Insuficiencia estructural del control humano significativo en armas autónomas: fundamentos para una reorientación normativa post-antropocéntrica

Structural Insufficiency of Meaningful Human Control in Autonomous Weapons: Foundations for a Post-Anthropocentric Normative Reorientation

Insuficiência estrutural do controle humano significativo em armas autônomas: fundamentos para uma reorientação normativa pós-antropocêntrica

Fernando A. Ramos-Zaga[1]
DOI: 10.5752/P.2317-773X.2025v13n2p149-165

Enviado em: 22 de setembro de 2025
Aceito em: 03 de março de 2026

RESUMEN

El avance acelerado de los Sistemas de Armas Letales Autónomas (LAWS) ha generado una transformación radical en las lógicas del conflicto armado, planteando desafíos ético-normativos que cuestionan la continuidad de la agencia humana en decisiones letales automatizadas. Frente a este escenario, el presente estudio examina críticamente la noción de Control Humano Significativo (MHC) como dispositivo normativo central en la regulación de LAWS, con el objetivo de construir un marco filosófico que fundamente éticamente su exigencia, explore su viabilidad en entornos militares avanzados y proponga una reconceptualización que articule agencia, dignidad y responsabilidad. Los hallazgos evidencian que el MHC, en su formulación actual, presenta limitaciones estructurales derivadas de la opacidad de los sistemas algorítmicos, la imprevisibilidad conductual y la dilución de la responsabilidad moral, lo cual imposibilita su función como garante ético en la guerra automatizada. En conclusión, se plantea un giro normativo que interpela la arquitectura jurídica y técnica existente, exigiendo la adopción de marcos regulatorios internacionales capaces de salvaguardar principios morales fundamentales en un contexto marcado por la creciente mediación algorítmica de la violencia.

Palabras clave: Sistemas de armas letales autónomas, control humano significativo, brecha de responsabilidad, guerra algorítmica, ética militar, inteligencia artificial.

1. Abogado. Docente investigador a tempo completo. Universidad Privada del Norte, Lima, Perú. https://orcid.org/0000-0001-6301-9460. fernandozaga@gmail.com

ABSTRACT

The rapid advancement of Lethal Autonomous Weapon Systems (LAWS) has fundamentally altered the dynamics of armed conflict, raising profound ethical and normative challenges that call into question the continuity of human agency in automated lethal decision-making. In response to this scenario, the present study critically examines the notion of Meaningful Human Control (MHC) as a central normative device in the regulation of LAWS. The objective is to construct a philosophical framework that ethically grounds the demand for MHC, assesses its operational viability in advanced military environments, and proposes a reconceptualization that integrates agency, dignity, and moral responsibility. The findings reveal that MHC, in its current formulation, suffers from structural limitations rooted in algorithmic opacity, behavioral unpredictability, and the fragmentation of moral responsibility, which collectively undermine its role as an ethical safeguard in autonomous warfare. The study concludes by proposing a normative shift that challenges existing legal and technical architectures, advocating for the establishment of binding international regulatory frameworks capable of upholding fundamental moral principles in a context increasingly shaped by the algorithmic mediation of violence.

Keywords: Lethal autonomous weapon systems, meaningful human control, responsibility gap, algorithmic warfare, military ethics, artificial intelligence.

RESUMO

O avanço acelerado dos Sistemas de Armas Letais Autônomas (LAWS) tem provocado uma transformação radical nas lógicas do conflito armado, suscitando desafios ético-normativos que questionam a continuidade da agência humana em decisões letais automatizadas. Diante desse cenário, o presente estudo examina criticamente a noção de Controle Humano Significativo (MHC) como dispositivo normativo central na regulação dos LAWS, com o objetivo de construir um marco filosófico que fundamente eticamente sua exigência, explore sua viabilidade em contextos militares avançados e proponha uma reconceitualização que articule agência, dignidade e responsabilidade. Os resultados evidenciam que o MHC, em sua formulação atual, apresenta limitações estruturais decorrentes da opacidade dos sistemas algorítmicos, da imprevisibilidade comportamental e da diluição da responsabilidade moral, o que inviabiliza sua função como garantidor ético na guerra automatizada. Conclui-se, portanto, que é necessário um giro normativo que interpele a arquitetura jurídica e técnica existente, exigindo a adoção de marcos regulatórios internacionais capazes de salvaguardar princípios morais fundamentais em um contexto marcado pela crescente mediação algorítmica da violência.

Palavras-chave: Sistemas de armas letais autônomas, controle humano significativo, lacuna de responsabilidade, guerra algorítmica, ética militar, inteligência artificial.

1 INTRODUCCIÓN

En los últimos años, el desarrollo exponencial de tecnologías aplicadas a la defensa ha reconfigurado los parámetros tradicionales de la guerra, inaugurando una nueva era marcada por la creciente autonomización de los sistemas armamentísticos. La reciente inversión de 600 millones de euros liderada por Daniel Ek, fundador de Spotify, en la empresa Helsing, especializada en sistemas de guerra basados en inteligencia artificial (Browne, 2025), ilustra de forma paradigmática la militarización

de la inteligencia artificial, impulsada tanto por actores estatales como por consorcios tecnológicos privados. El viraje hacia una guerra tecnológicamente mediada plantea preguntas urgentes sobre la legitimidad del uso autónomo de la fuerza y sobre los límites éticos de la delegación de decisiones letales a entidades no humanas, particularmente en un escenario donde los marcos jurídicos tradicionales y las estructuras morales heredadas resultan insuficientes para aprehender la complejidad del fenómeno emergente.

La aparición de los Sistemas de Armas Letales Autónomos (LAWS, por sus siglas en inglés) ha catalizado un debate cuyo eje central gira en torno a la noción de Control Humano Significativo (MHC), introducida en 2013 como respuesta normativa ante la opacidad decisional de los sistemas autónomos (Horowitz, 2016). No obstante, múltiples estudios han advertido sobre la ambigüedad conceptual del MHC y su limitada aplicabilidad en contextos operacionales reales, marcados por la alta velocidad, la complejidad ambiental y la impredecibilidad algorítmica (Santoni De Sio; Van Den Hoven, 2018). A ello se suma el desafío ontológico planteado por la autonomía algorítmica, cuya lógica causal-determinista entra en tensión con la estructura intencional y relacional que define la acción moral humana (Matthias, 2004).

