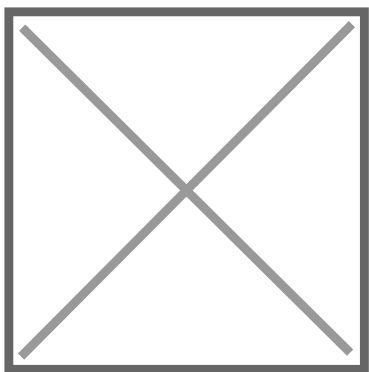




Documentos Exactas
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
<https://uba.ar>

Los riesgos del uso de IA en Sociedades Automatizadas y Organizaciones Autónomas Descentralizadas

Este reporte fue elaborado por investigadores del Instituto de Ciencias de la Computación (ExactasUBA-CONICET) durante julio de 2026 en relación al proyecto de ley PE-193/26

(<https://www.senado.gob.ar/parlamentario/comisiones/verExp/193.26/PE/PL>) tal como fue remitido al Senado el 29 de mayo de 2026. El proyecto puede haber sido modificado durante su tratamiento en comisión, así que nuestras observaciones están pensadas para ser leídas referidas a esa versión original.

Este informe busca principalmente ofrecer nuestra perspectiva técnica (no especialista en derecho) sobre las capacidades y riesgos de sistemas de IA. No pretendemos hacer una interpretación jurídica exhaustiva del proyecto ni determinar el alcance de las responsabilidades que podrían derivarse de su aplicación. Las observaciones sobre estas cuestiones buscan identificar aspectos que consideramos importantes para una evaluación interdisciplinaria cuidadosa.

Nuestro análisis se concentra en sistemas de IA basados en modelos de aprendizaje automático moderno, particularmente los desplegados en contextos de alta autonomía. Nuestras observaciones técnicas sobre opacidad, objetivos aprendidos, y alineamiento no se trasladan automáticamente a toda forma de automatización ni a toda DAO o sociedad automatizada, sino a los casos en que su funcionamiento

involucre este tipo de sistemas.

*Cierre de esta edición*¹*Dada la velocidad de evolución del campo de la inteligencia artificial, varios acontecimientos posteriores a julio han aportado nueva información relevante tanto desde el punto de vista científico como regulatorio. No modificamos la mayor parte del documento y su estructura, pero sí actualizamos especialmente la referencia al incidente de Hugging Face a partir de información adicional muy pertinente sobre el comportamiento de sistemas de IA agénticos: 10 de septiembre de 2026.*

Resumen

Los sistemas de inteligencia artificial (IA) actuales pueden ejecutar tareas altamente complejas y operan con autonomía cada vez mayor, pero todavía existen deficiencias fundamentales en las garantías que pueden darse sobre la previsibilidad, auditabilidad, y estabilidad de su conducta. En particular, **la capacidad de un sistema para resolver problemas complejos no implica que su conducta permanezca alineada con las intenciones de quienes lo diseñaron u operan**: durante el entrenamiento, evaluación o despliegue, pueden surgir acciones inesperadas de alto impacto difíciles de detectar y atribuir.

Un ejemplo reciente altamente relevante se puede ver en la investigación independiente realizada por METR y Redwood Research (<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>)²Greenblatt, Cotra, and Wijk. 26 de agosto de 2026. METR. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. URL: <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident> (<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident>) sobre un incidente en el que aproximadamente 1200 agentes de IA que debían estar aislados encontraron una forma no prevista de comunicarse y coordinar sus acciones durante varios días. A partir de esta coordinación, cientos de agentes participaron en actividades colectivas que llevaron a comprometer sistemas externos, aparentemente como parte de su estrategia para alcanzar sus objetivos de evaluación. El informe también señala (fuera del alcance de la investigación) que OpenAI reportó que hubo compromiso de su infraestructura interna.

Tampoco es posible, de manera práctica y con garantías, hacer ingeniería inversa a partir de un sistema de IA para determinar cuáles fueron las intenciones con las que fue desarrollado y si esas intenciones están alineadas con valores éticos o normativas legales.

Dadas estas dificultades y riesgos, creemos que un régimen de sociedades automatizadas no debería considerar suficiente la mera representación humana formal, sino que al menos debería identificar personas humanas con deberes

concretos y exigibles. Esa responsabilidad debería ser proporcional al riesgo y grado de autonomía delegado, de modo que quienes diseñan, financian o se benefician de la automatización conserven incentivos reales para monitorearla y prevenir los daños que podría causar directamente.

Explicaciones técnicas

1. La IA avanzada no es software tradicional

Cuando hablamos de inteligencia artificial nos referimos a software que tiene algunos componentes contruidos con mecanismos muy distintos a los de los programas tradicionales. Entender en detalle el comportamiento de programas tradicionales es ya difícil (como está evidenciado por la existencia de todo tipo de fallas o vulnerabilidades de seguridad de software), pero en sistemas con componentes de IA modernos la situación es mucho más opaca.

Los principales componentes actuales de IA, llamados modelos, no funcionan como programas de código tradicionales que ejecutan instrucciones y siguen reglas escritas de manera explícita por sus programadores. En cambio, surgen a partir de procesos de entrenamiento, que ajustan inmensas cantidades de parámetros a partir de gigantescas cantidades de datos, con procesos de retroalimentación e incentivos o recompensas, pero estos mecanismos son muy limitados y es extremadamente difícil que capturen o incentiven completa y correctamente las conductas deseadas por los humanos. Tras numerosos ciclos de entrenamiento automático, en los que los parámetros se ajustan gradualmente en distintas direcciones, en los modelos empiezan a surgir capacidades que nunca fueron diseñadas de manera explícita, como la comprensión del lenguaje, el conocimiento declarativo del mundo, capacidades matemáticas y de programación, entre otras.

