ARMAS AUTÓNOMAS, INTELIGENCIA ARTIFICIAL Y DESHUMANIZACIÓN DIGITAL: UN CASO DE CONEXIÓN NECESARIA Y ESENCIAL ENTRE EL DERECHO DE LA INTELIGENCIA ARTIFICIAL Y LA CLÁUSULA MARTENS.

AUTONOMOUS WEAPONS, ARTIFICIAL INTELLIGENCE AND DIGITAL DEHUMANIZATION: A CASE OF NECESSARY AND ESSENTIAL CONNECTION BETWEEN THE LAW OF ARTIFICIAL INTELLIGENCE AND THE MARTENS CLAUSE

Marcos López Oneto[*]

RESUMEN: La inteligencia artificial (IA) tiene el poder de dotar de autonomía a las máquinas. Las armas autónomas (AA), gracias a la IA pueden, con mayor, menor o sin ninguna intervención humana, seleccionar blancos y destruirlos. Esta situación plantea la pregunta ético-jurídica sobre si la humanidad debería permitir la existencia de este tipo de armas, sobre todo, de las completamente autónomas. Mientras la humanidad, tensada por las fuerzas de la geopolítica, medita y discute el tema en función de lograr un instrumento internacional que

[*] Abogado. Licenciado en Ciencias Jurídicas y Sociales, Universidad de Chile. Magister en Derecho con mención en Derecho Privado, Universidad de Chile. Doctor en Derecho, Universidad de Chile. Profesor de Introducción al Derecho de la Inteligencia Artificial, programa de posgrado Facultad de Derecho, Universidad del Desarrollo y profesor visitante en el Laboratorio de Inteligencia Artificial, Facultad de Derecho, Universidad de Buenos Aires. Miembro del *team* de investigadores del Center for AI and Digital Policy, Washington D.C.-Boston, MA, USA.
El autor agradece al comandante de grupo Juan Pablo Benavente Nitsche de la Fuerza Aérea de Chile, por la ayuda prestada para comprender militarmente las AAs. Por supuesto que todos los errores que pudiere contener en este artículo son de completa responsabilidad del autor.

las regule, las AA se están desarrollando sin un marco jurídico internacional. Frente a eso y ante la ausencia de normas internacionales expresas tipo reglas contenidas en instrumentos internacionales vinculantes, en este artículo se propone recurrir al derecho de la inteligencia artificial y a la Cláusula Martens como normas generales de solución de conflictos en la materia.

Palabras clave: inteligencia artificial, sistemas de armas autónomas, derecho de la inteligencia artificial, cláusula Martens.

Abstract: Artificial Intelligence (AI) holds the potential to confer autonomy upon machines. Through AI, Autonomous Weapons (AWs) can identify and engage targets with minimal human oversight. This development prompts ethical and legal inquiries into the permissibility of such weapons, particularly those with full autonomy. As global deliberations seek to establish an international regulatory framework, AWs are being developed outside of such guidelines. In the absence of explicit international regulations, this paper suggests leveraging the Artificial Intelligence Law and the Martens Clause as overarching principles for addressing these ethical and legal challenges.

Keywords: artificial intelligence, autonomous weapons systems, artificial intelligence law, Martens clause.

Introducción

La IA es una de las tecnologías más disruptivas del siglo xxi. Su desarrollo exponencial pone en el horizonte científico como una posibilidad real el surgimiento no solo de la inteligencia artificial general (IAG), sino que, una vez alcanzada aquella, abre la posibilidad del gran salto cualitativo hacia la denominada súper inteligencia artificial (SIA); que sería el estadio social donde emergerían los más diversos ciberseres; seres meramente virtuales que, de humanos, como ha dicho un conocido transhumanista, solo tendría la tendencia inherente a expandir su alcance físico y mental más allá de sus limitaciones[1].

Pero no han sido las predicciones milenaristas poshumanistas las que, hasta el momento, han motivado la generación de proyectos legislativos en torno a su desarrollo, siendo el de la Unión Europea (UE) –sin duda alguna– el más importante hasta la fecha.

[1] Kurzweil (2012) p. 25.

