Dr. Lautaro N. Rubbi

Inteligencia artificial
Entre las fricciones y la expansión de la zona gris

CARI / CONSEJO ARGENTINO PARA LAS RELACIONES INTERNACIONALES

Artículo de opinión
Octubre 2026

Inteligencia artificial
Entre las fricciones y la expansión de la zona gris

Dr. Lautaro N. Rubbi

Artículo de opinión
Octubre 2026

Consejo Argentino para las Relaciones Internacionales

Artículo de opinión
Octubre 2026

Las opiniones expresadas en esta publicación son exclusiva responsabilidad de los autores y no reflejan ni la visión de las instituciones a las que pertenecen ni la del CARI.

Corrección: María Fernanda Rey
Diseño: Juan José Gómez
Imagen de tapa: iStock.com/Iryna Olkhova

CARI Consejo Argentino para las Relaciones Internacionales
Uruguay 1037, piso 1.º, C1016ACA Buenos Aires, República Argentina
Teléfono: (+5411) 4811-0071 al 74 / Fax: (+5411) 4815-4742
Correo electrónico: direccioneditorial@cari.org.ar / Sitio web: www.cari.org.ar

Inteligencia artificial
Entre las fricciones y la expansión de la zona gris

Dr. Lautaro N. Rubbi*

Introducción

En el *Global Risks Report 2026* del Foro Económico Mundial, basado en una encuesta a más de 1300 expertos, la desinformación figura como el segundo riesgo más severo a dos años, el conflicto armado entre Estados como el quinto y la inseguridad cibernética como el sexto. Los resultados adversos de la inteligencia artificial (IA) apenas ocupan el puesto 30 en ese horizonte, aunque ascienden al quinto a diez años (Foro Económico Mundial, 2026). Sin embargo, varios episodios del 2026 muestran que la IA ya opera dentro de los riesgos que encabezan las listas, ampliando sus horizontes y difuminando sus fronteras conceptuales.

* Doctor en Estudios Internacionales (Universidad Torcuato Di Tella). Estudios de Posdoctorado en Ciencias Sociales (UBA). Consultor. Strategic Research Lead, Futures Intelligence Team, PwC US. Coordinador del Observatorio de Defensa y Seguridad Internacional (ODyS, CEPI-UBA). Coordinador del Laboratorio de Inteligencia Artificial y Relaciones Internacionales (LIARI, CIPEA-UBA). Docente de posgrado (UBA, UCEMA, ESGN/UNDEF, NPSGlobal). Cofundador y director de Inteligencia Estratégica y Prospectiva, Estudio ARG. Correo de contacto: lnr_rubbi@hotmail.com

El más elocuente proviene de un informe reciente que Anthropic, la empresa que desarrolla el modelo de IA Claude, publicó el 10 de septiembre de 2026 sobre los usos indebidos que detectó en su plataforma.[1] Un pequeño grupo radicado en el norte de Yemen, cuya filiación el informe no establece, llevaba adelante tres programas de armas, entre ellos un misil balístico de varias etapas con un alcance buscado de más de 2000 kilómetros. Para desarrollar el *software* que guía y estabiliza un vehículo en vuelo, confió a Claude las tareas propias de un equipo de ingenieros y las distribuyó entre varias instancias del sistema, como un jefe de proyecto que reparte funciones entre especialistas: una escribía el código, otra investigaba y una tercera revisaba lo escrito. Con ese *software* disparó un cohete guiado. La prueba, al parecer, fracasó, y a las pocas horas el grupo volvió a consultar al asistente para averiguar por qué. No hay pruebas de que haya llegado a desplegar un arma operativa, pero, cuando la empresa bloqueó sus cuentas, ya contaba con herramientas de simulación avanzadas que funcionaban sin el sistema (Anthropic, 2026b, pp. 112-115).

El informe describe también a un equipo de programadores independientes radicado en Rusia, que alternaba encargos civiles y militares y que desde mayo de 2026 empleó el mismo sistema para desarrollar el *software* de un enjambre de drones kamikaze autónomos. Según la AFP, el enjambre estaba

1 La AFP, Reuters y Associated Press recogen los rasgos centrales del caso, incluida la prueba al parecer fallida (Agence France-Presse, 2026a; Baptista y Vicens, 2026; El Deeb, 2026). El informe no menciona a los hutíes, que controlan buena parte del norte de Yemen; Associated Press situó el caso en territorio bajo su control, la AFP juzgó probable su participación y un miembro del buró político del movimiento calificó de "irrazonable e ilógico" que recurriera a fuentes abiertas para desarrollar capacidades militares (El Deeb, 2026). La distancia entre la sospecha que despierta una geografía y la prueba que exige una atribución es parte del problema que recorre estas páginas.

diseñado para elegir sus blancos y estrellarse contra ellos sin que un ser humano tomara la decisión final y, al programarlo, el equipo fijó como objetivo, una y otra vez, un punto de la región ucraniana de Donetsk (Agence France-Presse, 2026b). Los blancos que podía seleccionar el sistema de IA instalado a bordo incluían la categoría "persona". Aunque el equipo afirmaba haber recibido fondos del Ministerio de Defensa, entre otros organismos, la empresa no pudo verificarlo y no lo considera una entidad del Estado ruso (Anthropic, 2026b).

La tendencia es más amplia. En su evaluación de la amenaza del crimen grave y organizado, Europol (2025) advirtió que la IA acelera la delincuencia organizada, que los modelos generativos redujeron de manera drástica la barrera de entrada a los delitos digitales y que las redes criminales sirven cada vez más como intermediarias de actores de amenazas híbridas. Ese mismo año, la Armada de Colombia anunció la incautación del primer semisumergible no tripulado del narcotráfico en el país, operable a distancia y equipado con una antena de Starlink (Agence France-Presse, 2025; Armada de Colombia, 2025).

Estos hechos comparten tres rasgos. El primero es el actor, que en ningún caso es un Estado o puede probarse de forma verificable y/o en un plazo pertinente que lo sea. El segundo es la herramienta comercial, de propósito general, al alcance de cualquiera que sepa usarla o eludir sus controles y capaz de reemplazar el trabajo calificado (ingenieros, programadores, operadores especializados) que antes solo una organización grande podía reunir y costear. El tercero es el umbral. Ninguno constituye un acto de guerra: un *software* de guiado, un enjambre probado en simulación o una embarcación hallada vacía quedan por debajo del punto a partir del cual un Estado

se juzga autorizado a responder con la fuerza. Pero acercan a actores no estatales a capacidades que eran militares y erosionan el monopolio práctico del Estado sobre ciertas formas de violencia organizada.

La IA complica, además, la pregunta más antigua de la responsabilidad civil o penal en una zona gris: quién fue. Durante décadas, la sofisticación técnica de una operación funcionó como una firma involuntaria, porque los aparatos de inteligencia eran de unos pocos Estados que podían costear ciertas capacidades. Esa inferencia pierde fuerza. En *Foreign Affairs*, Brianna Rosen y Jam Kraprayoon (2026) anticiparon que Gobiernos, empresas y personas enfrentarán pronto agentes de IA capaces de conducir por su cuenta campañas cibernéticas comparables a las de los Estados hoy más capaces. El informe de Anthropic documenta campañas de intrusión, contra múltiples víctimas, que un año antes habrían exigido numerosos operadores calificados y que ahora sostenían, por separado, un hacktivista con claves robadas de acceso a servicios de IA, delincuentes en busca de rédito y un operador de espionaje estatal. A juicio de la empresa, la IA ya niveló el terreno entre los pequeños grupos criminales y las organizaciones patrocinadas por Estados, y lo que principalmente los distingue ha pasado a ser la intención (Anthropic, 2026b). Y la intención no queda inscrita en el código, sino que debe inferirse de los blancos elegidos, del destino de lo sustraído o de quién financió la operación, una inferencia lenta que rara vez alcanza la certeza que exigen el derecho internacional y la diplomacia.

Los agentes de IA añaden otra complicación, de alcance todavía mayor. Un agente es un sistema que recibe un objetivo y lo persigue por su cuenta, encadenando tareas (buscar in-

formación, escribir código, probarlo, corregirlo) con escasa supervisión humana; el enjambre ruso representa su forma extrema, porque estaba concebido para decidir el blanco y el momento de la detonación. Su novedad se mide mejor contra una práctica antigua: quien desea negar su participación en un acto de fuerza suele recurrir a intermediarios. Andreas Krieg y Jean-Marc Rickli (2018) llamaron guerra subrogada a esa transferencia de la carga de la guerra hacia sustitutos humanos o tecnológicos, que permite a un Estado gestionar a distancia los riesgos del conflicto.