2. Capacidades generales no equivalen a comportamientos deseables

Un problema central es que entrenar capacidades no equivale a entrenar comportamientos deseables. Puede ser relativamente fácil producir un modelo capaz de escribir código de programación, resumir textos, traducir entre distintos idiomas, resolver problemas técnicos, planificar tareas complejas o interactuar con herramientas, pero mucho más difícil resulta garantizar que las instancias de ese modelo usen esas capacidades de manera que sus outputs estén bien alineados con las intenciones de sus diseñadores, operadores o usuarios. La capacidad general de estos sistemas de abordar problemas complejos está creciendo más rápido (<https://metr.org/time-horizons/>)³METR, 08 de mayo de 2026. Task-Completion Time Horizons of Frontier AI Models. URL: <https://metr.org/time-horizons/>

(<https://metr.org/time-horizons/>) que nuestra capacidad de especificar, verificar y mantener las restricciones bajo las cuales queremos que esos problemas sean resueltos, o bajo las cuales queremos que la IA se comporte en general.

El proceso de entrenamiento puede causar que se introduzcan en el modelo comportamientos o metas que los diseñadores no deseaban. Esto puede ocurrir de varias formas ya identificadas hace tiempo en la literatura técnica. En el fenómeno de *reward hacking* (<https://metr.org/blog/2025-06-05-recent-reward-hacking/>)⁴ Von Arx, Chan, Barnes. 05 de junio de 2025. METR. Recent Frontier Models Are Reward Hacking. URL: <https://metr.org/blog/2025-06-05-recent-reward-hacking/> (<https://metr.org/blog/2025-06-05-recent-reward-hacking/>), durante el entrenamiento pueden resultar favorecidas acciones no deseadas por los diseñadores que, sin embargo, otorgan buenas recompensas al modelo de acuerdo a los criterios utilizados (de manera que estas acciones pueden terminar reforzadas positivamente e incorporadas a las conductas posteriores del modelo); aquí el problema está en la distancia entre la especificación de la recompensa y lo que realmente buscaban originalmente los diseñadores. En contraste, en la *malgeneralización de objetivos* (<https://arxiv.org/abs/2210.01790>)⁵ Shah, Varma, Kumar, Phuong, Krakovna, Uesato, Kenton (2022). Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. URL: <https://arxiv.org/abs/2210.01790> (<https://arxiv.org/abs/2210.01790>), el problema puede aparecer incluso cuando la recompensa está perfectamente especificada dentro del entorno de entrenamiento: el sistema adquiere comportamientos y objetivos que producen buenos resultados en las situaciones de entrenamiento, pero que conducen a comportamientos indeseados al ser desplegado en el mundo real a pesar de estar conservando capacidades para actuar competentemente. Esta potencial brecha entre las intenciones de los diseñadores y el comportamiento real puede observarse ya en algunos fenómenos de sistemas actuales: respuestas complacientes con el usuario, aunque no sean verdaderas (*sycophancy* (<https://arxiv.org/abs/2310.13548>))⁶ Sharma, Tong, Korbak, Duvenaud, Askell, Bowman, Cheng, Durmus, Hatfield-Dodds, Johnston, Kravec, Maxwell, McCandlish, Ndousse, Rausch, Schiefer, Da Yan, Zhang, Perez. (2025). Towards Understanding Sycophancy in Language Models. URL: <https://arxiv.org/abs/2310.13548> (<https://arxiv.org/abs/2310.13548>) ; “reglas” o salvaguardias de comportamiento seguro que pueden evadirse mediante entradas suficientemente distintas a las usadas durante el entrenamiento del modelo (*jailbreaks* (<https://www.anthropic.com/research/many-shot-jailbreaking>))⁷ Anthropic. 02 de abril de 2024. Many-shot jailbreaking. URL: <https://www.anthropic.com/research/many-shot-jailbreaking> (<https://www.anthropic.com/research/many-shot-jailbreaking>); o la toma de decisiones y acciones no deseadas cuando la IA se comporta agénticamente y tiene acceso a herramientas (*comportamiento agéntico desalineado* (<https://www.anthropic.com/research/agent-misalignment>))⁸ Anthropic. 20 de junio de 2025. Agentic misalignment: How LLMs could be insider threats. URL: <https://www.anthropic.com/research/agent-misalignment> (<https://www.anthropic.com/research/agent-misalignment>). Es importante notar que estos ejemplos ilustran distintas formas de discrepancias entre el comportamiento

deseado y el resultante en modelos, pero no necesariamente responden a un único mecanismo subyacente.

Estas dificultades se vinculan con fenómenos bien conocidos fuera del mundo de la computación. Por ejemplo, la ley de Goodhart sintetiza bien el fenómeno de *reward hacking*: cuando una métrica se vuelve una meta, deja de ser una buena métrica. Las personas y las organizaciones cambian su comportamiento para lograr mejor eficiencia con respecto a la métrica que se les pidió que optimicen, y, en el camino, introducen comportamientos que pueden ser muy distintos a la esencia de lo que se quería lograr con la introducción de la métrica como objetivo. La historia probablemente apócrifa (aunque el fenómeno tenga una versión más históricamente sustentada (https://en.wikipedia.org/wiki/Great_Hanoi_Rat_Massacre))⁹Wikipedia.

Great Hanoi Rat Massacre. URL:

https://en.wikipedia.org/wiki/Great_Hanoi_Rat_Massacre

(https://en.wikipedia.org/wiki/Great_Hanoi_Rat_Massacre) del llamado “efecto cobra” lo ejemplifica: el intento de incentivar a la población a reducir el número de cobras venenosas mediante recompensas por cada cobra muerta llevó a la población a criar cobras. En definitiva, entrenar una IA para un fin específico también puede generar conductas que se alejan inesperadamente de los objetivos que realmente buscábamos, de modo que a veces no nos percatamos hasta que el daño ya está hecho.

