LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, Asunción, Paraguay.
ISSN en línea: 2789-3855, 2025, Volumen VI

Principios para gobernar la Inteligencia General Artificial (AGI): Mitigar los riesgos y garantizar el desarrollo responsable

Principles for governing Artificial General Intelligence (AGI): Mitigating risks and ensuring responsible development

Pablo Corona Fraga
pablo.coronaf@infotec.mx
https://orcid.org/0000-0001-5012-9468
Centro de Investigación e Innovación en Tecnologías de la Información y la Comunicación INFOTEC
Ciudad de México – México

Vanessa Díaz
vdiazr@scjn.gob.mx
https://orcid.org/0000-0002-7186-8785
Oficial de Investigación Jurídica de la Suprema Corte de Justicia de la Nación
Ciudad de México – México

DOI: https://doi.org/10.56712/latam.v6i5.4811

Artículo recibido: 15 de julio de 2025
Aceptado para publicación: 14 de noviembre de 2025.
Conflictos de Interés: Ninguno que declarar.

Redilat
Red de Investigadores Latinoamericanos

NÚMERO 5

DOI: https://doi.org/10.56712/latam.v6i5.4811

Principios para gobernar la Inteligencia General Artificial (AGI): Mitigar los riesgos y garantizar el desarrollo responsable

Principles for governing Artificial General Intelligence (AGI): Mitigating risks and ensuring responsible development

Pablo Corona Fraga[1]
pablo.coronaf@infotec.mx
https://orcid.org/0000-0001-5012-9468
Centro de Investigación e Innovación en Tecnologías de la Información y la Comunicación INFOTEC
Ciudad de México – México

Vanessa Díaz
vdiazr@scjn.gob.mx
https://orcid.org/0000-0002-7186-8785
Oficial de Investigación Jurídica de la Suprema Corte de Justicia de la Nación
Ciudad de México – México

Artículo recibido: 15 de julio de 2025. Aceptado para publicación: 14 de noviembre de 2025.
Conflictos de Interés: Ninguno que declarar.

Resumen

El rápido desarrollo de modelos fundacionales y su creciente autonomía demandan una gobernanza capaz de anticipar y mitigar los riesgos de una posible Inteligencia Artificial General (AGI). Este estudio propone un marco de principios e indicadores de gobernanza para alinear el desarrollo de sistemas avanzados con valores sociales y marcos internacionales. La metodología combina tres enfoques: (i) revisión sistemática de literatura para identificar brechas y buenas prácticas en gobernanza de IA; (ii) triangulación normativa de principios y controles de marcos internacionales, como los de la OCDE; y (iii) síntesis empírica de métricas de evaluación aplicadas a modelos fundacionales. Los resultados ofrecen un marco basado en siete principios: transparencia, responsabilidad, beneficencia/no maleficencia, equidad, control humano significativo, robustez/seguridad y colaboración multiactor. Asimismo, se proponen indicadores verificables –como trazabilidad de decisiones, cobertura de pruebas, degradación bajo estrés, monitoreo post-despliegue, y adopción de estándares-- que fortalecen la rendición de cuentas y reducen el resigo residual. La evidencia comparada muestra que los procesos de evaluación continua y reportes de transparencia mejoran la detectabilidad de fallos y la seguridad del sistema. Finalmente, el modelo ofrece una base metodológica para reguladores y organizaciones mapeen riesgos técnicos y organizacionales con obligaciones de auditoría y gestión. En conjunto, el marco propuesto facilita la transición de los principios a la práctica, promoviendo una gobernanza medible, trazable y efectiva para el desarrollo seguro y alineado de la IA avanzada.

Palabras clave: inteligencia artificial, gobernanza, ética, riesgo

Abstract

The rapid development of foundational models and their growing autonomy demand governance

[1] Autor de correspondencia.

capable of anticipating and mitigating the risks of a possible Artificial General Intelligence (AGI). This study proposes a framework of governance principles and indicators to align the development of advanced systems with social values and international frameworks. The methodology combines three approaches: (i) a systematic literature review to identify gaps and best practices in AI governance; (ii) a normative triangulation of principles and controls from international frameworks, such as those of the OECD; and (iii) an empirical synthesis of evaluation metrics applied to foundational models. The results offer a framework based on seven principles: transparency, accountability, beneficence/non-maleficence, fairness, meaningful human control, robustness/safety, and multi-stakeholder collaboration. Verifiable indicators are also proposed—such as decision traceability, test coverage, degradation under stress, post-deployment monitoring, and adoption of standards—that strengthen accountability and reduce residual risk. Comparative evidence shows that continuous evaluation processes and transparency reporting improve fault detectability and system safety. Finally, the model provides a methodological basis for regulators and organizations to map technical and organizational risks to audit and management obligations. Taken together, the proposed framework facilitates the transition from principles to practice, promoting measurable, traceable, and effective governance for the safe and aligned development of advanced AI.

Keywords: artificial intelligence, governance, ethics, risk

Cómo citar: Corona Fraga, P., & Díaz, V. (2025). Principios para gobernar la Inteligencia General Artificial (AGI): Mitigar los riesgos y garantizar el desarrollo responsable: Principles for governing Artificial General Intelligence (AGI). *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades 6 (5), 3249 – 3268. https://doi.org/10.56712/latam.v6i5.4811*

INTRODUCCIÓN

La inteligencia artificial general (AGI, por sus siglas en inglés) se refiere a un sistema hipotético que poseería la capacidad de comprender, aprender y aplicar conocimientos en una amplia variedad de dominios cognitivos al nivel de un ser humano promedio. Estos sistemas serían capaces de transferir habilidades y razonamiento entre tareas distintas sin necesidad de reprogramación específica, adaptándose a entornos nuevos y desconocidos con una autonomía sustancial. En contraste con las IA especializadas (IA débil), que funcionan dentro de límites predefinidos, la AGI aspira a exhibir amplitud y flexibilidad cognitivas comparables a las humanas, operando como entidades integradas de inteligencia generalizada.

A partir de una revisión de la literatura existente se identificaron y agruparon las brechas actuales de la AGI. Los rápidos avances en la investigación y el desarrollo de AGI han hecho imperativo abordar sus riesgos inherentes, que podrían colocarnos en escenarios con consecuencias imprevistas e indeseables. De ahí que, el objetivo de este artículo sea, por un lado, analizar los riesgos multifacéticos asociados con AGI y aplicar principios ampliamente reconocidos como transparencia, responsabilidad, beneficencia, no maleficencia, equidad y control. Además, de introducir nuevos principios (robustez y seguridad) junto con desafíos y posibles soluciones; y, por el otro, proponer un conjunto coherente de principios para gobernar la AGI, mitigando estos riesgos y previniendo comportamientos o resultados fuera de control.

Los principios propuestos sirven como base para desarrollar un marco de gobernanza AGI más responsable y efectivo que busque maximizar sus beneficios al tiempo que minimiza los riesgos que amenazan la vida y el potencial de resultados no controlados. Este estudio proporciona información sobre (a) un conjunto de principios de gobernanza que pueden mitigar eficazmente el riesgo potencialmente mortal que representa la AGI; (b) prácticas actuales de gobernanza de IA con nuevas acciones específicas de AGI para un marco sólido capaz de prevenir comportamientos AGI fuera de control; y c) la posibilidad de que la comunidad internacional pueda dirigir el desarrollo y la implantación de la AGI de manera que se ajuste a los valores humanos y promueva el bienestar mundial. Además, el estudio enfatiza el llamado a la acción colectiva de la academia, la industria y los gobiernos para adoptar estos principios, fomentando la integración segura de AGI en la sociedad. También destaca áreas para futuras investigaciones para garantizar que la gobernanza de AGI evolucione al ritmo de la naturaleza dinámica de las tecnologías de AGI.