La persistencia de una brecha epistemológica y normativa entre las capacidades técnicas de los LAWS y las exigencias morales del *ius in bello* pone de manifiesto una zona crítica de indeterminación conceptual. La actual arquitectura del MHC, centrada en fases operacionales específicas, desconoce el carácter distribuido y sistémico de la responsabilidad en entornos de guerra algorítmica (Taddeo; Floridi, 2018). En tal sentido, la idea de una "certificación humana significativa" ha sido propuesta como alternativa que podría preservar la autoría moral sin sacrificar la eficiencia técnica (Winfield; Jirotka, 2018). Sin embargo, dicha reconceptualización plantea interrogantes profundos acerca de la trazabilidad de la intención humana en redes decisionales algorítmicas evolutivas y distribuidas, lo cual exige una revisión crítica de las nociones tradicionales de control, causalidad y responsabilidad.

El presente análisis se justifica por la urgencia de construir un marco filosófico-normativo que permita evaluar, fundamentar y, eventualmente, redefinir el concepto de MHC en un contexto donde las decisiones sobre la vida y la muerte ya no dependen exclusivamente de agentes humanos identificables. Lejos de constituir una discusión académica aislada, esta problemática tiene implicaciones directas sobre la legitimidad del uso de la fuerza, la arquitectura normativa del derecho internacional y, en última instancia, sobre los fundamentos éticos que sostienen la convivencia global. En un momento histórico caracterizado por la aceleración tecnológica y la erosión de las formas tradicionales de soberanía, la ausencia de un paradigma coherente que preserve la agencia humana ante el avance de la autonomía algorítmica representa una amenaza a los principios fundacionales del orden jurídico-moral contemporáneo.

Las implicaciones prácticas de esta investigación son múltiples. En primer lugar, el desarrollo de un marco normativo robusto en torno al MHC puede ofrecer orientaciones concretas para legisladores, diseñadores

de sistemas de defensa y organismos internacionales encargados de la regulación de armamento. En segundo lugar, puede contribuir a delimitar los márgenes éticamente aceptables de la innovación tecnológica en contextos de seguridad, evitando que los imperativos de eficiencia desplacen completamente los criterios de legitimidad moral. Finalmente, una redefinición del MHC podría servir como fundamento para sistemas de certificación internacional que articulen principios éticos, criterios operacionales y requisitos legales en torno al uso de tecnologías autónomas en la guerra.

La relevancia de esta reflexión se acentúa aún más cuando se la vincula con los desafíos contemporáneos que atraviesan tanto el campo de la gobernanza tecnológica como el del orden internacional. El creciente protagonismo de empresas privadas en la producción de armamento inteligente, la presión geopolítica por alcanzar la supremacía algorítmica y la normalización social del uso de drones autónomos en conflictos armados configuran un panorama en el cual las decisiones normativas actuales tendrán efectos estructurales en las décadas venideras. En este sentido, la discusión sobre el MHC no puede desligarse de las tensiones más amplias que afectan a las democracias liberales, a la arquitectura institucional del derecho internacional y a las formas emergentes de subjetividad en la era de la inteligencia artificial.

En ese sentido, el objetivo del presente artículo es analizar críticamente la noción de Control Humano Significativo en sistemas armamentísticos autónomos, con el fin de desarrollar un marco normativo-filosófico que permita fundamentar éticamente su exigencia, evaluar su viabilidad operativa en contextos militares tecnológicamente avanzados y proponer una reconceptualización sistemática que articule agencia, dignidad humana y responsabilidad moral en el entorno de la guerra algorítmica. Se espera que los resultados de este trabajo contribuyan a esclarecer los fundamentos éticos que deben orientar el desarrollo y despliegue de los LAWS, ofreciendo herramientas conceptuales y normativas que permitan preservar la centralidad de la agencia moral humana en un mundo progresivamente automatizado.

2 GENEALOGÍA DE LA AUTOMATIZACIÓN DE LA VIOLENCIA

La trayectoria histórica y conceptual de las armas automatizadas permite apreciar cómo la tecnología ha modificado progresivamente la estructura de la guerra. Durante la Guerra Fría, con el desarrollo de misiles antibalísticos guiados por radar, se consolidó un interés por delegar funciones críticas en las máquinas, aunque sin desplazar la centralidad del juicio humano. En tal contexto, la lógica de disuasión nuclear descansaba en la capacidad decisional de líderes políticos y militares, quienes conservaban el control último sobre la acción letal (Freedman, 1989). Ello evidenciaba un perfeccionamiento del apoyo técnico orientado a la eficiencia, más que una auténtica autonomía decisional.

Con la irrupción de la ciberguerra en los años noventa se produjo un giro significativo, ya que el escenario de confrontación se trasladó al espacio informacional. En este ámbito, los ataques contra infraestructuras digitales funcionaban primordialmente como formas de sabotaje y

no como guerra en sentido estricto, debido a su limitada capacidad de destrucción masiva (Arquilla; Ronfeldt, 1993; Libicki, 1995; Rid, 2013). De ese modo, se reforzaba la premisa de que el factor humano seguía siendo insustituible, pues la decisión sobre el momento, el lugar y la intensidad de la intervención permanecía en manos de actores políticos y militares.

Posteriormente, la evolución tecnológica introdujo un cambio cualitativo de mayor envergadura. El tránsito desde sistemas de apoyo hacia sistemas autónomos modificó sustancialmente la concepción de agencia en el campo bélico. Mientras la automatización remitía a la ejecución mecánica de instrucciones, la autonomía abrió la posibilidad de que la máquina actuara y decidiera sin supervisión directa (Scharre, 2018; Singer, 2009). La incorporación de inteligencia artificial y sensores avanzados intensificó esa transformación, acelerando el ritmo del combate a velocidades que superaban las capacidades cognitivas humanas (M. L. Cummings, 2021). Como consecuencia, surgió la noción de irresponsabilidad estructural, ya que cuando las máquinas operan con mínima intervención humana, la atribución de responsabilidad se vuelve difusa o incluso impracticable (Sparrow, 2007).

En la actualidad, los conflictos armados han convertido el debate en experiencia tangible. La guerra en Ucrania se ha configurado como un laboratorio de experimentación con drones kamikaze y sistemas híbridos que combinan innovación comercial y despliegue militar, lo que refleja la convergencia entre ambas esferas (Kunertova, 2023). De manera similar, en Gaza, los sistemas de selección automatizada de objetivos como "Lavender" han generado cuestionamientos éticos y jurídicos, debido a la dificultad de distinguir de forma confiable entre combatientes y civiles, con posibles vulneraciones al Derecho Internacional Humanitario (Gusterson, 2024). A su vez, en Yemen, la continuidad de ataques con drones de bajo costo ha configurado un patrón de guerra dronizada, aunque con eficacia relativa que atenúa la idea de autonomía plena en el terreno (Boyle, 2013; Kallenborn; Bleek, 2018).