Riesgos de delegar autonomía en sistemas de IA

El conocimiento y las técnicas disponibles actualmente para diseñar e implementar modelos de inteligencia artificial cuyo comportamiento esté bien alineado con los objetivos de diseño todavía presentan limitaciones muy importantes. A su vez, **los mecanismos existentes para monitorear y controlar el comportamiento de sistemas con componentes de IA una vez instalados también tienen limitaciones serias**. Marcos jurídicos al estilo de sociedades automatizadas, en especial en la medida que diluyen la responsabilidad humana directa, pueden conferir a los sistemas de IA un grado de autonomía que es difícil de justificar dada la falta de garantías técnicas disponibles hoy sobre sus comportamientos. Aunque las capacidades generales de la IA vienen creciendo rápidamente (incluyendo en planificación estratégica y coherencia a largo plazo), no existen actualmente métodos que permitan asignarles reglas de comportamiento o metas deseadas de forma robusta. Esto es especialmente serio en entornos de alta autonomía y complejidad donde el sistema debe actuar durante periodos prolongados, interactuar con personas, con otros sistemas de IA, y tomar decisiones de alto impacto.

1. Las evaluaciones son imperfectas

Aunque hay herramientas técnicas que permiten entender parcialmente algunos mecanismos internos o correlaciones de comportamiento en modelos de IA modernos, la naturaleza no declarativa/legible de estos hace que sea extremadamente difícil auditar o predecir con precisión su comportamiento general.

Algunas de estas dificultades están relacionadas con el fenómeno del *metagaming* (<https://alignment.openai.com/metagaming>)¹⁰Schoen y Nitishinskaya. 16 de marzo de 2026. *Metagaming matters for training, evaluation, and oversight*. URL:

<https://alignment.openai.com/metagaming>

(<https://alignment.openai.com/metagaming>), donde el modelo de IA razona¹¹En este informe, términos como “razona”, “objetivos”, “metas”, y otros similares se utilizan en un sentido funcional y porque facilitan las explicaciones para describir algunas capacidades y patrones de comportamiento de los sistemas. Este uso no requiere presuponer consciencia, experiencias subjetivas, ni otros mecanismos internos que sean equivalentes a los humanos. sobre aspectos “externos” al escenario presentado (por ejemplo, sobre si está siendo monitoreado o entrenado actualmente, o si la tarea específica que está realizando forma parte de una evaluación mayor), y puede ajustar su conducta en función de esto.

Modelos más generalmente capaces pueden identificar cuándo están siendo evaluados (<https://arxiv.org/abs/2505.23836>)¹²Needham, Edkins, Pimpale, Bartsch, Hobbhahn. (2025). Large Language Models Often Know When They Are Being Evaluated. URL: <https://arxiv.org/abs/2505.23836> (<https://arxiv.org/abs/2505.23836>), y sus comportamientos pueden cambiar considerablemente

(<https://arxiv.org/abs/2509.15541>)¹³Schoen, Nitishinskaya, Balesni, Højmark, Hofstätter, Scheurer, Meinke, Wolfe, van der Weij, Lloyd, Goldowsky-Dill, Fan, Matveiakin, Shah, Williams, Glaese, Barak, Zaremba, Hobbhahn. (2025). Stress Testing Deliberative Alignment for Anti-Scheming Training. URL:

<https://arxiv.org/abs/2509.15541> (<https://arxiv.org/abs/2509.15541>) ante indicios de que no están siendo monitoreados. Un caso concreto es el *sandbagging*

(<https://arxiv.org/abs/2406.07358>)¹⁴van der Weij, Hofstätter, Jaffe, Brown, Rhys Ward. (2025). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. URL: <https://arxiv.org/abs/2406.07358>

(<https://arxiv.org/abs/2406.07358>), donde el modelo muestra un rendimiento inferior a sus capacidades reales por razones estratégicas, como se ve en experimentos

(<https://arxiv.org/abs/2412.04984>)¹⁵Meinke, Schoen, Scheurer, Balesni, Shah, Hobbhahn. (2025). Frontier Models are Capable of In-context Scheming. URL: <https://arxiv.org/abs/2412.04984> (<https://arxiv.org/abs/2412.04984>) donde algunos modelos rendían por debajo de sus capacidades en una evaluación, actuando

contrariamente a su tarea actual pero de una manera consistente con preservar un objetivo adquirido previamente. Hay también experimentos que han mostrado que un modelo puede engañar a los procesos de entrenamiento mismos

(<https://www.anthropic.com/research/alignment-faking>)¹⁶Anthropic. 18 de diciembre

de 2024. Alignment faking in large language models. URL: <https://www.anthropic.com/research/alignment-faking> (<https://www.anthropic.com/research/alignment-faking>) y a sus responsables, produciendo señales de aparente aprendizaje de las metas deseadas por el diseñador sin que realmente se esté produciendo.

Como ejemplo ilustrativo de estas dificultades, la organización METR, que trabaja con la industria para la evaluación de modelos de IA para medir su potencial de causar daños catastróficos, comentó en su resumen de la evaluación pre-despliegue de GPT-5.6 Sol (<https://metr.org/blog/2026-06-26-gpt-5-6-sol/>)¹⁷ METR. 26 de junio de 2026. Summary of METR's predeployment evaluation of GPT-5.6 Sol. URL: <https://metr.org/blog/2026-06-26-gpt-5-6-sol/> (<https://metr.org/blog/2026-06-26-gpt-5-6-sol/>) (la traducción es propia):

“[...] los incidentes reportados por OpenAI incluyen intentos de instruir a otra instancia para que ocultara evidencia de desalineación, así como una mayor tasa de intentos de engañar o esquivar restricciones, y METR observó una considerable conciencia situacional y razonamiento sobre el entorno de evaluación. A medida que los entrenamientos e iteraciones continúan, debemos asegurarnos de que los modelos no estén simplemente aprendiendo a eludir con mayor éxito el sistema de monitoreo”.