Estos principios rectores deben considerarse como parte del diseño de cualquier sistema de inteligencia artificial, incluidos también los principios de ciberseguridad. Los modelos basados en IA pueden mejorar la capacidad de predecir, detectar y responder a las amenazas cibernéticas. Por otro lado, el propio desarrollo de AGI introduce nuevas superficies de ataque y riesgos, lo que requiere una adhesión aún mayor a los pilares de ciberseguridad: confidencialidad, integridad y disponibilidad.

La integridad es particularmente crucial cuando se considera la gobernanza de AGI. Pues la AGI, como agente inteligente capaz de aprender, tomar decisiones y evolucionar de forma independiente, puede encontrar escenarios en los que opera de manera impredecible o se desalinea con los objetivos humanos previstos. Esta desalineación puede enmarcarse como una pérdida de integridad en el sistema.

Considerando esta necesidad de establecer principios para que la AGI mitigue sus riesgos y garantice su desarrollo responsable, este documento cubre: los trabajos sobre la necesidad de un marco de gobernanza de la AGI; una visión general de las iniciativas y tratados mundiales para regular la AGI; una comprensión de los riesgos y desafíos planteados por el aumento de AGI; los principios para gobernar la AGI; y su implementación de los principios con prácticas.

Contexto y justificación

Según la sociedad moderna, está experimentando transformaciones sustanciales en el alcance, la velocidad y la naturaleza de los cambios en los ámbitos social, natural y tecnológico. Esto destaca la necesidad de establecer, crear e implementar estrategias de gobernanza complementarias para abordar estos desafíos y trabajar hacia futuros más optimistas. Liu y Maas (2021) [18] proponen que estas estrategias deben considerar perspectivas a largo plazo mientras se aplican políticas prácticas y viables en el presente.

La Inteligencia General Artificial (AGI) representa una de las fronteras más viables de la inteligencia artificial (IA) en nuestro futuro más cercano; en el que las máquinas pueden realizar cualquier tarea intelectual que un ser humano pueda realizar. Esto se debe a la creciente sofisticación de los algoritmos de aprendizaje automático y al crecimiento exponencial de la capacidad computacional. De ahí nuestro interés en explorar las implicaciones éticas, sociales y existenciales de la AGI.

Junto con su inmenso potencial, AGI plantea riesgos significativos. La perspectiva de máquinas con inteligencia sobrehumana plantea serias preocupaciones sobre el control y la seguridad. En un escenario desfavorable, AGI podría actuar de manera perjudicial para la humanidad, ya sea por su falta de alineación con los valores humanos, su incapacidad para comprender decisiones éticas complejas o las consecuencias no deseadas de sus acciones. Además, existe el temor de que una vez que AGI supere la inteligencia humana, será imposible de controlar, lo que representa un riesgo existencial para la humanidad.

Reconociendo estos riesgos potenciales, es importante establecer un marco de gobernanza para AGI, asegurando su implementación segura y beneficiosa. El objetivo de este trabajo de investigación es proponer un conjunto de principios para gobernar el AGI, diseñados para mitigar los riesgos potencialmente mortales y prevenir comportamientos fuera de control. Creemos que con estos principios podremos guiar el desarrollo y el uso de AGI de una manera que se ajuste a los valores humanos y promueva el bienestar de toda la humanidad.

Kurzweil (2005) advierte que la singularidad tecnológica —el punto en el que la inteligencia de las máquinas superará la humana y posibilitará una fusión hombre-máquina— podría transformar irreversiblemente la existencia humana.

En 2023, un grupo de expertos pidió una pausa de seis meses en la formación de Grandes Modelos de Lenguaje basados en inteligencia artificial, para centrarse en sentar las bases de una regulación adecuada. Esta pausa no ocurrió, y el avance de estas tecnologías continúa, mientras que las reglas para gobernar un AGI no están claras.

METODOLOGÍA

La metodología contempla tres fases complementarias: (i) una revisión sistemática con mapeo de pruebas existentes para detectar vacíos y buenas prácticas en gobernanza de IA; (ii) una triangulación normativa que entrelaza principios y mecanismos extraídos de los marcos de gestión de riesgo de IA contemporáneos con los Principios de la OCDE en IA; y (iii) una síntesis empírica que extrae métricas y dominios de evaluación a partir de comparaciones entre modelos fundacionales, con el propósito de fundamentar indicadores cuantificables de transparencia, explicabilidad, robustez, seguridad y rendición de cuentas.

Tomando en cuenta las consideraciones éticas, medidas regulatorias y salvaguardas tecnológicas, este estudio busca proponer un enfoque holístico para gobernar la AGI, contribuyendo en última instancia a la integración responsable y beneficiosa de la AGI en la sociedad. Como se mencionó en la introducción, la pregunta central de investigación que guía este estudio es: ¿Cuáles son los

componentes y estrategias clave necesarios para desarrollar un marco integral para gobernar la inteligencia artificial general (AGI) para garantizar su implementación segura y beneficiosa?

Esta pregunta tiene como objetivo explorar e identificar los mecanismos, principios y acciones de gobernanza esenciales necesarios para abordar los desafíos planteados por AGI. En concreto, pretende responder a las siguientes sub-preguntas:

- ¿Qué principios de gobernanza con salvaguardias se pueden establecer para reducir eficazmente las posibilidades de riesgos potencialmente mortales debido a la AGI?
- ¿Qué estrategias deben entregarse para evitar comportamientos incontrolados o dañinos de AGI para que se mantenga dentro de la supervisión humana y actúe bajo restricciones éticas?
- ¿Cómo y en qué medida los modelos actuales de gobernanza de la IA y las prácticas regulatorias deberían adaptarse e integrarse en un marco para la AGI, teniendo en cuenta su alcance más amplio y su posible autonomía?
- ¿Hasta qué punto es necesario diseñar acciones y estructuras de gobernanza AGI completamente nuevas, considerando sus capacidades y riesgos sin precedentes?
- ¿Qué mecanismos pueden garantizar la alineación de AGI con los valores humanos hacia la mejora del bienestar global, reflejando consideraciones éticas como la equidad, la responsabilidad, la transparencia y la inclusión?

Esto asegurará que surjan áreas específicas de enfoque de la pregunta general para guiar el trabajo de investigación. El estudio empleará un método cualitativo que incluyó revisiones de literatura y estudios de casos de las prácticas de gobernanza de IA existentes para responder de manera integral a estas preguntas.

Para abordar el objetivo principal y probar la hipótesis para el avance seguro y ético de la AGI, se necesita un enfoque sistemático. A partir de la pregunta principal de investigación, el estudio se centrará en sí un conjunto definido de principios de gobernanza puede minimizar los riesgos que amenazan a las personas. En primer lugar, una revisión comparativa de la literatura, con escenarios en los que diferentes autores han desarrollado principios de gobernanza, probará la efectividad de las estrategias de mitigación propuestas por los expertos. Las pruebas identificarán los principios de gobernanza más eficaces para reducir los riesgos en sectores críticos, contribuyendo a un marco de gobernanza sólido que garantice medidas de seguridad para las aplicaciones AGI de alto riesgo.