Desde un punto de vista analítico, resulta clave examinar los distintos grados de intervención humana. El modelo *human-in-the-loop* asegura que el operador conserve el control directo sobre cada decisión letal, con lo cual mantiene una trazabilidad ética básica (Horowitz, 2016). A su vez, el modelo *human-on-the-loop* permite la actuación automática bajo supervisión diferida, lo que abre un espacio controvertido en materia de responsabilidad (Scharre, 2018). En el extremo, el modelo *human-out-of-the-loop* delega en la máquina la totalidad del ciclo decisional y elimina cualquier posibilidad de control humano significativo (Asaro, 2012). De ese modo, se configura un continuo que va desde la supervisión plena hasta la autonomía absoluta, con consecuencias jurídicas y éticas de gran alcance.

El panorama descrito evidencia una mutación estructural de la guerra. La agencia bélica ya no se define exclusivamente en términos humanos, sino como producto de ensamblajes socio-técnicos que redistribuyen la capacidad de decidir y ejecutar acciones (Bode et al., 2024). Surge entonces un desfase entre la práctica militar y los marcos normativos, concebidos en un contexto donde el combatiente humano ocupaba un

lugar indiscutible. Tal desfase abre espacios de impunidad y erosiona la legitimidad del Derecho Internacional Humanitario (Winter, 2022).

Con todo, la discusión incorpora también una dimensión menos pesimista. Se ha planteado que la autonomía algorítmica podría, en condiciones específicas, contribuir a reducir violaciones al derecho de la guerra, dado que las máquinas carecen de emociones y sesgos que con frecuencia inducen a los combatientes humanos a excesos en el uso de la fuerza (Arkin, 2009). Se configura así una tensión entre quienes conciben la autonomía como un riesgo ontológico y normativo y quienes la valoran como una herramienta correctiva frente a la falibilidad humana. Reconocer tal tensión obliga a asumir la complejidad del fenómeno y a resistir la tentación de respuestas absolutas, puesto que el vínculo entre tecnología, ética y guerra exige un análisis abierto, crítico y matizado.

En ese marco, el concepto de "control humano significativo" (MHC, por sus siglas en inglés) se ha propuesto como una tentativa de reintroducir la agencia moral en contextos automatizados. La sección siguiente examinará su génesis conceptual, su potencial normativo y sus limitaciones estructurales, con el fin de valorar si puede consolidarse como principio operativo eficaz o si, por el contrario, se trata de una ficción regulativa que encubre la inercia de una automatización ya desbordada. Lo que está en juego no se limita a la regulación de una nueva tecnología, sino a la posibilidad de preservar una ética de la guerra en un mundo donde la decisión de matar podría ejecutarse sin que nadie, realmente, decida.

3 EL CONTROL HUMANO SIGNIFICATIVO COMO RESPUESTA NORMATIVA INSUFICIENTE

Desde su irrupción en el debate internacional sobre los sistemas de armas autónomos letales, el concepto de *control humano significativo* (MHC, por sus siglas en inglés) se ha consolidado como un artefacto normativo de notable plasticidad discursiva. Su emergencia puede interpretarse como una tentativa colectiva de reinstalar la figura humana en un escenario tecnológico que, de manera progresiva, desplaza su agencia decisional y su capacidad deliberativa. A lo largo de su desarrollo conceptual, se advierte una tensión persistente entre la voluntad de mantener el juicio humano como eje ético de la acción bélica y las limitaciones estructurales, epistémicas y morales que impone la creciente sofisticación de los sistemas algorítmicos.

Es necesario señalar que la primera formulación sustantiva del MHC se articuló en el seno de la organización Article 36 durante las discusiones de la Convención sobre Ciertas Armas Convencionales en 2013. En aquel contexto, se trataba de una propuesta concebida no solo como mecanismo regulador, sino también como una respuesta ética frente a la amenaza de deshumanización del uso de la fuerza (Klonowska, 2022). Desde esa perspectiva inicial, el MHC buscaba establecer un umbral mínimo de intervención humana que preservara la legitimidad moral y jurídica de la decisión letal. En consecuencia, más que un cuerpo doctrinal coherente, se ofrecía un principio guía suficientemente abierto para propiciar un consenso básico entre actores con posiciones divergentes.

Posteriormente, la consolidación del MHC en la agenda internacional respondió a la acción concertada de organizaciones de la sociedad civil y coaliciones transnacionales. De hecho, iniciativas como la *Campaign to Stop Killer Robots* y *Human Rights Watch* ejercieron presión en foros multilaterales con el objetivo de impulsar la adopción de un estándar normativo vinculante que garantizara un umbral mínimo de intervención humana en el uso de la fuerza (Roff, 2014; Solovyeva; Hynek, 2023). Sin embargo, la postura de potencias con alta inversión en sistemas autónomos, como Estados Unidos, Rusia e Israel, apuntó hacia aproximaciones más flexibles, alineadas con la noción de " juicio humano apropiado" y con énfasis en la interoperabilidad técnica y la eficacia táctica (Boulanin; Verbruggen, 2017; Department of Defense, 2012). Tal divergencia cristalizó en la Convención sobre Ciertas Armas Convencionales, donde los intentos de avanzar hacia un tratado vinculante enfrentaron la resistencia de Estados reticentes a limitar su superioridad estratégica (Payne, 2021).

En ese orden de hechos, la emergencia de posturas intermedias complejizó aún más el panorama. La Unión Europea y algunos de sus Estados miembros, entre ellos Alemania y Países Bajos, reconocieron la necesidad de preservar la agencia humana, pero prefirieron impulsar mecanismos de carácter político antes que jurídicamente obligatorios (Ekelhof, 2019). Como resultado, los puntos de fricción giraron en torno a la naturaleza vinculante del MHC, a la definición operativa del umbral de intervención humana y a la tensión entre criterios ético-jurídicos y exigencias de viabilidad militar. El MHC se configuró así como un campo de competencia discursiva en el que confluyen racionalidades heterogéneas, tanto éticas como jurídicas y geopolíticas (Santoni De Sio; Van Den Hoven, 2018).

Ahora bien, la vaguedad inicial que facilitó la aceptación política del MHC se ha transformado con el tiempo en un obstáculo epistemológico y normativo. La ambigüedad de su formulación permitió su captura por marcos estratégicos disímiles, diluyendo así su potencial regulador (Jani□atová; Mlejnková, 2021). El contraste con el enfoque estadounidense, centrado en la noción de "niveles apropiados de juicio humano", resulta ilustrativo. Mientras que el MHC aspira a preservar la agencia ética y la responsabilidad humana, la postura del Departamento de Defensa de Estados Unidos privilegia la interoperabilidad técnica y la eficacia táctica, relegando la cuestión moral a un plano secundario (Department of Defense, 2012). De ello se desprende una fractura profunda entre los imperativos éticos de control y las exigencias operativas de eficiencia militar.

