 기본 기술을 진정으로 이해하고 있다. AI 기반 시스템은 종종 이해하기가 어려우며, 때로는 그러한 시스템의 개발자와 소비자 및 정책 입안자를 포함한 다른 이해관계자 사이에 대규모 정보 비대칭이 발생한다. AI를 위한 효과적인 거버넌스 시스템은 적용의 영역에서 다른 표현과 맥락으로 AI 현상에 대한 우리의 집단적 이해를 향상시키기 위한 메커니즘을 통합할 필요가 있다.

규범적 합의 도출. 현재의 정책과 지배구조 논쟁은 주로 AI와 관련된 위험과 과제에 초점이 맞춰져 있다. 그러나 지속 가능한 개발 목표의 맥락에서 AI의 사용에 대한 논의가 보여주는 것처럼, AI는 또한 사회에 엄청난 잠재적 이익을 제공한다(참조). 거버넌스 모델은 특히 AI 시스템 설계에 트레이드오프가 관여하는 경우, 서로 다른 이해관계자 간의 비용 편익 분석과 규범적 합의 구축을 위한 공간을 개방해야 한다. 또한 미래 거버넌스 모델은 전후 사정과 지역학 간의 규범적 차이를 다루고 다른 프레임워크와 접근법 간의 상호운용성을 제공할 필요가 있다.

정부의 부조화. AI 기술, 기반 기술 및 바람직하지 않은 것에 대한 사회적 합의를 공유하는 경우에도 앞에서 언급한 실질적인 문제를 해결하기 위해 효과적이고 합법적인 수단(전략, 접근 방식, 도구 등)을 설계하는 것은 AI 생태계에서의 불확실성과 복잡성과 같은 조건을 고려할 때 어려운 과제이다. 그러나 더 큰 잠류는 디지털 시대의 법과 정책 결정에 대한 전통적인 접근법에도 제한을 두고 있다. 종합해보면, AI의 미래 거버넌스 모델에 대한 이러한 구조적 과제와 관련 설계 요구사항은 단순한 국가 중심, 명령 및 제어 규제 체계에서 벗어나 인터넷, 나노기술 거버넌스 또는 유전자 구동 거버넌스처럼 다양한 분야에서 출현하는 거버넌스에 대

한 보다 복잡한 접근 방식을 지향한다. 미래 AI 거버넌스 모델의 정확한 윤곽은 여전히 유동적이지만 능동 매트릭스 이론, 다중심 거버넌스, 하이브리드 규제 및 메쉬 규제와 같은 고급 거버넌스 모델은 그러한 미래 거버넌스 체제를 어떻게 설계할 수 있는지에 대한 영감과 개념적 지침을 제공할 수 있다. 다음 절에서는 이러한 많은 모델에 공통되는 한 가지 특징, 즉 계층화된 거버넌스의 형태로 구현된 모듈화의 개념을 강조하는데, 이는 앞서 언급한 실질적인 문제를 해결하고 해결함으로써 관련된 모든 행위자 간의 공동 책임이 된다. 이러한 새로운 모델은 AI의 개발과 배치가 공백 상태에서 이루어지지 않기 때문에 적용 가능한 법률과 정책, 특히 인권에 대한 기존의 제도적 프레임워크에 위치해야 하며 상호작용해야 한다는 점에 유의해야 한다.

계층화 모델

모듈화는 복잡한 시스템을 관리하는 주요 메커니즘 중 하나입니다. 모듈화는 상호의존성이 높은 작업과 그렇지 않은 작업을 식별하여 분석해야 하는 상호의존성의 수를 줄이는 것을 목표로 한다. 계층화는 전체 시스템의 다른 부분이 병렬 계층으로 배열되는 모듈화의 특별한 형태를 나타낸다. 계층화의 자주 인용되는 예는 1970년대 후반의 개방형 시스템 상호접속(OSI) 참조 모델이다. 레이어드 모델의 또 다른 예는 David Clark에 의해 제안되었다. 그것은 4개의 레이어가 있는 모델을 사용하여 사이버 공간의 특성을 나타내기 위해서이다. 첫째, 사이버 경험에 참여하는 사람들, 둘째, 사이버 공간에서 저장, 전송, 변환되는 정보, 셋째, 서비스를 구성하는 논리적인 구성 요소.넷째, 논리적 요소를 지원하는 물리적 기반입니다. AI 시스템의 규모, 이질성, 복잡성, 기술 자율성의 정도 등은 정책, 법률, 규제에 대한 새로운 사고를 요구한다. 우리는 세 개의 레이어가 있는 분석 모델을 사용하여 AI 거버넌스의 복잡한 특성을 포착하려고 한다. 위에서 아래로 상호 작용하는 레이어는 다음과 같습니다.

- 사회적, 법적 계층
- 윤리적 계층
- 윤리적 계층과 사회적 계층을 지원하는 기술적 기반 계층

그림 1: AI 거버넌스를 위해 계층화된 모델. 상호작용 계층(사회와 AI 애플리케이션 사이에 위치)은 사회적, 법적, 윤리적, 윤리적 계층을 지원하는 기술적 기반 계층이다.

