De la IA generativa al colapso epistémico: entornos epistémicamente hostiles en contextos de defensa de alto riesgo

From generative AI to epistemic collapse: Epistemically hostile environments in high-risk defense contexts

Alger Sans Pinillos

Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS)
Barcelona, España
alger.sanspinillos@bsc.es
ORCID: https://orcid.org/0000-0002-8817-7286

Francisco Andrés Pérez

Universidad de Salamanca, Salamaca, España
franc@usal.es
ORCID: https://orcid.org/0000-0002-1675-0717

Resumen

Este artículo analiza cómo los sistemas de inteligencia artificial (IA) empleados en contextos de defensa de alto riesgo pueden contribuir a la configuración de entornos epistémicamente hostiles. A partir de una reconstrucción conceptual de distintos modos de mediación epistémica asociados a la IA —sistemas basados en reglas, aprendizaje automático e IA generativa—, se examina cómo estas tecnologías reconfiguran la relación entre información, juicio humano y articulación entre hechos y valores en procesos de decisión bajo incertidumbre. El trabajo sostiene que la presencia de un humano en el bucle no garantiza por sí sola control, responsabilidad ni legitimidad, pues la supervisión humana solo resulta significativa cuando preserva condiciones efectivas de comprensión, revisión crítica, desacuerdo y veto. Sobre esta base, se propone la noción

Recepción: 6-4-2026 | Aceptado: 18-5-2026
Publicado: 30-6-2026

Citación:
Sans Pinillos, Alger y Francisco Andrés Pérez. "De la IA generativa al colapso epistémico: entornos epistémicamente hostiles en contextos de defensa de alto riesgo". *Estudios del Discurso* 12.1 (2026), pp. 1-31.

DOI: https://doi.org/10.30973/esdi.2026.12.1.259

de *colapso epistémico local* para describir situaciones en las que la mediación algorítmica degrada la capacidad humana de interpretar contextos, cuestionar recomendaciones y asumir responsabilidad por las decisiones adoptadas. Se concluye que una gobernanza responsable de la IA en defensa debe evaluarse no solo por la precisión o eficiencia de los sistemas, sino también por su impacto sobre la calidad epistémica del entorno sociotécnico en el que se ejerce el juicio.

Palabras clave: inteligencia artificial; humano en el bucle; supervisión humana significativa; hechos y valores; colapso epistémico local

Abstract

This article examines how artificial intelligence (AI) systems used in high-risk defense contexts may contribute to the emergence of epistemically hostile environments. Drawing on a conceptual reconstruction of different forms of epistemic mediation associated with AI—rule-based systems, machine learning, and generative AI—it analyzes how these technologies reshape the relationship between information, human judgment, and the articulation of facts and values in decision-making under uncertainty. The article argues that the mere presence of a human in the loop does not by itself guarantee control, accountability, or legitimacy, since human oversight is meaningful only when it preserves effective conditions for understanding, critical review, disagreement, and veto. On this basis, it introduces the notion of local epistemic collapse to describe situations in which algorithmic mediation degrades the human capacity to interpret contexts, question recommendations, and assume responsibility for decisions. The article concludes that responsible AI governance in defense should be assessed not only in terms of accuracy or efficiency, but also in terms of its impact on the epistemic quality of the sociotechnical environment in which judgment is exercised.

Keywords: artificial intelligence; human in the loop; meaningful human oversight; facts and values; local epistemic collapse

1. Introducción

El desarrollo reciente de la inteligencia artificial (en adelante, IA)[1] ha supuesto un salto cualitativo y cuantitativo que puede considerarse un cambio de paradigma en la velocidad de procesamiento, la capacidad de clasificación, predicción y generación, así como en la escala con la que la información, las imágenes, los textos, los diagnósticos y las recomendaciones pueden producirse, distribuirse y consumirse. Este aumento de capacidades no solo ha ampliado las posibilidades funcionales de los sistemas digitales, sino que también ha alterado profundamente las condiciones en las que los agentes humanos acceden a la información, la interpretan y la convierten en base para la toma de decisiones. Informes recientes sobre el desarrollo global de la IA apuntan precisamente a esta asimetría: las capacidades, la adopción y el despliegue de estos sistemas avanzan más rápido que los marcos de gobernanza, evaluación y medición disponibles para gestionarlos de manera responsable (Sajadieh et al. 2). Precisamente por ello, el impacto de la IA no puede evaluarse únicamente en términos de potencia o eficiencia técnica, sino como una forma específica de mediación que reconfigura el entorno cognitivo en el que se ejerce el juicio y se toman decisiones.

Esta capacidad de reconfiguración constituye también el punto de partida para el riesgo de que la información deje de ofrecer un apoyo efectivo al juicio o, dicho de otro modo, de que los resultados generados no aporten valor epistémico significativo. El riesgo aparece cuando esa reconfiguración del entorno cognitivo deja de operar como apoyo al juicio y comienza a condicionar qué se considera relevante, qué se presenta como plausible y qué posibilidades de revisión crítica permanecen efectivamente abiertas para el agente. Es en este desplazamiento —del apoyo al condicionamiento del juicio— donde comienza a perfilarse lo que aquí entenderemos por un entorno epistémicamente hostil: aquellos contextos en los que la mediación tecnológica deteriora las condiciones bajo las cuales los agentes pueden identificar qué información es relevante, interpretar adecuadamente una situación y mantener abierta la posibilidad de revisión crítica.

1 Utilizamos aquí la expresión “inteligencia artificial” en un sentido amplio para referirnos a distintos modelos avanzados de IA, entre ellos, de manera destacada, los modelos generativos y otros sistemas automatizados de apoyo a la decisión, en la medida en que comparten la capacidad de reconfigurar el entorno cognitivo en el que se ejerce el juicio. Las diferencias entre estos sistemas serán precisadas en la sección 2.

La hostilidad, por tanto, no se reduce a la circulación de falsedades o engaños, sino que afecta al modo mismo en que se estructuran cognitivamente la atención, la plausibilidad e incluso la propia autoridad, entendida como la facultad de otorgar potestad y crédito a las decisiones. Esta dimensión remite a la confianza epistémica, en la medida en que la distribución social del conocimiento depende de relaciones de deferencia, autoridad y fiabilidad que no pueden reducirse a razones estrictamente individuales (Origgi 61). En este sentido, un entorno epistémicamente hostil es aquel en el que la reorganización del entorno informacional erosiona las condiciones de posibilidad del conocimiento y de la comprensión. Por ello, su relevancia no es únicamente epistémica, sino también ética, en la medida en que compromete la capacidad de los agentes para deliberar, justificar y responder por sus decisiones; y operativa, porque dicha degradación emerge del propio entorno sociotécnico en el que esas decisiones se toman.

Una de las manifestaciones más visibles de esta ambivalencia se observa en el ámbito civil,[2] donde la expansión de la IA ha ido acompañada de fenómenos ampliamente discutidos, como la proliferación de desinformación, charlatanería (*bullshit*), teorías conspirativas, propaganda, noticias falsas (*fake news*) y ultrafalsos (*deepfakes*), todos ellos asociados a la erosión de la confianza social, la degradación del debate público y la creciente dificultad para distinguir entre información fiable, manipulación y ruido. Estos fenómenos responden a estrategias deliberadas de manipulación orientadas a fines comerciales o políticos. Tales estrategias son desplegadas tanto por grandes plataformas —nacionales o transnacionales— como por actores estatales que operan sobre sus propias poblaciones. En determinados contextos, especialmente cuando dichas prácticas son atribuidas a actores externos y reinterpretadas desde el prisma de la seguridad y la defensa, estas estrategias son formalizadas institucionalmente como manipulación e interferencia informativa extranjeras (FIMI, por sus siglas en inglés: *foreign information manipulation and interference*), en el marco de la denominada guerra cognitiva, entendida como una forma de conflicto orientada a alterar procesos cognitivos, explotar sesgos y afectar la toma de decisiones mediante herramientas informacionales y cibernéticas (Claverie y Du Cluzel 2-1; Matthews y Lamensch 14).

Sin embargo, la relevancia del problema no se limita al ámbito civil. En contextos de defensa, esta misma lógica puede desplegarse de manera más directamente

2 A lo largo del artículo se emplea la distinción, habitual en el ámbito de la defensa, entre los ámbitos civil y militar.

orientada al ecosistema militar, afectando la percepción situacional, la confianza, la moral o la capacidad de coordinación del personal y de las estructuras de mando. La propia OTAN, a través de su organismo científico-técnico, reconoce que las tecnologías capaces de alterar el comportamiento humano pueden dirigirse tanto a civiles como a militares (NATO STO 4). En consonancia con esta ampliación del ámbito de efectos, las operaciones de guerra cognitiva buscan producir efectos cognitivos no solo sobre poblaciones, sino también sobre dirigentes políticos y mandos militares (Deppe y Schaal 02, 04, 10). Desde un punto de vista más general, la OTAN caracteriza las amenazas informacionales como actividades intencionales, dañinas, manipulativas y coordinadas, desarrolladas por actores estatales y no estatales con el propósito de debilitar y dividir a la Alianza, a sus miembros y a sus socios (NATO 2025).