En paralelo, a medida que el MHC adquirió mayor visibilidad en foros internacionales, se le atribuyeron funciones progresivamente más ambiciosas. Así, de mero instrumento de supervisión pasó a convertirse en símbolo de resistencia ética frente a la automatización del juicio, llegando incluso a proyectársele la capacidad de garantizar el cumplimiento del derecho internacional humanitario y de prevenir la fragmentación de la responsabilidad en contextos tecnológicamente mediados (Santoni De Sio; Van Den Hoven, 2018). Sin embargo, esta sobrecarga funcional condujo a un fenómeno de inflación normativa, donde las expectativas asignadas superaron con creces su factibilidad operativa, generando una disonancia entre aspiraciones éticas y limitaciones técnicas.

En el plano estructural, el MHC se sostiene en tres pilares fundamentales: información, acción y responsabilidad. No obstante, un análisis riguroso de cada dimensión revela inconsistencias significativas. En lo que respecta al principio de información, la hipótesis de que el agente humano puede procesar datos relevantes en tiempo real resulta insostenible. Se ha demostrado que la capacidad cognitiva humana se ve rápidamente desbordada en entornos de elevada complejidad, lo que reduce al operador a un rol simbólico (M. M. Cummings, 2014). Además, disponer de información no equivale a comprenderla en términos semánticos, pues solo una inteligibilidad contextualizada permite fundamentar decisiones normativas (Floridi, 2011).

En cuanto al principio de acción, los marcos tradicionales han tendido a reducir la agencia humana a una intervención puntual sobre el sistema. Sin embargo, tal visión ignora las condiciones estructurales que determinan la posibilidad misma de actuar. En esa dirección, se ha planteado que la agencia incluye no solo la decisión inmediata, sino también los procesos de diseño anticipado y las restricciones ex ante que configuran el marco de acción (Taddeo; Floridi, 2018).

El principio de responsabilidad constituye, con todo, el pilar más problemático. La dispersión de atribuciones a lo largo del ciclo de vida del sistema genera lo que se ha denominado el problema de las "muchas manos" (Matthias, 2004). La dificultad de imputar responsabilidad compromete tanto la justicia retributiva como la confianza en los marcos éticos y legales vigentes. Surge así la pregunta sobre quién debe responder: el diseñador del algoritmo, el aprobador político o el operador en campo.

En consecuencia, los modelos clásicos de interacción hombre-máquina muestran limitaciones análogas. El modelo *human-in-the-loop*, que exige la autorización humana antes de ejecutar una acción, se ha revelado ineficaz, ya que la presión situacional y la familiaridad con el sistema conducen a un automatismo de la confianza (Skitka et al., 1999). De manera similar, el modelo *human-on-the-loop* otorga al operador un papel de supervisión diferida, pero la vigilancia sin capacidad efectiva de intervención equivale a una responsabilidad ilusoria (Denning, 2011). En el extremo, el modelo *human-out-of-the-loop* representa la consagración de la alienación moral, al romper la cadena que conecta juicio humano y acción operativa (Sparrow, 2007). Incluso se ha sostenido que la autonomía técnica no constituye una opción ideológica, sino una imposición funcional derivada de la velocidad de los sistemas (Scharre, 2018).

Todo ello permite constatar la paradoja estructural que atraviesa al MHC: a mayor autonomía técnica, menor margen para la responsabilidad moral. En tal sentido, se ha planteado que el MHC funciona menos como una solución normativa que como un dispositivo discursivo destinado a reconciliar lo irreconciliable, es decir, la eficiencia algorítmica con la deliberación ética (Danaher, 2016; Matthias, 2004).

En definitiva, el análisis del MHC no puede reducirse a cuestiones técnicas de implementación, sino que exige una reflexión crítica sobre las mutaciones normativas que introduce la automatización en el ámbito de la agencia moral. A medida que los sistemas autónomos reducen la posibilidad de intervención significativa, también se erosiona la relación

normativa entre intención, acción y responsabilidad. La discusión se orienta, por tanto, hacia la exploración de la brecha de responsabilidad, en la cual la opacidad algorítmica, la pérdida de atribución causal y la disolución del vínculo ético con el Otro configuran un impasse normativo que interpela los fundamentos antropocéntricos del derecho y de la ética contemporánea.

4 LA BRECHA DE RESPONSABILIDAD COMO IMPASSE ONTOLÓGICO

La creciente incorporación de sistemas autónomos en contextos bélicos plantea interrogantes éticos y ontológicos que desbordan las estructuras normativas existentes. En particular, la opacidad y el carácter emergente de los comportamientos algorítmicos marcan un punto de inflexión en la manera de entender la acción, la responsabilidad y el vínculo moral con el otro. La opacidad algorítmica no remite únicamente a una cuestión técnica, sino que conlleva implicaciones filosóficas profundas relacionadas con la posibilidad misma de comprender y atribuir agencia a sistemas no humanos.

Al respecto, cabe señalar que la arquitectura de las redes neuronales profundas, en tanto estructuras no lineales y no simbólicas, impide rastrear causalmente los procesos internos que conducen a una decisión específica. En esa dirección, se ha destacado que la opacidad epistémica[2] excede la comprensión humana incluso de los propios diseñadores, lo cual supone una ruptura con los supuestos clásicos de transparencia operativa sobre los que se apoyan la responsabilidad jurídica y la atribución ética (Greif, 2022).

Asimismo, la imposibilidad de establecer una secuencia causal legible entre insumos y resultados constituye un obstáculo insalvable para el tipo de racionalidad que históricamente ha orientado la atribución moral. De acuerdo con la argumentación contemporánea, cuando se pierde la capacidad de reconstruir nexos causales no se enfrenta únicamente una limitación de la interfaz humano-máquina, sino una crisis del modelo mismo de acción deliberativa (Shea, 2023). En el caso de los sistemas de armas letales autónomas, esta falta de trazabilidad implica la desaparición de un agente en sentido fuerte, es decir, de un sujeto que decida en conocimiento de causa. La consecuencia directa es la erosión del fundamento normativo de la rendición de cuentas.

Ahora bien, el problema no se reduce a la opacidad epistémica. Existe también un nivel de ininteligibilidad estructural que agrava el desfase entre sistemas técnicos y comprensión humana. Se ha sostenido que este fenómeno no alude únicamente a la falta de acceso técnico, sino a la ausencia de coherencia semántica que imposibilita al operador humano atribuir sentido a las decisiones algorítmicas (Zednik, 2021). En otros términos, el obstáculo no radica solo en no conocer cómo funciona el sistema, sino en la incapacidad de interpretar sus acciones en términos significativos para

2. El término hace referencia a la imposibilidad de comprender de manera exhaustiva los procesos mediante los cuales un sistema algorítmico complejo genera resultados decisionales, lo que compromete la trazabilidad requerida para la atribución de responsabilidad ética y jurídica.

la experiencia humana. Esa brecha semántica afecta directamente la capacidad de anticipación y de justificación de los operadores.