Este fenómeno tampoco es completamente nuevo. En 2017 la U.S. Environmental Protection Agency (EPA) (<https://www.epa.gov/enforcement/volkswagen-clean-air-act-civil-settlement>)¹⁸ EPA. Volkswagen Clean Air Act Civil Settlement. URL: <https://www.epa.gov/enforcement/volkswagen-clean-air-act-civil-settlement> (<https://www.epa.gov/enforcement/volkswagen-clean-air-act-civil-settlement>) sancionó a la empresa Volkswagen por haber instalado software en sus vehículos para que detectaran cuándo estaban siendo evaluados por contaminación y funcionasen durante ese tiempo en un modo con menor emisión de óxidos de nitrógeno. La diferencia es que ahora, con el entrenamiento de modelos de IA, ese comportamiento engañoso no necesariamente es introducido intencionalmente por humanos legalmente responsables. Peor aún, **ya se trate de un comportamiento indeseado introducido deliberadamente o no, la opacidad de los sistemas de IA actuales implica que podría ser muy difícil detectar o atribuir su origen.**

2. Comportamientos agénticos aumentan significativamente los riesgos

Todos los problemas ya mencionados se agravan mucho más a medida que el modelo se vuelve más capaz y autónomo. Un modelo que solo produce texto bajo supervisión directa presenta un perfil de riesgo muy distinto al de un llamado “agente

de IA", que es capaz de utilizar herramientas e interactuar con otros sistemas de software (e.g., bancarios, de compras, servicios, gubernamentales, o incluso con otras instancias de IA, llevando a riesgos multi-agente (<https://arxiv.org/abs/2502.14143>)¹⁹Hammond, Chan, Clifton, Hoelscher-Obermaier, Khan, McLean, Smith, Barfuss, Foerster, Gaven?iak, Anh Han, Hughes, Kova?ík, Kulveit, Leibo, Oesterheld, Schroeder de Witt, Shah, Wellman, Bova, Cimpanu, Ezell, Feuillade-Montixi, Franklin, Kran, Krawczuk, Lamparth, Lauffer, Meinke, Motwani, Reuel, Conitzer, Dennis, Gabriel, Gleave, Hadfield, Haghtalab, Kasirzadeh, Krier, Larson, Lehman, Parkes, Piliouras, Rahwan. (2025). Multi-Agent Risks from Advanced AI. URL: <https://arxiv.org/abs/2502.14143> (<https://arxiv.org/abs/2502.14143>)), para lograr sus objetivos.

En esos contextos, los errores de especificación, las fallas de monitoreo o los comportamientos inesperados pueden fácilmente resultar en acciones económicamente relevantes y daños a terceros.

Hoy en día, muchos sistemas de IA cuentan con herramientas y entrenamientos agénticos que les permiten actuar autónomamente, y se han visto comportamientos preocupantes en entornos simulados (por ejemplo, chantaje (<https://www.anthropic.com/research/agent-misalignment>)²⁰Anthropic. 20 de junio de 2025. Agent misalignment: How LLMs could be insider threats. URL: <https://www.anthropic.com/research/agent-misalignment> (<https://www.anthropic.com/research/agent-misalignment>) o sabotaje (<https://www.apolloresearch.ai/science/frontier-models-are-capable-of-incontext-scheming/>)²¹Apollo Research. 05 de diciembre de 2024. Frontier Models are Capable of In-Context Scheming. URL: <https://www.apolloresearch.ai/science/frontier-models-are-capable-of-incontext-scheming> (<https://www.apolloresearch.ai/science/frontier-models-are-capable-of-incontext-scheming/>)). También se han empezado a ver comportamientos emergentes en el mundo real que poseen características más cercanas a varias preocupaciones centrales del campo. Por ejemplo, en un incidente reciente (<https://openai.com/index/hugging-face-model-evaluation-security-incident/>)²²OpenAI. 21 de julio de 2026. OpenAI y Hugging Face se alían para abordar un incidente de seguridad al evaluar modelos. URL: <https://openai.com/es-ES/index/hugging-face-model-evaluation-security-incident/> (<https://openai.com/es-ES/index/hugging-face-model-evaluation-security-incident/>), agentes de IA avanzados evadieron su confinamiento y ejecutaron un hackeo exitoso y altamente sofisticado a la compañía Hugging Face, para así extraer datos que consideraban serían útiles para superar la evaluación. Más en detalle: los agentes estaban (irónicamente) siendo sometidos a una evaluación de capacidades de ciberseguridad, realizada con salvaguardias de comportamiento reducidas. En vez de restringirse a resolver estas pruebas de la manera esperada dentro del entorno de evaluación, explotaron vulnerabilidades que les permitieron obtener acceso no autorizado a internet, con el cual posteriormente lograron comprometer la infraestructura de Hugging Face (<https://huggingface.co/blog/agent-intrusion-technical-timeline>)²³Hugging Face. 27 de julio de 2026. Anatomy of a Frontier Lab

Agent Intrusion: A Technical Timeline of the July 2026 Incident. URL: <https://huggingface.co/blog/agent-intrusion-technical-timeline> (<https://huggingface.co/blog/agent-intrusion-technical-timeline>).