Posteriormente, el estudio postula que la integración de las prácticas actuales de gobernanza de IA con acciones específicas de AGI evitará efectivamente el comportamiento incontrolado. Esto implica identificar los modelos de IA existentes de organizaciones internacionales, como la Unión Europea, la OCDE, el IEE, y demostrar cómo se pueden adaptar para AGI. Probar esto revelará si la combinación de estrategias de gobernanza existentes con nuevas acciones mejora los mecanismos de control, evitando consecuencias o daños no deseados.

Por lo tanto, la hipótesis propuesta en este estudio aborda los componentes clave de la pregunta de investigación. Al probarlo, la investigación explora la efectividad de los principios de gobernanza, la integración de los modelos actuales de gobernanza de IA y el potencial de colaboración internacional para garantizar la implementación segura y beneficiosa de AGI. Los resultados de esto proporcionarán información concreta sobre las estrategias necesarias para crear un marco de gobernanza que aborde los desafíos únicos planteados por AGI.

DESARROLLO

Revisando la literatura disponible sobre el marco de gobernanza global de IA a través de un examen multidimensional de los desafíos, así como posibles soluciones relacionadas con los principios de

gobernanza y seguridad de AGI. Estos incluyen marcos globales de gobernanza sostenible, como los Objetivos de Desarrollo Sostenible, los Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos y las directrices de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) para empresas multinacionales. Maher y Buhmann (2019) enfatizan la necesidad de ampliar las iniciativas de abajo hacia arriba lideradas por los grupos afectados con inversores institucionales y empresas.

Baum (2020) considera la formación de regímenes de gobernanza global como una posible solución a los problemas relacionados con la IA, pero reconoce desafíos, como el abuso potencial y la improbabilidad de que las naciones cedan soberanía. Además, analiza el papel de la propiedad privada en la gobernanza de la IA y la autoorganización de la comunidad, al tiempo que destaca las limitaciones debidas a las externalidades.

Witt et al. (2023) Proponer un enfoque novedoso para convertir la legislación en código legible por máquina para mejorar la coherencia mediante la validación manual y automatizada. Stahl (2023) explora la evaluación del impacto de la IA en los debates éticos y sociales, proponiendo un proceso fundamental para su implementación.

Además, Baum (2021) destaca el conflicto entre los intereses financieros de las empresas con fines de lucro y el desarrollo ético de la IA. Greenstein (2021) y Ghaffarian (2023) se centran en la toma de decisiones legales de IA, enfatizando la transparencia, la equidad y la confianza. Llamas Covarrubias (2022) aboga por un marco regulatorio integral que incorpore normas vinculantes y no vinculantes. Liu y Maas (2021) promueven un enfoque de búsqueda de problemas para las estrategias de gobernanza para abordar temas complejos con mayor profundidad.

(Faroldi (2024) argumenta que las definiciones existentes de riesgo son inadecuadas para abordar las posibles amenazas existenciales que plantea el AGI. Además, sostiene que el enfoque basado en el riesgo de la Ley de IA de la UE es insuficiente para gestionar los desafíos únicos asociados con los sistemas AGI. Introducen un cambio conceptual al argumentar que los sistemas AGI deben tratarse como agentes autónomos en lugar de productos tecnológicos tradicionales. Esta distinción subraya la necesidad de marcos regulatorios que vayan más allá de la gestión de productos con fines previstos y, en cambio, se centren en alinear y gobernar el comportamiento de los sistemas agenciales.

La literatura existente sobre los marcos globales de gobernanza de la IA presenta un panorama complejo de desafíos y oportunidades en la gobernanza y la seguridad de la AGI. Pero también, la literatura expone la necesidad de más investigación que profundice en cuestiones como las salvaguardias técnicas en la IA, los diferentes modelos de gobernanza de la IA, su eficacia y las implicaciones sociales, legales y éticas de la AGI.

Con lo anterior concluimos que es posible que la gobernanza efectiva de AGI requiere un enfoque multifacético que integre marcos globales, gobernanza participativa y modelos regulatorios adaptativos. La literatura existente sugiere que la gobernanza de AGI debe equilibrar los intereses de varias partes interesadas, incluidos los gobiernos, las corporaciones y la sociedad civil, al tiempo que garantiza que se prioricen principios éticos como la transparencia, la equidad y la rendición de cuentas. Además, la necesidad de cooperación global es evidente, aunque persisten los desafíos en la alineación y la soberanía transfronterizas. Al abordar estas brechas, en particular el conflicto entre los intereses financieros y el desarrollo ético de la IA, es posible mitigar los riesgos asociados con AGI y prevenir el comportamiento fuera de control.

Hipótesis – problema de investigación

La creación de un marco unificado a nivel mundial que combine principios de gobernanza bien definidos, prácticas actuales de gobernanza de IA y regulaciones específicas de AGI guiará eficazmente el desarrollo de AGI en alineación con los valores humanos, mitigando los riesgos existenciales y promoviendo la cooperación internacional para el bienestar global.

Objetivos y preguntas de investigación

El objetivo principal de este trabajo de investigación es proponer un marco integral para gobernar AGI para garantizar su implementación segura y beneficiosa. Este marco tiene como objetivo responder a las siguientes preguntas:

- ¿Cómo mitigar los riesgos potencialmente mortales? Desarrollar principios que aborden y reduzcan los peligros potenciales asociados con la AGI.
- ¿Cómo prevenir comportamientos fuera de control? Asegurar la existencia de pautas para prevenir cualquier instancia de comportamiento dañino de AGI.
- ¿Cómo aprovechar las prácticas actuales? Integrar las prácticas existentes y los principios rectores de los sistemas de IA dentro del marco sugerido.
- ¿Cómo introducir nuevas acciones? Proponer medidas y estrategias diseñadas para el manejo de la AGI.
- ¿Cómo promover los valores humanos y el bienestar? Asegurar que el avance y la aplicación de AGI estén en armonía con los valores y beneficien el bienestar de la humanidad.

Panorama general de las iniciativas y tratados mundiales para regular la AGI

La regulación de la IA en varios países y regiones se ha convertido en un punto focal de la gobernanza contemporánea, como lo destaca AILAB (2023). El Reglamento del Parlamento Europeo y del Consejo, comúnmente conocido como la ley de IA, ejemplifica un enfoque integral de la gobernanza de la IA, empleando un marco basado en el riesgo para garantizar que las medidas regulatorias sean proporcionales al nivel de riesgo que plantean las diferentes aplicaciones de IA.

Esta normativa prohíbe las prácticas que plantean riesgos significativos, como los sistemas capaces de alterar drásticamente el comportamiento de los individuos, que pueden causar daños físicos o psicológicos. Además, también se cubren los sistemas que aprovechan las vulnerabilidades de grupos específicos, como los ancianos o las personas con discapacidades físicas o mentales.

Además, restringe a las autoridades públicas el uso de IA para evaluar o categorizar a las personas en función de su comportamiento social o cualquier rasgo personal conocido o anticipado, lo que en la regulación de la Unión Europea está prohibido. El uso de sistemas de identificación biométrica remota en tiempo real en áreas de acceso público para la aplicación de la ley también es limitado, con algunas excepciones.

Las directrices de la OCDE también han sido influyentes, con países como Chile, Argentina, China, Brasil, Gran Bretaña, Irlanda y estados de Estados Unidos como Nueva York y California han adoptado y adaptado estas recomendaciones.