Ahora bien, sería un error reducir la hostilidad epistémica en defensa a operaciones exógenas de manipulación o interferencia. Incluso en ausencia de campañas hostiles externas, pueden producirse efectos parcialmente análogos a los que preocupan en el marco de la FIMI cuando la mediación algorítmica reorganiza la información disponible de tal manera que dificulta identificar qué resulta relevante, interpretar adecuadamente el contexto o mantener abierta la posibilidad de revisión crítica. En este último caso, la hostilidad epistémica no procede de una operación deliberada de influencia, sino que emerge como un riesgo estructural derivado de la propia arquitectura sociotécnica de la decisión, con efectos potencialmente desestabilizadores no solo sobre el juicio individual, sino también sobre las condiciones institucionales de confianza y validación del conocimiento (Miller 46).

Esta distinción permite introducir una precisión adicional especialmente relevante para el argumento de este trabajo. No toda forma de hostilidad epistémica debe entenderse como una amenaza en sentido estricto, es decir, como una acción coordinada orientada deliberadamente destinada a erosionar la cohesión de una organización o comunidad. Cuando efectos similares emergen sin una operación hostil externa identificable y son consecuencia del modo en que sistemas legítimos de apoyo a la decisión estructuran, priorizan y presentan la información, lo que aparece ya no es primariamente una amenaza, sino un riesgo inherente al propio entorno sociotécnico. El problema radica entonces en la posibilidad de que determinadas arquitecturas sociotécnicas generen, incluso sin fallos técnicos manifiestos ni intención manipulativa, condiciones como la opacidad, el cierre interpretativo o la dependencia cognitiva, capaces de erosionar la revisión crítica y, con ello, las bases epistémicas del juicio responsable.

La diferencia crucial radica en que el riesgo remite a una lógica de vulnerabilidad emergente, derivada de sistemas complejos, acoplados y sometidos a condiciones persistentes de incertidumbre (Battistelli y Galantino 65, 73-74; Ebert et al. 436 n. 6). Esta segunda acepción permite incorporar los riesgos epistémicos no intencionales. Dicho de otra manera, un sistema puede clasificar correctamente, priorizar de acuerdo con sus parámetros y presentar recomendaciones coherentes y, aun así, contribuir a un entorno en el que el agente humano dispone de una menor capacidad efectiva para identificar anomalías, reconsiderar presupuestos, contrastar alternativas o ejercer un veto significativo.

En este sentido, la hostilidad epistémica es estructural, es decir, una propiedad relacional del ensamblaje sociotécnico en el que una herramienta opera. Su emergencia depende de la interacción entre modelos algorítmicos, interfaces, protocolos institucionales, cadenas de mando, umbrales de confianza, circunstancias temporales y expectativas organizativas. Por ello, su evaluación exige desplazar la atención desde la intención del actor hacia las condiciones bajo las cuales se sitúa el juicio humano. La cuestión central no es, entonces, quién pretende influir, sino qué arquitectura decisional hace más probable que ciertas formas de comprensión, duda o desacuerdo dejen de estar disponibles en la práctica.

Definida de este modo, la noción de entorno epistémicamente hostil permite situar el análisis de la IA en el terreno de los riesgos estructurales. Esta perspectiva resulta especialmente importante en el ámbito de la defensa porque obliga a complementar las lógicas tradicionales de protección frente a amenazas con formas de gobernanza orientadas a la anticipación, la revisión institucional y la preservación de capacidades críticas de juicio bajo incertidumbre. En lo que sigue, el artículo se centrará precisamente en esta segunda acepción: la producción endógena de hostilidad epistémica en entornos de defensa mediados por sistemas de IA.

Para desarrollar este argumento, el artículo se organiza en tres momentos. En primer lugar, la sección 2 precisa el tipo de mediación epistémica introducida por distintas modalidades de IA, mostrando cómo los sistemas basados en reglas, los modelos de aprendizaje estadístico y la IA generativa reconfiguran de manera diferenciada la relación entre información, juicio humano y articulación entre hechos y valores. Esta sección permite mostrar que el problema no reside únicamente en la precisión técnica de los sistemas, sino también en el modo en que contribuyen a definir qué cuenta como relevante, plausible o justificable en un entorno decisional. En segundo lugar, la

sección 3 traslada este marco analítico al ámbito de la defensa, entendido como un caso límite de decisión bajo incertidumbre, presión temporal y consecuencias potencialmente irreversibles. Allí se examina cómo la mediación algorítmica puede transformar patrones estadísticos en criterios prácticos de acción, producir cierres interpretativos mediante recomendaciones plausibles y erosionar las condiciones de una supervisión humana significativa. Finalmente, la sección 4 extrae las consecuencias normativas del análisis y sostiene que una gobernanza responsable de la IA en defensa debe orientarse no solo a mejorar la eficiencia o la fiabilidad técnica, sino también a preservar la calidad epistémica del entorno sociotécnico en el que se ejerce el juicio responsable.

2. Inteligencia artificial, mediación epistémica y conexión entre hechos y valores en contextos de decisión automatizada

Para los fines de este trabajo, es necesario analizar cómo distintos modelos de IA han configurado formas diferenciadas de mediación epistémica. Esta precisión es importante porque los problemas que hoy plantean los sistemas de IA en contextos de toma de decisiones automatizadas no aparecen de forma repentina con la irrupción de la IA generativa ni pueden reducirse a errores técnicos, sesgos estadísticos o problemas de precisión.

Una mirada breve a la historia conceptual de la IA permite observar una tensión persistente entre dos proyectos complementarios: por un lado, comprender la cognición humana modelándola computacionalmente; por otro, construir sistemas artificiales capaces de realizar tareas que, en condiciones humanas, asociaríamos con la inteligencia (Haugeland 164). En este contexto, la Conferencia de Dartmouth de 1956 consolidó un supuesto que resultaría decisivo para el desarrollo posterior de ambos programas: el ideal de formalización. Como es bien sabido, dicho proyecto partía de la hipótesis de que ciertos aspectos de la inteligencia humana podían describirse con suficiente precisión formal como para ser simulados por una máquina (McCarthy et al. 2, 5-6, 8-10). La IA simbólica clásica representa la expresión más clara de este enfoque. En ella, el conocimiento se concibe como algo susceptible de codificarse mediante hechos, reglas y procedimientos inferenciales explícitos. La principal ventaja epistémica de este modelo reside en su trazabilidad: al menos en principio, es posible reconstruir

qué regla ha sido aplicada, qué premisas han sido consideradas y por qué se ha alcanzado una determinada conclusión. De este modo, se preserva un vínculo identificable entre inferencia, justificación y decisión.

Sin embargo, esta concepción del conocimiento como representación formalizada contrasta con desarrollos posteriores en ciencias cognitivas. En las últimas décadas, la comprensión ha dejado de concebirse exclusivamente como procesamiento interno de representaciones para entenderse como un fenómeno situado, corporizado y enactivo, inseparable de las interacciones prácticas con un entorno material, social y normativo (Varela et al. 147-184; Newen et al. 3-13). Desde este marco, comprender no consiste únicamente en producir resultados correctos o formalmente adecuados, sino en participar completamente en prácticas situadas de interpretación, acción y evaluación. La comprensión aparece así vinculada a formas de involucramiento con el mundo que exceden la mera producción de salidas de sistema (*outputs*) coherentes.

En la práctica, la mayor parte de los sistemas actualmente desplegados pertenecen al segundo de los proyectos descritos anteriormente: clasifican, predicen, recomiendan, generan contenido o apoyan la toma de decisiones humanas en dominios específicos, sin disponer de comprensión situada ni de una relación enactiva con el mundo (Russell y Norvig 30-39). Sin embargo, esta limitación no elimina su relevancia epistemológica. Al contrario, el hecho de que sistemas carentes de comprensión situada puedan producir resultados útiles, plausibles y operativamente influyentes obliga a examinar de qué manera intervienen en los procesos mediante los cuales ciertos hechos adquieren relevancia, determinados valores se traducen en criterios operativos y algunos cursos de acción aparecen como justificados.

La cuestión central, por tanto, no es únicamente si los sistemas de IA procesan datos de forma más rápida o eficiente que los humanos, sino cómo median la relación entre información, evidencia, valoración y decisión. En contextos de alto riesgo, especialmente en el ámbito de la defensa, esta mediación no puede evaluarse exclusivamente en términos de rendimiento técnico, sino también en función del tipo de entorno epistémico y normativo que contribuye a configurar.

Una manera de entender esta imbricación entre mediación epistémica, juicio responsable y conexión entre hechos y valores proviene de la crítica pragmatista a las dicotomías clásicas, las cuales han contribuido a estructurar —y, en ocasiones, limitar— la reflexión epistemológica contemporánea (Sans Pinillos, “Neglected pragmatism” 1110-1116). Entre ellas, la distinción entre hechos y valores resulta especialmente

relevante para el presente argumento. Que la separación tajante entre descripción y prescripción es filosóficamente inestable puede advertirse al examinar las propias prácticas orientadas a la producción, validación y circulación del conocimiento, las cuales incorporan criterios valorativos tales como la relevancia, la simplicidad, la plausibilidad, la economía y la aceptabilidad (Longino 7; Putnam 28-45). De manera similar, existen situaciones caracterizadas por su opacidad epistémica —como la incertidumbre o la ignorancia— que solo pueden gestionarse mediante el recurso a consideraciones de carácter axiológico.