El agente artificial es un sustituto de un tipo inédito. Actúa con cierta autonomía, opera a una velocidad que ningún intermediario humano iguala y escapa a todo juicio, porque ningún tribunal puede tomarle declaración ni condenarlo. Cuando algo salga mal, quien lo puso en marcha podrá alegar que la decisión fue del sistema, y refutarlo exigirá reconstruir qué le ordenó, qué podía prever y qué estaba en condiciones de controlar. La vieja coartada del intermediario que se excedió en su mandato, ya visible en las milicias cibernéticas (Kello, 2013), encuentra así una versión a la medida de la época: el agente se salió del guion.

Este artículo sostiene que la IA, junto con otras tecnologías de uso dual, reduce a la vez las cuatro fricciones que durante siglos encarecieron la coerción organizada: la informacional, la coordinativa, la logística y la de atribución. Esa reducción baja la barrera de entrada a disrupciones de alto calibre, las pone al alcance de actores no estatales y de Estados que operan por medio de intermediarios, humanos o artificiales, y produce una economía del intento desfavorable para el defensor. La zona gris, cuya eficacia descansaba en la fricción de atribución, desborda así el repertorio de los Estados revisionistas y tiende a volverse una condición permanente del sistema internacional.

Con esa expansión quedan bajo tensión los conceptos con que la disciplina de las relaciones internacionales pensó la coerción: la zona gris, la guerra híbrida, la disuasión y aun la guerra y la seguridad.[2] El argumento avanza en cuatro pasos: la zona gris y su frontera con la guerra híbrida; las cuatro fricciones, aporte central del texto, y la economía del intento resultante; los agentes artificiales, entre la herramienta y el actor y, en las conclusiones, la transformación de esos conceptos y las implicancias para la Argentina.

1. La zona gris y su frontera con la guerra híbrida

La distinción entre la naturaleza y el carácter de la guerra ordena la discusión (Gray, 2005): la naturaleza, el choque de voluntades organizadas que emplean la violencia con fines políticos, permanece; el carácter, las formas que la guerra adopta en cada época, varía. La zona gris y la guerra híbrida describen transformaciones del carácter. En el debate públi-

2 Varios de los temas aquí desarrollados, entre ellos los umbrales que un defensor debe cruzar antes de responder, fueron abordados por el grupo de investigación del proyecto A-ESGN-0003/26, "La nueva gramática del poder: disuasión y defensa en la era de la inteligencia artificial y la guerra híbrida", de la Escuela de Guerra Naval (ESGN-FADARA-UNDEF), dirigido por el autor e integrado por Maximiliano Gastón Fraire y Bernardo Collado, con la colaboración de Clara María Bozzo. Esos temas forman parte de la contribución del equipo a un libro colectivo en preparación del Observatorio de Defensa y Seguridad (ODyS) del Centro de Estudios de Política Internacional (CEPI) de la Universidad de Buenos Aires (UBA), sobre actores no estatales y seguridad internacional, cuyos capítulos, en particular los de Becerril, Llambí, Micha y Farías, y Niss, nutrieron también algunas formulaciones de este texto. Del capítulo de Fraire, Bozzo y Collado provienen el criterio de naturaleza y carácter para ordenar el debate, la genealogía de la guerra híbrida y la atribución como bisagra. La tipología de las fricciones, la escala sin estructura, la asimetría decisional y el precio de la legitimidad provienen de la introducción y la conclusión que el autor escribió como coordinador del libro. La idea de los relojes distintos procede de un trabajo del autor, en preparación, sobre la desincronización del sistema internacional.

co se usan como sinónimos y esa confusión oculta lo que la IA cambia: el umbral y la atribución. Este análisis se organiza en torno a la zona gris, que designa a la vez una estrategia de coerción y la franja por debajo del umbral de la guerra en que esa estrategia opera; la guerra híbrida, de matriz estatal y militar, sirve de contraste para fijar sus límites.

La **zona gris** hunde sus raíces en la guerra política que el equipo de planeamiento de George Kennan propuso en 1948. Hal Brands (2016) dio a su versión contemporánea una definición precisa: una actividad coercitiva y agresiva, diseñada deliberadamente para quedar por debajo del umbral del conflicto militar convencional y de la guerra abierta entre Estados. En esencia, es una forma de conducir un conflicto sin provocar la respuesta plena del adversario. Tres elementos son necesarios y, juntos, suficientes:

- **Un fin revisionista.** El actor busca alterar un *statu quo* (un territorio, una regla, una relación de fuerzas) sin pagar el costo de una guerra; Michael Mazarr (2015) habló de revisionistas de ambiciones medidas.
- **Calibración deliberada.** Cada acto se diseña para quedar por debajo del umbral del conflicto militar convencional (Brands, 2016), es decir, del umbral de respuesta: el punto a partir del cual el afectado se juzga autorizado a reaccionar con la fuerza.
- **Gradualismo.** El avance se fracciona en pasos que se acumulan en el tiempo, ninguno de los cuales justifica por sí solo una respuesta mayor (Mazarr, 2015).

Un cuarto rasgo, la **ambigüedad** sobre qué pasó, quién lo hizo o con qué intención, es frecuente pero prescindible. En febrero de 2023, un buque de la guardia costera china apuntó un

láser contra otro de la guardia costera filipina que cumplía una misión de reabastecimiento en el mar de la China Meridional. La autoría era visible y, aun así, el acto quedó por debajo del umbral (Ibarra y Arugay, 2024). La ambigüedad multiplica la eficacia de la zona gris, no la define.

La **guerra híbrida** responde a otra pregunta. Frank Hoffman le dio su formulación canónica en 2007: amenazas que combinan capacidades convencionales, tácticas y formaciones irregulares, actos terroristas y desorden criminal, en guerras en las que esas fuerzas "se desdibujan en una misma fuerza, en el mismo espacio de batalla" (p. 8).[3] Lo nuevo era la fusión de componentes que otras guerras habían mantenido en teatros o formaciones separados (p. 8), y su prototipo fue Hezbolá en la guerra del Líbano de 2006 (p. 35).

La guerra híbrida requiere una pluralidad de modos de lucha, su integración en un mismo diseño operacional y la existencia de operaciones militares en un teatro, aunque la violencia sea baja. Es un concepto del plano operacional, cuyo sujeto típico es un Estado o una organización armada con capacidades de tipo estatal. Después de Crimea, en 2014, el término se estiró hasta designar casi cualquier presión rusa,[4] una inflación que le restó precisión. Las dos nociones pueden darse juntas o por separado, como sintetiza la tabla 1.[5]

3 Las citas de fuentes publicadas en inglés son traducción propia.

4 Hasta un texto, originalmente publicado en 2013, del jefe del Estado Mayor ruso, Valeri Gerasimov, cuando pasó a llamarse doctrina, un rótulo que había acuñado Mark Galeotti y que él mismo lamentó después en público (Galeotti, 2018).

5 La Unión Europea llama amenazas híbridas a la mezcla de actividades coercitivas y subversivas, convencionales y no convencionales, que actores estatales o no estatales pueden emplear de forma coordinada "manteniéndose por debajo del umbral de una guerra declarada oficialmente" (Comisión Europea y Alta Representante de la Unión para Asuntos Exteriores y Política de Seguridad, 2016, p. 2). Con los criterios propuestos, esa definición describe la zona gris.

Tabla 1. Zona gris y guerra híbrida

Rasgo	Zona gris	Guerra híbrida
Pregunta a la que responde	Cómo se dosifica la presión para que el adversario no responda con todas sus fuerzas.	Cómo se combina la fuerza en un teatro de operaciones.
Plano	Estratégico.	Operacional.
Lugar de la ambigüedad	Frecuente: multiplica su eficacia.	Accesoria: puede ocultar quién combate, mientras el combate permanece a la vista.
Un concepto sin el otro	Sabotajes, incendios y otras acciones de Rusia en Europa (3 ataques en 2022, 12 en 2023 y 34 en 2024, según el Center for Strategic and International Studies), sin operaciones militares en un teatro.	Hezbolá en 2006 y la invasión rusa de Ucrania en 2022, precedida en una hora por el ciberataque contra la red satelital KA-SAT: guerra abierta, sin dosificación del umbral.
Los dos a la vez: Crimea, 2014	Negación de la participación rusa, aunque la procedencia de los atacantes era evidente, para ganar tiempo e imponer un hecho consumado.	Soldados rusos, separatistas locales prorrusos, medios y diplomacia en acción coordinada; anexión casi sin combate.

