ISSN 1870-4069

Agentes autónomos descentralizados: Convergencia de grandes modelos de lenguaje y Blockchain para la toma de decisiones automatizadas, auditables y colaborativas

Jorge Polanco Roque

Instituto Tecnológico y de Estudios Superiores de Monterrey, México

jorge.polanco@tec.mx

Resumen. La conjunción de la inteligencia artificial (IA), especialmente a partir de modelos de lenguaje de gran tamaño (LLMs), y la tecnología de cadena de bloques (blockchain) está impulsando una transformación profunda de las estructuras de gobernanza y de los procesos de toma de decisiones en entornos distribuidos. En este contexto, surgen los Agentes Autónomos Descentralizados (AADs) como unidades de software capaces de razonar sobre datos registrados en contratos inteligentes, coordinarse con otros actores de la red y desencadenar acciones tanto on-chain como off-chain sin requerir intermediarios centralizados. Este trabajo presenta una propuesta conceptual para el desarrollo de AADs, articulando los componentes tecnológicos clave —infraestructura en la nube, interacción con redes blockchain, diseño de contratos inteligentes y oráculos— y su integración con modelos de IA que dotan a los agentes de capacidades cognitivas avanzadas. Se incluye un estudio de caso sobre votaciones descentralizadas, así como un marco preliminar de validación que aborda vulnerabilidades técnicas, riesgos de seguridad algorítmica y amenazas emergentes como ataques por inyección de prompts. Finalmente, se discuten los desafíos regulatorios y éticos, y se trazan líneas de investigación futura en torno a la gobernanza automatizada y transparente.

Palabras clave: Agentes autónomos descentralizados, blockchain, inteligencia artificial, modelos de lenguaje, smart contracts, validación de seguridad, auditoría algorítmica, prompt injection, gobernanza descentralizada, oráculos, web 3.0.

Decentralized Autonomous Agents: Convergence of Large Language Models and Blockchain for Automated, Auditable, and Collaborative Decision-Making

Abstract. The conjunction of artificial intelligence (AI), especially from large language models (LLMs), and blockchain technology is driving a profound transformation of governance structures and decision-making processes in distributed environments. In this context, decentralized autonomous agents (DAAs) emerge as software units capable of reasoning on data recorded in smart

contracts, coordinating with other network actors, and triggering both on-chain and off-chain actions without requiring centralized intermediaries. This paper presents a conceptual proposal for the development of DAAs, articulating the key technological components—cloud infrastructure, interaction with blockchain networks, smart contract and oracle design—and their integration with AI models that equip agents with advanced cognitive capabilities. A case study on decentralized voting is included, as well as a preliminary validation framework that addresses technical vulnerabilities, algorithmic security risks, and emerging threats such as prompt injection attacks. Finally, regulatory and ethical challenges are discussed, and future research directions for automated and transparent governance are outlined.

Keywords: Decentralized autonomous agents, blockchain, artificial intelligence, language models, smart contracts, security validation, algorithmic auditing, prompt injection, decentralized governance, oracles, web 3.0.

1. Introducción

La proliferación de tecnologías como la IA y blockchain ha posibilitado la construcción de ecosistemas digitales con un alto grado de autonomía y fiabilidad. La IA, impulsada por métodos de aprendizaje profundo y de aprendizaje automático tradicional, ha incrementado sus capacidades gracias a los denominados modelos de lenguaje de gran tamaño, que permiten extraer patrones y razonar de manera contextual a partir de un vasto conjunto de datos (Wang et al., 2019). Al mismo tiempo, la tecnología blockchain ofrece un entorno distribuido, inmutable y trazable para la ejecución de transacciones y el despliegue de aplicaciones descentralizadas, lo que reduce la dependencia de entidades centrales e incrementa la confianza y la transparencia (Nakamoto, 2008; Tapscott & Tapscott, 2016).

En la convergencia de estas dos corrientes tecnológicas se sitúan los Agentes Autónomos Descentralizados. Dichos agentes adquieren un rol activo en la red, recabando información on-chain y off-chain para procesarla mediante diversos tipos de algoritmos—particularmente, mediante LLMs— y emitir decisiones o recomendaciones que pueden plasmarse directamente en contratos inteligentes. Esto abre un abanico de oportunidades en la gobernanza distribuida, la gestión de cadenas de suministro, el arbitraje de disputas, la detección de fraude y la orquestación de recursos en organizaciones virtuales. Al mismo tiempo, surgen importantes preguntas acerca de la escalabilidad de estas soluciones, la responsabilidad legal en entornos completamente descentralizados y la forma en que se garantizará la privacidad de los datos personales (Yaga et al., 2018).

Este artículo se estructura en nueve secciones para analizar la convergencia entre inteligencia artificial y blockchain en la construcción de Agentes Autónomos Descentralizados. Tras esta introducción, la segunda sección expone los fundamentos de ambas tecnologías, con énfasis en los modelos de lenguaje de gran tamaño y su papel en la toma de decisiones en entornos distribuidos.

En la tercera sección, se presenta un marco conceptual para el desarrollo de AADs, detallando la infraestructura necesaria, la conexión con blockchain mediante proveedores como Infura, la gobernanza descentralizada y la integración con oráculos.

La cuarta sección introduce un estudio de caso sobre votaciones descentralizadas, donde un AAD con capacidades cognitivas avanzadas supervisa la emisión de votos y detecta irregularidades.