Un ejemplo que permite comprender esta relación entre incertidumbre, juicio y valores epistémicos es el caso arqueológico del Disco de Festo.[3] Su descubrimiento resulta especialmente ilustrativo porque permite aproximarse a la experiencia del cambio de paradigma desde la perspectiva de quienes participan en el descubrimiento científico, y no únicamente desde la reconstrucción retrospectiva de un marco teórico ya consolidado (Feyerabend, *Science in a free society* 18).[4] En este punto, la crisis no se experimenta como una transición ordenada entre teorías, sino como una situación en la que los criterios disponibles dejan de orientar suficientemente la acción. Emociones como la sorpresa, la inquietud, la frustración o la angustia no son meros factores psicológicos externos al conocimiento, sino indicadores de una ruptura en la relación entre hechos, expectativas y posibilidades de intervención. En este sentido, la sorpresa, la ignorancia o el interés señalan que el marco disponible ha dejado de estabilizar la experiencia y activan procesos de revisión, exploración y búsqueda de nuevas interpretaciones (Vallverdú y Sans Pinillos 2; Sans Pinillos, "Unpacking bad expectations" 265). En tales circunstancias, valores como la relevancia, la fecundidad, la economía, la

3 El Disco de Festo es un artefacto, *prima facie*, minoico de arcilla, habitualmente datado en torno al segundo milenio a. C., cuyas inscripciones siguen sin haber sido descifradas de manera concluyente (Balistier 9). Su interés filosófico no reside aquí en el contenido concreto de sus signos, sino en el tipo de situación epistémica que produce: un objeto materialmente disponible, pero metodológicamente opaco, que resiste tanto la identificación directa como la analogía con objetos conocidos. En este sentido, puede entenderse como una *affordance* inerte: un objeto que no ofrece por sí mismo oportunidades claras de acción cognitiva dentro del marco interpretativo disponible (Sans Pinillos, "Horror vacui" 174, 184-185).

4 La facilidad con la que a menudo se conceptualiza el descubrimiento en términos de justificación puede entenderse a partir de la distinción propuesta por Feyerabend entre la perspectiva del observador y la del participante. La primera corresponde a filósofos de la ciencia, historiadores o analistas que reconstruyen retrospectivamente un descubrimiento, muchas veces desde un paradigma ya estabilizado. La segunda remite a quienes experimentaron una crisis científica desde dentro, sin disponer todavía de los criterios que más tarde permitirán ordenar, justificar o narrar el cambio como un descubrimiento o una transformación paradigmática. Por ello, acontecimientos que más tarde serán reconstruidos como paradigmáticos pueden vivirse inicialmente como fenómenos desconcertantes, caóticos, amenazantes o comprometedores (Feyerabend, "On the improvement" 401).

plausibilidad, la confianza o la responsabilidad metodológica funcionan como criterios de orientación práctica cuando los hechos disponibles no bastan por sí solos para determinar el curso de acción.

El Disco de Festo resulta particularmente interesante porque permite distinguir entre dos grados de desconocimiento. Por un lado, la ignorancia relativa, propia de situaciones en las que faltan datos dentro de un marco interpretativo estable. Por otro, una forma más radical de ignorancia, en la que el objeto o fenómeno investigado no ofrece oportunidades claras de acción cognitiva. Es decir, el Disco de Festo constituye un enigma para la arqueología no solo porque no haya sido descifrado de manera concluyente, sino porque todavía no existen herramientas suficientemente robustas para integrarlo de forma estable en un marco interpretativo compartido (Sans Pinillos y Magnani 201-202). Lo relevante para este trabajo no es el caso arqueológico en sí mismo, sino la forma en que los agentes gestionan la necesidad de avanzar en la investigación cuando carecen de referencias teóricas plenamente consolidadas, analogías firmes o criterios metodológicos suficientemente estabilizados.

En estos casos, los investigadores pueden verse obligados a formular hipótesis alejadas de los supuestos ordinarios de su disciplina. Tales hipótesis no se sostienen únicamente en aquello que la práctica científica permite afirmar como hecho, sino también en valores epistémicos y disciplinarios que orientan la decisión de seguir investigando: la relevancia atribuida al enigma, la fecundidad posible de una línea de interpretación, la responsabilidad metodológica ante una anomalía persistente o la convicción de que la resistencia del objeto revela un límite del marco disponible. Dicho de otro modo, en situaciones de ignorancia radical, el juicio científico se sostiene sobre una tensión entre lo que puede decirse que es el caso y lo que, desde ciertos valores epistémicos y disciplinarios, el investigador considera que todavía merece ser investigado (Sans Pinillos, “Horror vacui” 174-175, 188-189).

Esta tensión permite trasladar el problema al caso de la IA en entornos de defensa. La cuestión no consiste únicamente en determinar si los sistemas automatizados identifican correctamente determinados datos, sino en si preservan o degradan la capacidad humana para mantener abierta una articulación crítica entre hechos, valores y posibilidades de acción.

2.1. Tres modos de mediación epistémica: regla, patrón y plausibilidad

Una forma útil de entender la evolución epistemológica de la IA consiste en distinguir tres modos de mediación entre sistemas automatizados, información y juicio humano. Esta distinción no pretende ofrecer una periodización rígida ni una historia exhaustiva de la IA, sino identificar tres formas mediante las cuales los sistemas artificiales han transformado las condiciones bajo las que se produce, valida y utiliza el conocimiento en contextos de toma de decisiones.

Un ejemplo temprano permite situar el problema. En el contexto de la Guerra Fría, los primeros programas de traducción automática entre el ruso y el inglés partían de la premisa de que el lenguaje podía formalizarse como un sistema cerrado de reglas gramaticales y correspondencias sintácticas. Sin embargo, el informe ALPAC de 1966 mostró los límites de esta expectativa: una traducción funcional no dependía solo de reglas, sino también de la ambigüedad, el contexto, el conocimiento implícito y el uso situado del lenguaje (ALPAC 19-20, 76-77, 122). Este episodio es relevante porque anticipa una tensión que atraviesa toda la historia de la IA: la distancia entre formalizar una tarea y comprender el entorno práctico en el que esa tarea adquiere sentido. A efectos analíticos, esta tensión puede reconstruirse mediante tres modos históricos y funcionales de la IA que no se suceden linealmente, sino que coexisten y configuran distintas formas de mediación epistémica.

El **Modo 1** corresponde a la IA simbólica y a los sistemas basados en reglas. En este modelo, dominante durante las primeras décadas de la disciplina, el conocimiento se concibe como algo que puede representarse explícitamente mediante hechos, conceptos, reglas condicionales y procedimientos inferenciales. Los sistemas expertos, los demostradores de teoremas y las arquitecturas basadas en lógica formal ejemplifican esta aproximación. Su principal fortaleza epistémica reside en la trazabilidad: al menos en principio, es posible reconstruir qué regla se ha aplicado, qué premisas se han considerado y por qué se ha alcanzado una determinada conclusión.

Esta trazabilidad sitúa al agente humano en una posición relativamente sólida de control sobre el sistema. Si el dominio está bien delimitado, las reglas son explícitas y las condiciones de aplicación permanecen estables, el operador puede revisar el procedimiento y evaluar sus resultados. Sin embargo, los límites del Modo 1 aparecen precisamente cuando el entorno es ambiguo, cambiante o incompleto. En tales casos, no todas las variables relevantes pueden anticiparse ni formalizarse de antemano. Así pues,

la IA simbólica permite hacer visible aquello que puede codificarse, pero también tiende a invisibilizar aquello que queda fuera del marco formal. En entornos de defensa, esta limitación es decisiva, pues interpretar una amenaza, evaluar una intención hostil o valorar la proporcionalidad de una acción no son operaciones puramente sintácticas, sino prácticas situadas en marcos institucionales, normativos y operativos.

El **Modo 2** corresponde al aprendizaje automático y al aprendizaje profundo. Aquí el conocimiento ya no se introduce principalmente mediante reglas explícitas, sino que se infiere a partir de datos. El sistema aprende regularidades, correlaciones y patrones mediante el entrenamiento, ajustando sus parámetros internos para clasificar, detectar, predecir o recomendar. La validez operativa deja de apoyarse en la reconstrucción del razonamiento y pasa a evaluarse mediante métricas de rendimiento como la precisión, la sensibilidad, la tasa de error, la capacidad de generalización o la reducción de la pérdida.

Este desplazamiento transforma profundamente los criterios de aceptabilidad epistémica. Un modelo puede considerarse exitoso, aunque sus inferencias no puedan reconstruirse paso a paso en términos humanos. La inteligibilidad cede terreno frente al desempeño. Ello ha permitido avances notables en dominios como la visión artificial, el diagnóstico médico, la detección de anomalías o el análisis de grandes volúmenes de datos. Sin embargo, el mecanismo subyacente no es la comprensión situada, sino la identificación de patrones en espacios de alta dimensionalidad. El sistema, por ejemplo, no *sabe* qué es un tumor, una amenaza o una conducta sospechosa; simplemente *aprende* a asociar ciertas configuraciones de datos con etiquetas, resultados o probabilidades.

Estas etiquetas han sido previamente definidas en contextos institucionales específicos y condicionan qué diferencias resultan operativamente relevantes (Sans Pinillos y Costa, secc. 3). La dimensión valorativa de estas clasificaciones se vuelve especialmente visible cuando categorías como discapacidad, vulnerabilidad o incapacitación se trasladan entre distintos dominios institucionales. Mientras que en contextos asistenciales pueden orientar prácticas de cuidado o provisión de ayuda, en entornos de defensa, emergencia o conflicto pueden convertirse en criterios de vigilancia, priorización, restricción o exposición diferencial al riesgo (Farnós et al. 1, 8).