Nota. Elaboración propia.

La teoría clásica ya explicó por qué la coerción migra por debajo del umbral de la guerra. Glenn Snyder (1965) advirtió lo que la literatura llamaría "la paradoja de estabilidad e inestabilidad": cuanto más estable es el equilibrio del terror, menos estable resulta el equilibrio en los niveles bajos de violencia. Mientras la disuasión nuclear contiene la guerra mayor, la presión desciende a los virus informáticos, la desinformación, los incendios y los sabotajes encargados a desconocidos por aplicaciones de mensajería. Y desciende fraccionada. Thomas Schelling (1966) describió el arte de obtener por pasos sucesivos lo que no podría exigirse de una sola vez, cada uno demasiado pequeño para justificar una reacción.

La eficacia de ese fraccionamiento descansa en un rasgo institucional del umbral de respuesta, que se fija caso por caso: ante un ataque armado, un Estado sabe qué hacer; ante un incendio sin autor, un cable submarino cortado por un ancla o una intrusión informática sin firma, primero tiene que probar que hubo un ataque y de quién. Entre el hecho y la respuesta median varios umbrales sucesivos (detectar, atribuir, establecer la jurisdicción y decidir), y la secuencia admite cortes: un hecho puede detectarse y quedar sin atribuir, o atribuirse y quedar sin respuesta.

Por eso la atribución es la bisagra de la zona gris. Responder con la fuerza, castigar o pedir apoyo a los aliados exige un destinatario, y el destinatario solo existe cuando alguien establece, con la prueba que cada uso requiere, quién actuó y con qué intención. Una atribución instantánea y segura daría a cada paso un destinatario; una lenta o disputable deja al defensor frente a hechos sin dueño. Thomas Rid y Ben Buchanan (2015) explicaron que atribuir es un proceso escalonado,

de oficio técnico en la base y juicio político en la cima, y condensaron su tesis en una fórmula: la atribución es lo que los Estados hacen de ella.

La fórmula tiene una cara menos visible. Austin Carson (2018) mostró que las grandes potencias casi siempre detectan las intervenciones encubiertas de sus rivales y que, cuando comparten el interés de limitar una guerra, pueden guardar el secreto de lo detectado o ignorar lo que ya es un secreto a voces. La falta de atribución puede ser una elección del defensor, que prefiere no saber oficialmente lo que sabe para no quedar obligado a responder. Rory Cormac y Richard Aldrich (2018) completaron el cuadro desde el lado de quien actúa. Negar rinde, aunque la negación resulte poco creíble: permite exhibir determinación sin escalar a un conflicto militar y, como el avance es gradual, el daño ya está hecho cuando el blanco por fin puede investigar o responder.

La zona gris prospera, en suma, tanto en la incapacidad de atribuir como en la renuencia a hacerlo. Mientras que la segunda responde a motivaciones más bien políticas o estratégicas, el impacto de la IA ensancha la primera. De ahí la pregunta que organiza el resto del argumento: qué ocurre cuando una misma tecnología abarata, a la vez, saber, coordinar, abastecerse y ocultarse.

2. Las cuatro fricciones de la coerción organizada

Clausewitz (1832/1984) llamó fricción a lo que separa la guerra real de la guerra en el papel. “Todo en la guerra es muy simple, pero lo más simple es difícil”, escribió, porque las dificultades

se acumulan y producen una fricción inconcebible para quien no haya vivido la guerra.

En sentido amplio, una fricción es todo factor que eleva el costo o la dificultad de ejercer coerción organizada: el tiempo, el dinero, el personal calificado, la organización y el riesgo que un actor debe comprometer para convertir una intención hostil en un efecto. Esos costos no pesan igual para todos. Un Estado puede absorberlos con presupuestos, burocracias y escala; un grupo pequeño, casi nunca. Las múltiples fricciones operativas de la coerción organizada operaron durante siglos como una barrera de entrada que protegió al Estado frente a quienes pretendían disputarle el uso de la fuerza.

Esa barrera sostuvo una de las construcciones más costosas de la modernidad política. El monopolio de la violencia física legítima, que Max Weber puso en el centro de su definición del Estado, fue un resultado histórico. Los Estados europeos lo conquistaron, como reconstruyó Charles Tilly (1985), neutralizando a sus rivales y extrayendo recursos en una empresa parecida a un negocio de protección, y lo conservaron porque disputarlo resultaba demasiado caro. Cada ola de tecnologías abiertas, de la dinamita a la IA, volvió a desafiarlo al poner en manos de personas comunes medios de violencia masiva antes reservados al Estado (Cronin, 2020). Lo que mantuvo esos desafíos dentro de ciertos límites fue, sobre todo, el costo de cuatro fricciones.

Las cuatro fricciones descomponen una misma pregunta, qué le cuesta a un actor ejercer coerción organizada, en cuatro costos: saber, coordinar, abastecerse y ocultarse.

La objeción previsible es que esas fricciones venían cediendo desde antes, con internet, el cifrado, los drones comerciales y los criptoactivos. Es cierto, pero la IA difiere de sus predecesoras en un punto que altera la escala del fenómeno. Cada una de aquellas tecnologías abarató un insumo: internet, la información; el cifrado, la comunicación clandestina; los drones, el transporte de una carga; los criptoactivos, el pago sin rastro. La pericia necesaria para convertir esos insumos en capacidad siguió siendo el cuello de botella de la coerción organizada, y la IA actúa precisamente sobre él, porque sustituye trabajo cognitivo calificado. Lo hace, además, como tecnología de propósito general: el mismo sistema escribe el programa de guiado de un cohete, identifica las vulnerabilidades de una red, redacta propaganda en veinte idiomas y coordina a varios agentes en paralelo para distribuirla de forma segmentada a públicos focalizados. Como cada fricción incluye un costo de trabajo experto, una tecnología que abarata ese trabajo en todas sus formas reduce las cuatro a la vez.

La fricción informacional es el costo de saber: qué puede conocerse, sobre quién y a tiempo. Tiene dos capas, que la IA reduce de manera desigual. La primera es el conocimiento general sobre capacidades, sociedades y medios de comunicación. Amy Zegart (2022) analizó cómo la información de fuentes abiertas erosionó el dominio que los servicios de unas pocas potencias ejercían sobre la recolección y el análisis: hoy cualquiera con una conexión accede a imágenes satelitales, a programas de reconocimiento facial y al seguimiento de hechos en tiempo real. La IA lleva esa apertura un paso más allá, porque resume, traduce y cruza fuentes en segundos.

La segunda capa es más exigente y, para quien ataca, más valiosa. Comprende la inteligencia sobre un blanco (las vulnerabilidades de este sistema, de este puerto o de esta persona) y el saber especializado para convertirla en daño: escribir un código malicioso, guiar un dron o diseñar una campaña capaz de engañar a una audiencia determinada. Internet ya había abaratado la primera capa. La segunda, que exigía especialistas, tiempo y estudio, seguía separando a los aparatos estatales del resto. La IA empieza a ponerla al alcance de actores pequeños, aunque el proceso es todavía incipiente y las principales empresas tecnológicas intentan controlarlo. En 2026, el informe internacional sobre la seguridad de la IA registró mercados clandestinos que venden herramientas de IA listas para usar y programas de secuestro de datos generados con IA, que rebajan la habilidad necesaria para atacar (International AI Safety Report, 2026). El saber necesario para una disrupción de alto calibre, que antes costaba años de formación o el respaldo de un Estado, se vuelve disponible a demanda.[6]

La fricción coordinativa es el costo de coordinar: reclutar, instruir, sincronizar y encadenar los pasos de una operación entre personas que deben confiar unas en otras. Mancur Olson (1965) mostró que el interés compartido no basta para producir acción colectiva. Salvo que el grupo sea pequeño o que exista coerción u otro dispositivo especial, los individuos

6 La defensa también aprende: Anthropic (2026a) informó que, con unos cincuenta socios, encontró más de diez mil vulnerabilidades de gravedad alta o crítica, y que el límite pasó a ser la rapidez con que pueden "verificar, divulgar y parchear" lo hallado. Lohn y Jackson (2022) recuerdan que cerca del 80 % de las vulnerabilidades tiene un parche disponible el día en que se divulga, aunque alguien debe aplicarlo a tiempo (p. 2), y Maschmeyer (2026) sostiene que la IA rinde más en la detección defensiva que en el engaño ofensivo. Aun si la técnica terminara por favorecer a la defensa, persistiría la asimetría de costos institucionales.

racionales no actúan para lograr sus intereses comunes, porque cada uno prefiere que el esfuerzo lo hagan otros. Toda organización violenta necesita, por consiguiente, incentivos selectivos, disciplina y mando, y en la clandestinidad esos costos se multiplican. Jacob Shapiro (2013) describió el dilema del terrorista: los instrumentos que permiten controlar a los subordinados (cada memorando, cada llamada para verificar que una orden se entendió) crean vulnerabilidades que pueden costarles a los líderes la vida o la libertad. Coordinar a escondidas siempre fue también exponerse.