La quinta sección explora mejoras tecnológicas y propuestas algorítmicas para optimizar el desempeño de los AADs, incluyendo mecanismos de consenso basados en reputación, Zero-Knowledge Proofs y estrategias de escalabilidad mediante sharding temático. En la sexta sección, se analizan los desafíos normativos y éticos, con un enfoque en la responsabilidad legal de las decisiones automatizadas y su compatibilidad con marcos regulatorios.

La séptima sección introduce un marco preliminar de validación para Agentes Autónomos Descentralizados, donde se analizan vulnerabilidades técnicas asociadas a contratos inteligentes, riesgos derivados del comportamiento de modelos de lenguaje de gran tamaño, y posibles vectores de ataque como la manipulación de prompts. Se discuten estrategias de mitigación basadas en verificación formal, auditoría algorítmica, mecanismos híbridos con supervisión humana, y protección frente a inyecciones de instrucciones maliciosas en el flujo de interacción entre la IA y la blockchain. Finalmente, la octava sección presenta conclusiones y futuras líneas de investigación, abordando el impacto potencial de los AADs en la gobernanza descentralizada y la economía digital.

Este trabajo se inscribe dentro del enfoque de conceptualización técnica, con el objetivo de proponer un marco de referencia estructurado para el diseño e implementación de Agentes Autónomos Descentralizados. A diferencia de trabajos empíricos que validan implementaciones específicas, aquí se busca integrar elementos tecnológicos, arquitectónicos y normativos en una visión coherente que permita guiar desarrollos futuros. El objeto de estudio son los AADs entendidos como entidades de software capaces de ejercer agencia en entornos descentralizados, y el problema abordado es la ausencia de un marco articulado que contemple tanto sus capacidades técnicas como sus implicaciones en sistemas de gobernanza distribuida.

La propuesta se sustenta en antecedentes de investigación sobre modelos de lenguaje (Brown et al., 2020; Weidinger et al., 2022), infraestructuras blockchain (Wood, 2014; Wang et al., 2019) y estructuras de gobernanza algorítmica (Hassan & De Filippi, 2021), contribuyendo a articular una síntesis que permita mapear desafíos técnicos, éticosy regulatorios.

2. Fundamentos de Blockchain, votaciones descentralizadas e IA

La tecnología blockchain se fundamenta en un registro inmutable y distribuido, donde los nodos validan transacciones mediante protocolos de consenso criptográficos. Este paradigma, introducido con Bitcoin (Nakamoto, 2008), evolucionó con redes como Ethereum, que incorporan contratos inteligentes para la ejecución programática de transacciones (Buterin, 2014). Aunque el consenso inicial se basaba en prueba de trabajo (Proof-of-Work), alternativas como la prueba de participación (Proof-of-Stake) han mejorado la escalabilidad y reducido el consumo energético (Wood, 2014).

No obstante, los costos de transacción y la latencia han impulsado soluciones de segunda capa, como sidechains y rollups, que optimizan la eficiencia de la red principal (Wang et al., 2019).

Paralelamente, la inteligencia artificial ha avanzado desde modelos de aprendizaje automático tradicionales hasta arquitecturas neuronales profundas capaces de procesar grandes volúmenes de datos. En este contexto, los modelos de lenguaje de gran tamaño han demostrado un desempeño sobresaliente en comprensión y generación de texto, síntesis de información y razonamiento contextual (Zhang et al., 2019). Estas arquitecturas —como GPT-4— se entrenan con conjuntos masivos de datos en infraestructuras escalables y pueden ajustarse para tareas específicas de todo tipo.

La convergencia entre IA y blockchain radica en la capacidad de los modelos de lenguaje para automatizar y optimizar procesos en redes descentralizadas, mientras la blockchain garantiza transparencia, auditabilidad e integridad de los datos (Tapscott & Tapscott, 2016). Para lograr esta integración, parte del procesamiento de IA se delega a infraestructuras en la nube o entornos de computación distribuida, mientras la blockchain registra los resultados esenciales, asegurando inmutabilidad y consenso.

Un campo donde esta integración cobra especial relevancia es el de las votaciones descentralizadas. La inmutabilidad y trazabilidad de la blockchain permiten un registro confiable de cada voto y garantizan la transparencia en los resultados.
Sin embargo, la incorporación de agentes basados en modelos de lenguaje introduce una capa de interpretación y análisis que amplía las capacidades de supervisión y detección de irregularidades en el proceso electoral de una “DAO”, por ejemplo.

Las DAOs (Organizaciones Autónomas Descentralizadas) son estructuras de gobernanza basadas en contratos inteligentes, donde las decisiones se toman colectivamente mediante un sistema de votación con tokens. Una de las primeras experiencias de gobernanza algorítmica mediante DAOs fue documentada por DuPont (2019) en su estudio histórico sobre 'The DAO', ilustrando los riesgos técnicos y organizacionales que pueden surgir cuando sistemas autónomos gestionan recursos colectivos.

Su transparencia e inmutabilidad han permitido su adopción en sectores como finanzas descentralizadas (DeFi), inversión colectiva y gestión de comunidades, eliminando intermediarios y fortaleciendo la autonomía organizativa.