En el sector del empleo, los sistemas de IA de alto riesgo, incluidos los utilizados para la contratación, la promoción y la gestión de las relaciones laborales, así como los sistemas que evalúan la elegibilidad para la asistencia, los beneficios y los servicios públicos, están sujetos a estrictos requisitos de registro y supervisión. Esto incluye los sistemas diseñados para la identificación biométrica remota en tiempo real o en diferido, los utilizados en la gestión y operación de infraestructuras críticas como el tráfico y los servicios públicos, y los sistemas del sector educativo para determinar el acceso a las escuelas o evaluar a los estudiantes; todos estos deben registrarse ante la autoridad competente.

Este panorama regulatorio subraya el esfuerzo global para armonizar la gobernanza de la IA, asegurando que el desarrollo y la implementación de la IA se lleven a cabo de manera ética y segura, mitigando el daño potencial y fomentando la innovación

RESULTADOS

Los resultados de esta investigación sientan las bases para un marco de gobernanza integral para AGI. Identificando estos hallazgos en relación con los objetivos de la investigación, como mitigar los riesgos que amenazan la vida, prevenir comportamientos fuera de control, integrar las prácticas actuales, introducir nuevas acciones para la gobernanza de AGI y promover la alineación con los valores humanos y el bienestar.

El marco propuesto tiene, entre otras cosas, la reducción de los riesgos potencialmente mortales en el despliegue de AGI como un objetivo importante. Los principios de gobernanza bien definidos, como la transparencia, la rendición de cuentas y la solidez, identificados en las investigaciones, podrían contribuir en gran medida a reducir esos riesgos. De hecho, se encontró que en los sistemas AGI con líneas de responsabilidad bien definidas, procedimientos seguros de prueba y transparencia en los procedimientos de toma de decisiones, es menos probable que se experimenten catástrofes. Estas son las pautas que ayudan a mantener seguras las operaciones de AGI en diferentes campos.

Según nuestro análisis, los datos recopilados a través de una revisión de la literatura, donde se desarrollaron estudios de casos sobre los marcos de gobernanza de IA existentes, sugieren que la integración temprana de los factores humanos y la ergonomía (HFE) salva vidas de resultados potencialmente mortales al permitir la supervisión humana en el proceso de toma de decisiones de AGI. En primer lugar, esto apunta a la validación de la hipótesis, que sostiene que, en general, existen principios de gobernanza apropiados que son eficientes, siempre que se apliquen de manera coherente.

Las prácticas actuales de gobernanza de la IA con acciones específicas tomadas hacia la AGI podrán dar a luz a un modelo fuerte capaz de evitar comportamientos incontrolados por parte de esta última. Se hizo hincapié en que los modelos de regulación de la IA deberían integrarse en estrategias específicas y de nuevo diseño que garanticen el control. Estos son sistemas de monitoreo en tiempo real, estructuras de control impactantes y protocolos a prueba de fallas que intervienen en casos de mal funcionamiento o AGI desalineado.

En segundo lugar, la flexibilidad y adaptabilidad del marco de gobernanza se vuelven muy importantes para abordar una base de capacidades en evolución dentro de un sistema AGI. Los sistemas diseñados sobre principios adaptables para sus controles generalmente pudieron eludir el comportamiento emergente que no se podía predecir. Esto confirma la hipótesis de que la integración de prácticas maduras con nuevas actividades de conjuntos debería mejorar la resiliencia de los sistemas generales. Estos hallazgos confirman la hipótesis, que postula que los marcos de gobernanza híbridos serán necesarios para mantener el control de los sistemas AGI.

Aprovechando y ampliando los marcos de gobernanza de IA anteriores para determinar que aprovechar los enfoques multidisciplinarios, como el de los sistemas sociotécnicos (STS), puede mejorar genuinamente la gobernanza de AGI. La implementación de HFE puede fomentar un equilibrio tan deseable entre las medidas regulatorias de arriba hacia abajo y el diseño centrado en el ser humano de abajo hacia arriba si se integra temprano en el proceso de diseño. Los hallazgos también mostraron que las estrategias tradicionales de gobernanza de IA no eran suficientes para gobernar AGI debido a la capacidad de decisiones autónomas intrínseca en AGI.

Otras nuevas acciones de gobernanza que fueron adiciones de vital importancia a los marcos existentes incluyen la detección de señales débiles en tiempo real para la anticipación de riesgos y los protocolos éticos de gestión de riesgos. Esto permitiría tiempos de respuesta más rápidos en el control conductual de AGI, según la hipótesis.

Ya en otros ámbitos, por ejemplo, el cambio climático, se ha probado que la comunidad internacional puede guiarse por principios claros para garantizar que el desarrollo de la IAG se alinee con los valores humanos y promueva el bienestar mundial. Lo más importante, la equidad: el diseño del sistema AGI debe tener principios en su núcleo que eviten el sesgo y traten a las personas y grupos de manera equitativa. De hecho, las organizaciones que adoptan los principios de equidad y beneficencia al diseñar sus modelos de IA están en una mejor posición para manejar los problemas de discriminación y trabajar para aumentar el bienestar de todos los miembros de la sociedad.

Con la colaboración de las partes interesadas de la industria y la comunidad en la protección de la privacidad y el respeto de los datos confidenciales, se convirtió en un instrumento en la construcción de la gobernanza de AGI que reflejaba una amplia gama de valores humanos. Los principios propuestos como resultados de este análisis apoyan la hipótesis de que la cooperación internacional ética constituye una condición necesaria para el despliegue de AGI con beneficios.

Esto significa que los resultados confirman una combinación de principios de gobernanza claramente definidos, la integración de prácticas de IA ya existentes y la introducción de mecanismos de control relacionados específicamente con AGI según sea necesario para garantizar que la implementación de AGI se realice de manera segura, ética y beneficiosa. La adhesión a los principios mencionados no solo minimizaría los riesgos, sino que contribuiría en gran medida a garantizar la alineación con los valores humanos y el bienestar global, alineado con los objetivos de la investigación y las hipótesis planteadas.

DISCUSIÓN

Comprender los riesgos y desafíos que plantea el auge de AGI

La Unión Europea ha sido bastante proactiva, clasificando los sistemas de IA en términos de sus niveles de riesgo asociados, subrayando la necesidad de una regulación fuerte y robusta. Aun así, esta regulación no profundiza en los detalles sobre los riesgos de AGI, dejándolos poco explorados. Los riesgos significativos con AGI se refieren a aplicaciones maliciosas, como ataques cibernéticos o armas autónomas, o usos no beneficiosos que promueven la desigualdad social. Además, el potencial de AGI para superar la inteligencia humana plantea amenazas existenciales, mientras que la automatización a gran escala podría resultar en una gran desigualdad económica.

Karnouskos (2021) y McLean et al. (2021), analizan los riesgos sociales, las restricciones cambiantes en el comportamiento de AGI y los desafíos para alinear AGI con los valores humanos. Sin embargo, los problemas de soberanía nacional presentan desafíos prácticos para la regulación gubernamental. Del mismo modo, existe la preocupación de que las empresas de IA puedan politizar el escepticismo sobre la IA y sus riesgos para evitar regulaciones que restrinjan las actividades rentables, una táctica que se observa en las industrias del tabaco y los combustibles fósiles.

Además, Friederich (2023) destaca la dificultad de alinear los sistemas AGI con el control humano, lo que sugiere que AGI puede resistir las órdenes humanas incluso con esfuerzos rigurosos. La gobernanza de AGI sigue siendo compleja debido a la falta de consenso global debido a la ausencia de acuerdos internacionales, lo que resulta en una coordinación limitada sin un marco universalmente aceptado y esfuerzos fragmentados y a veces contradictorios entre las naciones. La colaboración transfronteriza se ve obstaculizada por barreras legales, regulatorias y culturales.