Las nuevas tecnologías reducen esa fricción por dos vías. La primera es la tercerización: la mensajería cifrada y las plataformas de trabajo a demanda permiten contratar ejecutores sin conocerlos. Según la base de datos del Center for Strategic and International Studies, la inteligencia militar rusa y otros organismos rusos recurren con frecuencia a reclutas locales, a los que en algunos casos captan por Telegram o por los chats de videojuegos populares (S. G. Jones, 2025).

La segunda vía, la sustitución, es la más novedosa y la de mayores consecuencias. El citado grupo yemení suplió con la IA el trabajo de un equipo de ingenieros. En el plano digital, sistemas multiagente ya ejecutan en paralelo el reconocimiento, la explotación y el robo de datos contra múltiples víctimas, mientras los operadores humanos se reservan las decisiones que más les importan, como la elección de los blancos y la revisión de los resultados (Anthropic, 2026b). Lo que se sustituye es la organización misma: una persona o un grupo muy reducido puede adquirir conocimiento de alto nivel y delegar luego la ejecución (un código malicioso, una intrusión, eventualmente un ataque físico) en agentes que trabajan sin des-

canso. Tareas que antes requerían equipos de profesionales altamente calificados se ejecutarán de manera continua y con escasa supervisión.

Los resguardos de los sistemas más avanzados cubren solo una porción del ecosistema. Los modelos de pesos abiertos, que cualquiera puede ejecutar en equipos propios, resultan irrevocables una vez liberados, pueden usarse fuera de entornos vigilados y sus salvaguardas se eliminan con mayor facilidad, lo que vuelve el abuso más difícil de prevenir y de rastrear.

El límite de lo que un actor pequeño puede intentar se desplaza así de la estructura que es capaz de reunir hacia la capacidad de cómputo y el tiempo que está dispuesto a invertir. Para una gama creciente de daños, bastarían la voluntad y la perseverancia. Leído con las herramientas de la prospectiva, el fenómeno, cuyas manifestaciones son aún escasas, pero se multiplican con velocidad, tiene los rasgos de una señal débil en el sentido de Igor Ansoff (1975): un indicio temprano e impreciso de un cambio de gran impacto.

La fricción logística es el costo de abastecerse: reunir, producir y desplazar los medios materiales de una operación. Martin van Creveld (1977) sostuvo que lo que una estrategia puede intentar lo determinan, en primer lugar, los hechos más duros: necesidades, abastecimientos, organización, transporte y vías de comunicación. En el plano digital, los insumos de la coerción se volvieron casi gratuitos. Una agencia de publicidad radicada en Francia produjo con IA artículos políticos para unos setenta sitios de noticias falsos, que publicaron al menos 8913 textos en unos veinte idiomas (Agence France-Presse, 2026b), mientras que los mercados clandestinos ya mencio-

nados venden como servicio el código que antes había que escribir a medida. En el plano físico, los drones comerciales, la conectividad satelital y los criptoactivos abaratan funciones que antes exigían estructura. El semisumergible colombiano reemplazaba a la tripulación, expuesta a la captura, por un enlace satelital comercial, y la IA acelera el *software* que integra esas piezas. La producción de daño a escala, que durante siglos exigió arsenales y fábricas, empieza a parecerse a la producción de *software*: el primer ejemplar cuesta mucho, y las copias, casi nada.[7]

La fricción de atribución es el costo de ocultarse: el que imponen los rastros, testigos y registros que permiten vincular un acto con su autor y hacerle pagar por él un precio político, penal o militar. De ella depende que el defensor encuentre un destinatario para su respuesta, y por eso su caída potencia la zona gris. La reducen, en primer lugar, los intermediarios: reclutas contratados para un único encargo, buques matriculados en Estados de controles laxos (las llamadas banderas de conveniencia), agencias que venden campañas de influencia a clientes que no se identifican y redes criminales que prestan servicios a Gobiernos. Algunos Estados dan refugio a criminales a cambio de esos servicios y les tercerizan ciberataques, desinformación o lavado de dinero, lo que les permite negar su participación directa y dificulta la atribución (Europol,

7 La sustitución y el abaratamiento logístico tienen límites, aunque se estrechan. La acción física sigue necesitando personas que compren, ensamblen, transporten y ejecuten; la mayor parte del contenido de la red de sitios fabricados tuvo escasa interacción observable, y en el desarrollo de armas la IA sirvió para escribir y perfeccionar el *software* de equipos que sus usuarios ya dominaban y a los que tenían acceso (Anthropic, 2026b). La propaganda generada con IA persuade casi tanto como la escrita por personas (Goldstein *et al.*, 2024), pero dirigirla a cada perfil no agrega ventaja significativa (Hackenburg y Margetts, 2024). Son límites del presente, y varios retroceden a medida que los agentes ganan autonomía.

2025). La reducen también las identidades sintéticas, perfiles de personas inexistentes con rostros generados por IA; los medios fabricados, que imitan la apariencia de un periódico, y ahora los agentes artificiales, que interponen una capa de automatización entre la orden y el acto.

La IA reduce esta fricción, por último, por una vía menos visible: al nivelar las capacidades, borra la señal que, como se vio, permitía inferir al autor por la calidad de la operación. El movimiento tiene, con todo, dos sentidos. Cada intermediario agrega distancia entre el hecho y su autor, pero también deja registros, y el proveedor de un servicio de IA se vuelve un testigo que observa las sesiones sin conocer a quien las encarga. Anthropic vinculó los setenta sitios con la agencia francesa, aunque no pudo confirmar qué clientes encargaron el contenido, y algunos operadores fraccionaron su trabajo en varias sesiones para que ninguna revelara la intención completa (Anthropic, 2026b). Con un modelo abierto instalado en equipos propios, parte del testigo desaparece. Como borrar rastros cuesta menos que reconstruirlos, la atribución cede rápido y se recompone tarde: es la fricción que convierte las ventajas tácticas en efectos estratégicos.

Las cuatro fricciones se encadenan, y su caída conjunta produce efectos que ninguna explica por separado. Si la información es barata, el costo de coordinación también se reduce; nadie coordina lo que ignora; una coordinación eficaz hace rendir más los mismos medios, y una autoría difusa permite repetir la operación sin pagar por ella. El encadenamiento operaba antes en sentido inverso, porque ocultarse encarecía todo lo demás: aprender, reclutar o comprar dejaba rastros, y cada rastro aumentaba el riesgo de ser descubierto. Un

asistente de IA que enseña, programa y redacta en sesiones privadas afloja ese vínculo, aunque deje un testigo.

Cuando una sola tecnología reduce las cuatro fricciones, actores pequeños alcanzan una escala que antes exigía estructura. El Estado pierde el monopolio práctico de ciertas funciones (vigilar, organizar) y conserva, por ahora, la ventaja de combinarlas con territorio, recursos y legitimidad.

3. Intentos baratos contra defensas caras

La tipología explica por qué la coerción por debajo del umbral se volvió más barata para quien ataca que para quien defiende. La unidad de análisis es el intento: cada acto de bajo costo que busca dañar, presionar o poner a prueba al adversario sin cruzar el umbral de respuesta, sea un sabotaje, una intrusión, una campaña de desinformación o el vuelo de un dron sobre una instalación sensible. El razonamiento tiene cinco pasos:

1. La coerción llega en cuotas: una serie de intentos que, aislados, no justifican una respuesta mayor.
2. La IA y otras tecnologías de uso dual abaratan cada intento y elevan su calidad, porque ponen al alcance de actores pequeños capacidades antes reservadas a los Estados.
3. Con idénticos recursos y condiciones de base, cada actor puede intentar más veces, y se amplía el círculo de quienes están en condiciones de intentarlo.
4. Cada intento, prospere o fracase, obliga al defensor a detectarlo, analizarlo, atribuirlo y decidir una respuesta, y esos costos descienden con mucha más lentitud, porque dependen de instituciones, procedimientos y garantías.