En este contexto, los LLM pueden actuar como supervisores inteligentes, analizando dinámicas de participación y detectando anomalías en tiempo real. Un modelo entrenado y/o contextualizado en gobernanza descentralizada puede evaluar si una propuesta contradice decisiones previas, identificar estrategias de manipulación por grupos minoritarios o detectar patrones de votación atípicos. Además, estos agentes pueden generar reportes en lenguaje natural, facilitando la toma de decisiones informadas por parte de la comunidad.

Más aún, los LLM pueden desempeñar un papel activo en la votación, operando como representantes algorítmicos dentro de una DAO. Un agente de este tipo podría analizar el historial de gobernanza, evaluar propuestas según criterios predefinidos y emitir votos en representación de usuarios que deleguen su decisión en la IA. Esto resulta particularmente útil en entornos donde la toma de decisiones requiere análisis técnico o económico complejo, permitiendo que un Agente Autónomo Descentralizado vote con base en información estructurada y en tiempo real.

Al combinar el razonamiento avanzado de los LLM con la seguridad y descentralización de la blockchain, es posible construir sistemas de gobernanza más eficientes y autónomos. En lugar de depender exclusivamente de reglas predefinidas o de la intervención manual de los votantes, las DAOs pueden evolucionar hacia modelos

de decisión más dinámicos, donde la IA no solo detecte irregularidades, sino que también proponga estrategias de mitigación, facilite la deliberación colectiva y participe activamente en la toma de decisiones.

Además, la participación de estos agentes refuerza la confianza en los procesos electorales, ya que su supervisión sigue reglas codificadas en contratos inteligentes, auditable por la comunidad y sin intervención de una autoridad central. Esto no solo mejora la resistencia de las votaciones frente a manipulaciones, sino que también permite la adaptación dinámica de las reglas de gobernanza conforme evoluciona la comunidad, consolidando así un modelo más robusto y participativo.

La integración de modelos de lenguaje de gran tamaño con blockchain redefine la gobernanza descentralizada al combinar automatización, transparencia y análisis avanzado en la toma de decisiones. Esta sinergia no solo fortalece la resiliencia de los sistemas de votación descentralizados, sino que también permite la evolución hacia modelos de gobernanza más dinámicos, eficientes y adaptativos, reduciendo riesgos de manipulación y mejorando la toma de decisiones en comunidades autónomas.

3. Arquitectura de los agentes autónomos descentralizados

El concepto de Agentes Autónomos Descentralizados surge como un intento de unificar la lógica de los agentes autónomos de IA y la ejecución segura en blockchain en un único marco de referencia. La necesidad de integrar inteligencia autónoma con contratos autoejecutables en registros distribuidos ha sido explorada por autores como Christidis y Devetsikiotis (2016), quienes sentaron las bases para la fusión de blockchain, contratos inteligentes y agentes autónomos en aplicaciones como IoT. Más recientemente, Kuznetsov et al. (2024) analizaron los retos vinculados a la integración de IA y tecnología blockchain, haciendo énfasis en los desafíos de seguridad, escalabilidad y confiabilidad en contextos descentralizados.

El surgimiento de Agentes Autónomos Descentralizados debe entenderse como un paso evolutivo posterior al desarrollo de Organizaciones Autónomas Descentralizadas y a los primeros experimentos de gobernanza algorítmica (DuPont, 2017; Wright & De Filippi, 2015). Mientras que las DAOs delegan la ejecución de acuerdos a contratos inteligentes, los AADs extienden esta capacidad mediante la incorporación explícita de mecanismos de razonamiento autónomo basados en IA, aproximándose a un modelo de agencia algorítmica distribuida. Esta visión se alinea con los lineamientos propuestos por Kuznetsov et al. (2024), quienes describen la arquitectura de sistemas híbridos donde la autonomía algorítmica y la inmutabilidad del registro distribuido permiten una toma de decisiones automatizada, segura y verificable.

Un AAD se concibe como una entidad de software que posee llaves criptográficas para firmar transacciones, extrae datos tanto de la cadena (on-chain) como de fuentes externas (off-chain), y razona sobre dichos datos mediante algoritmos de inteligencia artificial, con el objetivo de desplegar acciones automatizadas. En términos de arquitectura, la infraestructura para un AAD puede organizarse en capas que facilitan su diseño y despliegue:

- En la **Capa de Datos y Conectividad**, el AAD se vincula con proveedores de nodos, como Infura, para interactuar con la red principal sin necesidad de mantener un nodo completo, lo que reduce la complejidad de configuración y

mantenimiento. Además, esta capa integra servicios de almacenamiento como InterPlanetary File System (IPFS) o bases de datos en la nube para manejar datos masivos que no se almacenan directamente en la blockchain. El AAD también establece enlaces con oráculos, que suministran información externa, por ejemplo, datos de mercado o identidad digital, y permiten la verificación de eventos del mundo real.