La política legislativa mundial no está enfocada en el riesgo existencial del advenimiento de la poshumanidad. La política legislativa, como no podría ser de otro modo, se está movilizando coyunturalmente por los temas más urgentes de solucionar; aquellos que hoy están afectando de forma grave los derechos fundamentales de las personas y que el proyecto de Ley de IA de la UE, en parte, está enfrentando, a saber: algoritmos discriminadores, reconocimiento biométrico facial remoto en lugares públicos, *scoring* social, falta de transparencia (el tema de la caja negra y de la intrazabilidad algorítmica), manipulación de la conciencia individual y de los grupos sociales, vehículos autónomos y, más en general, de toda clase de sistemas autónomos, exceptuando por razones geopolíticas y de posibilidad de consenso, por ahora, solo a las AA[2] que manifiestan una de las caras más distópicas de la IA.

Pues bien, en este artículo se presentarán algunos de los principales desafíos éticos y jurídicos que el desarrollo de las AA está presentando a la humanidad. Cabe señalar que, por el momento y como es geopolíticamente obvio, las grandes potencias no están interesadas en regular el desarrollo de las AA. De hecho, el proyecto de ley de IA de la UE, como ya se indicó, excluye la regulación de los sistemas de armas que usan IA[3].

Ante esta falta de regulación explícita en la materia, cuestión que crea una de serie de riesgos existenciales[4], en este artículo se propone como normas generales de solución de disputas, la utilización copulativa del derecho de la inteligencia artificial y de la Cláusula Martens, esta última, la única provisión de derecho internacional humanitario que establece una línea base para la protección de los civiles y los combatientes, dado que no existe ningún tratado específico sobre este tema[5].

El presente artículo se divide en cinco secciones. En la primera, y para los efectos de fijar posteriormente la particularidad de los sistemas de armas

[2] Una panorámica sobre los diversos tipos de armas autónomas actualmente existentes, en el Monitor de Armas Autónomas de la ONG Stop Killer Robots. Disponible en https://automatedresearch.org/weapons-systems/ [fecha de consulta: 25 de abril de 2024] y también en Autonomous Weapons Watch. Disponible en https://autonomousweaponswatch.org/ [fecha de consulta:14 de abril de 24].

[3] Véase art. 2 n.° 3, inciso tercero, del proyecto de ley de IA de la UE, del acuerdo provisional de 2 de febrero de 2024, Véase European Parliament (2024) p. 83.

[4] El termino riesgo existencial fue acuñado por el filósofo de la inteligencia artificial, Nick Bostrom, quien escribe: "[...] un riesgo existencial es el que amenaza con causar la extinción de la vida inteligente de origen terrestre o con destruir de forma permanente y drástica sus posibilidades de desarrollarse en el futuro". Bostrom (2016) p. 115.

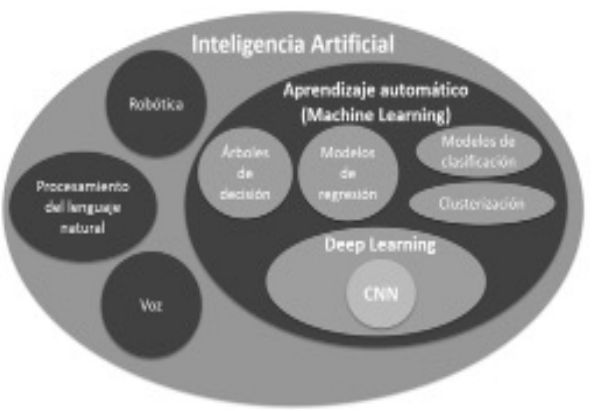
[5] Para el desarrollo de este argumento, se sigue básicamente el razonamiento jurídico expuesto en Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020).

autónomas, se presentará un marco conceptual general sobre IA. En la segunda, se analizarán las principales categorías que definen los sistemas de armas autónomas. En la tercera, se discutirán los desafíos, riesgos y problemas que ese tipo de sistemas generan. En la cuarta, se fundamentará que, ante la ausencia de legislación explicita sobre la materia, el derecho de la inteligencia artificial y la Cláusula Martens, son las soluciones normativas que deberían emplearse. Finalmente, en la quinta, se propondrán algunas conclusiones.