Desde la perspectiva de este artículo, el problema no reside únicamente en la opacidad técnica del Modo 2, sino también en sus implicaciones para la relación entre hechos y valores. Los sistemas de aprendizaje automático aprenden regularidades a

partir de datos producidos en contextos previos de observación, clasificación y decisión. Sin embargo, dichos datos no son neutrales: incorporan decisiones institucionales, sesgos, omisiones, categorías previas y condiciones concretas de producción. Por ello, pueden reproducir o amplificar patrones injustos incluso cuando funcionan conforme a criterios técnicos aparentemente adecuados (Sans Pinillos y Casacuberta 320-321). En entornos de defensa, esto implica que un sistema puede aprender a identificar amenazas, priorizar señales o recomendar acciones a partir de datos incompletos, sesgados u operacionalmente contaminados. La automatización no elimina el sesgo humano; lo estabiliza, lo acelera y lo reviste de autoridad técnica.

El **Modo 3** corresponde a la IA generativa y a los sistemas basados en modelos fundacionales, entendidos como modelos entrenados a gran escala sobre conjuntos amplios de datos y adaptables a una gran diversidad de tareas posteriores (Bommasani et al. 3). Este modo no rompe por completo con el Modo 2, ya que sigue basándose en el aprendizaje estadístico, las redes neuronales y el entrenamiento con grandes volúmenes de datos. Lo que cambia es el tipo de salida producida. Mientras que el Modo 2 clasifica, detecta o predice, el Modo 3 genera objetos simbólicos: textos, imágenes, sonidos, código, simulaciones, planes, explicaciones o recomendaciones. Sus antecedentes pueden situarse en modelos generativos como las GANs (Goodfellow et al. 2-3), aunque el salto decisivo en el procesamiento del lenguaje natural se produjo con la arquitectura *transformer*, que permitió representar relaciones contextuales de manera flexible y escalable (Vaswani et al. 2-6).

El rasgo epistemológicamente decisivo del Modo 3 no es solo que la máquina genere contenido, sino que produzca interpretaciones plausibles. Una IA generativa puede redactar informes, resumir documentación, proponer hipótesis, construir una justificación, simular deliberaciones o formular recomendaciones tácticas mediante un lenguaje que los usuarios reconocen como razonable. En ese punto, el sistema no solo entrega información: proporciona una forma de sentido. La interfaz comunicativa puede convertirse así en una fuente aparente de autoridad epistémica.

Esta transformación es especialmente relevante para el argumento de este trabajo. La plausibilidad no equivale necesariamente a la verdad, la comprensión ni la responsabilidad. Un sistema generativo puede producir enunciados correctos, pero también errores convincentes; puede organizar información relevante, pero también fijar marcos interpretativos prematuros; puede ampliar la capacidad humana de análisis, pero también sustituir gradualmente el esfuerzo deliberativo por formas de

dependencia cognitiva. En contextos de defensa, caracterizados por la presión temporal, la incertidumbre, la opacidad técnica y las consecuencias morales graves, el Modo 3 intensifica el riesgo de que una recomendación coherente sea tratada como si incorporara comprensión contextual y justificación normativa.

Por tanto, los tres modos no deben entenderse como una simple sucesión técnica ni como etapas que se reemplazan unas a otras. Cada uno produce una forma específica de mediación epistémica y modifica de manera distinta la relación entre información, conocimiento y juicio humano. El Modo 1 formaliza reglas y procedimientos, favoreciendo la trazabilidad de las inferencias, aunque al precio de simplificar el contexto. El Modo 2 aprende patrones a partir de datos y mejora el rendimiento en tareas complejas, pero incrementa la opacidad y la dependencia respecto de conjuntos de datos previamente estructurados. El Modo 3 genera contenidos e interpretaciones plausibles, ampliando las capacidades comunicativas y analíticas de los usuarios, aunque también corre el riesgo de adquirir una autoridad epistémica desproporcionada respecto de su capacidad real de comprensión. Lo relevante para este trabajo es que, en los tres casos, la IA no actúa como un instrumentos neutral de procesamiento de información. Por el contrario, participa activamente en la configuración del entorno epistémico al contribuir a determinar qué aparece como relevante, qué se presenta como plausible, qué alternativas resultan visibles y qué posibilidades de revisión crítica permanecen efectivamente disponibles para los agentes humanos.

2.2. De los modos de IA a la arquitectura del juicio

La distinción anterior no pretende todavía describir en detalle lo que más adelante denominaremos *colapso epistémico local*, sino delimitar la arquitectura epistémica en la que dicho riesgo puede emerger: el modo en que cada forma de IA redistribuye la relación entre información, interpretación, responsabilidad y posibilidad de revisión. Por ello, la integración de sistemas de IA en entornos de defensa no puede evaluarse únicamente en términos de capacidad técnica. Los sistemas basados en reglas, los modelos de aprendizaje estadístico y los sistemas generativos no solo realizan tareas distintas, sino que redistribuyen de manera diferenciada la carga epistémica entre máquinas, operadores e instituciones. En consecuencia, la pregunta central no es solo qué hace el sistema, sino qué tipo de entorno decisional contribuye a configurar.

En el Modo 1, el riesgo principal se vincula con la rigidez de la formalización. Aunque el sistema sea trazable, su eficacia depende de que las reglas disponibles

logren capturar los elementos relevantes del contexto. En el Modo 2, el problema se desplaza hacia la opacidad estadística y la dependencia de datos acumulados en contextos previos de observación, clasificación y decisión. En el Modo 3, el riesgo se intensifica porque el sistema no solo clasifica o predice, sino que produce explicaciones, escenarios y recomendaciones dotadas de coherencia narrativa. La mediación deja entonces de operar exclusivamente sobre el acceso a los hechos y comienza a intervenir en la construcción de su sentido práctico.

Esta evolución tiene una implicación directa para la noción de juicio responsable. En contextos de alto riesgo, *decidir* no consiste simplemente en procesar más información ni en escoger la opción técnicamente óptima. Implica *identificar* qué hechos son relevantes, qué incertidumbres deben mantenerse abiertas, qué valores están en juego, qué consecuencias son aceptables y qué responsabilidad se asume al actuar. Cuando un sistema automatizado filtra información, prioriza señales, genera hipótesis o presenta una recomendación como plausible, interviene precisamente en ese espacio de articulación entre hechos, valores y acción.

Desde esta perspectiva, la IA no debe entenderse solo como una herramienta externa al juicio humano, sino como un componente de una arquitectura sociotécnica que puede fortalecer o debilitar las condiciones bajo las cuales ese juicio se ejerce. Puede ampliar capacidades, reducir la sobrecarga informativa y acelerar la detección de patrones. Pero también puede favorecer cierres interpretativos prematuros, incrementar la dependencia cognitiva, ocultar incertidumbres relevantes o desplazar criterios normativos hacia parámetros operativos.

Este punto permite conectar la tipología de los tres modos con la tesis general del artículo. Un entorno epistémicamente hostil no surge únicamente cuando circula información falsa o cuando un sistema falla de manera manifiesta. También puede emerger cuando la arquitectura decisional organiza la información de tal manera que los operadores pierden capacidad efectiva para reconocer anomalías, cuestionar recomendaciones, revisar presupuestos o ejercer un veto significativo. En este sentido, un sistema puede funcionar de acuerdo con sus parámetros técnicos y, aun así, degradar las condiciones necesarias para el ejercicio de un juicio responsable.

La IA generativa ocupa aquí una posición especialmente delicada. Al producir explicaciones, informes, simulaciones o recomendaciones en lenguaje natural, puede presentar como ya articulada una relación entre hechos, valores y cursos de acción que el operador debería poder reconstruir críticamente por sí mismo. Bajo presión

temporal, una interpretación generada de forma convincente puede reducir la disposición a considerar señales débiles, hipótesis alternativas o fricciones contextuales. El riesgo no es solo que la IA sustituya al humano, sino que reorganice el entorno decisional de tal manera que la intervención humana quede reducida a una validación meramente formal.

De ahí que la supervisión humana no pueda entenderse como una simple presencia dentro de la cadena decisional. Para que la noción de humano en el bucle posea contenido normativo y epistémico, el operador y, en contextos complejos, los distintos humanos implicados en el proceso (Sans Pinillos, "CELL" sec. 3.4) deben conservar condiciones reales de comprensión, duda, revisión y veto. En rigor, el juicio humano siempre ha formado parte de los procesos de decisión. Por ello, el problema no consiste en añadir un humano al bucle como garantía formal, sino en preservar las circunstancias bajo las cuales los distintos agentes humanos implicados pueden ejercer un juicio significativo dentro de una arquitectura decisional ya mediada por sistemas IA. Esto exige que dichos sistemas no sean evaluados únicamente por su precisión o eficiencia, sino también por su impacto sobre la calidad epistémica del entorno sociotécnico en el que se toman decisiones. La cuestión decisiva no es si el humano permanece en el bucle, sino si el propio bucle conserva las condiciones necesarias para que el juicio humano siga siendo significativo.