- En la Capa de Inteligencia Artificial, se emplean modelos de lenguaje de gran tamaño para procesar y razonar sobre grandes volúmenes de información. Estos modelos se entrenan en infraestructuras de computación escalables (GPU, TPU) y luego se implementan en servidores o contenedores que puedan comunicarse con la capa on-chain mediante interfaces de programación. El LLM puede especializarse en la detección de fraude, el análisis de transacciones financieras, la clasificación de propuestas en la DAO o cualquier otra tarea requerida por la organización descentralizada.
- En la Capa de Blockchain y Contratos Inteligentes, se definen las reglas de operación de la red y las condiciones en las que el AAD puede tomar acciones específicas. Los contratos inteligentes, escritos en lenguajes como Solidity (en el caso de Ethereum), recogen la lógica necesaria para transferir activos, restringir comportamientos indebidos y gestionar eventos relevantes para el AAD. De esta manera, la validación de cada acción del agente se registra en la red, generando transparencia y facilitando el escrutinio público o privado.
- En la Capa de Gobernanza Descentralizada, la comunidad o los participantes que ostenten tokens de gobernanza votan las actualizaciones de políticas, los parámetros de IA y las potenciales sanciones a comportamientos maliciosos. El AAD puede tener derecho de voto si la comunidad así lo dispone, o puede asumir un papel de auditor a fin de identificar anomalías y proponer sanciones. Este mecanismo de participación abierta y verificable crea incentivos para la contribución honesta y dificulta la manipulación por parte de uno o pocos actores.
- Finalmente, en la Capa de Interfaz y Aplicaciones Híbridas, los usuarios finales, desarrolladores o empresas interactúan con el AAD y con la red. Esta capa puede incluir tableros de control, formularios de votación o servicios de suscripción a eventos. Combina tecnologías web 2.0 tradicionales (por ejemplo, servidores en AWS, front-ends JavaScript) con el acceso a la blockchain a travéss de librerías como web3.js o ethers.js, con la finalidad de presentar la información de la forma más accesible posible.

Esta arquitectura escalable y modular se sustenta en la sinergia entre la nube y la cadena de bloques, aprovechando la potencia de cómputo fuera de la cadena para entrenar y ejecutar la IA, mientras la blockchain garantiza la inmutabilidad, la trazabilidad y la gobernanza colectiva (Zyskind et al., 2015).

4. Caso de uso: Votaciones descentralizadas asistidas por LLMs

Un Agente Autónomo Descentralizado potenciado por un modelo de lenguaje de gran tamaño actúa como un sistema de monitoreo dinámico dentro de un proceso de

votación, contrastando las dinámicas actuales de participación con distribuciones históricas modeladas estadísticamente. A partir de este análisis, el AAD puede identificar patrones atípicos o desviaciones significativas y reportarlas formalmente mediante alertas registradas on-chain, desencadenando mecanismos de auditoría automática o intervención comunitaria. Este enfoque permite incorporar capacidades de supervisión avanzadas en estructuras de gobernanza descentralizada, siguiendo la lógica de resiliencia algorítmica delineada por Christidis y Devetsikiotis (2016).

La metodología para implementar este caso incluye la creación de un contrato inteligente que dicta el plazo de la votación y almacena los resultados de forma cifrada. El AAD recibe notificaciones en tiempo real de las transacciones vinculadas a la votación, gracias a la conexión con un nodo Ethereum vía Infura, y a un oráculo que extrae metadatos sobre los votantes. El LLM analiza la coherencia de cada voto respecto a la base histórica de comportamientos normales, enviando alertas al contrato inteligente cuando encuentra discrepancias o indicios de manipulación. En ese momento, la DAO puede suspender temporalmente la votación y requerir aclaraciones o auditorías, implicando nuevamente al LLM para evaluar evidencia adicional.

Los resultados de estas pruebas pudieran demostrar que la participación de un AAD supervisando la votación potencia la confianza de la comunidad, al brindar detección temprana de fraude y total transparencia de las acciones del agente, auditables en la cadena de bloques. No obstante, la experimentación pudiera señalar que el incremento del número de votos y de transacciones genera un aumento significativo de costos de gas, lo cual motiva la adopción de soluciones en sidechains o en capas de escalabilidad para gestionar la afluencia masiva de participantes de AAD.

5. Propuestas de mejoras tecnológicas y algorítmicas

La implementación práctica de Agentes Autónomos Descentralizados que emplean modelos de lenguaje de gran tamaño enfrenta retos en confiabilidad, privacidad y escalabilidad. Para abordarlos, existen diversas estrategias que permiten llevar estos agentes a entornos de producción de manera más segura y eficiente.

Un primer paso es ampliar los mecanismos de consenso más allá de la prueba de participación (Proof-of-Stake), incorporando métricas de reputación o confiabilidad generadas por la IA. Este enfoque requiere que cada nodo aporte un historial de comportamiento (p. ej., participación en votaciones, precisión en la detección de fraude o calidad de datos en procesos de aprendizaje federado). El LLM evalúa este historial y genera una puntuación de reputación dinámica, la cual se integra en el consenso para penalizar nodos maliciosos y recompensar contribuciones positivas.

En cuanto a la privacidad, el uso de cifrado homomórfico y Zero-Knowledge Proofs (ZKPs) (Zyskind et al., 2015) permite que los AADs trabajen con datos sensibles sin desencriptarlos. Por ejemplo, el agente podría validar la legitimidad de una transacción o marcarla como fraudulenta mediante pruebas criptográficas que demuestren la corrección de su análisis sin revelar su contenido. Aunque esta aproximación eleva la complejidad y el costo computacional, resulta esencial en aplicaciones donde la confidencialidad es prioritaria.

Para escalar estas operaciones, la fragmentación de la red en subcadenas especializadas o sharding temático (Wang et al., 2019) distribuye la carga de cómputo.

Una subcadena puede dedicarse al entrenamiento y la ejecución del modelo IA, mientras la cadena principal registra únicamente el hash de los resultados y las actualizaciones clave. Esto aligera la congestión y reduce los costos de transacción, pero exige un protocolo de interoperabilidad multicadena, que permita a los AADs leer y escribir en las distintas subcadenas y, a su vez, publicar en la red principal los resultados finales de su procesamiento.