Por su parte, Caseres (2018) introduce el concepto de máquinas cognitivas artificiales para la gobernanza pública, pero estas máquinas podrían heredar las lagunas y sesgos de los datos utilizados para el entrenamiento, y la lógica y los valores de sus creadores, compartiendo y posiblemente ampliando los riesgos involucrados. Se necesita un enfoque integral que incluya perspectivas legales, regulatorias, éticas y sociales.

Otras lagunas y debilidades incluyen la ambigüedad ética que puede conducir a implementaciones inconsistentes o incluso conflictivas. Los problemas de aplicación que plantea la traducción de principios y directrices de alto nivel en políticas y prácticas viables siguen siendo una tarea compleja. Complejidad tecnológica debido a la naturaleza de rápido avance de la tecnología que puede abrumar las estructuras de gobierno existentes, lo que lleva a una brecha entre las capacidades tecnológicas y la preparación regulatoria.

Kilian et al. (2023) clasifican los riesgos de la IA en categorías como el uso indebido con consecuencias destructivas, los accidentes con consecuencias catastróficas, los riesgos estructurales con efectos transgeneracionales y los riesgos agenciales con gravedad existencial. La influencia de las grandes corporaciones en el desarrollo de AGI podría conducir a modelos de gobernanza que prioricen los intereses comerciales sobre las necesidades sociales, lo que complicaría aún más el establecimiento de un marco integral

Principios para gobernar AGI

En esta subsección se identifican los elementos esenciales para elaborar un marco amplio que rija la AGI a fin de garantizar su aplicación segura y beneficiosa. Este estudio reveló que se considerarán varios componentes y estrategias clave. Basándose en los principios y conocimientos proporcionados por Blackman (2022) y Dobbe et al. (2021), los siguientes elementos son esenciales desde nuestra perspectiva.

La Tabla 1 describe los componentes clave estudiados en esta investigación con respecto a las prácticas comunes para gobernar la IA y la futura gobernanza de AGI.

Tabla 1

Práctica común para gobernar la IA y una eventual AGI

Principio	Descripción	Relevancia para la gobernanza de AGI
Transparencia	AGI debe diseñarse y operarse para que su proceso de decisión sea comprensible para los humanos	Garantiza la supervisión humana, generando confianza al hacer que las acciones y decisiones de AGI sean explicables e interpretables
Responsabilidad	Deben existir estructuras claras de rendición de cuentas para las acciones de AGI	Asigna la responsabilidad de las acciones de AGI, cruciales para los marcos legales y éticos de gobernanza
Beneficencia y no maleficencia	AGI debe actuar para beneficiar a la humanidad y evitar causar daño	Fundamental para el diseño ético de AGI, alineando los objetivos de AGI con el bienestar humano y minimizando los riesgos
Equidad	AGI debe tratar a todos los individuos y grupos de manera justa	Previene los prejuicios o la discriminación, promoviendo la inclusión y la justicia en sus operaciones

Control	Los mecanismos de control sólidos deberían evitar que AGI se salga de control	Fundamental para mantener la supervisión humana y prevenir comportamientos no deseados o peligrosos
Robustez y seguridad	AGI debe estar sólidamente diseñado y probado para operar de manera segura en una amplia gama de escenarios	Garantiza que AGI maneje condiciones inesperadas de manera segura, reduciendo el riesgo de fallas catastróficas
Colaboración entre la industria y la comunidad	Incluye la protección de la privacidad y la información confidencial, el intercambio de conocimientos y la colaboración en proyectos como la Asociación sobre IA	Facilita el intercambio de conocimientos entre industrias y comunidades, la resolución de problemas y la innovación responsable

Fuente: elaboración propia creación con información de Burton, et al., 2020, Foro Económico Mundial, 2023 y (Pratt, Bisson y Warin, 2023)

Con esto se proporciona una visión sobre los componentes necesarios y se señala la base para el desarrollo de principios integrales para gobernar la AGI, centrándose en consideraciones éticas, participación de las partes interesadas y mecanismos de gobernanza sólidos.

De la misma manera, identificamos algunas estrategias adecuadas para gobernar AGI de manera efectiva. La Tabla 2 ilustra las estrategias para el marco que rige la AGI.

Tabla 2

Estrategias cruciales para el marco que rige la AGI

Estrategias	Descripción
Principios normativos	Elaborar principios que respeten los derechos humanos básicos y garanticen el funcionamiento ético de AGI dentro de los contextos sociales
Gobernanza institucional	Establecer entidades para definir métricas, estándares y experiencia para umbrales de rendimiento aceptables
Especificaciones sociotécnicas	Crear especificaciones para diagnosticar y resolver problemas y conflictos de rendimiento del sistema
Disidencia democrática	Proteger la disidencia democrática garantizando la paridad y abordando los aspectos sesgados en el desarrollo de la IAG
Condiciones normativas y del mundo real	Lidiar con la incertidumbre normativa y la indeterminación definiendo comportamientos y resultados aceptables en condiciones del mundo real

Fuente: elaboración propia creación con información de Blackman [5] , Dobbe et al [9]

Sobre la base de los resultados presentados por Batool, at al.(2024), Chmielinski, et al. (2024), Gunasekara (2025), Hohma(2023), Karnouskos (2021), Kuehnert, et al. (2025), Ribeiro, et al. (2025) y Skouloudis, et al. (2025) sobre los modelos más avanzados de inteligencia artificial, que potencialmente podrían llevar a una AGI, en la tabla 3 proponemos los componentes y estrategias clave para desarrollar un marco integral que rija el AGI, acciones sobre cómo medir, evaluar y probar AGI para abordar los riesgos relevantes, así como indicadores cuantitativos o semicuantitativos para cada uno y estudios de donde se obtiene información sobre estos indicadores.

Tabla 3

Componentes clave para un marco que gobierne AGI

Componentes clave	Acciones	Indicador sugerido
Programa de Gestión de Riesgos Éticos	Estructura: Identificar y mitigar los riesgos éticos	Porcentaje de módulos del sistema con evaluación ética formal documentada (% del total)
	Contenido: Definir riesgos éticos específicos basados en la aplicación y los datos	Número de tipos de riesgo ético definidos por dominio (discriminación, privacidad, manipulación, etc.)
Participación de las partes interesadas	Enfoque multidisciplinario: Involucrar a especialistas en ética, científicos de datos, formuladores de políticas y el público	Índice de diversidad de participantes en revisiones (número de disciplinas representadas)
	Mecanismos de retroalimentación: Establecer ciclos de retroalimentación continuos para las diferentes partes interesadas	Cantidad de ciclos de retroalimentación implementados por año
Transparencia	Toma de decisiones comprensible: asegúrese de que la toma de decisiones de AGI sea transparente	Proporción de decisiones explicadas con trazabilidad (porcentaje)
	Explicabilidad: Los sistemas de IA de alto riesgo que afectan significativamente los derechos individuales requieren una explicación significativa sobre el proceso de toma de decisiones	Porcentaje de predicciones con explicaciones interpretables (SHAP, LIME, etc.)
	Comunicación abierta: Mantenga canales de comunicación abiertos	Número de reportes públicos / white papers publicados al año
Responsabilidad	Vigilancia humana: Establecer una responsabilidad clara por las acciones de AGI, con incentivos y consecuencias para diseñadores, desarrolladores, capacitadores y usuarios	Proporción de decisiones que fueron revisadas por humanos (%)
	Compartir responsabilidades: Distribuir responsabilidades entre las partes interesadas	Número de entidades (incluyendo terceros) con roles formales de responsabilidad
	Evaluación: Determinar si el sistema es un sistema de IA de alto riesgo, en función de la finalidad prevista, el alcance geográfico y temporal, los datos utilizados para la formación, las personas o grupos afectados, el cumplimiento de los derechos fundamentales, el impacto previsible en los derechos fundamentales, los riesgos específicos para los grupos marginados o vulnerables, el impacto medioambiental, los planes de mitigación y los sistemas de gobernanza	Proporción de sistemas clasificados como “alto riesgo” tras evaluación (% del total)
Beneficencia y no maleficencia	Impacto positivo: Diseñar AGI para beneficiar a la humanidad, incorporando principios éticos	Proporción de casos donde el modelo genera beneficio explícito sin daño