El desfase se acentúa aún más en presencia de comportamientos emergentes. Los sistemas adaptativos, al integrar retroalimentación continua y aprendizaje autónomo, generan respuestas inéditas que no remiten a ninguna intención humana explícita. En tal dinámica, el origen de la acción se traslada desde el sujeto humano hacia la lógica interna del sistema técnico. En consecuencia, se diluye el nexo entre decisión, intención y sujeto, es decir, se pierde el punto de anclaje que sostiene los marcos normativos basados en la imputabilidad individual. Lo que emerge es una agencia distribuida, no reducible a una entidad unitaria, lo que plantea desafíos conceptuales para una ética aún anclada en el sujeto racional.

En ese escenario, adquiere plena relevancia la noción de "brecha de responsabilidad", formulada como una desconexión estructural entre el agente y la acción. La imposibilidad de imputar responsabilidad moral no se explica por una falla accidental, sino por la delegación de decisiones con implicancias letales a sistemas autónomos que operan sin supervisión continua ni previsibilidad suficiente (Matthias, 2004). Frente a tal panorama, resulta inviable pensar en soluciones jurídicas o administrativas que simplemente redistribuyan la culpa. Se ha argumentado que la fractura entre delegación técnica y agencia moral constituye una aporía insoluble, en la que toda tentativa de atribuir responsabilidad queda frustrada por la ausencia de un agente deliberativo identificable (Danaher, 2016).

A partir de lo anterior, se hace evidente que el paradigma antropocéntrico de imputabilidad entra en crisis. El modelo clásico de responsabilidad, basado en intención individual, previsibilidad y control, resulta incompatible con sistemas algorítmicos cuyas decisiones no emergen de una subjetividad consciente ni de una secuencia causal gobernable. De hecho, se ha mostrado que dichos sistemas eluden los requisitos mínimos para la aplicación del modelo kantiano de sujeto moral (List, 2021). De ello se desprende la urgencia de un giro conceptual. Así, se ha sostenido la necesidad de abandonar una ética centrada en el individuo racional y adoptar un enfoque postantropocéntrico, capaz de concebir la agencia como cualidad emergente de estructuras sociotécnicas complejas (Coeckelbergh, 2020). En un marco semejante, la responsabilidad no puede atribuirse a un agente singular, sino que debe entenderse como una propiedad relacional distribuida entre distintos niveles de intervención humana y técnica.

La crisis de imputabilidad se hace especialmente visible en el plano del encuentro ético con el otro. Desde la filosofía de Lévinas, se ha insistido en que el rostro del Otro constituye la interpelación ética por excelencia, una presencia que exige respuesta moral (Lévinas, 1969). Sin embargo, la mediación algorítmica elimina ese encuentro: la víctima de una decisión automatizada ya no aparece como un rostro humano, sino como un objeto codificado, una probabilidad estadística o una firma térmica. El resultado es un proceso de objetivación que reduce la alteridad a una variable operativa y anula la posibilidad de respuesta ética.

Del mismo modo, la reducción de la alteridad afecta también al agente. Tal como se ha advertido, la acción política se constituye en un

espacio compartido de pluralidad, donde los seres humanos interactúan como agentes libres (Arendt, 2018). La guerra algorítmica destruye ese espacio, sustituyendo la acción deliberativa por cálculo instrumental. El operador ya no actúa, sino que ejecuta protocolos; ya no delibera, sino que calibra parámetros. En ese tránsito, el sujeto desaparece como agente ético, subsumido en una lógica que privilegia la eficiencia por sobre la deliberación.

Este fenómeno ha sido descrito como el triunfo de la racionalidad instrumental, es decir, una forma de acción guiada exclusivamente por la optimización de medios (Habermas, 1984). En la distinción entre "sistema" y "mundo de la vida", el primero coloniza al segundo e impone su lógica sobre ámbitos previamente sustentados en la generación de sentido intersubjetivo. Las armas autónomas operan dentro de ese dominio sistémico, colonizando el espacio militar y configurando un paradigma donde la compasión, el juicio y la responsabilidad quedan subordinados a la eficiencia algorítmica.

La consecuencia es una forma radical de alienación. Los agentes humanos se ven desplazados por procesos técnicos que no comprenden ni controlan, de modo que en lugar de ejercer juicio moral simplemente ajustan parámetros y cumplen funciones. Esa dinámica refleja la consumación de una tendencia más amplia: la instrumentalización de lo humano por la técnica. En tal contexto, la ética tradicional no solo resulta insuficiente, sino conceptualmente desfasada.

De lo anterior se desprende que la brecha de responsabilidad no puede cerrarse mediante reformas jurídicas ni con mejoras en la supervisión técnica. Se trata, más bien, de una aporía estructural que exige una revisión profunda de los marcos normativos y ontológicos con los que se ha concebido la responsabilidad. En este horizonte, el siguiente paso requiere una reconceptualización filosófica del control, orientada hacia formas anticipatorias de certificación, rediseños éticos de los sistemas y esquemas robustos de atribución distribuida. Solo una transformación integral permitirá recuperar un horizonte de legitimidad moral en escenarios donde la decisión humana ha sido desplazada por la automatización técnica.

5 RECONCEPTUALIZACIÓN NORMATIVO-FILOSÓFICA DEL CONTROL

La consolidación de una ética de la certificación significativa en el contexto de los sistemas autónomos letales constituye uno de los desafíos más urgentes y complejos para la filosofía moral contemporánea, en particular para la bioética en su vertiente tecnopolítica. En lugar de limitarse a marcos normativos centrados en la intervención reactiva o situacional, resulta imprescindible avanzar hacia una transformación estructural que reformule la supervisión ética de los sistemas automatizados, desplazando el eje desde el control humano inmediato hacia una responsabilidad anticipada, distribuida e integrada de manera sistémica. Desde una perspectiva crítica, la reevaluación de nociones como agencia, dignidad y responsabilidad no puede restringirse al plano conceptual, sino que debe proyectarse en su aplicabilidad institucional y tecnológica.

