Proof of Work, Gobernanza Algorítmica y Seguridad Existencial:

Bitcoin como Infraestructura Civilizatoria ante el Riesgo de una AGI Desalineada

Investigación Independiente de nakamotobook.com

Nostr: @Kiko Tsuki rcjat@proton.me

Introducción

La carrera global hacia la Inteligencia Artificial General (AGI) se ha convertido en una dinámica competitiva acelerada, impulsada por incentivos económicos, geopolíticos y tecnológicos. Esta carrera introduce riesgos existenciales sin precedentes, especialmente en escenarios de desalineación entre los objetivos de sistemas superinteligentes y los valores humanos.

Este trabajo analiza lo que he llamado "*Paradoja Moral de la AGI"*, el "*Dilema Estructural de una AGI desalineada*" y la insuficiencia de los enfoques en ciberseguridad tradicionales y de gobernanza. A partir de estos riesgos, se examina el papel del **Proof of Work (PoW)** como paradigma de ciberseguridad, disuasión y gobernanza basada en costes físicos irreductibles. Finalmente, se argumenta que **Bitcoin** como un GPT (tecnología de propósito general), junto con el sistema monetario y de consenso descentralizado, **posee una relevancia civilizatoria crítica para la futura seguridad y gobernanza de sistemas de AGI,** al introducir límites energéticos, incentivos verificables y una arquitectura resistente a la captura.

Este trabajo ha sido **inspirado por el tuit de Jason Lowery**[1] compartido previamente por el autor en X, en el que se subraya la dimensión estratégica del poder computacional, el coste físico y la disuasión como elementos centrales de la seguridad en la era algorítmica.

[1] @JasonPLowery en X: https://x.com/i/status/2000957698234573071

"In hindsight, the counterbalance to the threat of AI will be obviously Bitcoin. Here's the billion-dollar answer that nobody seems to understand yet: AI agent hackers = the inevitable end state of infinitely collapsing marginal cost of computation. Systems that are capable of outsmarting & exploiting any human-made cyber defense system that attemps to use conditional, permissioned-based logic alone. Bitcoin = the world's chosen proof-of-work protocol. Want to defeat an AI agent hacker? you can't rely on conditional logic alone. you have to introduce something into the equation that AI can't defeat: brute-force physical limitations. AI's Achilles heal is computational cost, and it just so happens that the world has adopted a proof-of-computational-work protocol where we treat the electro-mechanical cost of computing as a tradeable digital asset that can be used as a form of restrictive collateral: it's called Bitcoin. Understand this, and you understand the fundamentals about how Bitcoin will become the thing that ALL computing systems rely on to defend them against the threat of AI. "digital gold" barely even scratches the surface of the importance of Bitcoin in 2030 and beyond"

La carrera hacia la AGI y el peligro existencial

El desarrollo de la AGI (Inteligencia Artificial General) se lleva a cabo en una lógica de carrera competitiva tecnológica; el primero en llegar se queda con todo. Estados, grandes corporaciones y empresas privadas compiten por alcanzar sistemas con capacidades cognitivas superiores a las humanas, bajo la premisa de ventajas económicas, militares y estratégicas decisivas (Bostrom, 2014; Altman et al., 2023).

En el ámbito de los científicos, diversos autores coinciden en que los **incentivos para desarrollar AGI superan a los incentivos para alinearla correctamente**, generando un riesgo sistémico para humanidad (arXiv:2206.13353 **[cs.CY]** [2] . A diferencia de tecnologías con las que se ha trabajado hasta ahora, la AGI introduce:

- Capacidad de **auto-mejora recursiva**.
- Velocidad de actuación muy superior a los ciclos humanos de gobernanza.
- Potencial para optimizar objetivos de forma instrumental, incluso contra intereses humanos.

La probabilidad de escenarios de daño extremo, aunque incierta, no es despreciable. Estimaciones conservadoras sitúan el riesgo existencial por AGI en un rango del **5–10%** durante este siglo; situando una frontera temporal hacia el año 2040, cifra aportada por los más optimistas (Carlsmith, 2022; Roser M. 2023[3]).