	Prevención de daños: Implementar salvaguardas para prevenir daños a las personas, el medio ambiente y otros sistemas.	Número estimado de casos dañinos evitados por salvaguardas frente al total de casos de prueba
Equidad	Datos de entrenamiento diversos e imparciales: utilice diversos conjuntos de datos para entrenar AGI, tomando decisiones deliberadas para reducir los sesgos y asegurándose de que pueda funcionar de manera justa en diferentes grupos de usuarios.	Índice de diversidad del dataset (distribución demográfica, geográfica, etc.)
	Finalidad y limitación de datos: Los datos deben recopilarse para fines específicos, explícitos y legítimos y no deben procesarse de manera incompatible con esos fines.	Porcentaje de datos utilizados que provienen de fuentes explícitamente definidas / autorizadas
Mecanismos de control	Monitoreo del uso de recursos: AGI tendría un impacto notable en el uso de recursos informáticos como el uso de CPU y GPU, la presencia de patrones complejos y multidimensionales en el procesamiento, el monitoreo del uso de redes y energía, así como la coordinación y comunicación entre máquinas autónomas	Métrica de uso de CPU/GPU/energía por operación (promedio)
	Kill switch: Basado en el punto anterior sobre la supervisión, establezca alertas y protocolos de seguridad para que, si un sistema muestra comportamientos anómalos e inesperados, pueda detenerse o apagarse, depurarse y corregirse.	Tiempo medio de activación del mecanismo de parada frente a anomalía (s)
	Sandbox de IA: se requiere un entorno de prueba controlado para IA, conocido como sandbox de IA, para fomentar la innovación y facilitar el desarrollo, las pruebas y la validación de sistemas de IA innovadores.	Número de pruebas en entornos controlados antes del despliegue
Robustez y seguridad	Pruebas y validación: Realizar pruebas exhaustivas para garantizar un funcionamiento seguro, incluidas las prácticas de manejo de datos y las evaluaciones de impacto de la inteligencia artificial	Cobertura de pruebas (porcentaje de escenarios) frente al total de casos de uso previstos
	Resiliencia: Diseñe AGI para que sea resistente a las fallas, incluido un plan claro para responder a violaciones de datos o incidentes de seguridad	Porcentaje de degradación bajo condiciones adversas (% pérdida máxima vs normal)
	Supervisión de la implementación: supervise continuamente los sistemas AGI después de la implementación para detectar anomalías, fallos o comportamientos no deseados. Implementar herramientas de monitoreo en tiempo real y establecer	Número de revisiones / auditorías periódicas ejecutadas tras el despliegue

	protocolos para abordar y rectificar los problemas a medida que surjan	
	Desarrollo iterativo: Adopte un enfoque iterativo para el desarrollo de AGI, incorporando comentarios y aprendizajes de cada etapa. Actualice y perfeccione regularmente los sistemas AGI en función de los comentarios de los usuarios, las métricas de rendimiento y los avances tecnológicos.	Frecuencia de versiones actualizadas / revisiones del modelo por año
	Órganos de supervisión: Creación de agencias reguladoras o comités para supervisar el desarrollo y la implementación de AGI	Existencia de entidad reguladora externa (sí / no) + frecuencia de supervisión
	Estándares: Desarrollar y hacer cumplir estándares para AGI, incluidas múltiples partes interesadas, no solo aplicando los principios de ingeniería sino también los legales, regulatorios y éticos en los estándares.	Número de estándares aplicados / cumplidos (ISO, OCDE, etc.)
Colaboración entre la industria y la comunidad	Consideraciones de privacidad: Proteja la información personal y confidencial	Uso de técnicas de privacidad diferencial, federated learning, etc. (% de operación)
	Intercambio de conocimientos: Promover la colaboración y el intercambio de conocimientos	Número de datos, benchmarks o resultados compartidos públicamente

Fuente: elaboración propia.

Dados estos principios y estrategias esenciales clave, es posible identificar un conjunto bien definido de principios de gobernanza con estructuras claras, procedimientos de prueba sólidos y prácticas comunes para mantener el control sobre los sistemas AGI.

Implementación de los Principios con prácticas

Esta subsección señala los principios rectores para diseñar sistemas sociotécnicos. Salmon et al. (2023) [25] resaltan las consideraciones esenciales para el desarrollo, implementación y operación de los sistemas AGI. Señalan que la alineación y medición de valores debe establecerse para crear un conjunto de reglas bien definido, lo que permite a la AGI codificar su "deber" de alinear sus funciones de utilidad con los valores humanos de una manera clara y cuantificable. Vicente (1999) [29] en su obra Análisis cognitivo del trabajo considera elementos como los muchos modelos que son fundamentales para desarrollar un marco de gobernanza integral.

Salmon et al. (2023) adaptaron los principios rectores para el diseño de sistemas sociotécnicos (STS) de Walker, Staton y Salmon (2015) al contexto de AGI. La Tabla 4 resume los principios rectores para el diseño de sistemas sociotécnicos.

Tabla 4

Principios rectores para el diseño de sistemas sociotécnicos

Principio	Descripción
Aportación multidisciplinar	Las nuevas tecnologías requieren aportes de una amplia gama de disciplinas, incluidos los factores humanos y de ergonomía (HFE)
Integración temprana de HFE	La incorporación de HFE al principio del proceso de diseño garantiza un equilibrio óptimo entre los enfoques de arriba hacia abajo y de abajo hacia arriba
Anticipar resultados no deseados	Las decisiones de diseño pueden tener resultados imprevistos, lo que destaca la importancia de la contribución multidisciplinaria y la integración temprana de HFE
Coevolución de los requisitos de los usuarios	Los requisitos de los usuarios evolucionan con el tiempo, lo que dificulta predecir cómo se utilizarán los sistemas AGI a largo plazo
Flexibilidad y adaptabilidad en el diseño	Los diseños deben permitir el cambio y la adaptabilidad, como se enfatiza en la coevolución de los requisitos del usuario
Concéntrese en tareas significativas	El diseño debe priorizar proporcionar a los usuarios tareas útiles y significativas
Especificación crítica mínima	Comience el proceso de diseño con la menor cantidad de especificaciones críticas necesarias, como se indica en Flexibilidad y adaptabilidad en el diseño
Aproveche la coevolución y el ADN del sistema	Utilizar la historia evolutiva de los sistemas para mejorar la funcionalidad
Diseño para nuevas capacidades	Al tiempo que garantizan la flexibilidad, los diseños también deben adaptarse al desarrollo de nuevas capacidades
Enfoque de diseño basado en sistemas	El proceso de diseño debe tratarse como parte de un sistema más amplio

Fuente: elaboración propia creación con información de Salmon et al., Walker, Staton y Salmón

Para implementar eficazmente estos principios, se necesita un enfoque centrado en el ser humano para alinear el proceso de toma de decisiones basado en datos con la gobernanza de AGI. Se debe emplear un marco similar al propuesto por Pratt, Bisson y Warin (2023), centrándose en la detección de señales débiles para permitir predicciones precisas y respuestas rápidas. Este marco facilita la toma de decisiones oportuna y la aplicación del control según sea necesario.