그림 1은 계층화된 거버넌스 모델을 표현한 것이다. 이 모델은 사회와 AI 애플리케이션 사이에 위치하게 될 것이다. 레이어에 매핑된 기기는 서로 다른 시대에 개발될 수 있다. 단기적으로 거버넌스 제안서는 AI 알고리즘의 표준과 원칙을 개발하는 데 집중할 것이다. 중장기적으로, 국가는 성숙한 AI 응용을 규제하기 위한 구체적인 법안을 마련할 수 있을 것이다. 이 모델은 AI 기반 과제와 기회에 대응하는 원칙, 정책, 규범, 법칙이 어떻게 결합되고 계층 내부와 계층 간에 함께 작동할 수 있는지를 보여주는 유용한 지침이 될 수 있다.

기술 계층

기술 계층은 AI 거버넌스 생태계의 토대이고, 생태계를 구성하는 알고리즘과 데이터이다. AI 시스템과 자율 시스템은 물리적 시스템(자율주행 자동차 및 상용 로봇 등)이나 소프트웨어 시스템(형사사법이나 의료 진단 시스템, 지능형 개인 보조 시스템 등)에 관계없이 데이터와 알고리즘에 의존한다. '데이터, 책임감 있게(Data, Responsibly)'라는 주제로 열린 닥스툴 세미나(Dagstuhl Seminar)에서 책임 있는 알고리즘과 관련하여 제안된 사회적 영향 진술에 대한 원칙이 만들어졌다. 사회적 영향을 미치는 책임 알고리즘에 대해 제안된 원칙은 책임, 설명 가능성, 정확성, 감사성 및 공정성이다. 데이터 거버넌스(data governance)로 알려진 AI 알고리즘에 의한 데이터 수집, 사용 및 관리는 인종, 피부색, 국적, 종교, 생물학적 성별, 사회적 성별, 성적 지향, 장애 또는 집안과 관련된 차별에 대한 공정성과 보호를 도모하는 원칙을 따라야 한다.

윤리 계층

기술 계층 외에도, 우리는 모든 유형의 AI 애플리케이션 및 시스템에 적용되는 높은 수준의 윤리적 우려를 명확히 설명할 수 있었다. 윤리적 원칙의 발전을 위한 중요한 원천 중 하나는 인권 원칙이다. AI 윤리 규범 출현의 또 다른 예는 AI와 자율 시스템에 대한 IEEE 일반 원칙이다. 알고리즘에 의한 행동들은 윤리적인 기준과 원칙에 따라 평가될 수 있다. 예를 들어, AI 애플리케이션이 보험회사의 데이터를 분석하고 특정 그룹의 사람들에게 더 높은 보험료를 부과할 때, 성별이나 나이와 같은 변수를 기준으로 하는 이러한 의사결정 애플리케이션은 항상 동등하거나 공정한 처우를 해야 한다는 윤리적 원칙을 위반하는 것이다.

사회계층과 법률계층

사회 및 법률 계층은 기관을 구성하고 AI와 자율 시스템을 규제하는 책임을 배분하는 과정을 진행할 수 있다. 그 예로, 매튜 쉬어러(Matthew Scherer)는 AI를 정의할 수 있는 권력을 가지고 있고, 연구자가 엄격한 책임을 지지 않고 특정 환경에서 AI 연구를 수행할 수 있는 예외를 만들고 AI 인증 절차를 수립할 수 있는 정책결정 기구에 대해 설명한다. AI를 규제하기 위한 구체적인 규범의 출발점 중 하나는 인권을 포함한 일반적인 국가 및 국제법 체계 외에 윤리 및 기술 계층에서 벗어나는 원칙과 기준이 될 수 있다. 계층화된 모델은 AI와 자율 시스템에 적합한 동작을 정의하는 것을 목표로 AI 거버넌스에 대해 생각하는 프레임워크를 제공한다.

마치며

AI 및 알고리즘 의사결정 시스템을 위한 거버넌스 구조를 구현하는 것은 여러 계층에서 발생할 수 있으며 여러 가지 접근법이 수반된다. 본문에서는 특정 AI 애플리케이션이 존재하는 위험이 실질적이고 구체적인 경우에만 고려된다는 점을 고려하여 이러한 계층 중 일부를 설명하였다. 거버넌스 프로세스는 시장 지향 솔루션에서 정부 기반 구조에 이르기까지 광범위하며 국가 또는 국제적으로 적용될 수 있다. 지역 수준에서 좋은 예시는, 유럽 연합 내의 모든 개인에 대한 데이터 보호를 강화하고 통합하기 위한 광범위하고 복잡한 규정인 일반 데이터 보호 규정(General Data Protection Regulation, GDPR)이다. 이 규정은 기업들이 알고리즘의 목적을 설명하고, 알고리즘이 자동화된 결정을 내릴 때 사용하는 데이터의 종류를 설명할 의무가 있는 (제한된)'설명권'을 의미한다. AI 고유의 국제법적 프레임워크가 없다면 복수 이해당사자 위원회의 형태를 취할 수 있는 글로벌 감독기구가 AI 시스템의 글로벌 원칙과 새로운 규범의 큐레이터가 될 수 있을 것이다.