3. De la asistencia a la supervisión nominal: IA, juicio humano y colapso epistémico local en defensa

3.1. Defensa como caso límite de hostilidad epistémica

La elección de los entornos de defensa como caso de estudio no responde únicamente a su relevancia estratégica, sino también a su carácter paradigmático como contextos epistémicamente hostiles. En operaciones militares de alto riesgo, el juicio se ejerce bajo condiciones que tensionan al máximo las capacidades cognitivas de los agentes: presión temporal extrema, información incompleta o contradictoria, ambigüedad de entorno y cambios contextuales abruptos, consecuencias potencialmente irreversibles y una marcada asimetría entre la urgencia de la decisión y las posibilidades de verificación. A ello se suma una arquitectura sociotécnica compleja, —cadenas de mando jerárquicas, sensores distribuidos, sistemas de conciencia situacional, reglas de

enfrentamiento y protocolos operativos— que ya media de forma significativa la relación entre los operadores y el entorno.

En este escenario, la incorporación de sistemas de IA se produce dentro de un ecosistema donde la identificación de elementos relevantes, la interpretación de señales débiles y la revisión crítica ya son tareas frágiles. Precisamente por ello, el ámbito de la defensa permite observar con especial claridad cómo la IA puede reorganizar los procesos de decisión: no solo ampliando las capacidades de análisis, sino también intensificando dinámicas de dependencia cognitiva, reforzando cierres interpretativos prematuros y reduciendo la sensibilidad ante anomalías o fricciones contextuales.

La IA ha pasado de desempeñar funciones de apoyo analítico a integrarse progresivamente en capacidades operativas críticas como la inteligencia, la vigilancia y el reconocimiento, el análisis de imágenes satelitales, la detección de amenazas, la ciberdefensa, la guerra electrónica, la planificación táctica y el apoyo a la toma de decisiones. En estos ámbitos, los sistemas de aprendizaje automático permiten procesar grandes volúmenes de datos, identificar patrones y generar alertas o recomendaciones que superan la capacidad humana de análisis en tiempo real. Sin embargo, esta ampliación funcional no es epistémicamente neutral. Al filtrar datos, priorizar señales y ordenar escenarios, dichos sistemas contribuyen a definir qué aparece como relevante para quienes deben decidir.

El problema se vuelve especialmente crítico cuando la IA participa en funciones vinculadas al uso de la fuerza. Los sistemas de *selección de objetivos* automatizados o asistidos por IA, por ejemplo, combinan datos procedentes de múltiples sensores con modelos predictivos para identificar o priorizar objetivos, constituyendo un punto de convergencia entre la eficiencia operativa y el riesgo ético, especialmente cuando actúan como sistemas de apoyo a la decisión en procesos vinculados con el empleo de la fuerza (Pratzner 87).

La transición desde la detección de patrones hacia la recomendación de acciones introduce una forma de mediación algorítmica que puede alterar la relación entre información, juicio y decisión. En tales casos, no basta con preguntar si el sistema mejora la velocidad o la precisión del análisis; resulta igualmente necesario preguntarse si preserva las condiciones requeridas para evaluar el contexto, distinguir adecuadamente entre objetivos legítimos e ilegítimos, aplicar criterios de proporcionalidad y precaución y, en última instancia, atribuir responsabilidades por las decisiones adoptadas (Schmitt y Widmar 134).

Como se argumentará en las conclusiones, una vía central para mitigar estos riesgos consiste en incorporar principios de IA responsable desde el diseño (*responsible AI by design*), no como un añadido posterior, sino desde las fases de generación, selección y curación de datos, así como en el diseño de interfaces, protocolos de revisión y condiciones efectivas de supervisión humana. En conjunto, estos desarrollos muestran que la IA no solo incrementa la capacidad operativa de los sistemas de defensa, sino que reconfigura los procesos mediante los cuales se identifican amenazas, se interpretan situaciones y se decide actuar. La cuestión decisiva, entonces, no es exclusivamente tecnológica, sino también epistémica y ética: ¿en qué medida la mediación algorítmica altera las condiciones bajo las cuales los agentes humanos pueden ejercer un juicio responsable en contextos donde las consecuencias del error pueden resultar irreversibles?

3.2. Del patrón estadístico a la decisión normativa

La dificultad no reside solo en los dilemas normativos asociados al uso de la fuerza, sino también en la fragilidad epistémica de los sistemas que median esas decisiones. Los modelos basados en aprendizaje estadístico y generación plausible no acceden a la verdad operativa del contexto en sentido fuerte, sino que optimizan patrones a partir de datos, etiquetas o criterios de evaluación previamente definidos. Su capacidad de generalización depende de la calidad, representatividad y estabilidad de esos datos, lo que introduce vulnerabilidades como sesgos sistemáticos, errores fuera de distribución, alucinaciones en modelos generativos o manipulaciones deliberadas mediante *data poisoning* (envenenamiento de datos). En contextos relativamente estables, estos sistemas pueden alcanzar niveles muy elevados de eficacia; sin embargo, en entornos dinámicos, ambiguos o adversariales, pequeñas desviaciones en los datos de entrada o en el contexto pueden producir errores difíciles de detectar oportunamente.

Aquí aparece uno de los problemas centrales para el argumento de este trabajo: el rendimiento estadístico no equivale ni a comprensión contextual ni a justificación normativa. Los sistemas de aprendizaje automático identifican regularidades a partir de datos acumulados en contextos previos de observación, clasificación y decisión, pero esos datos no son neutrales. Incorporan decisiones institucionales, sesgos, omisiones, categorías previas y condiciones concretas de producción. Por ello, los algoritmos de *machine learning* (aprendizaje automático) pueden reproducir e incluso amplificar patrones injustos presentes en el mundo social, aun cuando operen de acuerdo con criterios técnicos aparentemente correctos (Sans Pinillos y Casacuberta 320-321).

La dificultad se agrava porque estos sistemas pueden simular sensibilidad moral y, sin embargo, verse afectados por variaciones superficiales en la formulación, el encuadre o la presentación del caso, lo que dificulta distinguir entre sensibilidad a consideraciones normativamente relevantes y una mera respuesta a artefactos de la entrada (*input*) (Oh y Demberg, 2-4, 6, 10-11, 18). Más aún, desde la perspectiva de la conexión entre hechos y valores, el problema no consiste solamente en que el sistema pueda equivocarse acerca de los hechos, sino en que puede transformar regularidades descriptivas en criterios prescriptivos para la acción. Aquello que ha ocurrido con frecuencia puede pasar a funcionar como fundamente de lo que debe hacerse. En términos filosóficos, esto reproduce una variante de falacia naturalista: inferir indebidamente un deber ser a partir de lo que es (Moore I B §10), un riesgo especialmente sensible cuando se intenta traducir valores humanos a reglas algorítmicas o preferencias cuantificables (Sans Pinillos, "Apuntes" 146).

Esta cuestión adquiere una gravedad específica en el ámbito de la defensa. Si un sistema aprende a identificar amenazas a partir de datos acumulados incompletos, sesgados u operacionalmente contaminados, puede reforzar patrones previos de sospecha, vigilancia o intervención. La automatización no elimina entonces el sesgo humano; lo estabiliza, lo acelera y lo reviste de autoridad técnica. El problema no es únicamente la posibilidad de error, sino la consolidación de una determinada economía de la relevancia: qué señales importan, qué patrones merecen atención, qué anomalías se descartan y qué riesgos pasan a ser aceptables.

Por ello, la mediación algorítmica resulta especialmente problemática cuando interviene en decisiones que requieren evaluación normativa. Decidir qué hacer con una información implica determinar qué valores están en juego, qué daños deben evitarse, qué incertidumbre resulta tolerable, qué riesgos pueden asumirse y qué responsabilidades se activan. En este punto, uno de los errores más importantes en ética de la IA consiste en confundir valores con preferencias. Una preferencia expresa una elección dentro de un marco dado; un valor, en cambio, orienta la evaluación de ese marco (Sans Pinillos y Casacuberta 320-321). Reducir valores a preferencias implica transformar una cuestión normativa en un problema de agregación estadística, gestión probabilística de riesgos o ajuste entre variables entre, cuando precisamente estos procesos siguen requiriendo una justificación ética sobre cómo se distribuyen los riesgos, quién queda expuesto a ellos y qué valores terminan privilegiándose en la práctica (Goodall 817; Himmelreich 678).

La distinción es fundamental en los contextos de defensa. Valores como el respeto a la dignidad humana, junto con principios como la distinción, la proporcionalidad, la precaución, la necesidad militar o responsabilidad de mando, no funcionan como preferencias ordenables en una escala de utilidad, sino como criterios normativos que estructuran la justificación de la acción. Cuando un sistema los traduce en pesos, umbrales o *rankings* operativos, puede generar la apariencia de que la cuestión normativa ha quedado resuelta cuando, en realidad, solo ha sido operacionalizada.[5] La decisión aparece entonces como si derivara directamente de los datos, cuando depende de elecciones previas relativas a qué medir, cómo clasificar, qué optimizar, qué excluir y qué riesgos considerar aceptables.

3.3. Plausibilidad generativa y cierre deliberativo

La IA generativa profundiza el desplazamiento descrito en la sección 2.2 porque no solo clasifica o predice, sino que también produce objetos simbólicos dotados de carga interpretativa, tales como textos, explicaciones, escenarios o recomendaciones. Esta capacidad no incorpora comprensión situada, pero sí transforma el modo en que la información se presenta, se articula y adquiere relevancia operativa, especialmente cuando las salidas de sistema resultan plausibles, coherentes y accionables en contextos de decisión.