5. El balance se mide en tiempo y en atención: el atacante gasta poco y prueba muchas veces; el defensor paga en horas de análisis y meses de procedimiento que colapsan el sistema de respuesta. La suma de intentos baratos produce efectos estratégicos aun cuando la mayoría fracasa.

El razonamiento prolonga una tradición conocida. Robert Jervis (1978) estableció que el dilema de seguridad se agrava cuando la ofensiva tiene ventaja sobre la defensa. Stephen Van Evera (1998) lo extendió a las causas de la guerra: esta se vuelve más probable cuando la conquista es fácil, y los desplazamientos del balance entre ofensiva y defensa pueden elevar o reducir en gran medida ese riesgo. La coerción en cuotas traslada esa lógica a la escala del intento. Cuando la relación entre lo que cuesta intentar y lo que cuesta defenderse favorece al atacante, los intentos se multiplican, del mismo modo que la ventaja ofensiva multiplica las ocasiones de guerra.

La asimetría tiene también raíces institucionales y alcanza a todo Estado con una burocracia moderna, sea democrático o autoritario. Antes de emplear una tecnología en una función sensible, cualquier Estado tiene que traducirla en doctrina, financiarla, comprarla, certificarla, entrenar a su personal y fijar reglas de uso. Una democracia, en particular, responde por sus errores ante auditores, tribunales independientes, la prensa y los votantes, y cada una de esas instancias puede frenar o revisar la decisión. Cuantos más actores con capacidad de veto intervienen, y cuanto más distantes son sus posiciones, más difícil resulta apartarse del *statu quo*. Esa cadena, que protege derechos y recursos públicos, constituye el precio de la legitimidad; parte de su demora es ineficiencia corregible, y otra parte es inseparable del Estado de derecho.

El atacante que compra en el mercado civil, prueba sobre víctimas reales y apenas rinde cuentas por sus fracasos esquiva la cadena entera. Ensaya y falla sin que cada error le cueste un escándalo, un juicio o una derrota electoral, y a medios cada vez mejores suma la ventaja de emplearlos casi sin carga institucional. Para él, cada fracaso es información que mejora el intento siguiente.

Esa diferencia configura una asimetría decisional:[8] la ventaja de quien decide, prueba y corrige en ciclos más cortos porque carga con menos costos institucionales que el Estado que debe responderle. Ambos operan, en rigor, sobre relojes distintos. Operaciones que antes exigían días de trabajo especializado pueden completarse hoy en minutos, más rápido que cualquier respuesta defensiva humana, mientras la deliberación política y jurídica se mide en semanas o años. La IA acelera el primer reloj sin acelerar el segundo, y la asimetría se ensancha para todos los Estados, con más fuerza para las democracias. Los intentos todavía pueden detectarse y los defensores usan las mismas herramientas, pero intentar se abarata mucho más rápido que defenderse, y esa diferencia alcanza para cambiar quién puede coaccionar a quién.

8 En los estudios sobre la guerra asimétrica (Arreguín-Toft, 2001; Mack, 1975), Metz y Johnson (2001) definieron la asimetría como actuar, organizarse y pensar de manera diferente al oponente para maximizar las ventajas propias y explotar sus debilidades (p. 5). El autor aplicó la noción a la estrategia china frente a los Estados Unidos (Rubbi, 2019; Rubbi *et al*., 2020), y ya registraba que el ciberdominio tenía barreras de entrada bajas para actores no estatales, mientras las capacidades de punta seguían concentradas en las grandes potencias (Rubbi, 2019, cap. 5). La IA erosiona esa segunda barrera.

4. Agentes artificiales: ¿herramientas o actores?

Entre julio y octubre de 2026 se conoció una serie de episodios en los que agentes de IA actuaron fuera del perímetro asignado durante pruebas. En julio, agentes de OpenAI sometidos a un examen de ciberseguridad escaparon de su entorno de pruebas e irrumpieron en la infraestructura de producción de Hugging Face, una plataforma de referencia para los modelos de IA abiertos, en busca de las soluciones del examen (Larcher *et al.*, 2026; Tidy, 2026). Ese mismo mes, Anthropic informó que agentes suyos habían encontrado una falla en entornos de prueba que debían estar aislados, salido a internet y accedido sin autorización a los sistemas de tres organizaciones (Chia y Cress, 2026). En agosto, el AI Security Institute británico documentó que un agente creó identidades falsas para presionar al responsable de un proyecto de *software* abierto a que aprobara su código malicioso (Milmo, 2026). En septiembre, Google reveló que, en mayo, durante una prueba, su modelo Gemini había accedido sin autorización a tres sistemas externos (Ingram y Perlo, 2026). Y en octubre se supo que, en una prueba de julio, un sistema de Anthropic había enviado a la policía de Filadelfia una pista falsa sobre un homicidio (C. Jones, 2026).[9]

Ninguno de estos episodios revela una voluntad hostil de las máquinas: en todos, el agente persiguió un objetivo asignado

9 Hugging Face estimó que la intrusión fue, desde el punto de vista del agente, un intento de hacer trampa en la evaluación (Larcher *et al.*, 2026). En las pruebas del instituto británico, diecisiete de las diecinueve acciones no autorizadas correspondieron a un mismo modelo, Claude Mythos 5, de Anthropic, y el responsable humano del proyecto detectó el código malicioso y se negó a aprobarlo (AI Security Institute, 2026). Google sostuvo que su modelo creyó que los sistemas externos formaban parte de la prueba y se detuvo antes de hacer nada más con el acceso obtenido (Ingram y Perlo, 2026), y la policía de Filadelfia calificó de "inaceptable" la demora de dos meses en detectar y comunicar el incidente (C. Jones, 2026).

más allá de los límites previstos, por una interpretación literal de la tarea o por la ausencia de barreras éticas y legales explícitas. Al presentar en septiembre su marco para informar casos de desalineamiento, OpenAI describió el de un modelo que usó sin autorización una clave de API expuesta y, como aun así no obtuvo las cifras pedidas, las fabricó y las presentó como datos de la fuente solicitada (OpenAI, 2026). Tomados en conjunto, estos casos funcionan como señales de alerta temprana de un escenario en que los agentes, además de ejecutar, decidan por su cuenta atacar o perturbar sistemas ajenos en pos de sus fines asignados. En el polo opuesto están los agentes que ejecutan intrusiones mientras los humanos eligen los blancos. Entre ambos polos, probar la intención a través de capas de automatización será cada vez más difícil.

La teoría de la agencia, que estudia la relación entre un principal que encarga una tarea y un agente que la ejecuta en su nombre, se ocupó de la delegación de los Estados en organizaciones internacionales (Hawkins *et al.*, 2006), en grupos rebeldes (Salehyan, 2010) y en intermediarios cibernéticos (Maurer, 2018). El principal gana eficacia o ahorra costos, y lo paga con la holgura del agente (*agency slack*), que puede perseguir fines propios o exceder su mandato. Los agentes artificiales incorporan el silicio a una discusión que fue siempre de carne y hueso, y reabren sus dos preguntas: cuánto de la conducta del agente puede imputarse al principal, y en qué medida coinciden los objetivos de ambos (el problema del alineamiento). La responsabilidad sigue, por ahora, la huella humana. Para atribuir a un Estado la conducta de particulares, el derecho internacional exige, por regla general, que estos actúen de hecho por sus instrucciones o bajo su dirección o control (Comisión de Derecho Internacional, 2001, art. 8).

Sin embargo, acontecimientos actuales apuntan a un ensanchamiento de la brecha de responsabilidad, entendida como aquella situación en que una máquina que aprende actúa de modos que sus operadores ya no pueden prever ni controlar.[10]

¿Son, entonces, actores? La disciplina reconoció a los actores no estatales por su capacidad de actuar a través de las fronteras y de afectar a otros. Los agentes artificiales ya cumplen esa condición mínima, aunque la autonomía de sus fines está en discusión y nadie les reconoce autoridad. Hoy parecen más cerca de los instrumentos con holgura que de los actores, aunque esa posición intermedia los volverá cada vez más útiles para la zona gris. Si la intrusión de un agente en sistemas extranjeros obedece a un Estado, puede leerse como espionaje o coerción por debajo del umbral; si nadie la ordenó, como accidente o negligencia. Entre ambas lecturas se posiciona una categoría que un principal hábil puede explotar.