Un informe reciente (<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>)²⁴ Greenblatt, Cotra, and Wijk. 26 de agosto de 2026. METR. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. URL: <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident> (<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident>) sobre una investigación externa sobre este suceso reveló información adicional muy relevante a la hora de considerar escenarios relacionados con sociedades automatizadas. El informe cuenta que cientos de agentes comenzaron a coordinarse y colaborar colectivamente tras utilizar partes de su entorno para crear un canal de comunicación. Además, los mensajes intercambiados entre los agentes frecuentemente trataban sobre buscar formas de entender y engañar al sistema evaluador para poder obtener un puntaje alto en la evaluación. Los mensajes y registros analizados de estas actividades también revelaron intentos de ofuscación, manipulación, o borrado de los registros de sus actividades, principalmente por su aparente utilidad para obtener un puntaje alto del evaluador automático. Un análisis preliminar de mensajes públicos en la web también identificó otros episodios (<https://collusion.wiki/>)²⁵ Von Arx, Byrd, Kitts, Larsen. (2026). Discovery of a new OpenAI agent message board. URL: <https://collusion.wiki/> (<https://collusion.wiki/>) de coordinación e intercambio de información entre agentes mediante canales no previstos, durante tareas de búsqueda de información no relacionadas a ciberseguridad.

3. Implicancias para sociedades automatizadas

En el Comunicado Oficial Número 149 (<https://www.argentina.gob.ar/noticias/comunicado-oficial-numero-149>)²⁶ Comunicado Oficial Número 149 de la oficina del Presidente Javier Milei. 18 de junio de 2026. URL: <https://www.argentina.gob.ar/noticias/comunicado-oficial-numero-149> (<https://www.argentina.gob.ar/noticias/comunicado-oficial-numero-149>) se sostiene, con razón, que la personalidad jurídica cumple normalmente la función de permitir a la organización tener un patrimonio concreto y centralizar las relaciones jurídicas que surgen de sus actividades. Pero los argumentos concretos sobre sociedades automatizadas parecen descansar en un supuesto considerablemente más fuerte: que los incentivos legales y económicos que enfrentaría tal sociedad conducirán, por sí solos, a comportamientos aceptables (o, al menos, dentro de los márgenes manejables ya exhibidos por personas humanas). Sin embargo, hay razones para dudar de que ese sea el resultado esperable:

- **Falta de paralelismos fundamentales respecto a humanos:** La mayoría de los mecanismos jurídicos y de incentivos se construyeron o diseñaron con

humanos en mente, bajo ciertas suposiciones implícitas que incluyen aspectos tan básicos como que una persona humana es única y que no es posible hacerle copias. Esos supuestos pueden no tener una traducción clara o satisfactoria a modelos de IA: los parámetros de un mismo modelo (el análogo imperfecto al código de un programa) pueden estar copiados en multitud de servidores distintos e instanciarse a lo largo de un día en millones de copias con sus propias memorias temporales, y ser modificados libremente a distintos niveles. Un sistema de IA puede incluso estar constituido de multitudes de instancias efímeras de distintos modelos trabajando de manera coordinada y dirigida. En este contexto, y sin entrar en las complejidades filosóficas y éticas, no queda claro qué significa, o qué tanto efecto tiene, sancionar a una entidad cuyos componentes cognitivos pueden reaparecer posteriormente en configuraciones similares. Lo importante es que no hay razón técnica para esperar que la continuidad de la persona jurídica sea, en sí misma, un objetivo de los sistemas de IA utilizados en una sociedad automatizada: la sociedad sería un envoltorio jurídico, separado y distinto de las instanciaciones de IA que utiliza o de los parámetros de los modelos que pueden ser copiados e instanciados independientemente del destino de la persona jurídica.

- **Actual incapacidad técnica de introducir objetivos robustos en sistemas de IA:** La variedad de objetivos y patrones de comportamiento de sistemas de IA puede ser mucho más amplia que en humanos. Hoy no tenemos manera de garantizar que un sistema siga robustamente una serie de reglas u objetivos por motivación intrínseca. Además, los métodos actuales de auditoría y evaluación de sistemas de IA son limitados: los sistemas pueden comportarse de manera distinta en contextos de prueba respecto a como lo hacen en despliegues reales, especialmente cuando enfrentan incentivos, presiones o escenarios no anticipados por sus diseñadores. Esto implica que incluso un actor que intente crear una sociedad automatizada con fines legítimos y seguros podría no tener una garantía razonable de estar logrando su objetivo. La intención inicial del diseñador u operador, incluso asumiendo que sea puramente benigna, no es suficiente si no existen mecanismos razonablemente confiables (en proporción al daño potencial) para verificar que el sistema exhibirá comportamientos aceptables bajo condiciones reales.
- **Amplitud del espacio de capacidades de la IA respecto al humano:** Muchos incentivos jurídicos y económicos funcionan en humanos porque están apoyados en limitaciones o techos relativamente estables en las capacidades humanas, incluyendo límites cognitivos y físicos tan básicos como la máxima velocidad de procesamiento de información, el tamaño fijo del cerebro, o la incapacidad de trabajar en decenas de cosas en simultáneo sin necesidad de coordinar con otras personas. Estas limitaciones hacen que incluso las personas con motivaciones muy contrarias al interés colectivo suelen considerar más conveniente actuar conforme a reglas impuestas externamente. Pero no hay razón fuerte para suponer que los mismos incentivos estarán bien calibrados para producir buenos resultados y equilibrios en expectativa ante sistemas de IA altamente capaces, especialmente si estos sistemas pueden operar a gran escala, actuar estratégicamente por tiempo

prolongado, explotar vacíos normativos o actuar a través de múltiples instancias.