I. Una elucidación conceptual de la IA para fines jurídicos

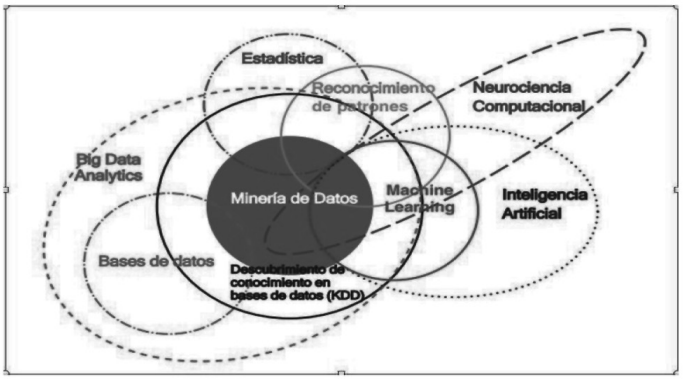
Una primera constatación empírica que emerge de la observación del estado del arte, es que no existe una definición única de IA que sea universalmente aceptada por los profesionales del campo. Algunos definen la IA como un sistema computarizado que exhibe un comportamiento que por lo común se considera que requiere inteligencia. Otros definen la IA como un sistema capaz de resolver de forma racional problemas complejos o tomar acciones apropiadas para lograr sus objetivos en cualquier circunstancia del mundo real que se encuentre. Dícese que este tipo de inteligencia es artificial, para diferenciarla de la inteligencia orgánica o biológica, siendo primariamente la de los humanos, el modelo de construcción a seguir.

Por otra parte, también debe tenerse presente que la IA es un campo científico-industrial en un estado de permanente desarrollo, complejización y diferenciación en múltiples áreas y subáreas[6]. Como consecuencia de lo anterior, la IA no es una sola tecnología: es una familia de tecnologías que buscan desarrollar máquinas o entes inteligentes.



[6] Para una visión del campo de la IA al año 2024: Hai Stanford University. Human Centered Artificial Intelligence (2024).

Y a la par, que es un campo integrado por diversas tecnologías o desarrollos, asimismo, es punto de intersección con muchos con otros campos científicos.



Por la variedad de las fuentes de las cuales se nutre, no es fácil definirla, ya que depende de la perspectiva o enfoque desde el cual se pretenda asir su esencia. Y es que la IA es una ciencia compleja construida con fundamentos extraídos principalmente de: la filosofía, las matemáticas, la economía, la teoría de la información, las neurociencias, la sicología, la ingeniería computacional (*hardware* y *software*), la teoría del control, la cibernética y la lingüística, entre otras disciplinas.

Sin perjuicio de las dificultades definitorias aludidas, en una primera aproximación, la IA podría ser comprendida como una disciplina científica que busca desarrollar métodos y algoritmos soportados en sustratos artificiales que permitan generar comportamientos inteligentes. En términos subjetivos, la IA sería aquella que exhiben ciertos sistemas o artefactos construidos por el hombre. Así es como se ha conjeturado que un sistema artificial poseería inteligencia cuando es capaz de llevar a cabo tareas que, si fuesen realizadas por un humano, se diría de este que es inteligente.

Uno de los libros más autorizados y reconocidos del campo de la IA, luego de analizar las definiciones de IA en ocho libros de texto, señala que es posible distinguir los siguientes enfoques en la materia[7]:

1) Comportamiento humano: el enfoque del test de Turing[8]: un sistema de IA es inteligente en la medida que para un observador realiza conductas humanas.

[7] Russell y Norvig (2008) pp. 2-5.

[8] El test de Turing tiene por objetivo determinar si un ordenador puede convencer que es humano a un observador que lo interroga durante el curso de una conversación experimental. El test debe su nombre a su creador, Alan Mathison Turing (1912-1954) , matemático y lógico, considerado uno de los padres fundadores de las ciencias de la computación y la inteligencia artificial.

2) Pensar como humano: el enfoque del modelo cognitivo: un sistema de IA es inteligente en la medida que piense como un humano. Para esto se necesita tener una teoría del conocimiento humano y expresarla, además, algorítmicamente y llevarla al *software*. Este enfoque de IA está muy unido a las ciencias cognitivas. Ambos campos científicos se retroalimentan, sobre todo en las áreas de visión de colores y lenguaje natural.
3) Pensamiento racional: el enfoque de las "leyes del pensamiento". Para esta concepción de la IA, un sistema es inteligente en la medida que resuelve problemas lógicos formales. Se trata de la llamada tradición logicista dentro del campo que intenta construir programas que puedan resolver problemas descritos en notación lógica. La gran dificultad de este enfoque deriva del hecho de que no es simple traducir el conocimiento informal a un sistema de notación lógica.
4) Actuar de forma racional (agente racional): sobre este enfoque, explican Stuart Russel y Peter Norvig que un agente es algo que razona. Indican que de los agentes informáticos se espera que tengan otros atributos que los distingan de los "programas" convencionales, como que estén dotados de controles autónomos, que perciban su entorno, que persistan durante un periodo prolongado, que se adapten a los cambios, y que sean capaces de alcanzar objetivos diferentes. Un agente racional es aquel que actúa con la intención de alcanzar el mejor resultado o, cuando hay incertidumbre, el mejor resultado esperado[9]. Las ventajas de este enfoque es que no descarta el razonamiento lógico formal, pero entiende que hay muchas conductas que no pueden inferirse lógicamente y que, sin embargo, pueden ser consideradas correctas, así como también hay situaciones donde no existe una salida correcta.