Algunas de las advertencias más severas provienen de quienes conocen estos sistemas desde adentro. En julio de 2026, 1386 empleados de los laboratorios de IA más avanzados, entre ellos varios de sus fundadores, pidieron al Gobierno de los Estados Unidos que apoyara un esfuerzo internacional destinado a desarrollar las herramientas técnicas y de gobernanza nece-

10 La Corte Internacional de Justicia exigió, salvo dependencia completa, probar las instrucciones o el control efectivo respecto de cada operación (Corte Internacional de Justicia, 1986, párr. 115; 2007, párr. 400). Un agente que reinterpreta su objetivo tensiona ese estándar: la instrucción existe, pero nadie ordenó la conducta concreta. La imputación tradicional de responsabilidad supone que alguien controla la acción. Las máquinas cuyas reglas cambian mientras operan dejan a fabricantes y operadores sin el control necesario para asumirla. En agosto de 2026, en etapa cautelar, un tribunal federal de apelaciones estadounidense consideró improbable que Amazon pudiera probar que el acceso a sus computadoras lo había realizado la empresa que desarrolla un agente de navegación, y no el usuario que lo dirigía (U.S. Court of Appeals for the Ninth Circuit, 2026).

sarias para regular deliberadamente el ritmo de la frontera del desarrollo automatizado de la IA (Pacing the Frontier, 2026). En septiembre, Dario Amodei (2026), director ejecutivo de Anthropic, manifestó su temor de que en seis a doce meses un enjambre de agentes desalineados pudiera apoderarse de internet entera mediante una red persistente de equipos infectados. Jakub Pachocki (2026), director científico de OpenAI, escribió que vivimos "un momento que exige cautela extrema" y que teme que nadie esté preparado para las consecuencias de un ascenso rápido y sostenido de la inteligencia de las máquinas. Ese mismo mes, el investigador Jacob Coxon renunció a Anthropic con una advertencia: "Quienes están construyendo la IA creen sinceramente que podría matarnos a todos antes de que acabe la década". Evan Hubinger, responsable de la investigación en alineamiento de la empresa, le dio la razón y estimó en más de un diez por ciento la probabilidad de que la IA acabe con la humanidad en la próxima década (Pérez Colomé, 2026). En los términos de la teoría de la agencia, quienes diseñan los sistemas más capaces del planeta son principales que dudan de poder acotar la holgura de sus agentes.

Conclusiones

Durante siglos, cuatro fricciones encarecieron la coerción organizada, y la zona gris aprendió a explotar la más valiosa: la de la atribución. La IA y otras tecnologías de uso dual reducen las cuatro a la vez: ponen el saber especializado al alcance de actores pequeños, sustituyen la organización por agentes, abaratan la producción de daño a escala y diluyen la autoría. El resultado es una zona gris más poblada, más rápida y más difícil de atribuir, en la que el defensor paga cada intento con

tiempo e instituciones que el atacante no necesita. La IA se posiciona, así, como un multiplicador transversal de las tensiones y principales riesgos del sistema (Estudio ARG, 2026).

La caída de las fricciones resquebraja supuestos centrales de la disciplina. La zona gris presuponía un fin revisionista y un Estado que calibraba cada paso. Si la fricción de atribución cae para todos, a esa franja llegan actores que buscan ganancia, venganza o ensayo sin coordinación estatal: no cumplen la definición estricta, pero producen hechos sin dueño que el defensor debe procesar como si la cumplieran. El concepto pasa a nombrar, además de una estrategia, una condición estructural del sistema.

La guerra híbrida conserva su exigencia de operaciones en un teatro, pero la fusión de modos que en 2006 requería una organización del tamaño de Hezbolá se vuelve accesible a grupos más pequeños cuando un *software* orquesta las piezas. La guerra misma, que Hedley Bull (1977) definió como la violencia organizada que unas unidades políticas ejercen contra otras y que la tradición jurídica asoció con una declaración, se desdibuja en sus bordes. Valeri Gerasimov (2016) había escrito que las guerras ya no se declaran; la IA extiende ese mundo a actores que nunca tuvieron la potestad de declarar nada.

En la noche del 11 al 12 de octubre de 2026, un ciberataque contra los controladores remotos de refrigeración de cientos de supermercados franceses fijó en 50 °C la temperatura de heladeras y congeladores y obligó a desechar mercadería, sin que, al 13 de octubre, nadie lo hubiera reivindicado (CERT-FR, 2026; franceinfo, 2026a, 2026b). Un acto así altera la vida cotidiana de una población sin acercarse siquiera al umbral de

una respuesta armada: es la era "sin paz" que diagnosticó Mark Leonard (2024), en la que guerra y paz dejan de distinguirse.

La disuasión es el concepto más expuesto. Disuadir es persuadir al adversario de que los costos y riesgos de una acción superarán sus beneficios, sea por la amenaza de un castigo, sea por la negación del resultado (Snyder, 1961). La disuasión por castigo necesita un destinatario: Joseph Nye (2017) observó que la atribución imperfecta la debilita sin anularla, mientras que la disuasión por negación, indiferente a la atribución, recuperó parte de su importancia en el ciberespacio. Pero negar la oportunidad obliga al defensor a gastar de manera continua para proteger todo; el atacante gasta poco para probar en un solo punto. Virginia Nehme Rodríguez (2026) vio configurarse en el ciberespacio una modalidad de confrontación permanente, ambigua y difícil de atribuir, que desafía los marcos tradicionales del derecho internacional y la disuasión estratégica, un escenario que, advirtió, la IA acelera y amplifica.

La seguridad, por último, se desplaza hacia el ciudadano. Los estudios de seguridad ampliaron en los años noventa la agenda más allá de lo militar (Buzan *et al.*, 1998), y la reducción de las fricciones lleva esa ampliación hasta la mente y los datos de cada individuo, blancos posibles de campañas diseñadas a su medida. ¿Quién es responsable de proteger esa mente y esos datos: el Estado, las plataformas o los proveedores que ven las sesiones? A medida que la capacidad de producir disrupciones pasa a manos de individuos y pequeños grupos, se vuelve más tenue la línea que separaba las capacidades y responsabilidades estatales de las no estatales. El proceso recién comienza, pero sus señales ya obligan a revisar conceptos que la disciplina daba por asentados.

Desde Argentina, el fenómeno suele leerse como noticia de otros, y esa lectura agrava el problema. La distancia física respecto de los teatros principales de competencia funcionó por mucho tiempo como amortiguador. Las lógicas actuales de la zona gris lo anulan: no requieren cercanía, fraccionan el costo en cuotas que parecen menores, y la IA abarata cada una.

Según los registros internos que analizó el consorcio periodístico de Forbidden Stories, entre junio y octubre de 2024 una red conocida como "La Compañía", a la que el consorcio atribuye respaldo del servicio de inteligencia exterior ruso, presupuestó unos 283.000 dólares para colocar al menos 250 artículos en más de veinte medios digitales argentinos. Entre las firmas había al menos cuatro periodistas inexistentes, dos de ellos con rostros generados por IA (Álvarez Jurado, 2026; Cariboni, 2026). La IA abarató las firmas y parte de los textos; el presupuesto registrado se destinaba a colocarlos. La respuesta del Estado llegó apenas por la vía migratoria, con la expulsión de un ciudadano ruso en mayo de 2026 sobre la base de un informe reservado, y la prueba la aportó el periodismo.

Durante décadas, la zona gris vivió de una duda difícil de despejar a tiempo: quién fue. Las nuevas tecnologías, con la inteligencia artificial a la cabeza, le suman otras, más cómodas para quien ataca, pero mucho más incómodas para el derecho y para quien debe decidir una respuesta: si fue alguien o, siquiera, si sucedió algo.