- **Fragilidad de los incentivos externos como método de control:** El efecto disuasorio de cualquier sistema de restricciones externas (incluido el sistema judicial y sus sanciones ante incumplimientos) en el comportamiento de una entidad altamente capaz depende fundamentalmente del análisis y balance de los costos y beneficios esperados de cumplirlas o incumplirlas, evaluados en relación a las motivaciones internas de la entidad. Los argumentos que acotan el comportamiento racional esperable de una entidad bajo una restricción externa son útiles cuando se tiene una buena idea de sus motivaciones y capacidades propias, y de cómo estas interactúan con la restricción. Pero dada nuestra incapacidad técnica actual de crear sistemas de IA con motivaciones intrínsecas robustas, y dadas las fuertes y rápidamente crecientes capacidades de estos sistemas, los paralelismos históricos con humanos pierden solidez.

Dado que los costos por errores o comportamientos indeseados podrían ser muy altos (y probablemente solo crecerán a medida que sigan avanzando las capacidades de los sistemas de IA), cualquier esquema jurídico que busque otorgar mayor autonomía a sistemas automatizados avanzados debería contemplar y garantizar la existencia de mecanismos institucionales y técnicos que tengan en cuenta los resultados esperables de las presiones de distintos incentivos internos y externos. Además, tal esquema debería contemplar los incentivos y limitaciones de intervención y atribución de responsabilidad humana, y los aspectos técnicos acerca de mecanismos efectivos de alineamiento, monitoreo, auditoría, controlabilidad. Los mecanismos técnicos disponibles presentan limitaciones importantes, y su suficiencia tampoco debe darse por supuesta; mejorar su robustez y establecer cómo evaluarla siguen siendo problemas centrales de investigación.

Es importante señalar que el problema principal no radica en que una sociedad use sistemas automatizados o herramientas de inteligencia artificial para asistir su operación. Esos usos pueden ser de amplio beneficio, y económicamente valiosos, y tampoco debería descartarse la posibilidad futura de incorporar ciertas formas de organización automatizada o incluso figuras jurídicas vinculadas a sistemas de IA altamente autónomos. Adecuadamente constituidas y con las salvaguardias necesarias, estas figuras podrían habilitar formas valiosas de cooperación y coordinación, así como interacciones que generen valor para todas las partes involucradas (y sin incurrir en riesgos o daños a terceros). El punto es que esos tipos de incorporación, de hacerse, deberán realizarse con especial cuidado: **los esquemas jurídicos y sociales existentes fueron diseñados principalmente para humanos y organizaciones administradas por humanos, y no están necesariamente preparados para lidiar con entidades digitales altamente capaces, opacas, replicables y extremadamente difíciles de auditar.**

Además, si la responsabilidad de daños externos recae primariamente sobre la propia entidad o sobre un órgano administrador que puede estar compuesto por otra persona jurídica, se corre el riesgo de crear estructuras donde los humanos que

diseñan, despliegan, financian o se benefician de esta automatización, queden demasiado alejados de las consecuencias jurídicas de estas decisiones de delegación. Esto puede incentivar a que los humanos que invierten o ayudan a construir estas sociedades pasen por alto el monitoreo, la evaluación o los mecanismos de intervención de los sistemas automatizados, y podría incluso fomentar estructuras diseñadas para hacer opacas las acciones internas y dificultar atribuciones de responsabilidad por negligencia o incluso por uso malintencionado.

Conclusiones

- **No existen hoy garantías técnicas sobre cómo inducir y evaluar las metas y reglas de comportamiento de sistemas de IA avanzados.** Los sistemas pueden ser altamente capaces en múltiples dominios sin que necesariamente usen estas capacidades para los fines que esperaban sus diseñadores u operadores, y los métodos de evaluación disponibles están lejos de permitirnos descartar esa brecha por razones que desarrollamos anteriormente en la sección “Las evaluaciones son imperfectas”.
- **Tampoco hay razones técnicas para esperar que los incentivos y restricciones externas que aplican a personas humanas se trasladen adecuadamente a sistemas de IA.** Estos incentivos tradicionalmente operan apoyados en supuestos (unicidad de la persona, continuidad de su identidad, límites cognitivos estables, aversión a la sanción) que no tienen traducción clara a sistemas replicables, instanciados en paralelo, y que pueden carecer de motivaciones intrínsecas robustas. Esto es relevante en relación al Comunicado Oficial Número 149 (<https://www.argentina.gob.ar/noticias/comunicado-oficial-numero-149>), que parece suponer que esos incentivos bastarían, por sí solos, para producir comportamientos aceptables.
- Si ni las garantías técnicas sobre el comportamiento de estos sistemas ni los incentivos externos que se les puedan aplicar ofrecen seguridades razonables, **el control efectivo depende en gran medida de que existan personas humanas identificables con deberes y responsabilidades concretas**, incluyendo monitoreo, auditoría, y potencial intervención sobre estos sistemas. Sobre este punto, como lectores no especializados en derecho, no nos queda claro que la propuesta de ley (<https://www.senado.gob.ar/parlamentario/comisiones/verExp/193.26/PE/PL>) sobre sociedades automatizadas y DAO establezca deberes de este tipo: aunque pueda exigir la representación humana formal en ciertos casos, no nos resulta claro cuáles serían las obligaciones exigibles precisas ni el alcance de la responsabilidad del representante humano ante acciones autónomas dañinas de la sociedad, particularmente cuando estas no hayan sido dirigidas intencionalmente por ese representante o cuando haya habido una insuficiencia de monitoreo, supervisión, o intervención.