En contextos de defensa, esta plausibilidad puede tener efectos directos sobre el juicio. Un sistema generativo puede resumir documentación, organizar un escenario, atribuir relevancia a determinadas señales, sugerir intenciones del adversario —por ejemplo, a partir de etiquetas culturales, sociológicas, históricas u operativas—, construir narrativas causales y recomendar cursos de acción. En ese punto, la IA deja de mediar únicamente el acceso a los hechos para intervenir también en la configuración de su sentido práctico. No se limita a indicar que algo es probable; puede presentar una situación como urgente, una hipótesis como razonable, una acción como proporcionada o una alternativa como secundaria. La frontera entre asistencia informacional y orientación del juicio se vuelve, así, mucho más porosa.

5 La operacionalización de un principio normativo en forma de instrucción, umbral o regla ejecutable no implica por sí misma su justificación moral. El riesgo consiste en que dicho principio deje de funcionar como un criterio crítico de evaluación y pase a operar como un procedimiento técnico cuya aplicación parezca suficiente, desplazando la deliberación contextual que originalmente justificaba su validez.

Esta dinámica reactualiza el problema de la apariencia de competencia en un nuevo contexto. Un sistema puede producir respuestas formalmente correctas, lingüísticamente coherentes o pragmáticamente útiles sin que ello implique comprensión situada, experiencia del mundo o capacidad genuina de justificación (Haas et al. 565). En escenarios de alto riesgo, esta diferencia es crucial, pues la plausibilidad comunicativa puede modificar la relación entre información y juicio. Una recomendación presentada de forma coherente, ordenada y aparentemente razonable puede reducir la disposición del operador a cuestionar los supuestos implícitos sobre los que se sustenta, especialmente bajo presión temporal.

Esta pérdida de agencia deliberativa puede verse reforzada por sesgos ampliamente documentados en contextos de apoyo automatizado a la decisión, como el sesgo de automatización, la delegación acrítica o el sesgo de autoridad (Skitka et al. 991-995, 998-1004; Parasuraman y Riley 237-240, 248-250). En estos casos, el agente tiende a aceptar, priorizar o dejar de cuestionar una recomendación debido a su forma de presentación, a su aparente coherencia o a su procedencia técnica, más que a partir de un examen independiente de sus fundamentos. El riesgo, por tanto, no comienza únicamente cuando la IA se equivoca, sino cuando su forma de presentación induce a atribuirle una autoridad epistémica que transforma indebidamente el espacio de decisión.

Este punto conecta directamente con la tesis general del artículo. El problema no radica únicamente en que el sistema pueda alucinar, errar o fabricar información. Más profundamente, consiste en que puede organizar hechos y valores en una narrativa suficientemente plausible como para desactivar la deliberación crítica. Bajo presión temporal, una interpretación generada de forma convincente puede reducir la atención a anomalías, señales débiles o hipótesis alternativas. La hostilidad epistémica aparece entonces cuando el entorno decisional deja de favorecer la revisión crítica y comienza a inducir una convergencia acelerada hacia la interpretación ofrecida por el sistema.

3.4. Humano en el bucle y colapso epistémico local

El punto anterior permite precisar el problema institucional del humano en el bucle. En defensa, la incorporación de sistemas de IA no elimina necesariamente al agente humano, pero puede transformar su papel dentro de entornos caracterizados por presión temporal, incertidumbre estructural, jerarquías operativas y consecuencias morales de alto impacto. En estos contextos, la transición desde sistemas basados en reglas hacia sistemas de inferencia estadística y, más recientemente, hacia sistemas generativos

capaces de producir recomendaciones, interpretaciones y cursos de acción puede desplazar progresivamente las condiciones efectivas del juicio humano. El agente no desaparece de la cadena decisional, pero puede dejar de ser una fuente primaria de interpretación y decisión para convertirse en una instancia de validación de resultados producidos por sistemas cuya lógica interna resulta, en gran medida, opaca.

En este contexto, la noción ampliamente difundida del humano en el bucle aparece como una garantía insuficiente. La mera inserción formal del humano en la cadena de decisión puede ocultar una dinámica en la que las capacidades fundamentales para el juicio —comprensión del contexto, identificación de lo relevante, capacidad de cuestionamiento, autoridad para disentir— se ven progresivamente erosionadas. Esta crítica no implica negar la importancia del control humano, sino precisar sus condiciones de posibilidad: cuando se asignan a los humanos funciones críticas dentro de un sistema, es necesario analizar si el diseño sociotécnico permite realmente atender, comprender, evaluar y aplicar la información recibida, en lugar de presuponer que la presencia humana basta para evitar el fallo (Cranor 1-2, 8-9).

En este sentido, conviene distinguir entre el control supervisor humano (*human supervisory control*) y el control humano significativo (*meaningful human control*). La noción de control supervisor humano, desarrollada por Sheridan en el marco de la telerrobótica y la automatización, permite entender que la función humana en sistemas automatizados no consiste necesariamente en ejercer un control directo y continuo sobre cada operación, sino en supervisar sistemas capaces de ejecutar tareas con distintos grados de autonomía, manteniendo capacidades de planificación, monitorización, diagnóstico e intervención (Sheridan 1-3). Ahora bien, en contextos de defensa de alto riesgo, esta supervisión solo adquiere relevancia normativa cuando se transforma en control humano significativo: esto es, cuando preserva una capacidad efectiva de comprensión, evaluación crítica, intervención y veto. El problema no es, por tanto, asegurar una presencia humana formal, sino garantizar que la arquitectura sociotécnica conserve las condiciones bajo las cuales la supervisión puede seguir siendo epistémica y moralmente relevante. Solo así el control humano puede entenderse no como una agencia sustraída por la automatización, sino como una agencia apoyada, refinada y ampliada mediante marcos sociotécnicos que redistribuyen responsabilidades y aumentan la probabilidad de resultados moralmente valiosos (Floridi et al. 689, 692).

En este sentido, la orientación centrada en el ser humano no debe entenderse como una apelación genérica a la centralidad humana, sino como una exigencia

epistémica: preservar un punto de vista situado capaz de evaluar valores, responsabilidades y límites de la autonomía técnica sin convertir esa centralidad en antropocentrismo acrítico (Rodilosso 2-6, 10). Cuando el entorno decisional está estructurado por sistemas que filtran, priorizan y generan información de forma autónoma, el operador ya no interviene sobre una situación bruta, sino sobre una representación previamente configurada por la máquina. El riesgo no es, por tanto, la automatización completa de la decisión, sino el desanclaje epistémico del juicio humano: una situación en la que las decisiones se adoptan sin una conexión suficientemente robusta con los marcos contextuales, normativos y valorativos que deberían justificarlas.

En contextos operativos, esta dinámica puede derivar en formas de dependencia cognitiva estructural. Una vez que un sistema ha sido integrado en los flujos de trabajo y ha demostrado utilidad operativa, su uso tiende a normalizarse, reduciendo progresivamente la disposición a cuestionar sus resultados o a operar sin él. En el caso de la IA generativa, esta tendencia se intensifica porque las recomendaciones pueden presentarse mediante explicaciones coherentes, narrativas familiares y formas lingüísticas que inducen una atribución excesiva de competencia epistémica. Sistemas que "hablan como nosotros" pueden ser tratados como si también "comprendieran como nosotros".

Como se ha señalado, en ámbitos como la *selección de objetivos* asistida por IA, esta normalización puede convertir la verificación humana en un proceso crecientemente procedimental o simbólico, próximo a lo que algunos análisis describen como ratificación mecánica (*rubber stamping*), especialmente cuando la velocidad y escala de los sistemas de apoyo a la decisión reducen el margen efectivo para la comprobación crítica (Bruun y Bo 15).

Este fenómeno adquiere especial relevancia en defensa, donde las decisiones no solo deben ser eficaces, sino también justificables en términos éticos, jurídicos y políticos. La delegación progresiva de funciones cognitivas en sistemas que operan sobre correlaciones estadísticas o generan conocimiento plausible sin garantías de comprensión situada introduce el riesgo de que las decisiones humanas queden desvinculadas de los valores que deberían guiarlas. En escenarios extremos, esta dinámica puede desembocar en lo que denominaremos colapso epistémico local: una situación en la que el agente humano conserva formalmente la autoridad decisional, pero pierde las condiciones prácticas necesarias para interpretar, evaluar críticamente y, en última instancia, rechazar las recomendaciones del sistema.

Este colapso puede identificarse cuando concurren tres condiciones. En primer lugar, cuando la representación algorítmica pasa a estructurar de forma dominante qué cuenta como información relevante. En segundo lugar, cuando las incertidumbres, alternativas o anomalías dejan de ser visibles, inteligibles o practicables para el operador. Finalmente, cuando el veto humano permanece formalmente disponible, pero resulta material, institucional o cognitivamente improbable. La literatura sobre sesgo en IA militar refuerza esta preocupación: los sesgos presentes en sistemas de apoyo a la decisión y en armas autónomas no constituyen únicamente problemas técnicos, sino riesgos capaces de afectar la identificación de objetivos, la detección de personas u objetos protegidos y, en consecuencia, el cumplimiento de principios fundamentales del derecho internacional humanitario, como la distinción, la proporcionalidad y la precaución (Bruun y Bo 10-18).