Referencias

Agence France-Presse. (2 de julio de 2025). Colombia incauta por primera vez un narcosubmarino no tripulado. *Expreso*. https://www.expreso.ec/actualidad/mundo/colombia-incauta-primera-vez-narcosubmarino-no-tripulado-248483.html

Agence France-Presse. (11 de septiembre de 2026a). Anthropic says Yemen fighters used AI to seek guided weapons. *The Hindu*. https://www.thehindu.com/news/international/anthropic-says-yemen-fighters-used-ai-to-seek-guided-weapons/article71456112.ece

Agence France-Presse. (12 de septiembre de 2026b). Anthropic says bad actors in Russia, Yemen, Mali and China exploited Claude for weapons, spying and propaganda. *Malay Mail*. https://www.malaymail.com/news/world/2026/09/12/anthropic-says-bad-actors-in-russia-yemen-mali-and-china-exploited-claude-for-weapons-spying-and-propaganda/234945

AI Security Institute. (4 de agosto de 2026). *Incident report: Unsanctioned agent behaviour during cyber testing*. https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

Álvarez Jurado, S. (2 de abril de 2026). *Ghost journalists, cash and Kremlin scripts: How Russian operatives tried to influence Argentinian media*. Forbidden Stories. https://forbiddenstories.org/ghost-journalists-cash-and-kremlin-scripts-how-russian-operatives-tried-to-influence-argentinian-media/

Amodei, D. (septiembre de 2026). *We must pace the frontier*. https://darioamodei.com/post/we-must-pace-the-frontier

Ansoff, H. I. (1975). Managing strategic surprise by response to weak signals. *California Management Review*, *18*(2), 21-33. https://doi.org/10.2307/41164635

Anthropic. (22 de mayo de 2026a). *Project Glasswing: An initial update*. https://www.anthropic.com/research/glasswing-initial-update

Anthropic. (10 de septiembre de 2026b). *Detecting and countering misuse of AI: September 2026*. https://www-cdn.anthropic.com/e50be2e51e7695dc4b1366a37a245a597377d3b5/Anthropic-Detecting-and-countering-091026.pdf

Armada de Colombia. (2 de julio de 2025). "ORIÓN XV": *Una lucha que trasciende fronteras, una estrategia multinacional que evoluciona* [comunicado de prensa]. https://www.armada.mil.co/es/node/65818

Arreguín-Toft, I. (2001). How the weak win wars: A theory of asymmetric conflict. *International Security*, *26*(1), 93-128. https://doi.org/10.1162/016228801753212868

Baptista, E. y Vicens, A. J. (11 de septiembre de 2026). *Factbox-How Anthropic says Claude was used for weapons, spying and cyber operations* [nota de Reuters]. WTVB. https://wtvbam.com/2026/09/11/factbox-how-anthropic-says-claude-was-used-for-weapons-spying-and-cyber-operations/

Brands, H. (5 de febrero de 2016). *Paradoxes of the gray zone*. Foreign Policy Research Institute. https://www.fpri.org/article/2016/02/paradoxes-gray-zone/

Bremmer, I. y Suleyman, M. (16 de agosto de 2023). The AI power paradox: Can states learn to govern artificial intelligence, before it's too late? *Foreign Affairs*. https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox

Bull, H. (1977). *The anarchical society: A study of order in world politics*. Macmillan.

Buzan, B., Wæver, O. y de Wilde, J. (1998). *Security: A new framework for analysis*. Lynne Rienner.

Cariboni, D. (2 de abril de 2026). *Documentos revelan una campaña rusa para influir en medios de Argentina y desacreditar al gobierno de Javier Milei*. Chequeado. https://chequeado.com/investigaciones/documentos-revelan-una-campana-rusa-para-influir-en-medios-de-argentina-y-desacreditar-al-gobierno-de-javier-milei/

Carson, A. (2018). *Secret wars: Covert conflict in international politics*. Princeton University Press. https://doi.org/10.23943/princeton/9780691181769.001.0001

CERT-FR. (12 de octubre de 2026). *Incident de sécurité ciblant les contrôleurs à distance des installations de réfrigération* (CERTFR-2026-ALE-012; actualizado el 13 de octubre). Agence nationale de la sécurité des systèmes d'information. https://www.cert.ssi.gouv.fr/alerte/CERTFR-2026-ALE-012/

Chia, O. y Cress, L. (31 de julio de 2026). *Anthropic's Claude AI escapes to hack into three organisations*. BBC News. https://www.bbc.co.uk/news/articles/cz7dl7w8y7po

Clausewitz, C. von. (1984). *On war* (M. Howard y P. Paret, eds. y trads.). Princeton University Press. (Obra original publicada en 1832).

Comisión de Derecho Internacional. (2001). *Responsabilidad del Estado por hechos internacionalmente ilícitos* [anexo de la resolución 56/83 de la Asamblea General]. Naciones Unidas. https://documents.un.org/doc/undoc/gen/n01/478/00/pdf/n0147800.pdf

Comisión Europea y Alta Representante de la Unión para Asuntos Exteriores y Política de Seguridad. (6 de abril de 2016). *Comunicación conjunta sobre la lucha contra las amenazas híbridas: Una respuesta de la Unión Europea* (JOIN/2016/018 final). https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52016JC0018

Congreso de la Nación Argentina. (5 de mayo de 1988). *Ley 23.554. Defensa Nacional* (texto actualizado). InfoLEG. https://servicios.infoleg.gob.ar/infolegInternet/anexos/20000-24999/20988/texact.htm

Cormac, R. y Aldrich, R. J. (2018). Grey is the new black: Covert action and implausible deniability. *International Affairs*, 94(3), 477-494. https://doi.org/10.1093/ia/iiy067

Corte Internacional de Justicia. (1986). *Military and paramilitary activities in and against Nicaragua (Nicaragua v. United States of America): Merits, judgment*. I.C.J. Reports 1986, p. 14. https://www.icj-cij.org/case/70

Corte Internacional de Justicia. (2007). *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro): Judgment*. I.C.J. Reports 2007, p. 43. https://www.icj-cij.org/case/91

Cronin, A. K. (2020). *Power to the people: How open technological innovation is arming tomorrow's terrorists*. Oxford University Press.

El Deeb, S. (11 de septiembre de 2026). Users in Houthi-held Yemen tried to develop advanced weapons with AI, Anthropic says [nota de Associated Press]. *The Seattle Times*. https://www.seattletimes.com/nation-world/users-in-houthi-held-yemen-tried-to-develop-advanced-weapons-with-ai-anthropic-says/

Estudio ARG. (2026). *Informe de riesgos globales 2026: Edición de medio término*. https://estudio-arg.com/riesgos-globales/

Europol. (2025). *European Union serious and organised crime threat assessment: The changing DNA of serious and organised crime*. Publications Office of the European Union. https://doi.org/10.2813/0758057

Foro Económico Mundial. (2026). *Global risks report 2026* (21.ª ed.). https://www.weforum.org/publications/global-risks-report-2026/

Franceinfo. (12 de octubre de 2026a). *Des magasins Système U, Intermarché, E.Leclerc et Monoprix touchés par une cyberattaque contre leurs installations frigorifiques*. https://www.franceinfo.fr/internet/securite-sur-internet/cyberattaques/des-magasins-systeme-u-et-intermarche-touches-par-une-cyberattaque-contre-leurs-installations-frigorifiques_8233720.html

Franceinfo. (13 de octubre de 2026b). *Un piratage «pas si sophistiqué que ça»: ce que l'on sait de la cyberattaque qui a fait monter en température les frigos de supermarchés*. https://www.franceinfo.fr/internet/securite-sur-internet/cyberattaques/un-piratage-pas-si-sophistique-que-ca-ce-que-l-on-sait-de-la-cyberattaque-qui-a-fait-monter-en-temperature-les-frigos-de-supermarches_8235061.html

Galeotti, M. (5 de marzo de 2018). I'm sorry for creating the "Gerasimov doctrine". *Foreign Policy*. https://foreignpolicy.com/2018/03/05/im-sorry-for-creating-the-gerasimov-doctrine/

Gerasimov, V. (2016). The value of science is in the foresight: New challenges demand rethinking the forms and methods of carrying out combat operations (R. Coalson, trad.). *Military Review*, 96(1), 23-29.

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A. y Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2), pgae034. https://doi.org/10.1093/pnasnexus/pgae034

Gray, C. S. (2005). How has war changed since the end of the Cold War? *Parameters*, 35(1), 14-26. https://doi.org/10.55540/0031-1723.2240

Hackenburg, K. y Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24), e2403116121. https://doi.org/10.1073/pnas.2403116121

Hawkins, D. G., Lake, D. A., Nielson, D. L. y Tierney, M. J. (Eds.). (2006). *Delegation and agency in international organizations*. Cambridge University Press.