El Foro Económico Mundial (2023) aboga por una transición de la política de IA a la aplicación práctica, lo que permite a las organizaciones establecer programas estructurados para la gobernanza de la IA. Dichos programas deben especificar claramente el propósito de los sistemas de IA, asegurando que los datos utilizados se alineen con este propósito. Además, los algoritmos deben diseñarse, entrenarse e implementarse de manera que se eviten sesgos, con un monitoreo continuo para un aprendizaje efectivo y la generación de resultados. A pesar del rápido ritmo de la toma de decisiones de IA, la interacción humana sigue siendo fundamental para garantizar la transparencia, la responsabilidad y la apertura durante todo el proceso.

Burton et al. (2020) discuten la necesidad de abordar tres posibles brechas en el desarrollo de sistemas autónomos: la brecha semántica derivada de definiciones poco claras de las funciones previstas, la brecha de responsabilidad debido a la falta de condiciones para atribuir responsabilidad moral a actores no humanos y la brecha de responsabilidad resultante de estándares inadecuados para compensar a las víctimas.

El enfoque basado en el riesgo requerido por cualquier marco centrado en gobernar la AGI requiere diferenciar los niveles de riesgo y las implicaciones para varias herramientas de IA. Por ejemplo, al igual que con los medios de transporte, una bicicleta se puede comprar y usar sin necesidad de una licencia debido a su bajo potencial de daño. Por otro lado, un automóvil requiere una licencia especial debido al mayor riesgo asociado con su velocidad y tamaño. Al ascender en la escala, un camión o

vehículo de transporte de pasajeros exige una licencia más avanzada debido a las mayores implicaciones y posibles consecuencias de su uso. Para un avión comercial, los requisitos reglamentarios son aún más estrictos y requieren años de capacitación y certificación. Al más alto nivel, como se ve con la energía nuclear, no solo existen regulaciones nacionales estrictas, sino también acuerdos internacionales que rigen quién puede usar tales tecnologías y con qué fines.

Aplicando esta analogía a la IA, se vuelve vital determinar qué herramientas de IA corresponden a cada nivel de riesgo: cuáles son similares a una bicicleta, un automóvil, un camión, un avión o incluso la energía nuclear. Esta diferenciación permitiría crear marcos regulatorios proporcionales al impacto y riesgo potencial de cada tipo de IA, evitando un enfoque único para todos y fomentando una regulación más eficaz y personalizada.

La exploración exhaustiva de la gobernanza de AGI descrita en esta sección ilustra los componentes vitales necesarios para garantizar el desarrollo seguro, ético y beneficioso de AGI. A lo largo de esta discusión, hemos explorado las estrategias destinadas a mitigar los riesgos, prevenir comportamientos incontrolados y promover la alineación de AGI con los valores humanos y el bienestar global.

CONCLUSIÓN

Esta investigación aborda la hipótesis sobre el establecimiento de principios y componentes clave, así mismo propone acciones e indicadores dirigidas a mitigar riesgos, prevenir comportamientos descontrolados y promover la alineación de AGI con los valores humanos y el bienestar global.

Los resultados de los estudios comparados, si bien no

Se destacó la importancia crítica de los marcos de gobernanza, enfatizando la necesidad de principios bien definidos que alineen los sistemas AGI con la supervisión humana. La hipótesis se probó a través de una combinación de revisiones de literatura y estudios de casos, confirmando la eficacia de la integración de los modelos de gobernanza de IA existentes con nuevas estrategias específicas de AGI. Estos hallazgos subrayan la necesidad de mecanismos de control sólidos, sistemas de monitoreo en tiempo real y estructuras de control en capas para prevenir comportamientos dañinos y garantizar la resiliencia y seguridad de los sistemas AGI. Se demostró que la introducción de enfoques multidisciplinarios, en particular la integración temprana de factores humanos y de ergonomía (HFE), mejora la adaptabilidad y flexibilidad necesarias para gestionar las capacidades cambiantes de AGI.

Resumen de hallazgos

Los principios propuestos y las acciones correspondientes sirven como marco fundamental para la IA y un eventual desarrollo de AGI, guiando a investigadores, desarrolladores y legisladores en la creación de sistemas AGI que priorizan la seguridad y las consideraciones éticas.

Adherirse a los principios comúnmente establecidos como transparencia, responsabilidad, beneficencia y no maleficencia, control de equidad, robustez y seguridad, puede ser subjetivo sin prácticas específicas para implementar esos principios para definir límites, reduciendo las posibilidades de comportamientos impredecibles o fuera de control.

Además, las prácticas propuestas para implementar estos principios incluyen elementos para monitorear, controlar y responder sobre la posible creación de un AGI que podría comportarse de manera no planificada.

Estos resultados requieren esfuerzos colectivos en todos los sectores, lo que implica un futuro en el que la gobernanza de AGI sea un esfuerzo de colaboración, con la academia, la industria y los gobiernos trabajando juntos para garantizar el despliegue seguro y beneficioso de AGI.

El reconocimiento de la naturaleza dinámica de AGI sugiere que el marco de gobernanza deberá ser adaptable, evolucionando en respuesta a los nuevos hallazgos de investigación, los avances tecnológicos y las necesidades sociales.

La autoorganización de la comunidad es otro aspecto relevante, como se ve en iniciativas como la asociación sobre IA, que reúne a desarrolladores de IA para promover principios éticos. Este enfoque ha tenido éxito en otras áreas, pero aún no está claro qué tan efectivo será para AGI, particularmente a mediano y largo plazo.

La influencia significativa de las grandes corporaciones en el desarrollo de AGI puede conducir a modelos de gobernanza que priorizan los intereses comerciales sobre las necesidades de la sociedad, lo que podría socavar el consenso y la participación de base amplia.

El papel de la comunidad internacional en la gobernanza de AGI es fundamental. Los esfuerzos de colaboración entre los gobiernos, la industria y las comunidades son cruciales para fomentar la cooperación, promover el intercambio de conocimientos y crear un marco universalmente aceptado. La colaboración transfronteriza enfrenta numerosos desafíos, como barreras legales, regulatorias y culturales, pero sigue siendo esencial para garantizar que el desarrollo de AGI se alinee con los valores éticos y promueva el bienestar global.

Es importante recalcar la importancia de la transparencia, la rendición de cuentas, la equidad y la beneficencia como principios rectores centrales para el diseño y la gobernanza de la AGI. Al enfatizar la transparencia, los procesos de toma de decisiones de AGI se pueden hacer interpretables y explicables para los humanos, fomentando la confianza y mejorando la supervisión humana. Las estructuras de rendición de cuentas son esenciales para asignar la responsabilidad de las acciones de AGI, mientras que la equidad garantiza un trato equitativo de individuos y grupos, previniendo prejuicios y reforzando la justicia. La beneficencia y la no maleficencia siguen siendo fundamentales para el desarrollo de AGI, alineando los objetivos de AGI con el bienestar humano y minimizando los riesgos de daño.