Un ejemplo práctico de este flujo incluye la fase de entrenamiento y evaluación en una subcadena, seguida de la generación de una prueba criptográfica que describe los resultados, y finaliza con el registro en la cadena principal, donde un contrato inteligente solo acepta los resultados acompañados de dicha prueba. En el caso de votaciones, la misma lógica puede garantizar la validez de las decisiones tomadas por el LLM, exigiendo evidencias de su análisis on-chain para salvaguardar la transparencia del proceso.

La auditoría algorítmica constituye otro pilar fundamental. La DAO puede configurar contratos inteligentes de auditoría que obliguen al AAD a proporcionar explicaciones resumidas de las decisiones más relevantes, ya sea la suspensión de un voto masivo o la detección de comportamientos anómalos. Si la comunidad considera insuficientes las justificaciones, puede votar sanciones o la revocación de los privilegios del agente. Este mecanismo no solo promueve la responsabilidad del AAD, sino que también aporta claridad y confianza a los participantes.

La integración técnica entre IA y blockchain requiere librerías o APIs que faciliten la comunicación bidireccional. La IA debe consultar datos on-chain con baja latencia, lo que implica la adopción de servicios como Infura o nodos locales configurados de forma óptima. Por otra parte, los resultados de la IA tienen que transformarse en transacciones firmadas por el propio agente, para lo que se requiere un módulo criptográfico de alta seguridad, ya sea implementado en enclaves de hardware o en contenedores con medidas de protección reforzadas.

En conjunto, la incorporación de mecanismos de reputación IA, cifrado homomórfico, ZKPs y sharding temático permite construir AADs escalables y seguros. La auditoría algorítmica, sumada a la gobernanza descentralizada, refuerza la confianza en la toma de decisiones automatizada e impulsa nuevos modelos de colaboración y gobernanza en la economía digital.

6. Perspectiva normativa y ética

La creciente autonomía de los Agentes Autónomos Descentralizados conlleva desafíos significativos en el ámbito legal y ético, pues redefine la manera en que se asigna la responsabilidad, se salvaguarda la privacidad y se gestionan los sesgos inherentes a los modelos de inteligencia artificial.

Un primer punto de tensión surge en torno a la asignación de responsabilidad cuando las decisiones son tomadas por un AAD que, en sentido estricto, no se halla bajo el dominio de una autoridad central. Este fenómeno se enmarca dentro de lo que Wright y De Filippi (2015) conceptualizaron como 'Lex Cryptographia', es decir, sistemas donde el código y las reglas algorítmicas sustituyen —al menos parcialmente— a las formas tradicionales de regulación legal. Tal como expusieron Wright y De Filippi (2015) en su análisis sobre Lex Cryptographia, la programación de reglas

autoejecutables en entornos distribuidos redefine las nociones clásicas de contrato, autoridad y responsabilidad, trasladándolas al dominio del código. Los AADs, al operar bajo estos principios, intensifican los dilemas regulatorios al eliminar progresivamente los intermediarios humanos que tradicionalmente serían los sujetos de derecho.

Los marcos legales vigentes, concebidos para organizaciones con personalidad jurídica y jerarquías definidas, carecen de pautas claras para determinar quién responde en caso de que las acciones de un AAD ocasionen daños o vulneren derechos. Las DAO pueden implementar mecanismos internos de resolución de disputas y votaciones correctivas, pero el valor legal de tales resoluciones ante las jurisdicciones tradicionales es incierto (Tapscott & Tapscott, 2016). En la práctica, se abre la posibilidad de que los tribunales requieran identificar a los desarrolladores o promotores de la DAO, extendiendo la responsabilidad a quienes hayan facilitado la puesta en marcha del agente autónomo. Esto contrasta con la intención de descentralizar la gobernanza y reducir la dependencia de figuras centrales, generando un vacío regulatorio que exige la adaptación de los marcos normativos.

Un segundo punto delicado está relacionado con la protección de datos personales y la privacidad. La inmutabilidad de la cadena de bloques desafía la aplicación de leyes como el Reglamento General de Protección de Datos (RGPD) en Europa, que establece derechos como el de supresión o rectificación de la información (Yaga et al., 2018).

Si bien la cadena de bloques garantiza la trazabilidad y la verificación de registros, esta característica choca con la posibilidad de eliminar completamente un dato o revertirlo de manera forense. Para abordar estas exigencias, se investigan técnicas criptográficas de "borrado selectivo" o chameleon hashes, que permiten enmendar transacciones bajo consenso de la red, aunque su adopción masiva conlleva profundos cambios en el diseño de los protocolos y puede diluir, en cierta medida, el principio de inmutabilidad. Otra línea de desarrollo se centra en la disociación y el cifrado homomórfico, para que el contenido sensible nunca se exponga en texto plano, aunque estos métodos suelen incrementar la complejidad de la red y los costos de cómputo.