Desde esta perspectiva, el colapso epistémico local no debe entenderse como la mera acumulación de errores, sesgos o fallos técnicos, sino como la degradación de las condiciones que permiten al agente humano actuar como sujeto de conocimiento responsable. El colapso no implica la desaparición del humano de la cadena decisional, sino su transformación en una instancia de validación dentro de un entorno donde la comprensión, la duda y el desacuerdo dejan de estar efectivamente disponibles. En paralelo, la noción de racionalidad limitada (Simon 196-205) adquiere una nueva dimensión: no se trata únicamente de limitaciones inherentes a las capacidades cognitivas del agente, sino de limitaciones producidas y amplificadas por la propia arquitectura sociotécnica en la que el juicio tiene valor.

De ahí que el problema no pueda resolverse únicamente mediante la inserción formal del humano en el bucle ni mediante mejoras incrementales de precisión, explicabilidad o supervisión procedimental. Lo que está en juego es la calidad epistémica del entorno sociotécnico en el que se toman decisiones. Un sistema puede ser rápido, robusto y estadísticamente eficiente y, aun así, generar un entorno hostil para el juicio si reduce la sensibilidad a señales débiles, bloquea hipótesis alternativas, presenta recomendaciones desprovistas incertidumbre inteligible o desplaza cuestiones normativas hacia parámetros operativos. Desde esta perspectiva, el control humano debe entenderse como una capacidad situada, dependiente del diseño del sistema, de la transparencia de los datos, de los procedimientos de verificación, de la formación del operador y de la autoridad institucional para suspender o rechazar una recomendación.

Por ello, la integración de IA en defensa debe evaluarse no solo por su rendimiento técnico, sino también por su impacto sobre las condiciones efectivas de comprensión, duda, revisión y veto. La cuestión decisiva no es si el humano permanece en el bucle, sino si el bucle conserva las condiciones necesarias para que el juicio humano siga siendo significativo. El riesgo más profundo no es que la IA sustituya abiertamente al ser humano, sino que reorganice el entorno de tal manera que el juicio humano sobreviva solo como trámite.

4. Conclusiones: preservar el juicio responsable en entornos epistémicamente hostiles

El objetivo de este artículo ha sido mostrar que la integración de sistemas de IA en contextos de defensa de alto riesgo no plantea únicamente problemas de precisión, fiabilidad técnica o atribución formal de responsabilidad. El problema más profundo es epistémico y normativo: determinados sistemas de IA pueden reorganizar las condiciones bajo las cuales los agentes humanos identifican información relevante, interpretan situaciones, conectan hechos con valores y revisan críticamente recomendaciones automatizadas en contextos caracterizados por incertidumbre e ignorancia. En este sentido, la hostilidad epistémica no debe entenderse solo como el resultado de campañas externas de manipulación, desinformación o interferencia, sino también como un riesgo estructural que puede emerger dentro de arquitecturas sociotécnicas legítimas y técnicamente funcionales.

Esta distinción entre amenaza y riesgo resulta central. Cuando la hostilidad epistémica procede de una operación deliberada de influencia, el problema se formula en términos de ataque, interferencia o manipulación. Pero cuando emerge de la propia organización del entorno decisional —a través de modelos opacos, interfaces persuasivas, protocolos rígidos, presiones temporales o cadenas de mando que dificultan la revisión—, el problema no es primariamente la intención hostil de un actor externo, sino la vulnerabilidad estructural del sistema. Un sistema puede clasificar correctamente, priorizar de acuerdo con sus parámetros y presentar recomendaciones coherentes y, aun así, contribuir a degradar las condiciones efectivas del juicio responsable.

La distinción presentada entre los tres modos de mediación epistémica permite precisar que el riesgo no reside en un único tipo de fallo técnico, sino en distintas formas

de reconfiguración del entorno decisional. Los sistemas basados en reglas pueden empobrecer el contexto; los modelos estadísticos pueden estabilizar sesgos y opacidades; y los sistemas generativos pueden producir una autoridad epistémica aparente mediante interpretaciones plausibles. En todos los casos, la IA no solo procesa información: contribuye a configurar qué aparece como relevante, qué se presenta como plausible y qué posibilidades de duda, revisión o veto permanecen abiertas para el agente humano.

El análisis realizado sobre la conexión entre hechos y valores permite entender por qué este problema no puede resolverse mediante una concepción meramente técnica de la decisión. En contextos de defensa, decidir no consiste simplemente en identificar datos correctos o maximizar una función de rendimiento, sino en determinar qué incertidumbres deben mantenerse abiertas, qué daños deben evitarse, qué riesgos son aceptables, qué principios normativos estructuran la acción y qué responsabilidades se activan. Por ello, cuando un sistema transforma patrones estadísticos en recomendaciones operativas, o valores normativos en pesos, umbrales o *rankings*, puede generar la apariencia de que una cuestión práctica ha sido resuelta técnicamente cuando, en realidad, solo ha sido operacionalizada.

Esta es la razón por la que la noción de *humano en el bucle* resulta insuficiente si se entiende meramente como un procedimiento. La presencia humana en la cadena decisional no garantiza por sí sola control, responsabilidad ni legitimidad. Para que la supervisión humana sea significativa, los operadores deben conservar condiciones reales de comprensión, duda, revisión, desacuerdo y veto. De otro modo, el ser humano puede permanecer formalmente en el bucle mientras su función práctica queda reducida a la validación de interpretaciones previamente estructuradas por la máquina.

La noción de colapso epistémico local introducida en este trabajo nombra precisamente esta degradación. No se trata de la desaparición del agente humano ni de su sustitución completa por sistemas automatizados, sino de una pérdida situada de las condiciones que hacen posible el ejercicio del juicio responsable. El agente sigue presente, pero su capacidad para comprender el contexto, identificar anomalías, cuestionar supuestos, conectar hechos con valores y rechazar una recomendación se ve comprometida estructuralmente. Este colapso no requiere fallos evidentes; puede producirse precisamente cuando los sistemas funcionan con suficiente eficacia, coherencia y plausibilidad como para desalentar la deliberación crítica.

De este análisis se desprende una exigencia normativa clara: la integración de la IA en la defensa debe evaluarse no solo por su precisión, robustez o eficiencia, sino

también por su impacto en la calidad epistémica del entorno sociotécnico en el que se toman decisiones. Una aproximación de IA responsable desde el diseño debe anticipar estos riesgos desde las fases de generación, curación y selección de datos, pero también mediante el diseño de interfaces, protocolos de uso, umbrales de confianza, procedimientos de revisión, formación de operadores y mecanismos institucionales de distribución de la autoridad. No basta construir sistemas más eficaces; es necesario preservar las condiciones bajo las cuales el desacuerdo humano sigue siendo posible.

En términos prácticos, esto implica diseñar sistemas que hagan visibles las carencias informativas, expongan alternativas, permitan rastrear supuestos relevantes, mantengan sensibilidad a señales débiles, eviten cierres interpretativos prematuros y garanticen un veto humano efectivo. También exige evitar que principios normativos como la dignidad humana, la distinción, la proporcionalidad, la precaución, la necesidad militar o la responsabilidad de mando sean tratados como meras preferencias cuantificables. Estos principios no son variables decorativas del cálculo, sino condiciones de justificación de la acción.

La tesis defendida, por tanto, es que la incorporación de la IA debe evaluarse desde una concepción más exigente de la supervisión humana. La pregunta decisiva no es si hay un humano en el bucle, sino si el bucle conserva los recursos epistémicos, materiales e institucionales necesarios para que el juicio humano siga siendo significativo. Cuando esos recursos se degradan, la IA no sustituye necesariamente al ser humano; produce algo más difícil de advertir y, por ello, más peligroso: reorganiza el entorno en uno epistémicamente hostil, donde la capacidad de comprender, cuestionar, deliberar y asumir responsabilidad por las decisiones adoptadas se erosiona progresivamente.

Financiación: Alger Sans Pinillos contó con el apoyo de la iniciativa “Generación D” (Red.es, Ministerio para la Transformación Digital y de la Función Pública) para la atracción de talento (C005/24-ED CV1), financiada por el programa NextGenerationEU de la Unión Europea a través del PRTR.

Declaración ética: las opiniones y puntos de vista expresados son exclusivamente responsabilidad de los autores y no reflejan necesariamente los de la Unión Europea ni los de la Comisión Europea. Ni la Unión Europea ni la Comisión Europea pueden ser consideradas responsables de ellos.

Referencias

ALPAC. *Language and machines: computers in translation and linguistics: a report by the automatic language processing advisory committee, division of behavioral sciences, National Academy of Sciences, National Research Council*. National Academy of Sciences, 1966.

Balistier, Thomas. *The phaistos disk: an account of its unresolved mystery*. Verlag, 2000.

Battistelli, Fabrizio y Maria Grazia Galantino. "Dangers, risks and threats: An alternative conceptualization to the catch-all concept of risk". *Current Sociology*, vol. 67, núm. 1, 2019, pp. 64–78. https://doi.org/10.1177/0011392118793675

Bommasani, Rishi, et al. "On the opportunities and risks of foundation models". arXiv, 2022. https://doi.org/10.48550/arXiv.2108.07258

Bruun, Laura y Mata Bo. *Bias in Military Artificial Intelligence and Compliance with International Humanitarian Law*. Stockholm International Peace Research Institute, 2025. https://doi.org/10.55163/NLWV5347

Claverie, Bernard y François Du Cluzel. "'Cognitive warfare': The advent of the concept of 'cognitics' in the field of warfare". *Cognitive warfare: the future of cognitive dominance*, editado por B. Claverie, B. Prébot, N. Buchler y F. Du Cluzel. NATO Science and Technology Organization, 2022.