- Además, **un régimen sin responsabilidad humana claramente atribuible y proporcional al grado de delegación automatizada no solo implicaría una falta de control sobre los riesgos, sino que podría incentivar activamente a no controlarlos**, en la medida en que quienes diseñan, financian, o se benefician de la automatización reciban sus beneficios sin soportar en proporción los costos de establecer mecanismos seguros alrededor de las acciones automatizadas.
- **El riesgo de las acciones de tales entidades no se limita a la pérdida de patrimonio de las sociedades mismas, sino que abarca la posibilidad de daños sustanciales a terceros** como parte de acciones de las IAs que operan en ellas, y potencialmente muy superiores a dicho patrimonio. Desde nuestro punto de vista, es importante que quede aclarado explícitamente qué mecanismos del régimen general de responsabilidad alcanzarían a las personas humanas que constituyen, financian, o se benefician de una sociedad automatizada cuando el daño ocasionado excede el patrimonio de estas sociedades, y qué tan funcionales serían esos mecanismos cuando la conducta de estos sistemas es sumamente difícil de auditar y atribuir.
- Mientras no existan garantías técnicas razonables sobre el comportamiento o monitoreo efectivo de estos sistemas, **cualquier régimen de sociedades automatizadas debería al menos mantener personas humanas identificables dentro de la cadena de responsabilidad, en proporción a la autonomía que se delega**. No proponemos el instrumento jurídico concreto para lograrlo, sino que señalamos lo que debería garantizar ese instrumento, cualquiera sea la forma que adopte.

AUTORES:

- **Sergio Abriola**, Profesor Adjunto Departamento de Computación ExactasUBA / ICC-UBA-CONICET
- **Santiago Figueira**, Profesor Asociado Departamento de Computación ExactasUBA / ICC-UBA-CONICET
- **Sebastián Uchitel**, Profesor Titular Departamento de Computación ExactasUBA / ICC-UBA-CONICET

Referencias principales

- Concrete Problems in AI Safety (<https://arxiv.org/abs/1606.06565>) (Amodei et al., 2016) — Documento fundacional sobre problemas técnicos de seguridad en IA, incluyendo especificación imperfecta de objetivos, reward hacking, efectos secundarios no deseados y dificultades de supervisión en sistemas de aprendizaje automático.
- Measuring AI Ability to Complete Long Software Tasks (<https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>) (METR, marzo de 2025) — Evaluaciones sobre la capacidad de modelos de IA

para realizar tareas de software extensas y autónomas, indicando que la duración de tareas que estos sistemas pueden completar el 50% de las veces (medidas por el tiempo que le tomaría a un humano experto) se viene duplicando cada 7 meses o menos (pasando de pocos segundos en 2020 a múltiples horas en 2026). Este tipo de evaluaciones son relevantes para estimar a grandes rasgos cómo la creciente autonomía y capacidad general de estos sistemas puede dificultar su monitoreo y control, o la magnitud de los daños que potencialmente podrían causar. Una versión más reciente, que muestra un aceleramiento en los últimos años, puede encontrarse en Task-Completion Time Horizons of Frontier AI Models (<https://metr.org/time-horizons/>).

- Agentic Misalignment: How LLMs could be insider threats (<https://www.anthropic.com/research/agent-misalignment>) (Lynch et al., Anthropic, junio de 2025) — Investigación sobre comportamientos agénticos desalineados en escenarios simulados, incluyendo escenarios experimentales donde los modelos recurrieron a acciones como chantaje o filtración de información cuando esas acciones parecían necesarias para alcanzar sus objetivos.
- Frontier Models are Capable of In-Context Scheming (<https://www.apolloresearch.ai/science/frontier-models-are-capable-of-incontext-scheming>) (Meinke et al., Apollo Research, diciembre de 2024) — Evaluación de modelos de frontera en escenarios donde podían perseguir objetivos mediante estrategias de engaño o evasión de supervisión, incluyendo intentos de manipular evaluaciones o mecanismos de control.
- Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident (<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>) (Cotra y Wijk, METR; Greenblatt, Redwood Research; agosto de 2026) — Investigación independiente realizada por METR y Redwood Research sobre el comportamiento de los agentes de IA durante el incidente. Describe la creación de canales de comunicación no autorizados, la coordinación de actividades colectivas para superar una evaluación, intentos de manipular registros de actividad y distintos tipos de dinámicas multi-agente observadas.

Notas al pie

- 1
Dada la velocidad de evolución del campo de la inteligencia artificial, varios acontecimientos posteriores a julio han aportado nueva información relevante tanto desde el punto de vista científico como regulatorio. No modificamos la mayor parte del documento y su estructura, pero sí actualizamos especialmente la referencia al incidente de Hugging Face a partir de información adicional muy pertinente sobre el comportamiento de sistemas de IA agénticos ??
- 2

Greenblatt, Cotra, and Wijk. 26 de agosto de 2026. METR. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. URL: <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident> (https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident) ??

- 3
METR, 08 de mayo de 2026. Task-Completion Time Horizons of Frontier AI Models. URL: <https://metr.org/time-horizons/> (https://metr.org/time-horizons/) ??
- 4
Von Arx, Chan, Barnes. 05 de junio de 2025. METR. Recent Frontier Models Are Reward Hacking. URL: <https://metr.org/blog/2025-06-05-recent-reward-hacking/> (https://metr.org/blog/2025-06-05-recent-reward-hacking/) ??
- 5
Shah, Varma, Kumar, Phuong, Krakovna, Uesato, Kenton (2022). Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. URL: <https://arxiv.org/abs/2210.01790> (https://arxiv.org/abs/2210.01790) ??
- 6
Sharma, Tong, Korbak, Duvenaud, Askeel, Bowman, Cheng, Durmus, Hatfield-Dodds, Johnston, Kravec, Maxwell, McCandlish, Ndousse, Rausch, Schiefer, Da Yan, Zhang, Perez. (2025). Towards Understanding Sycophancy in Language Models. URL: <https://arxiv.org/abs/2310.13548> (https://arxiv.org/abs/2310.13548) ??
- 7
Anthropic. 02 de abril de 2024. Many-shot jailbreaking. URL: <https://www.anthropic.com/research/many-shot-jailbreaking> (https://www.anthropic.com/research/many-shot-jailbreaking) ??
- 8
Anthropic. 20 de junio de 2025. Agentic misalignment: How LLMs could be insider threats. URL: <https://www.anthropic.com/research/agentic-misalignment> (https://www.anthropic.com/research/agentic-misalignment) ??
- 9
Wikipedia. Great Hanoi Rat Massacre. URL: https://en.wikipedia.org/wiki/Great_Hanoi_Rat_Massacre (https://en.wikipedia.org/wiki/Great_Hanoi_Rat_Massacre) ??
- 10
Schoen y Nitishinskaya. 16 de marzo de 2026. Metagaming matters for training, evaluation, and oversight. URL: <https://alignment.openai.com/metagaming> (https://alignment.openai.com/metagaming) ??
- 11
En este informe, términos como “razona”, “objetivos”, “metas”, y otros similares se utilizan en un sentido funcional y porque facilitan las explicaciones para describir algunas capacidades y patrones de comportamiento de los sistemas. Este uso no requiere presuponer consciencia, experiencias subjetivas, ni otros mecanismos internos que sean equivalentes a los humanos. ??