En cuanto a la ética de la IA, la naturaleza de los modelos de lenguaje de gran tamaño implica el riesgo de reproducir sesgos de los datos de entrenamiento, lo que puede derivar en discriminaciones o falsos positivos y perjudicar a ciertos sectores de la población. Los AADs que incorporen IA en su lógica decisoria enfrentan la necesidad de asegurar que dichos modelos cumplan criterios de justicia y equidad. Una manera de mitigar estos riesgos es la auditoría algorítmica, en la que se somete el comportamiento del modelo a evaluaciones continuas por parte de la comunidad o de equipos especializados. Sin embargo, la responsabilidad de corregir dichos sesgos recae en un colectivo difuso: los propietarios de tokens de gobernanza, los desarrolladores que implementan las actualizaciones y, en última instancia, la comunidad de nodos.

Si el AAD asume funciones críticas —por ejemplo, supervisar transacciones financieras o decidir políticas de un fondo de inversión—, estos sesgos podrían tener consecuencias sociales y económicas de gran alcance.

La transparencia de las operaciones constituye otro factor clave en la legitimidad de los AADs. La capacidad de un agente para firmar transacciones y desencadenar acciones en la blockchain sin una supervisión humana explícita puede generar preocupación si su lógica interna se percibe como una "caja negra". Para enfrentar esta inquietud, las DAO pueden exigir que el agente ofrezca descripciones mínimamente explicables de su razonamiento en los casos con mayor impacto. Herramientas de

explicabilidad (por ejemplo, mecanismos de extracción de reglas o resúmenes de atención en modelos de lenguaje) permiten a los participantes comprender las motivaciones del AAD y, de ser necesario, responder con votaciones que limiten su autonomía o exijan reentrenamientos. No obstante, esta aspiración de transparencia debe equilibrarse con la protección de datos y la propiedad intelectual, pues exponer todos los detalles del modelo podría revelar información confidencial o estratégica.

Otro eje de tensión concierne a la compatibilidad con los sistemas legales existentes. Muchas jurisdicciones no reconocen la figura de un agente autónomo sin personalidad jurídica, lo que complica su participación en contratos legalmente vinculantes. La figura del "contrato inteligente" en sí misma plantea cuestiones acerca de la ejecutabilidad legal de los acuerdos, especialmente si las cláusulas se ejecutan de manera automática sin posibilidad de reclamación ante un órgano jurisdiccional. Algunas iniciativas exploran la creación de organizaciones híbridas, donde la DAO se registra bajo formas legales tradicionales, lo que habilitaría una responsabilidad más clara en caso de disputas o demandas.

Finalmente, la responsabilidad en la actualización y el mantenimiento de estos agentes abre interrogantes relacionados con la continuidad de su desarrollo y la legitimidad de los cambios introducidos. Un AAD basado en IA no es estático: sus modelos pueden requerir reentrenamientos y/o contextualizaciones nuevas, ajustes de parámetros o cambios en el set de datos que se considera confiable. La DAO debe definir procesos transparentes para aprobar cada actualización y asegurar que los equipos técnicos cumplan con requisitos de calidad y validación. Si los procedimientos resultan ambiguos o demasiado restrictivos, se arriesga la obsolescencia del agente; en cambio, si se otorga un margen de maniobra excesivo, la gobernanza colectiva puede perder control en la gobernanza.

En conclusión, la autonomía de los AADs y su integración con modelos de inteligencia artificial generan un conjunto de desafíos que no se limitan al plano técnico, sino que se extienden al terreno legal, ético y socioeconómico. La posibilidad de que la DAO sirva como espacio de resolución de disputas y vigilancia colectiva ofrece oportunidades para repensar la responsabilidad y la equidad, pero también introduce incertidumbre sobre la eficacia de tales mecanismos fuera del ecosistema blockchain. En la medida en que las iniciativas tecnológicas avancen, resultará crucial articular diálogos con los reguladores y los actores de la sociedad civil, a fin de diseñar marcos normativos flexibles que reconozcan la naturaleza distribuida de estos agentes y, al mismo tiempo, protejan los derechos fundamentales y promuevan la innovación responsable.

7. Validación de la propuesta: Análisis de vulnerabilidades y consideraciones de seguridad

La validación rigurosa de los Agentes Autónomos Descentralizados resulta indispensable para garantizar su viabilidad en entornos de gobernanza descentralizada. Esta validación debe abordar tanto los riesgos inherentes a la infraestructura blockchain como las vulnerabilidades asociadas al uso de modelos de lenguaje de gran tamaño.

En primer lugar, la resiliencia de los contratos inteligentes que sustentan las operaciones de los AADs debe ser verificada mediante técnicas de verificación formal

y análisis estático de seguridad, utilizando herramientas como MythX o Slither (Feist, Grieco, & Groce, 2019). Estas técnicas permiten detectar vulnerabilidades comunes como reentrancy attacks, integer overflows o accesos indebidos, contribuyendo a mitigar riesgos sistémicos antes del despliegue.

En segundo término, el comportamiento del modelo de lenguaje debe someterse a evaluaciones de robustez adversarial (Goodfellow, Shlens, & Szegedy, 2015). La generación de ejemplos adversarios y la implementación de mecanismos de defensa como el adversarial training resultan esenciales para identificar debilidades en la inferencia del modelo, especialmente frente a inputs maliciosos diseñados para inducir comportamientos anómalos o sesgados.

Particular atención requiere la mitigación de ataques basados en ingeniería de prompts (Prompt Injection Attacks), donde un agente malicioso podría manipular las instrucciones internas del modelo para alterar su comportamiento previsto (Perez, Ranta, & Raffel, 2022; Carlini et al., 2021). Para reducir este riesgo, es necesario diseñar prompts robustos, implementar filtros de sanitización de entradas, emplear técnicas de prompt verification, y limitar las acciones críticas a aquellos outputs validados explícitamente mediante políticas de control (Weidinger et al., 2022).