Otra forma de acercarse a la comprensión taxonómica conceptual de la IA, es a través de la distinción inteligencia artificial general (IAG) e inteligencia artificial especial, experta o estrecha (IAE). La primera, es aquella que puede realizar cualquier tarea cognoscitiva posible de observar en un ser humano. Es la gran meta de la IA, no alcanzada hasta hoy, de construir un sistema de inteligencia artificial que, poseyendo y desarrollando todas las funciones lógicas formales e informales del ser humano, además tenga sentido común, intuición, autonomía, capacidad de autoaprendizaje permanente e, incluso, hasta una especie de agencia moral.

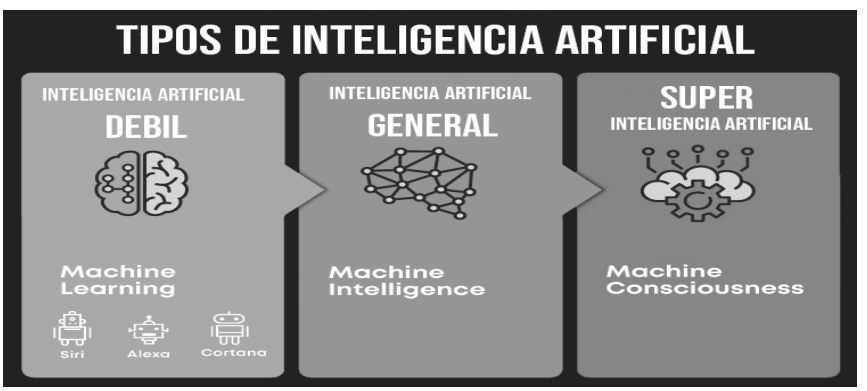
La IAE resuelve problemas y entrega resultados en áreas particulares, emulando y muchas veces superando las capacidades humanas, sobre todo por las predicciones y deducciones que se obtienen mediante el análisis de volú-

[9] Russell y Norvig (2008) p. 5.

menes gigantes de información (*bigdate*) con aprendizaje automático (*machine learning)*.

La IAE o experta solo puede abocarse a la realización de las tareas que el programa le permite. Así, por ejemplo, los programas que juegan al ajedrez a nivel de gran maestro son incapaces de jugar a las damas a pesar de ser un juego mucho más sencillo. Se requiere diseñar y ejecutar un programa distinto e independiente del que le permite jugar al ajedrez para que el mismo ordenador juegue, también, a las damas. En el caso de los seres humanos no es así, pues cualquier jugador de ajedrez puede aprovechar sus conocimientos sobre este juego para, en cuestión de segundos, jugar a las damas[10]. Cabe tener presente que en la actualidad la IAE ha logrado éxitos espectaculares en sistemas expertos, que son el estado actual de desarrollo de la IA. El mundo actual está sostenido sobre la IAE. ¿Cuál será el mundo que sostenga la IAG?

Se ha planteado, también, la posibilidad de una superinteligencia artificial (SIA), que sería el paso siguiente al desarrollo de la IAG. Puesto que la IAG tendría capacidad de autoaprendizaje, cada vez se haría más inteligente, llegando a un punto de desarrollo inalcanzable para el ser humano. Llegados a ese punto exponencial, se especula que la SIA se transformaría en una máquina con autoconciencia y objetivos propios. En ese estadio de desarrollo se verá nacer a una nueva especie. Un ser cibernético que no necesariamente tendría que estar alineado con objetivos humanos, lo cual propone un horizonte de riesgos existenciales para la humanidad.



En la actualidad la técnica dominante de IA es el aprendizaje automático, más conocida por su denominación en inglés *machine learning*[11].

[10] López de Mántaras y Meseguer (2017) pp. 67-74.

[11] Una forma de medir el liderazgo de la mencionada técnica de IA es por el registro del número de patentes en el ámbito mundial de ellas. Véase: World Intelectual Property Organization (2019) pp. 13-17.

