**Blockchain for explainable and trustworthy artificial intelligence**

[**https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1340**](https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1340)

**0. INTRODUCTION**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 이 논문에서는 **복잡한 AI 시스템에서의 중요한 결정이 신뢰할 수 있는 AI와 XAI 에이전트 간의 합의에 의해 도출**되는 프레임워크를 제안한다.   * 여기서 agent의 대다수가 주장하는 쪽이 옳다고 가정한다. * Trustworthy AI에는 **bias 또는 adversarial 공격으로부터 탄력성**이 있어야 한다는 요구사항이 있는데, 이것은 **블록체인 기술**을 이용하여 충족된다.   블록체인은 XAI 시스템에 다음과 같은 key feature를 제공한다.   |  |  | | --- | --- | | Transparency and visibility | **모든 트랜잭션은 publicly auditable하고 삽입만 가능**하며 **투명한 ledger**에 저장되어야 한다.   * Ledger state와 transaction은 **보안성이 높고 분산된 방법**으로 기록되어야 한다. | | Immutability | **Blockchain ledger는 타임스탬프 블록으로 구성**되어 있다.   * 각 블록은 **hash 암호로 보호**되어 있다. * 각 블록은 여러 transaction으로 구성되어 있고 해당 블록의 hash를 참조한다. | | Traceability and nonrepudiation | 블록체인에 참여하는 각 노드는 **각 트랜잭션 또는 함수 호출을 암호학적으로 서명**해야 한다.   * 각 **signed item은 블록체인을 가진 노드에 의해 인증**되어야 한다. * 트랜잭션은 immutable한 ledger의 일부분이 되고 **transacting node는 이것을 거부할 수 없다.** | | Smart contract | Smart contract (SC)는 **서로 다른 참여자 간의 상호작용을 지배**하고, 자동적이고 신뢰할 수 있으며 분산된 방법의 **contractual 또는 business logic의 실행을 허용**한다. |   따라서 black-box AI에 대한 해결 방법은 하나의 예측 시스템이 아닌, **prediction과 explanation을 제공하는 분산된 predictor를 이용**하는 것이다.   * 여기에는 **AI의 예측과 결정이 기록, 저장, 합성, 관리, 분산**되며 이것은 보안성이 높고 신뢰할 수 있다. |

**1. 블록체인과 AI 기술의 결합**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 블록체인은 각 AI agent에 분산된 AI 시스템을 **open-source and publicly accessible digital ledger를 이용하여 augment**할 수 있다.   * 이것은 **AI agent들이 서로 협력하여 합의를 도출**하고, 새로운 결정을 **추적 가능하고 수정이 어려운 블록**에 저장하게 한다. * 블록체인은 모든 AI agent에게 **AI의 기존 결정들에 대한 투명성과 가시성을 제공**하므로 각 AI agent가 수정 또는 거부할 수 없다. * **Programmable blockchain platform**은 분산된 AI application에 SCs에 기반한 프로그래밍 모델을 사용 가능하게 한다.   **<블록체인과 AI 시스템>**   |  |  | | --- | --- | | 블록체인 | AI agent와 데이터에 대한 각종 보안 공격으로부터 *분산(decentralization), 결정성(determinism), 불변성(immutability), 데이터의 진실성(integrity), 회복력(resilience)*을 제공한다. | | AI 시스템 | 일반적으로 *컴퓨팅 시스템과 데이터가 중앙 집중*되어 있고, 계속 증가하는 데이터를 처리해야 하고 이것은 결정에 영향을 미친다. |  * *블록체인과 AI 시스템의 결합*은 **합의에 기반한 분산된 데이터 및 결정 메커니즘**을 가지므로 **data security와 collective intelligence를 증가**시킨다. |

**2. 블록체인 기반의 Trustworthy AI 프레임워크**

|  |
| --- |
| 이 논문에서는 여**러 node 또는 predictor가 AI computational model을 작동시키고, AI의 결정에 대한 설명을 제공**하며, 이 predictor들에 대한 **reputation을 증가**시킨다.   * 이 접근 방법을 통해, 직접 또는 다른 model-agnostic explanation system을 이용하여 **각 predictor들의 예측에 대한 설명을 제공하는 시스템**을 구성한다. * 이 시스템은 *신뢰할 수 있는 AI 및 XAI node 또는 predictor*를 포함한다.   + 이것은 **계산**을 수행하고, blockchain immutable ledger에 있는 실행 결과와 결정을 기록하는 **다른 blockchain SC와 상호작용**한다. * **predictor는 SC들에게 결과를 제공**하는 역할을 한다. |