Adicionalmente, se propone la incorporación de mecanismos de validación híbrida (Human-in-the-Loop, HITL) (Amershi et al., 2014), en los cuales decisiones críticas tomadas por el AAD sean auditadas por validadores humanos antes de su ejecución definitiva. Este enfoque híbrido refuerza la confiabilidad general del sistema, equilibrando eficiencia automatizada con supervisión consciente.

La consistencia en las dinámicas de gobernanza también debe ser validada mediante simulaciones de ataque de gobernanza (governance attack simulations), que modelen escenarios de manipulación de votaciones, colusión de nodos o concentración de poder mediante la acumulación de tokens (Hassan & De Filippi, 2021).

Finalmente, se enfatiza la necesidad de establecer un esquema de auditoría algorítmica continua (Brundage et al., 2020), en el cual los AADs deban proporcionar explicaciones auditables de sus acciones relevantes, registradas on-chain, permitiendo así una trazabilidad robusta y un control comunitario efectivo.

En conjunto, estos mecanismos conforman un marco de validación preliminar que refuerza la robustez, confiabilidad y legitimidad de los Agentes Autónomos Descentralizados en escenarios productivos, sentando las bases para su adopción escalable y su integración segura en ecosistemas de gobernanza distribuida.

8. Conclusiones y futuras líneas de investigación

La construcción de Agentes Autónomos Descentralizados con modelos de lenguaje de gran tamaño en redes blockchain configura un nuevo paradigma de gobernanza y automatización que trasciende la lógica habitual de sistemas centralizados. Al permitir la participación directa de la IA en la ejecución de contratos inteligentes y la toma de decisiones colectivas, se abren vías innovadoras para la gestión de recursos, la supervisión de procesos y la detección temprana de anomalías. Esta convergencia de IA y blockchain, no obstante, exige equilibrios técnicos, normativos y éticos que se han discutido a lo largo del presente documento.

Por un lado, la optimización técnica demanda tanto soluciones de escalabilidad (sharding temático, rollups, sidechains) como la adopción de técnicas criptográficas avanzadas (cifrado homomórfico, Zero-Knowledge Proofs) para proteger la privacidad y respaldar el procesamiento seguro de datos sensibles. A su vez, la construcción de módulos de reputación y mecanismos de auditoría algorítmica fortalece la confiabilidad y la transparencia de los AADs, evitando la dependencia de una autoridad central y promoviendo la participación de la comunidad en la validación de acciones y resultados.

En el plano legal y ético, la autonomía creciente de los AADs invita a redefinir los esquemas de responsabilidad y protección de derechos. Es esencial aclarar los alcances de la "personalidad virtual", la vinculación legal de contratos inteligentes y el tratamiento de datos almacenados en una cadena inmutable. Asimismo, mitigar sesgos en la IA y asegurar la explicabilidad de las decisiones se vuelve crítico para la confianza y la aceptación social de estos sistemas, sobre todo cuando desempeñan funciones sensibles o de alto impacto.

De cara al futuro, se vislumbran múltiples líneas de investigación y desarrollo:

1. **Arquitecturas Multicadena y Aprendizaje Federado**: Profundizar en protocolos que conecten diversas redes orientadas a tareas específicas (por ejemplo, entrenamiento de IA, almacenamiento masivo, ejecución de contratos de gobernanza), para escalar las operaciones y reducir costos.

2. **Modelos de Gobernanza Reputacional**: Diseñar mecanismos de consenso dinámicos basados en indicadores de contribución y fiabilidad, calculados por la IA a partir de historiales de comportamiento de los nodos.

3. **Explicabilidad y Auditoría Algorítmica**: Consolidar metodologías para que los AADs ofrezcan descripciones claras de sus procesos de inferencia y razonamiento, en especial ante decisiones conflictivas, y posibilitar la participación activa de la comunidad en la corrección de sesgos.

4. **Identidad Digital y Protección de Datos**: Integrar soluciones como Self-Sovereign Identity (SSI) y técnicas de borrado selectivo para conciliar la inmutabilidad de la cadena con las demandas normativas y el derecho al olvido.

5. **Interacción con la Economía Real**: Evaluar el rol de los AADs en sectores como logística, finanzas tradicionales, seguros o administración pública, y estudiar cómo se adaptan los procedimientos legales para acoger a estos agentes en la práctica.

No obstante, el desarrollo de Agentes Autónomos Descentralizados enfrenta limitaciones críticas que deben ser abordadas en investigaciones futuras. Entre ellas destacan la validación exhaustiva de su robustez frente a ataques adversarios (Perez et al., 2022; Carlini et al., 2021), el diseño de mecanismos de gobernanza adaptativa capaces de resistir dinámicas de colusión, y la necesidad de protocolos de interoperabilidad multicadena que aseguren su escalabilidad práctica. La consolidación de estos agentes requerirá asimismo marcos regulatorios flexibles que equilibren la innovación técnica con la protección de derechos fundamentales, en línea con

propuestas recientes sobre gobernanza algorítmica responsable (Hassan & De Filippi, 2021).