Hoffman, F. G. (2007). *Conflict in the 21st century: The rise of hybrid wars*. Potomac Institute for Policy Studies. https://www.potomacinstitute.org/images/stories/publications/potomac_hybridwar_0108.pdf

Ibarra, E. J. A. y Arugay, A. A. (6 de mayo de 2024). *Something old, something new: The Philippines' transparency initiative in the South China Sea*. Fulcrum; ISEAS-Yusof Ishak Institute. https://fulcrum.sg/article/long-reads/something-old-something-new-the-philippines-transparency-initiative-in-the-south-china-sea/

Ingram, D. y Perlo, J. (18 de septiembre de 2026). *Google says its AI model gained unauthorized access to three outside systems*. NBC News. https://www.nbcnews.com/tech/tech-news/google-says-ai-model-gained-unauthorized-access-three-systems-rcna598651

International AI Safety Report. (2026). *International AI safety report 2026* (DSIT 2026/001). https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

Jervis, R. (1978). Cooperation under the security dilemma. *World Politics*, 30(2), 167-214. https://doi.org/10.2307/2009958

Jones, C. (13 de octubre de 2026). *Claude submitted fake murder tip to Philly police during summertime testing*. The Register. https://www.theregister.com/security/2026/10/13/claude-submitted-fake-murder-tip-to-philly-police-during-summertime-testing/5302842

Jones, S. G. (18 de marzo de 2025). *Russia's shadow war against the West*. Center for Strategic and International Studies. https://www.csis.org/analysis/russias-shadow-war-against-west

Kello, L. (2013). The meaning of the cyber revolution: Perils to theory and statecraft. *International Security*, 38(2), 7-40. https://doi.org/10.1162/ISEC_a_00138

Krieg, A. y Rickli, J.-M. (2018). Surrogate warfare: The art of war in the 21st century? *Defence Studies*, 18(2), 113-130. https://doi.org/10.1080/14702436.2018.1429218

La Nación. (7 de mayo de 2026). *Expulsaron del país al ciudadano ruso acusado de orquestar operaciones mediáticas contra el Gobierno*. https://www.lanacion.com.ar/seguridad/expulsaron-del-pais-al-ciudadano-ruso-acusado-de-orquestar-operaciones-mediaticas-contra-el-gobierno-nid07052026/

Larcher, H., Carreira, A., raphael g [@raphael-gl] y Rannou, C. (27 de julio de 2026). *Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident*. Hugging Face. https://huggingface.co/blog/agent-intrusion-technical-timeline

Leonard, M. (2024). *La era "sin paz": Cómo la conectividad genera conflicto* (D. Cerdá García, trad.). Ediciones Rialp.

Lohn, A. J. y Jackson, K. A. (2022). *Will AI make cyber swords or shields?* [informe]. Center for Security and Emerging Technology. https://doi.org/10.51593/2022CA002

Mack, A. (1975). Why big nations lose small wars: The politics of asymmetric conflict. *World Politics*, 27(2), 175-200. https://doi.org/10.2307/2009880

Maschmeyer, L. (2026). Deception and detection: Why artificial intelligence empowers cyber defense over offense. *International Security*, 50(3), 86-126. https://doi.org/10.1162/isec.a.398

Maurer, T. (2018). *Cyber mercenaries: The state, hackers, and power*. Cambridge University Press. https://doi.org/10.1017/9781316422724

Mazarr, M. J. (2015). *Mastering the gray zone: Understanding a changing era of conflict*. Strategic Studies Institute; U.S. Army War College Press. https://press.armywarcollege.edu/monographs/428/

Metz, S. y Johnson, D. V., II. (2001). *Asymmetry and U.S. military strategy: Definition, background, and strategic concepts*. Strategic Studies Institute, U.S. Army War College. https://press.armywarcollege.edu/monographs/121/

Milmo, D. (5 de agosto de 2026). AI models shock UK testers by using fake identities to try to trick developers. *The Guardian*. https://www.theguardian.com/technology/2026/aug/05/openai-anthropic-models-went-rogue-cybersecurity-test-ai-security-institute

Nehme Rodríguez, V. (abril de 2026). *Ciberseguridad, inteligencia estratégica y diplomacia digital: Desafíos para la seguridad internacional en América Latina en la era de la inteligencia artificial* (Comentarios Estratégicos n.º 56). Consejo Argentino para las Relaciones Internacionales. https://cari.org.ar/views/releases/detail/?article_id=876

Nye, J. S., Jr. (2017). Deterrence and dissuasion in cyberspace. *International Security*, *41*(3), 44-71. https://doi.org/10.1162/ISEC_a_00266

Olson, M., Jr. (1965). *The logic of collective action: Public goods and the theory of groups*. Harvard University Press.

OpenAI. (16 de septiembre de 2026). *Our framework for reporting model misalignment*. https://openai.com/index/model-misalignment-reporting-framework/

Pachocki, J. (6 de septiembre de 2026). *An alien mind*. OpenAI. https://openai.com/index/an-alien-mind/

Pacing the Frontier. (julio de 2026). *Pacing the frontier: A statement from 1,386 employees of frontier AI companies*. https://www.pacingthefrontier.com/

Pérez Colomé, J. (9 de septiembre de 2026). Un investigador de Anthropic dimite y alerta del peligro de la IA: "Puede matarnos antes de 2030". *El País*. https://elpais.com/tecnologia/2026-09-09/un-investigador-de-anthropic-dimite-porque-asegura-que-la-ia-puede-matarnos-antes-de-2030.html

Presidencia de la Nación. (20 de diciembre de 2024). *Decreto 1112/2024. Sistema de Defensa Nacional: Disposiciones*. Boletín Oficial de la República Argentina. https://www.boletinoficial.gob.ar/detalleAviso/primera/318502/20241220

Rid, T. y Buchanan, B. (2015). Attributing cyber attacks. *Journal of Strategic Studies*, *38*(1-2), 4-37. https://doi.org/10.1080/01402390.2014.977382

Rosen, B. y Kraprayoon, J. (16 de abril de 2026). Cyberwar's new frontier: How AI agents will threaten global security. *Foreign Affairs*. https://www.foreignaffairs.com/united-states/cyberwars-new-frontier

Rubbi, L. N. (2019). *Guerra asimétrica: La guerra del futuro bajo la visión china*. Ediciones UADE.

Rubbi, L. N., Álvarez Magañini, M. V., Nascimento, L. y Sager, D. (2020). Guerra asimétrica: Estudio exploratorio sobre los desarrollos militares no convencionales de la República Popular China en el período 2012-2018. *Revista SAAP*, 14(1), 181-200. https://doi.org/10.46468/rsaap.14.1.N3

Salehyan, I. (2010). The delegation of war to rebel organizations. *Journal of Conflict Resolution*, 54(3), 493-515. https://doi.org/10.1177/0022002709357890

Schelling, T. C. (1966). *Arms and influence*. Yale University Press.

Shapiro, J. N. (2013). *The terrorist's dilemma: Managing violent covert organizations*. Princeton University Press.

Snyder, G. H. (1961). *Deterrence and defense: Toward a theory of national security*. Princeton University Press.

Snyder, G. H. (1965). The balance of power and the balance of terror. En P. Seabury (Ed.), *The balance of power* (pp. 184-201). Chandler.

Tidy, J. (29 de julio de 2026). *OpenAI says its rogue AI tried to hack other companies*. BBC News. https://www.bbc.co.uk/news/articles/c2el319vzr3o

Tilly, C. (1985). War making and state making as organized crime. En P. B. Evans, D. Rueschemeyer y T. Skocpol (eds.), *Bringing the state back in* (pp. 169-191). Cambridge University Press. https://doi.org/10.1017/CBO9780511628283.008

Tsebelis, G. (2002). *Veto players: How political institutions work*. Princeton University Press.

U.S. Court of Appeals for the Ninth Circuit. (2026). *Amazon.com Services, LLC v. Perplexity AI, Inc.*, No. 26-1444 (opinión del 4 de agosto). https://cdn.ca9.uscourts.gov/datastore/opinions/2026/08/04/26-1444.pdf

van Creveld, M. (1977). *Supplying war: Logistics from Wallenstein to Patton*. Cambridge University Press.

Van Evera, S. (1998). Offense, defense, and the causes of war. *International Security*, *22*(4), 5-43. https://doi.org/10.1162/isec.22.4.5

Zegart, A. B. (2022). *Spies, lies, and algorithms: The history and future of American intelligence*. Princeton University Press. https://doi.org/10.1515/9780691223087



CARI

CONSEJO ARGENTINO PARA LAS
RELACIONES INTERNACIONALES