**3. 블록체인 프레임워크**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Frontend  **DApps** | 설명 가능하고 신뢰할 수 있는 AI 응용 프로그램을 위한 이 시스템은 **다양한 분산된 frontend application을 지원**한다.   * Frontend DApp의 인터페이스는 **AI 또는 XAI predictor의 수를 설정**하고 이 외에도 **파라미터를 다양하게 지정**할 수 있다. | | Backend  **Access Layer** | 다양한 **데이터 전송 프로토콜을 위한 인터페이스**를 가능하게 한다.   * DApps와 블록체인 간 직접 통신을 위한 Web3 등 | | Backend  **AI Layer** | 가장 중요한 부분으로, 데이터 프로세싱 및 데이터 마이닝을 통해 **신뢰할 수 있고 협력적이면서 동의에 기반한 결정**이 실행된다.   * *AI predictor ():* 데이터를 전통적인 **Black-box 모델**을 통해 처리하고 예측을 생성한다. * *XAI predictor ():* AI predictor와 데이터 처리 과정이 다르며, **decision outcome에 summary information을 추가**로 제공한다.   + XAI predictor는 다른 AI predictor와 협력하여 **설명 능력**과 함께 **설명 가능한 결정에 대한 요약**을 생성한다. | | Backend  **Support services** | **Registration service와 Reputation service**라는 2가지의 서비스 제공   * *Registration service:* 시스템 환경에 **참여자를 등록하고 관리**할 수 있게 한다. (사용자, AI 및 XAI 예측 서비스 제공자 등) * *Reputation service:* AI 및 XAI predictor의 **reputation을 관리**하며, **대다수의 결정과 다르게 결정한 predictor는 페널티**를 받는다. | | Backend  **Blockchain platform** | 서로 다른 SC들을 동작시키기 위한 **블록체인 네트워크**, AI와 XAI predictor들에 의해 보고된 결과 및 메타데이터를 저장하는 **분산 저장소**를 포함한다. | |

**4. 기본적인 작업 수행 과정**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Oracle registration and predictor selection | 여기서 제안한 블록체인 기반 프레임워크는 **publicly available한 prediction을 제공하는, 신뢰할 수 있는 oracle**을 허용한다.   * Oracle node는 여러 개의 AI 또는 XAI predictor를 host할 수 있다. | | Decision Making | 이 프레임워크는 **open public platform**이므로 **서로 다른 여러 oracle이 서로 다른 종류의 AI/XAI predictor를 사용**하여 같은 문제를 해결할 수 있다.   * 이때 AI 알고리즘의 계산 구조 등의 이유로 **서로 다른 AI/XAI predictor 간의 decision의 variation**이 중요하다. * 따라서 AI/XAI predictor를 다음과 같이 2가지로 구분한다.  |  |  | | --- | --- | | Deterministic predictor | 입력 및 학습 데이터에 의해 **결정적인 결론**(True/False 등)을 도출한다. | | Nondeterministic predictor | **비결정적인 결론**을 도출하며, 현존하는 대부분의 AI/XAI predictor가 이에 해당한다. | | | Aggregated decisions | Aggregator SC는 **predictor로부터 보고된 hash들을 비교**하고 **majority에 기반하여 알맞은 결정**을 도출한다. | |

**5. 실제 응용 사례**

|  |  |
| --- | --- |
| Medical image diagnosis | **Deep medicial image analytic 기술**을 분산된 disease diagnose system을 이용하여 발전시킬 수 있다. |
| Customer profiling | Social network 및 인터넷 포털 사이트 상의 **거대한 고객 데이터를 분석**하는 데 도움이 된다. |
| Tax auditing and fraud detection | 정부는 시민의 은행 트랜잭션, 수입원, 세금 납부 등에 대한 빅 데이터를 분석할 수 있으므로 **세금 납부 회피 및 관련 사기 행위를 탐지**할 수 있다. |
| Voting and election predictions | **허위 투표 또는 선거운동에서의 불법적이거나 돈을 이용한 프로모션 행위**를 탐지하고 설명할 수 있다. |
| Use cases for real-time AI applications | 신뢰할 수 있고 설명 가능한 실시간적인 예측 및 결정을 위하여, **자율주행 자동차에 적용하여 사고가 발생하기 전에 자동차의 결정에 대해 분석하고 설명**할 수 있는 등으로 활용할 수 있다. |