El paradigma dominante del control humano significativo, sustentado en fórmulas como *human-in-the-loop* o *human-on-the-loop*, ha presumido durante años que la proximidad temporal entre el ser humano y la acción letal garantiza la supervisión moral. Sin embargo, dicha presunción se demuestra insuficiente cuando se somete a un análisis técnico y normativo riguroso. La aceleración propia de los entornos algorítmicos reduce drásticamente los márgenes de intervención, imposibilitando la toma de decisiones reflexiva en tiempo real. De acuerdo con la evidencia empírica, limitaciones como la latencia perceptual, la sobrecarga cognitiva y la no linealidad del comportamiento sistémico restringen la efectividad de la supervisión (M. M. Cummings, 2014). A ello se añade el sesgo de automatización, según el cual los operadores tienden a aceptar de manera acrítica las recomendaciones de los sistemas, incluso cuando contradicen la información disponible (Skitka et al., 1999). Por tanto, vincular la agencia moral únicamente con la inmediatez temporal no solo resulta falaz desde un punto de vista filosófico, sino también impracticable en términos operativos.

Frente a tales limitaciones, la certificación significativa se configura como una alternativa epistemológica y normativamente más sólida. El objetivo no consiste en eliminar la supervisión humana, sino en redistribuirla de manera anticipatoria a lo largo del ciclo de vida del sistema, incorporando evaluaciones éticas *ex ante* que establezcan umbrales de aceptabilidad moral antes del despliegue operativo. En esta línea, la certificación actúa como un mecanismo de orquestación moral en el que convergen la ingeniería ética, la evaluación legal y la deliberación política (Winfield; Jirotka, 2018). De esta manera, la anticipación adquiere un carácter no solo técnico, sino también axiológico, puesto que implica definir desde el diseño cuáles conductas resultan moralmente tolerables y cuáles deben quedar descartadas.

La certificación supera, por tanto, la mera conformidad regulatoria. Mientras el cumplimiento legal suele adoptar un carácter formal, reactivo e incluso insuficiente, la certificación supone la configuración activa de valores en el sistema, mediante simulaciones éticas, algoritmos limitadores y validaciones orientadas por principios normativos. De este modo, la autoría moral no se diluye, sino que se reformula como una práctica distribuida, en la que cada etapa del ciclo, tales como diseño, implementación y evaluación, comporta responsabilidades diferenciadas, aunque interdependientes.

Sobre esta base, resulta fundamental identificar tres imperativos éticos que estructuren cualquier desarrollo de sistemas autónomos letales. En primer lugar, la preservación de la agencia debe asumirse como un nexo intencional irreductible. La acción moral requiere la existencia de una intención atribuible a un sujeto que pueda comprender, deliberar y responsabilizarse de sus actos. Cuando la automatización diluye esta intencionalidad mediante abstracción procedimental o delegación algorítmica, se erosiona la base misma de la responsabilidad moral (Anscombe, 1957). En términos más contemporáneos, la agencia puede entenderse como una capacidad situada de intervención con sentido (Bennett, 2005), de modo que todo diseño que impida rastrear la autoría humana de las decisiones incurre en una irresponsabilidad estructural éticamente inaceptable.

En segundo lugar, el respeto a la dignidad humana debe asumirse como un límite ontológico ineludible. No se trata de un mero atributo legal, sino de una condición constitutiva del ser moral. Desde una perspectiva kantiana, la dignidad radica en la capacidad de autolegislación racional, lo que prohíbe tratar a los individuos como medios para fines externos (Kant, 1996). Así, cualquier sistema que habilite la supresión de vidas humanas sin mediación moral transgrede este principio de manera radical. En esta dirección, se ha defendido la necesidad de un "diseño con dignidad" que integre restricciones morales no negociables en la arquitectura misma de los sistemas (Sparrow, 2016). El objetivo, por tanto, no consiste únicamente en prevenir abusos, sino en garantizar que la dignidad permanezca inviolable incluso en escenarios de automatización extrema.

Así, la atribución clara y distribuida de la responsabilidad constituye un imperativo ético decisivo. La complejidad de los sistemas tiende a diluir la culpa y a dispersar la rendición de cuentas, lo que exige instaurar un modelo de responsabilidad estructurada que defina obligaciones específicas en cada etapa del ciclo de vida del sistema (Floridi; Cowls, 2019), el cual debe institucionalizarse mediante mecanismos verificables de transparencia, auditabilidad y exigibilidad jurídica, de manera que la responsabilidad no quede reducida a un plano retórico, sino que se traduzca en prácticas concretas capaces de responder por los efectos generados.

Con base en lo anterior, resulta pertinente articular un marco post-MHC compuesto por tres dimensiones temporales: la anticipación ética en el diseño, los límites operativos en la implementación y la trazabilidad moral en la fase posterior al despliegue. Dichas dimensiones, aunque diferenciadas, convergen en la aspiración de reconstruir la agencia humana en contextos crecientemente automatizados. La dimensión *ex ante* implica incorporar, desde la génesis del sistema, criterios normativos que orienten su comportamiento futuro, lo que requiere la cooperación interdisciplinaria capaz de traducir principios éticos en restricciones funcionales y líneas de código (Taddeo; Blanchard, 2022).

Asimismo, durante la implementación se deben establecer límites operativos que prevengan desvíos moralmente problemáticos, ya sea mediante zonas de exclusión, umbrales de validación o mecanismos internos de autorregulación, como los denominados "gobernadores éticos" (Arkin, 2009). No obstante, dichos mecanismos solo resultan legítimos si los operadores humanos conservan siempre una capacidad real de anulación. Por otro lado, la dimensión *ex post* requiere sistemas de trazabilidad que permitan reconstruir las cadenas de decisión, identificar responsables y generar aprendizajes para nuevas iteraciones. Dicho propósito demanda no solo registros técnicos, sino también herramientas de inteligencia artificial explicable, auditorías independientes y estructuras institucionales adaptativas (Wood, 2024).

En definitiva, el modelo post-MHC no pretende erradicar la agencia humana, sino reconstruirla en condiciones de creciente complejidad técnica. Dicha reconstrucción debe ser ética, robusta y distribuida, de manera que preserve la integridad moral de la acción colectiva incluso cuando se medie por tecnologías altamente autónomas. No se trata de

sostener una ilusión de control, sino de reconfigurar las condiciones de posibilidad para una agencia deliberativa, responsable y, en última instancia, genuinamente humana.

6 CONCLUSIONES

El análisis realizado confirma que el marco antropocéntrico clásico ha perdido eficacia como fundamento normativo ante la creciente autonomía de los sistemas de armas letales. El desplazamiento ontológico producido por la automatización del conflicto no puede reducirse a una mera sofisticación técnica, sino que inaugura nuevas formas de agencia cuya opacidad desafía los esquemas normativos convencionales. El concepto de control humano significativo, aunque reiterado en diversos foros regulatorios, parece responder más a una necesidad simbólica que a una viabilidad práctica. Desde esta perspectiva, la legitimidad ética de los entornos bélicos automatizados no puede mantenerse sin una revisión estructural de los marcos valorativos que justifican el uso de la fuerza.