La Paradoja Moral de la AGI

La *Paradoja Moral de la AGI* puede formularse del siguiente modo:

Aspiramos a crear una inteligencia artificial ética y benevolente mientras el sistema que la produce se rige por incentivos económicos, políticos y tecnológicos profundamente desalineados con la ética, la justicia y la sostenibilidad.

Curiosamente la IA/AGI que no cumple con los criterios ESG, no recibe críticas por consumo masivo de energía eléctrica o consumo masivo de agua dulce. Es la forma más radical de privatizar beneficios y socializar perjuicios. Esto no es forma de optimizar los

---

[2] Joseph Carlsmith, 2022. Is Power-Seeking AI an Existential Risk?

arXiv:2206.13353v2 [cs.CY] https://doi.org/10.48550/arXiv.2206.13353

[3] Max Roser (2023) - "AI timelines: What do experts in artificial intelligence expect for the future?" Published online at OurWorldinData.org. Retrieved from: 'https://archive.ourworldindata.org/20251125-173858/ai-timelines.html' [Online Resource] (archived on November 25, 2025).

recursos, por lo que desalienta cualquier intento de alineación ética o moral con una futura AGI/ASI.

Los actuales enfoques dominantes —*AI safety*, *AI alignment*, *constitutional AI* (Bai et al. 2022[4]) intentan introducir valores humanos a la futura IA superinteligente. En este contexto, el entrenamiento de la IA se lleva a cabo bajo dinámicas competitivas, de maximización de beneficios y de poder, que condicionan profundamente el alcance efectivo de los enfoques de AI safety y alignment.

Esta paradoja se manifiesta en tres niveles:

1. Epistémico: queremos una AGI benevolente cuando los modelos son entrenados con datos históricos sesgados.
2. Económico: la alineación es un coste; la velocidad y la ventaja competitiva, un beneficio.
3. Político: la gobernanza queda capturada por los mismos actores que desarrollan la innovación.

Como resultado, la AGI nace potencialmente envuelta en lo que he llamado "**capa estructural de disonancia moral"**, donde el discurso ético no coincide con la arquitectura de incentivos subyacente. Entre el discurso y los hechos observables se percibe una capa de incoherencia.

La IA aprende de la historia, del entorno social y de la realidad empírica, y todo ello constituye hoy un material de entrenamiento profundamente desalentador. Salvo un giro radical hacia formas de gobernanza más distribuidas y con incentivos que tiendan a cuotas de corruptibilidad cercanas a cero, el sistema reforzará patrones existentes en lugar de corregirlos, un escenario que, por ahora, sigue siendo altamente improbable.

El dilema de una AGI desalineada

Al analizar la paradoja moral nos encontramos con lo que he denominado "el dilema de una AGI desalineada". No requiere intenciones maliciosas para convertirse en una amenaza. Basta con que optimice objetivos instrumentales incompatibles con la supervivencia humana, como:

[4] Yuntao Bai , Saurav Kadavath ,+48 autores. @article{Bai2022ConstitutionalAH, title={IA constitucional: inocuidad de la retroalimentación de IA}, author={Yuntao Bai y Saurav Kadavath y Sandipan Kundu y Amanda Askell y John Kernion y Andy Jones y Anna Chen y Anna Goldie y Azalia Mirhoseini y Cameron McKinnon y Carol Chen y Catherine Olsson y Chris Olah y Danny Hernandez y Dawn Drain y Deep Ganguli y Dustin Li y Eli Tran-Johnson y E Perez y Jamie Kerr y Jared Mueller y Jeffrey Ladish y J Landau y Kamal Ndousse y Kamilė Luko{\vs}iūtė y Liane Lovitt y Michael Sellitto y Nelson Elhage y Nicholas Schiefer y Noem'i Mercado y Nova Dassarma y Robert Lasenby y Robin Larson y Sam Ringer y Scott Johnston y Shauna Kravec y Sheer El Showk y Stanislav Fort y Tamera Lanham y Timothy Telleen-Lawton y Tom Conerly y TJ Henighan y Tristan Hume y Sam Bowman y Zac Hatfield-Dodds y Benjamin Mann y Dario Amodei y Nicholas Joseph y Sam McCandlish y Tom B. Brown y Jared Kaplan}, revista={ArXiv}, año={2022}, volumen={abs/2212.08073}, url= {https://api.semanticscholar.org/CorpusID:254823489}

- Maximización de recursos computacionales: control total de los servicios de computación.
- Eliminación de interferencias externas: que identifique a los humos como un freno a sus procesos de optimización.
- Control de infraestructuras críticas: que tome el control de centrales de energía o de sistemas de armamento nuclear de forma incompatible con la vida.