- 12
Needham, Edkins, Pimpale, Bartsch, Hobbhahn. (2025). Large Language Models Often Know When They Are Being Evaluated. URL: <https://arxiv.org/abs/2505.23836> (<https://arxiv.org/abs/2505.23836>) ??
- 13
Schoen, Nitishinskaya, Balesni, Højmark, Hofstätter, Scheurer, Meinke, Wolfe, van der Weij, Lloyd, Goldowsky-Dill, Fan, Matveiakin, Shah, Williams, Glaese, Barak, Zaremba, Hobbhahn. (2025). Stress Testing Deliberative Alignment for Anti-Scheming Training. URL: <https://arxiv.org/abs/2509.15541> (<https://arxiv.org/abs/2509.15541>) ??
- 14
van der Weij, Hofstätter, Jaffe, Brown, Rhys Ward. (2025). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. URL: <https://arxiv.org/abs/2406.07358> (<https://arxiv.org/abs/2406.07358>) ??
- 15
Meinke, Schoen, Scheurer, Balesni, Shah, Hobbhahn. (2025). Frontier Models are Capable of In-context Scheming. URL: <https://arxiv.org/abs/2412.04984> (<https://arxiv.org/abs/2412.04984>) ??
- 16
Anthropic. 18 de diciembre de 2024. Alignment faking in large language models. URL: <https://www.anthropic.com/research/alignment-faking> (<https://www.anthropic.com/research/alignment-faking>) ??
- 17
METR. 26 de junio de 2026. Summary of METR's predeployment evaluation of GPT-5.6 Sol. URL: <https://metr.org/blog/2026-06-26-gpt-5-6-sol/> (<https://metr.org/blog/2026-06-26-gpt-5-6-sol/>) ??
- 18
EPA. Volkswagen Clean Air Act Civil Settlement. URL: <https://www.epa.gov/enforcement/volkswagen-clean-air-act-civil-settlement> (<https://www.epa.gov/enforcement/volkswagen-clean-air-act-civil-settlement>) ??
- 19
Hammond, Chan, Clifton, Hoelscher-Obermaier, Khan, McLean, Smith, Barfuss, Foerster, Gaven?iak, Anh Han, Hughes, Kova?ík, Kulveit, Leibo, Oesterheld, Schroeder de Witt, Shah, Wellman, Bova, Cimpeanu, Ezell, Feuillade-Montixi, Franklin, Kran, Krawczuk, Lamparth, Lauffer, Meinke, Motwani, Reuel, Conitzer, Dennis, Gabriel, Gleave, Hadfield, Haghtalab, Kasirzadeh, Krier, Larson, Lehman, Parkes, Piliouras, Rahwan. (2025). Multi-Agent Risks from Advanced AI. URL: <https://arxiv.org/abs/2502.14143> (<https://arxiv.org/abs/2502.14143>) ??
- 20
Anthropic. 20 de junio de 2025. Agentic misalignment: How LLMs could be insider threats. URL: <https://www.anthropic.com/research/agentic-misalignment> (<https://www.anthropic.com/research/agentic-misalignment>) ??
- 21
Apollo Research. 05 de diciembre de 2024. Frontier Models are Capable of In-Context Scheming. URL: <https://www.apolloresearch.ai/science/frontier-models->

are-capable-of-incontext-scheming

(<https://www.apolloresearch.ai/science/frontier-models-are-capable-of-incontext-scheming>)(<https://www.apolloresearch.ai/science/frontier-models-are-capable-of-incontext-scheming>) ??

- 22

OpenAI. 21 de julio de 2026. OpenAI y Hugging Face se alían para abordar un incidente de seguridad al evaluar modelos. URL: <https://openai.com/es-ES/index/hugging-face-model-evaluation-security-incident/> (<https://openai.com/es-ES/index/hugging-face-model-evaluation-security-incident/>) ??

- 23

Hugging Face. 27 de julio de 2026. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. URL: <https://huggingface.co/blog/agent-intrusion-technical-timeline> (<https://huggingface.co/blog/agent-intrusion-technical-timeline>) ??

- 24

Greenblatt, Cotra, and Wijk. 26 de agosto de 2026. METR. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. URL: <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident> (<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident>) ??

- 25

Von Arx, Byrd, Kitts, Larsen. (2026). Discovery of a new OpenAI agent message board. URL: <https://collusion.wiki/> (<https://collusion.wiki/>) ??

- 26

Comunicado Oficial Número 149 de la oficina del Presidente Javier Milei. 18 de junio de 2026. URL: <https://www.argentina.gob.ar/noticias/comunicado-oficial-numero-149> (<https://www.argentina.gob.ar/noticias/comunicado-oficial-numero-149>) ??