El andamiaje teórico aquí propuesto introduce tres aportaciones decisivas para la ética de la tecnología. En primer lugar, la noción de brecha de responsabilidad permite problematizar la desconexión entre agencia técnica y responsabilidad moral como un problema estructural, no meramente resoluble mediante ajustes jurídicos. En segundo lugar, se propone el modelo de certificación significativa, que desplaza el foco desde la supervisión puntual hacia formas de responsabilidad anticipada e inscrita desde el diseño. Finalmente, se formulan tres principios éticos ineludibles, cuya función no es meramente declarativa sino operativa: la preservación de la agencia intencional, la inviolabilidad de la dignidad humana como límite ontológico y la asignación distribuida de responsabilidad. Estas premisas configuran un marco normativo integral frente a los desafíos éticos de la automatización letal.

La dimensión aplicada del marco planteado requiere reestructuraciones institucionales y metodológicas sustanciales. La incorporación de equipos transdisciplinarios desde el diseño inicial de los sistemas permitiría que las restricciones normativas se traduzcan en arquitecturas funcionales concretas. Paralelamente, se demanda la adopción de protocolos de evaluación ética ex ante, con énfasis en simulaciones moralmente sensibles y trazabilidad de decisiones técnicas. Estas transformaciones no pueden depender de iniciativas fragmentarias. Resulta imperativo avanzar hacia marcos regulatorios internacionales con fuerza vinculante, capaces de unificar criterios de certificación ética y evitar disonancias normativas que pongan en riesgo la gobernanza efectiva de estas tecnologías.

Los hallazgos invitan a una agenda investigativa abierta y plural. En el plano teórico, se requiere profundizar en la articulación entre agencia distribuida y responsabilidad colectiva, sin incurrir en esquemas que diluyan la rendición de cuentas individual. La filosofía moral enfrenta el desafío de construir una ética post-antropocéntrica con capacidad prescriptiva, sin recurrir a formas de relativismo que neutralicen su eficacia normativa. Desde un enfoque empírico, urge el desarrollo de metodologías que operacionalicen la certificación significativa en escenarios

tecnológicos concretos. Finalmente, se impone una reflexión política de corte comparativo que identifique modelos exitosos de gobernanza algorítmica, con el fin de extrapolar aprendizajes al dominio de la automatización militar.

El debate sobre sistemas autónomos letales desborda su especificidad técnica para convertirse en un punto de inflexión filosófico. La automatización de la violencia obliga a reconsiderar las condiciones bajo las cuales puede sostenerse una ética de la responsabilidad en contextos de mediación tecnológica extrema. La guerra algorítmica no solo plantea riesgos operativos, sino que expone una tensión más profunda entre racionalidad instrumental y juicio moral, que recorre otras esferas del presente. En ese marco, los conceptos elaborados para abordar la autonomía bélica ofrecen un potencial heurístico relevante para pensar la automatización en educación, salud, justicia o gestión ambiental. La defensa de la agencia humana, por tanto, no implica una negación nostálgica del progreso técnico, sino una afirmación crítica de su orientación ética.

Declaración sobre el uso de IA generativa

En la elaboración de este manuscrito es usó GPT en su versión 4o para poder corregir errores de redacción, tipográficos. Se aseguró que los resultados mantengan el tono e intención del autor.

REFERENCIAS

ANSCOMBE, G. E. M. Intention. *Proceedings of the Aristotelian Society*, v. 57, n. 1, p. 321–332, 1957. DOI: https://doi.org/10.1093/aristotelian/57.1.321.

ARENDT, H. *The Human Condition*. 2. ed. Edited by M. Canovan and D. Allen. Chicago: University of Chicago Press, 2018.

ARKIN, R. *Governing Lethal Behavior in Autonomous Robots*. 0 ed. [S. l.]: Chapman and Hall/CRC, 2009. DOI: https://doi.org/10.1201/9781420085952.

ARQUILLA, J.; RONFELDT, D. Cyberwar is coming! *Comparative Strategy*, v. 12, n. 2, p. 141–165, 1993. DOI: https://doi.org/10.1080/01495939308402915.

ASARO, P. On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, v. 94, n. 886, p. 687–709, 2012. DOI: https://doi.org/10.1017/S1816383112000768.

BODE, I.; HUELSS, H.; NADIBAIDZE, A.; QIAO-FRANCO, G.; WATTS, T. F. A. Algorithmic Warfare: Taking Stock of a Research Programme. *Global Society*, v. 38, n. 1, p. 1–23, 2024. DOI: https://doi.org/10.1080/13600826.2023.2263473.

BOULANIN, V.; VERBRUGGEN, M. *Mapping the development of autonomy in weapon systems*. [S. l.]: SIPRI, 2017. DOI: https://doi.org/10.13140/RG.2.2.22719.41127.

BOYLE, M. J. The costs and consequences of drone warfare. *International Affairs*, v. 89, n. 1, p. 1–29, 2013. DOI: https://doi.org/10.1111/1468-2346.12002.

BROWNE, R. Spotify's Daniel Ek leads investment in Defense startup Helsing [Noticias]. *CNBC*, 2025. Disponível em: https://www.cnbc.com/2025/06/17/spotifys-daniel-ek-leads-investment-in-defense-startup-helsing.html.

COECKELBERGH, M. *AI Ethics*. Cambridge: The MIT Press, 2020.

CUMMINGS, M. L. Rethinking the Maturity of Artificial Intelligence in Safety‐Critical Settings. *AI Magazine*, v. 42, n. 1, p. 6–15, 2021. DOI: https://doi.org/10.1002/j.2371-9621.2021.tb00005.x.

CUMMINGS, M. M. Man versus Machine or Man + Machine? *IEEE Intelligent Systems*, v. 29, n. 5, p. 62–69, 2014. DOI: https://doi.org/10.1109/MIS.2014.87.

DANAHER, J. The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, v. 29, n. 3, p. 245–268, 2016. DOI: https://doi.org/10.1007/s13347-015-0211-1.

DENNING, D. Cyber Conflict as an Emergent Social Phenomenon. In: HOLT, T. J.; SCHELL, B. H. (eds.). *Corporate Hacking and Technology-Driven Crime: Social Dynamics and Implications*. Hershey: IGI Global, 2011. DOI: https://doi.org/10.4018/978-1-61692-805-6.

DEPARTMENT OF DEFENSE. *Directive No. 3000.09 Autonomy in weapon systems*. 2012. Disponível em: https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.

EKELHOF, M. Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation. *Global Policy*, v. 10, n. 3, p. 343–348, 2019. DOI: https://doi.org/10.1111/1758-5899.12665.

FLORIDI, L. *The Philosophy of Information*. Oxford: Oxford University Press, 2011. DOI: https://doi.org/10.1093/acprof:oso/9780199232383.001.0001.