Cranor, Lorrie Faith. "A framework for reasoning about the human in the loop". *Proceedings of the 1st Conference on usability, psychology, and security*. USENIX Association, 2008, pp. 1-15.

Deppe, Christoph y Gary Schaal. "Cognitive warfare: A conceptual analysis of the NATO ACT cognitive warfare exploratory concept". *Frontiers in Big Data*, vol. 7, 2024, article 1452129. https://doi.org/10.3389/fdata.2024.1452129

Ebert, Philip, et al. "Varieties of risk". *Philosophy and Phenomenological Research*, vol. 101, núm. 2, 2020, pp. 432–455. https://doi.org/10.1111/phpr.12598

Farnós, Joan, et al. "Ethical prompting: Toward strategies for rapid and inclusive assistance in dual-use AI systems". *Frontiers in Artificial Intelligence*, vol. 8, 2025, 1646444. https://doi.org/10.3389/frai.2025.1646444

Feyerabend, Paul K. (1978). *Science in a free society*. Lowe & Brydone Ltd.

Feyerabend, Paul K. "On the improvement of the sciences and the arts, and the possible identity of the two". *Proceedings of the Boston Colloquium for the Philosophy*

of Science 1964/1966, editado por R. S. Cohen y M. W. Wartofsky. Springer, 1967, pp. 387–415. https://doi.org/10.1007/978-94-010-3508-8_22

Floridi, Luciano, et al. "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations". *Minds and Machines*, vol. 28, núm. 4, 2018, pp. 689–707. https://doi.org/10.1007/s11023-018-9482-5

Goodall, Noah J. "Away from trolley problems and toward risk management". *Applied Artificial Intelligence*, vol. 30, núm. 8, 2016, pp. 810–821. https://doi.org/10.1080/08839514.2016.1229922

Goodfellow, Ian J., et al. "Generative adversarial nets". *Advances in Neural Information Processing Systems*, vol. 27, 2014. https://doi.org/10.48550/arXiv.1406.2661

Haas, Julia, et al. "A roadmap for evaluating moral competence in Large Language Models". *Nature*, vol. 650, 2026, pp. 565–573. https://doi.org/10.1038/s41586-025-10021-1

Haugeland, John. *Artificial intelligence: The Very Idea*. MIT Press, 1989.

Himmelreich, Johannes. "Never mind the trolley: The ethics of autonomous vehicles in mundane situations". *Ethical Theory and Moral Practice,* vol. 21, 2018, pp. 669–684. https://doi.org/10.1007/s10677-018-9896-4

Longino, Helen E. "Beyond 'bad science'". *Science, Technology, & Human Values*, vol. 8, núm. 1, 1983, pp. 7–17.

Matthews, Kyle y Marie Lamensch. *Wired for war: how authoritarian states are weaponizing AI against the west*. Konrad-Adenauer-Stiftung e.V., 2025.

McCarthy, John, et al. "A proposal for the Dartmouth summer research project on artificial intelligence". 1955 http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Miller, Seumas. "Cognitive warfare: An ethical analysis". *Ethics and Information Technology*, vol. 25, article number 46, 2023. https://doi.org/10.1007/s10676-023-09717-7

Moore, George Edward. *Principia ethica*. Cambridge University Press, 2002.

NATO STO. "Cognitive warfare: NATO Chief Scientist research report". NATO, 2025. https://www.nato.int/content/dam/nato/webready/documents/sto/chief-scientist-report-cognitive-warfare.pdf

NATO. "NATO's approach to counter information threats". NATO, 2025. https://www.nato.int/en/what-we-do/wider-activities/natos-approach-to-counter-information-threats

Newen, Albert, et al. *The Oxford Handbook of 4E Cognition*. Oxford University Press, 2018.

Oh, Soyoung y Vera Demberg. “Robustness of large language models in moral judgements”. *Royal Society Open Science,* vol. 12, núm. 4, 2025, pp. 241229. https://doi.org/10.1098/rsos.241229

Origgi, Gloria. Is trust an epistemological notion? *Episteme*, vol. 1, núm. 1, 2004, pp. 61–72. https://doi.org/10.3366/epi.2004.1.1.61

Parasuraman, Raja y Victor Riley. “Humans and automation: Use, misuse, disuse, abuse”. *Human Factors*, vol. 39, núm 2, 1997, pp. 230–253. doi.org/10.1518/001872097778543886

Pratzner, Philip R. “The current targeting process”. *Targeting: the challenges of modern warfare,* editado por P. A. L. Ducheine, M. N. Schmitt y F. P. B. Osinga. T.M.C. Asser Press, 2016, pp. 77-97. https://doi.org/10.1007/978-94-6265-072-5_4

Putnam, Hilary. *The collapse of the fact/value dichotomy and other essays*. Harvard University Press, 2002.

Rodilosso, Ermelinda. “Epistemic vs. moral: A differentiated approach to human-centered AI ethics”. *Topoi,* 2025. https://doi.org/10.1007/s11245-025-10324-y

Russell, Stuart y Peter Norvig. *Artificial intelligence: a modern approach* (4th ed.). Pearson, 2021.

Sajadieh, Sha, et al. *The AI index 2026 annual report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2026.

Sans Pinillos, Alger y David Casacuberta. “Remarks on the possibility of ethical reasoning in an artificial intelligence system by means of abductive models”. *Model-based reasoning in science and technology: MBR 2018*, editado por Á. Nepomuceno-Fernández, L. Magnani, F. Salguero-Lamillar, C. Barés-Gómez y M. Fontaine. Springer, 2019, pp. 318–333. https://doi.org/10.1007/978-3-030-32722-4_19

Sans Pinillos, Alger y Lorenzo Magnani. “How do we think about the unknown? The self-awareness of ignorance as a tool for managing the anguish of not knowing”. *Embodied, extended, ignorant minds*, editado por S. Arfini y L. Magnani. Springer, 2022. https://doi.org/10.1007/978-3-031-01922-7_9

Sans Pinillos, Alger y Vincent Costa. “Más allá de los datos: la transformación digital del museo tradicional”. *Daimon Revista Internacional De Filosofía*, vol. 90, 2023, pp. 81–94. https://doi.org/10.6018/daimon.563231

Sans Pinillos, Alger. “Apuntes sobre los aspectos de valor prescriptivo del razonamiento abductivo”. *El jardín de senderos que se bifurcan y confluyen: filosofía, lógica y matemáticas*, editado por D. P. Fernandes y R. Lopez-Orellana. Universidad de Valparaíso, 2020, pp. 143–155.

Sans Pinillos, Alger. “CELL (Contextual Ethics Level Layer) as an architecture for contextual ethical governance in dual-use defense technologies”. *Clicking the pause: the role of transatlantic cooperation in artificial intelligence supervision*, editado por F. Andrés Pérez y R. García Alonso. NATO Science for Peace and Security Series E: Human and Societal Dynamics, 2026, pp. 145-165. https://doi.org/10.3233/NHSDP260044

Sans Pinillos, Alger. “Horror vacui: Characterizing the experience of paradigm crisis through Magnani’s EC-model of abduction”. *Scientific cognition, semiotics, and computational agents: essays in honor of Lorenzo Magnani*. Vol. 2, editado por S. Arfini. Springer, 2025, pp. 171-192. https://doi.org/10.1007/978-3-031-96688-0_9

Sans Pinillos, Alger. “Neglected pragmatism: discussing abduction to dissolute classical dichotomies”. *Foundations of Science*, vol. 27, 2022, pp. 1107–1125. https://doi.org/10.1007/s10699-021-09817-x

Sans Pinillos, Alger. “Unpacking bad expectations: a framework for anticipation and social justice in the eco-cognitive paradigm”. *Model-Based reasoning, abductive cognition, creativity. MBR 2023*, editado por E. Ippoliti, L. Magnani y S. Arfini. Springer, 2024, pp. 254-270. https://doi.org/10.1007/978-3-031-69300-7_15

Schmitt, Michael N. y Eric Widmar. “The law of targeting”. *Targeting: the challenges of modern warfare*, editado por P. A. L. Ducheine, M. N. Schmitt y F. P. B. Osinga. T.M.C. Asser Press, 2016, pp. 121–145. https://doi.org/10.1007/978-94-6265-072-5_6

Sheridan, Thomas B. *Telerobotics, automation, and human supervisory control*. MIT Press, 1992.

Simon, Herbert A. *Models of man: social and rational*. Wiley, 1957.

Skitka, Linda, et al. “Does automation bias decision-making?” *International Journal of Human-Computer Studies*, vol. 51, núm 5, 1999, pp. 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Vallverdú, Jordi y Alger Sans Pinillos. “The foundations of creativity: Human inquiry explained through the neuro-multimodality of abduction”. *Handbook of abductive cognition*, editado por Lorenzo Magnani. Springer, 2022, pp. 1-27. https://doi.org/10.1007/978-3-030-68436-5_71-1

Varela, Francisco J., et al. *The embodied mind: cognitive science and human experience*. MIT Press, 2017.

Vaswani, Ashish, et al. “Attention is all you need”. *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://doi.org/10.48550/arXiv.1706.03762 Đ