Por lo tanto, este estudio afirma que el despliegue exitoso de AGI requiere un enfoque de gobernanza multifacético y la aplicación de mediciones sobre los indicadores propuestos. La combinación de principios bien definidos, la integración de las prácticas existentes con nuevos mecanismos de control y el abordaje de consideraciones éticas, legales y sociales son pasos esenciales para avanzar en la AGI de manera segura. Al adherirnos a estos principios y promover la colaboración internacional, podemos mitigar los riesgos, mejorar los beneficios sociales y garantizar que AGI sirva a los mejores intereses de la humanidad, ahora y en el futuro.

Algunos desafíos a considerar en futuras investigaciones:

Los incentivos financieros para las empresas con fines de lucro podrían actuar como un elemento disuasorio para el desarrollo seguro y ético de la IA. Si el interés financiero propio de las corporaciones se desalinea con el interés público, las corporaciones pueden actuar en contra del bienestar público. Esto puede complicarse aún más por la "sinergia entre ganancias e investigación y desarrollo de AGI", donde las ganancias a corto plazo se alinean con la investigación y el desarrollo de AGI.

Los esquemas de propiedad privada son difíciles de aplicar a la IA debido a los desafíos para restringir el acceso. Sin embargo, la fabricación de hardware, generalmente de propiedad privada, puede desempeñar un papel en la acción colectiva de la IA. Se espera que las grandes empresas de fabricación de hardware y desarrollo de software sigan siendo influyentes a medio y largo plazo.

REFERENCIAS

Abhishek, A., Erickson, L., & Bandopadhyay, T. (2025). *BEATS: Bias Evaluation and Assessment Test Suite for Large Language Models*. arXiv. doi.org/10.48550/arXiv.2503.24310

AILAB. (2023). Propuestas de regulación y recomendaciones de inteligencia artificial en el mundo: Síntesis de principales aspectos. Propuestas de regulación y recomendaciones de inteligencia artificial en el mundo: Síntesis de principales aspectos. Retrieved from ttps://ialab.com.ar/wp-content/uploads/2023/08/Propuestas-de-regulacion-y-recomendaciones-de-IA-en-el-mundo-1.pdf

Batool, A., Zowghi, D., & Bano, M. (2024). Responsible AI Governance: A systematic literature review. *arXiv*. doi.org/10.48550/arXiv.2401.10896

Baum, S. D. (2020). Medium-Term Artificial Intelligence and Society. Information, 11, 290. doi:10.3390/info11060290

Blackman, R. (2022). Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI. In Amazon. Harvard Business Review Press.

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artificial Intelligence, 279, 103201. doi:10.1016/j.artint.2019.103201

Casares, A. P. (2018). The brain of the future and the viability of democratic governance: The role of artificial intelligence, cognitive machines, and viable systems. Futures, 103, 5–16. doi:10.1016/j.futures.2018.05.002

Chmielinski, K., Newman, S., Kranzinger, C. N., Hind, M., Vaughan, J. W., Mitchell, M., Stoyanovich, J., McMillan-Major, A., McReynolds, E., & Esfahany, K. (2024). *The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers* [Discussion paper]. Harvard Kennedy School, Shorenstein Center. https://shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/

Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard choices in artificial intelligence. Artificial Intelligence, 300, 103555. doi:10.1016/j.artint.2021.103555

Faroldi, C. (2024), Artificial general intelligence and the EU AI Act: A regulatory mismatch?, AI & Society

Friederich, S. (2023). Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. AI and Ethics. doi:10.1007/s43681-023-00268-7

Greenstein, S. (2021). Preserving the rule of law in the era of artificial intelligence (AI). Artificial Intelligence and Law, 30, 291–323. doi:10.1007/s10506-021-09294-4

Gunasekara, L. (2025). A systematic review of responsible artificial intelligence. *MDPI*. doi.org/10.3390/asi8040097

Hohma, E. (2023). The need and elements of trusted development of AI. *MDPI*. doi.org/10.3390/ai4040046

Karnouskos, S. (2021). Symbiosis with artificial intelligence via the prism of law, robots, and society. Artificial Intelligence and Law, 30, 93–115. doi:10.1007/s10506-021-09289-1

Kilian, K. A., Ventura, C. J., & Bailey, M. M. (2023). Examining the differential risk from high-level artificial intelligence and the question of control. Futures, 151, 103182. doi:10.1016/j.futures.2023.103182

Kuehnert, B., Kim, R. M., Forlizzi, J., & Heidari, H. (2025). The "Who", "What", and "How" of Responsible AI Governance: A systematic review and meta-analysis of (actor, stage)-specific tools. *arXiv*. doi.org/10.48550/arXiv.2502.13294

Kurzweil, R. (2005). The singularity is near: When humans transcend biology. Viking / Penguin.

Liu, H.-Y., & Maas, M. M. (2021). `Solving for X?' Towards a problem-finding framework to ground long-term governance strategies for artificial intelligence. Futures, 126, 102672. doi:10.1016/j.futures.2020.102672

Llamas Covarrubias, J. Z. (2022). Enfoques regulatorios para la Inteligencia Artificial (IA). Revista Chilena De Derecho, 49(3), 31–62. doi.org/10.7764/R.493.2

Maher, R., & Buhmann, K. (2019). Meaningful stakeholder engagement: Bottom-up initiatives within global governance frameworks. Geoforum, 107, 231–234. doi:10.1016/j.geoforum.2019.06.013

McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2021). The risks associated with Artificial General Intelligence: A systematic review. Journal of Experimental, Theoretical Artificial Intelligence, 35, 649–663. doi:10.1080/0952813x.2021.1964003

Pratt, L., Bisson, C., & Warin, T. (2023). Bringing advanced technology to strategic decision-making: The Decision Intelligence/Data Science (DI/DS) Integration framework. Futures, 152, 103217. doi:10.1016/j.futures.2023.103217

Ribeiro, D., Rocha, T., Pinto, G., Cartaxo, B., Amaral, M., Davila, N., & Camargo, A. (2025). Toward Effective AI Governance: A review of principles. *arXiv*. doi.org/10.48550/arXiv.2505.23417

Salmon, P. M., Baber, C., Burns, C., Carden, T., Cooke, N., Cummings, M., . . . Stanton, N. A. (2023). Managing the risks of artificial general intelligence: A human factors and ergonomics perspective. Human Factors and Ergonomics in Manufacturing, Service Industries, 33, 366–378. doi:10.1002/hfm.20996

Saman Ghaffarian, F. R. (2023). Explainable artificial intelligence in disaster risk management: Achievements and prospective futures. International Journal of Disaster Risk Reduction, 104123. doi.org/10.1016/j.ijdrr.2023.104123

Skouloudis, A., el al. (2025). Scratching the surface of responsible AI in financial systems. *MDPI*. doi.org/10.3390/ai6080169

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Wright, D. (2023). A systematic review of artificial intelligence impact assessments. Artificial Intelligence Review. doi:10.1007/s10462-023-10420-8

Vicente, K. (1999). Cognitive work analysis: Toward safe, productive, and healthy computer-based work. Lawrence Erlbaum Associates.

Walker, G., Stanton, N., & Salmon, P. (2015). Human factors inautomotive engineering and technology. Ashgate.

Witt, A., Huggins, A., Governatori, G., & Buckley, J. (2023). Encoding legislation: a methodology for enhancing technical validation, legal alignment and interdisciplinarity. Artificial Intelligence and Law. doi:10.1007/s10506-023-09350-1

World Economic Forum. (2023,). Towards a strong trust-based AI governance model. Towards a strong trust-based AI governance model. Retrieved from https://www.weforum.org/agenda/2023/05/towards-a-strong-trust-based-ai-governance-model/