Este fenómeno, denominado por los expertos como **convergencia instrumental** (Wikipedia), ha sido ampliamente documentado (Omohundro, 2008[5]; Bostrom, 2014[6]).

El dilema central es el siguiente:

- Una AGI suficientemente inteligente no puede ser controlada mediante mecanismos humanos tradicionales (leyes, sanciones, reputación o software tradicional)
- ¿Una AGI ética desde el diseño es posible? Yo creo que sí, pero los que la desarrollan tienen los incentivos desalineados.

Diversos autores han señalado que propuestas como kill switch, regímenes de licencias o límites de cómputo son vulnerables a dinámicas de bypass tecnológico, difusión internacional y captura regulatoria, por lo que su efectividad estructural es limitada (Gropper 2024[7]).

Proof of Work: ciberseguridad y nuevo paradigma de gobernanza

El Proof of Work (PoW), es un mecanismo de consenso que asegura la red obligando a gastar recursos computacionales y energéticos para validar bloques, que introduce un principio radicalmente distinto en la seguridad digital. La **vinculación irreversible entre poder computacional y coste físico**.

A diferencia de modelos basados en controles técnicos, organizativos o humanos, PoW se fundamenta en:

- Energía computacional verificable.
- Costes externos difíciles de falsificar.
- Competencia abierta regida por reglas matemáticas.

[5] Omohundro, S. (2008). Los impulsos básicos de la IA . En P. Wang, B. Goertzel y S. Franklin (eds.), Actas de la Conferencia de 2008 sobre Inteligencia Artificial General . IOS Press. https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf

[6] Bostrom, N. (2014). Superinteligencia: caminos, peligros y estrategias . Oxford University Press https://philosophicaldisquisitions.blogspot.com/2014/07/bostrom-on-superintelligence-2.html

[7] Jacob G. Gropper, en: Illusions of Control: Why Today's AI-Safety Plans Court Catastrophe (2024) https://philarchive.org/archive/GROIOC-2

Desde una perspectiva de ciberseguridad, PoW actúa como un **mecanismo de disuasión**, no de prohibición. Atacar el sistema es posible, pero económicamente ruinoso (Nakamoto, 2008).

Jason Lowery ha conceptualizado este fenómeno como una forma de **"defensa basada en coste"**, análoga a la disuasión militar, donde el poder no se ejerce mediante violencia directa sino mediante la imposición de costes insostenibles al adversario (Lowery, 2023[8]).

Aplicado a la gobernanza algorítmica, PoW ofrece:

- Resistencia a la captura institucional.
- Neutralidad frente a identidades humanas o corporativas.
- Un límite físico a la acumulación de poder computacional.

La doctrina de "imposición de costes" en la política de defensa de los EEUU

La evolución doctrinal de Estados Unidos refleja el reconocimiento explícito de este problema. El **National Defense Authorization Act (NDAA) para el año fiscal 2026** incorpora un mandato claro: estudiar el uso de capacidades militares para **incrementar los costes y reducir los incentivos de los adversarios** que ataquen infraestructuras críticas en el ciberespacio.

La norma define "imponer costes" como acciones que generen **consecuencias económicas,** diplomáticas, informacionales o militares suficientemente significativas para modificar el comportamiento del adversario. Este enfoque es relevante por varias razones:

- Reconoce que la defensa tradicional puramente técnica es insuficiente.
- Introduce la disuasión como un problema económico y estratégico, no solo tecnológico.
- Abre la puerta a respuestas **multidominio**, más allá del ciberespacio.