En síntesis, los AADs combinados con LLMs constituyen un avance significativo en la convergencia de la inteligencia artificial y la tecnología blockchain, al posibilitar la toma de decisiones autónoma bajo mecanismos de gobernanza colectiva y alta auditabilidad. La naturaleza descentralizada de estos entornos, unida a la creciente sofisticación de los agentes cognitivos, plantea retos y oportunidades que invitan a un debate multidisciplinario, donde confluyan la ingeniería, el derecho, la economía y la ética. Conforme se consoliden estas soluciones y se establezcan los marcos regulatorios apropiados, es probable que los AADs desempeñen un papel creciente en la configuración de la próxima generación de sistemas de gestión y organización social.

Referencias

1. Amershi, S., Cakmak, M., Knox, W.B.: Power to the People: The Role of Humans in Interactive Machine Learning. AI Magazine, 35(4), pp. 105–120 (2014) doi: 10.1609/aimag.v35i4.2513.
2. Bonneau, J., Miller, A., Clark, J.: SoK: Research Perspectives and Challenges for Bitcoin and Cryptocurrencies. In: IEEE Symposium on Security and Privacy, pp. 104–121 (2015) doi: 10.1109/SP.2015.14.
3. Brown, T.B., Mann, B., Ryder, N.: Language Models are Few-shot Learners. Advances in Neural Information Processing Systems, 33, pp. 1877–1901 (2020) doi: 10.48550/arXiv.2005.14165.
4. Brundage, M., Avin, S., Clark, J.: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims (2020) doi: 10.48550/arXiv. 2004.07213.
5. Buterin, V.: A Next-generation Smart Contract and Decentralized Application Platform. Ethereum White Paper, 3, pp. 1–36 (2013)
6. Carlini, N., Tramer, F., Wallace, E.: Extracting Training Data from Large Language Models. In: USENIX Security Symposium, pp. 1–19 (2021) doi: 10.48550/arXiv. 2012.07805.
7. Christidis, K., Devetsikiotis, M.: Blockchains and Smart Contracts for the Internet of Things. IEEE Access, 4, pp. 2292–2303 (2016) doi:10.1109/ACCESS. 2016.2566339.
8. ConsenSys Diligence. MythX: Smart Contract Security Analysis Service. https://mythx.io. (2020)
9. DuPont, Q.: Experiments in Algorithmic Governance: A History and Ethnography of 'The DAO,' a Failed Decentralized Autonomous Organization. En M. Campbell-Verduyn (Ed.), Bitcoin and Beyond: Cryptocurrencies, Blockchains, and Global Governance pp. 157–177 (2017) doi: 10.4324/9781315211909-8.
10. Feist, J., Grieco, G., Groce, A.: Slither: A Static Analysis Framework for Smart Contracts. In: Proceedings of the 2nd International Workshop on Emerging Trends in Software Engineering for Blockchain, pp. 8–15 (2019) doi: 10.1109/ WETSEB.2019.00008.
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. pp. 1–15 (2015) doi: 10.48550/arXiv.1412.6572.

12. Hassan, S., De Filippi, P.: Decentralized Autonomous Organizations: Beyond the Hype. Internet Policy Review, 10(2) (2021) doi: 10.14763/2021.2.1556.
13. Kuznetsov, O., Sernani, P., Romeo, L.: On the Integration of Artificial Intelligence and Blockchain Technology: A Perspective About Security. In: IEEE Access, 12, pp. 3881–3897 (2024) doi: 10.1109/ACCESS.2023.3349019.
14. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System. pp. 1–9 https://bitcoin.org/bitcoin.pdf (2008)
15. Perez, F., Ribeiro, I.: Ignore Previous Prompt: Attack Techniques for Language Models. pp. 1–21 (2022) doi: 10.48550/arXiv.2211.09527.
16. Tapscott, D., Tapscott, A.: Blockchain Revolution: How the Technology Behind Bitcoin and Other Cryptocurrencies is Changing the World. Penguin (2016)
17. Wang, S., Ouyang, L., Yuan, Y.: Blockchain-Enabled Smart Contracts: Architecture, Applications, and Future Trends. In: IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(11), pp. 2266–2277 (2019) doi: 10.1109/TSMC.2019.2895123.
18. Weidinger, L., Mellor, J., Rauh, M.: Ethical and Social Risks or Harm from Language Models (2021) doi: 10.48550/arXiv.2112.04359.
19. Wood, G.: Ethereum: A Secure Decentralised Generalised Transaction Ledger. Ethereum Project Yellow Paper, 151, pp. 1–32 (2014) https://ethereum.org/en/whitepaper/.
20. Wright, A., De Filippi, P.: Decentralized Blockchain Technology and the Rise of Lex Cryptographia. SSRN Electronic Journal, 34, pp. 41–52 (2015) doi: 10.2139/ssrn.2580664.
21. Yaga, D., Mell, P., Roby, N.: Blockchain Technology Overview. National Institute of Standards and Technology (2018) doi: 10.6028/NIST.IR.8202.
22. Zhang, Y., Kasahara, S., Shen, Y.: Smart Contract-Based Access Control for the Internet of Things. In: IEEE Internet of Things Journal, 6(2), pp. 1594–1605 (2019) doi: 10.1109/JIOT.2018.2847705.
23. Zyskind, G., Nathan, O., Pentland, A.: Decentralizing Privacy: Using Blockchain to Protect Personal Data. In: IEEE Security and Privacy Workshops, pp. 180–184 (2015) doi: 10.1109/SPW.2015.27.