FLORIDI, L.; COWLS, J. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 2019. DOI: https://doi.org/10.1162/99608f92.8cd550d1.

FREEDMAN, L. *The Evolution of Nuclear Strategy*. 2. ed. London: Macmillan for the International Institute for Strategic Studies, 1989.

GREIF, H. Models, Algorithms, and the Subjects of Transparency. In: MÜLLER, V. C. (ed.). *Philosophy and Theory of Artificial Intelligence 2021*. Cham: Springer International Publishing, 2022. v. 63, p. 27–37. DOI: https://doi.org/10.1007/978-3-031-09153-7_3.

GUSTERSON, H. It's all Lavender in Gaza. *Anthropology Today*, v. 40, n. 6, p. 1–2, 2024. DOI: https://doi.org/10.1111/1467-8322.12923.

HABERMAS, J. *The Theory of Communicative Action: Volume 1: Reason and the Rationalization of Society*. Tradução de T. McCarthy. Boston: Beacon Press, 1984.

HOROWITZ, M. C. The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons. *Daedalus*, v. 145, n. 4, p. 25–36, 2016. DOI: https://doi.org/10.1162/DAED_a_00409.

JANI□ATOVÁ, S.; MLEJNKOVÁ, P. The ambiguity of hybrid warfare: A qualitative content analysis of the United Kingdom's political–military discourse on Russia's hostile activities. *Contemporary Security Policy*, v. 42, n. 3, p. 312–344, 2021. DOI: https://doi.org/10.1080/13523260.2021.1885921.

KALLENBORN, Z.; BLEEK, P. C. Swarming destruction: Drone swarms and chemical, biological, radiological, and nuclear weapons. *The Nonproliferation Review*, v. 25, n. 5–6, p. 523–543, 2018. DOI: https://doi.org/10.1080/10736700.2018.1546902.

KANT, I. *The Metaphysics of Morals*. Edited by M. J. Gregor. Cambridge: Cambridge University Press, 1996.

KLONOWSKA, K. Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare. In: GILL, T. D.; GEIß, R.; KRIEGER, H.; MIGNOT-MAHDAVI, R. (eds.). *Yearbook of International Humanitarian Law, Volume 23 (2020)*. The Hague: T.M.C. Asser Press, 2022. v. 23, p. 123–153. DOI: https://doi.org/10.1007/978-94-6265-491-4_6.

KUNERTOVA, D. Drones have boots: Learning from Russia's war in Ukraine. *Contemporary Security Policy*, v. 44, n. 4, p. 576–591, 2023. DOI: https://doi.org/10.1080/13523260.2023.2262792.

LÉVINAS, E. *Totality and Infinity: An Essay on Exteriority*. Tradução de A. Lingis. 22nd reprint. Pittsburgh: Duquesne University Press, 1969.

LIBICKI, M. C. *What Is Information Warfare?* Strategic Forum, n. 28, p. 4. Washington, D.C.: Center for Advanced Concepts and Technology, Institute for National Strategic Studies, National Defense University, 1995.

LIST, C. Group Agency and Artificial Intelligence. *Philosophy & Technology*, v. 34, n. 4, p. 1213–1242, 2021. DOI: https://doi.org/10.1007/s13347-021-00454-7.

MATTHIAS, A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, v. 6, n. 3, p. 175–183, 2004. DOI: https://doi.org/10.1007/s10676-004-3422-1.

PAYNE, K. *I, Warbot: The Dawn of Artificially Intelligent Conflict*. London: Hurst, 2021.

RID, T. *Cyber War Will Not Take Place*. Oxford: Oxford University Press, 2013.

ROFF, H. M. The Strategic Robot Problem: Lethal Autonomous Weapons in War. *Journal of Military Ethics*, v. 13, n. 3, p. 211–227, 2014. DOI: https://doi.org/10.1080/15027570.2014.975010.

SANTONI DE SIO, F.; VAN DEN HOVEN, J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, v. 5, p. 15, 2018. DOI: https://doi.org/10.3389/frobt.2018.00015.

SCHARRE, P. *Army of None: Autonomous Weapons and the Future of War*. Reprint. New York: W. W. Norton & Company, 2018.

SHEA, N. Organized representations forming a computationally useful processing structure. *Synthese*, v. 202, n. 6, p. 175, 2023. DOI: https://doi.org/10.1007/s11229-023-04373-2.

SINGER, P. W. *Wired for War: The Robotics Revolution and Conflict in the Twenty-first Century*. Illustrated reprint. New York: Penguin Press, 2009.

SKITKA, L. J.; MOSIER, K. L.; BURDICK, M. Does automation bias decision-making? *International Journal of Human-Computer Studies*, v. 51, n. 5, p. 991–1006, 1999. DOI: https://doi.org/10.1006/ijhc.1999.0252.

SOLOVYEVA, A.; HYNEK, N. When stigmatization does not work: Over-securitization in efforts of the Campaign to Stop Killer Robots. *AI & SOCIETY*, v. 38, n. 6, p. 2547–2569, 2023. DOI: https://doi.org/10.1007/s00146-022-01613-w.

SPARROW, R. Killer Robots. *Journal of Applied Philosophy*, v. 24, n. 1, p. 62–77, 2007. DOI: https://doi.org/10.1111/j.1468-5930.2007.00346.x.

SPARROW, R. Robots and Respect: Assessing the Case Against Autonomous Weapon Systems. *Ethics & International Affairs*, v. 30, n. 1, p. 93–116, 2016. DOI: https://doi.org/10.1017/S0892679415000647.

TADDEO, M.; BLANCHARD, A. Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems—A Moral Gambit. *Philosophy & Technology*, v. 35, n. 3, p. 78, 2022. DOI: https://doi.org/10.1007/s13347-022-00571-x.

TADDEO, M.; FLORIDI, L. How AI can be a force for good. *Science*, v. 361, n. 6404, p. 751–752, 2018. DOI: https://doi.org/10.1126/science.aat5991.

WINFIELD, A. F. T.; JIROTKA, M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 376, n. 2133, p. 20180085, 2018. DOI: https://doi.org/10.1098/rsta.2018.0085.

WINTER, E. The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law. *Journal of Conflict and Security Law*, v. 27, n. 1, p. 1–20, 2022. DOI: https://doi.org/10.1093/jcsl/krac001.

WOOD, N. G. Explainable AI in the military domain. *Ethics and Information Technology*, v. 26, n. 2, p. 29, 2024. DOI: https://doi.org/10.1007/s10676-024-09762-w.

ZEDNIK, C. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, v. 34, n. 2, p. 265–288, 2021. DOI: https://doi.org/10.1007/s13347-019-00382-7.