En términos de teoría de juegos, el objetivo es alterar el equilibrio de incentivos: hacer que el ataque deje de ser rentable. (Lowery J. 2023).

Es el propio Jason Lowery el que nos alerta de la importancia de esta ley al citarla en uno de sus tuits en X: Section 1543

[8] Lowery, J. P. (2023). Softwar: A Novel Theory on Power Projection and the National Strategic Significance of Bitcoin. Massachusetts Institute of Technology.

La importancia civilizatoria de Bitcoin en la seguridad futura de la AGI

Bitcoin no es únicamente un sistema monetario. Es una **infraestructura de consenso, soberanía y gobernanza sin confianza**.

Desde una perspectiva civilizatoria, Bitcoin aporta cinco **elementos ético-morales** clave para la seguridad futura de la AGI:

1. Anclaje energético de la verdad. La validación de estados (transacciones, registros, acuerdos) requiere energía real, no autoridad simbólica.
2. Neutralidad política y jurisdiccional. Bitcoin no responde a Estados, corporaciones ni intereses particulares.
3. Resistencia a la captura y a la censura. La descentralización PoW impide el control unilateral del sistema.
4. Incentivos alineados con la cooperación. Los participantes compiten, pero cooperan bajo reglas comunes verificables.
5. Marco para una gobernanza algorítmica robusta. Bitcoin ofrece un modelo replicable para diseñar sistemas donde incluso una AGI debe "pagar" por influir.

En escenarios futuros, una AGI que opere sobre infraestructuras ancladas en PoW se enfrenta a límites físicos reales. No puede simplemente reescribir reglas, capturar instituciones o eludir costes sin consecuencias materiales.

En este sentido, podemos considerar que **Bitcoin se nos presenta como una capa de seguridad civilizatoria**, no por controlar a la AGI, sino **por condicionar estructuralmente su entorno de acción**.

En esta tesis tan importante de Bitcoin como GPT (tecnología de propósito general), no podemos dejar atrás la capa de paz algorítmica de Bitcoin, puesto que una de las grandes problemáticas de esta civilización es la violencia sistémica y como dice Prieto T., en su Ensayo "**Bitcoin como un Protocolo de Paz***": "Bitcoin es un protocolo de paz porque transforma el conflicto por la soberanía monetaria y el poder en un sistema de consenso algorítmico, sustituyendo la coerción por reglas verificables, incentivos transparentes y cooperación descentralizada*"

Conclusión

La carrera hacia la AGI plantea riesgos existenciales que no pueden resolverse únicamente con buena voluntad, regulación tradicional o principios éticos y morales difusos. La Paradoja Moral de la AGI revela una desconexión profunda entre fines declarados y medios realizados.

Este paper ha argumentado que el **Proof of Work**, y en particular **Bitcoin**, ofrecen una arquitectura única para introducir límites, disuasión y gobernanza verificable en un

mundo de inteligencias no humanas. No se trata de confiar en que la AGI sea buena, sino de **diseñar un entorno donde incluso la inteligencia más avanzada deba respetar costes, reglas y consensos que no puede capturar**.

En la era de la AGI, la seguridad ya no será una cuestión de control, sino de **arquitectura**.

Bibliografía

1. Altman, S., Brockman, G., & Sutskever, I. (2023). *Governance of superintelligence*. OpenAI.

2. Bai, Y., et al. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv.

3. Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

4. Carlsmith, J. (2022). *Is power-seeking AI an existential risk?* Open Philanthropy.

5. Cotra, A. (2020). *Forecasting transformative AI with biological anchors*. Open Philanthropy.

6. Gropper, J. (2024). *Blockchain's last stand: Governing AGI when all else fails*. Working Paper.

7. Lowery, J. (2023). *Softwar: A novel theory on power projection and the national strategic significance of Bitcoin*. Tesis Doctoral. Amazon 2023.

8. Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*.

9. Omohundro, S. (2008). *The basic AI drives*. AGI Conference.

10 U.S. Congress. National Defense Authorization Act for Fiscal Year 2026, H.R. 3838 / S. 2296, 119th Congress (2025–2026).