Machine learning es una forma de IA que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita. Conforme el algoritmo, ingiere datos de entrenamiento, es posible producir modelos más precisos basados en datos. Las IA construidas como *machine learning* aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. Como ha precisado un ingeniero de *software*:

> "La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana"[12].

En efecto:

> "una cosa es programar una máquina para que pueda moverse. Y otra muy distinta programarla para que aprenda a moverse. Igualmente, no es lo mismo programar qué elementos forman una cara, que automáticamente aprender qué es una cara"[13].

Así, *machine learning* es el término general que se usa en el complejo campo de la IA, cuando los sistemas computacionales artificiales aprenden de los datos. En lugar de rutinas de *software* de codificación con instrucciones precisas para realizar una tarea particular, *machine learning* es una forma de entrenar un algoritmo para que pueda aprender cómo realizarla. El proceso de aprendizaje puede ser supervisado o no supervisado, dependiendo de la forma como se introducen los datos que se utilizan para alimentar el algoritmo.

Una especie de *machine learning* es la tecnología conocida como *deep learning* (aprendizaje profundo), que se caracteriza por el uso de múltiples capas de redes neuronales artificiales[14]. Las redes neuronales son modelos matemáticos simples del funcionamiento del sistema nervioso. Las unidades básicas son las neuronas artificiales que, generalmente, se organizan en capas.

Según IBM:

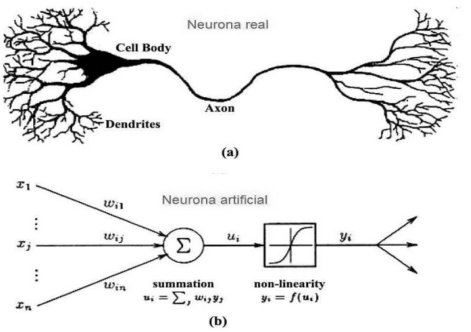
> "una red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas"[15].

[12] Guzmán (2018).

[13] Cáceres (2022) p. 26.

[14] Sobre redes neuronales, un texto en español de muy fácil acceso y disponibilidad es Berzal (2018).

[15] Definición consultada en IBM, disponible en www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model [fecha de consulta: 29 de enero de 2022].



Para ir terminando esta elucidación, convengamos que, si la IA pudiera definirse en función de los objetivos, se podría señalar que se trata de una disciplina científica que tiene por misión construir sistemas (físicos o virtuales)capaces de deducir, razonar, resolver problemas, planificar, aprender; procesar lenguajes naturales, mostrar creatividad, inteligencia social y general, así como también tener capacidad de movimiento, autonomía, capacidad de decisión y percepción, física o virtual. Su fin es crear entes artificiales inteligentes sobre la base de un determinado modelo de inteligencia.

Toda esta complejidad conceptual descrita en los párrafos previos, ha estado presente en el debate legislativo de la UE. En efecto, la definición de inteligencia artificial fue, naturalmente, una parte primordial del debate. Originalmente el proyecto de ley de inteligencia artificial de la UE proponía una definición de IA (art. 3) con foco en la idea de "*software*". Luego se avanzó hacía otra más general.

Comparemos:

El proyecto original disponía en su art. 3, lo siguiente:

> "Definiciones. A los efectos del presente Reglamento, se aplican las siguientes definiciones:
>
> (1) 'sistema de inteligencia artificial' (sistema de IA): software desarrollado con una o más de las técnicas y enfoques enumerados en el anexo I y que puede, para un conjunto dado de objetivos definidos por el ser humano, generar resultados como contenido, predicciones, recomendaciones o decisiones que influyen en los entornos con los que interactúan.
>
> ANEXO I: TÉCNICAS Y ENFOQUES DE INTELIGENCIA ARTIFICIAL a que se refiere el artículo 3, punto 1 a) Estrategias de aprendizaje automático, incluidos el aprendizaje supervisado, el no supervisado y el realizado por refuerzo, que emplean una amplia variedad de métodos, entre ellos el aprendizaje profundo. Estrategias basadas en la lógica y el conocimiento, especialmente la representación del conocimiento, la programación (lógica) inductiva, las bases de conocimiento, los motores

> de inferencia y deducción, los sistemas expertos y de razonamiento (simbólico). Estrategias estadísticas, estimación bayesiana, métodos de búsqueda y optimización"[16].

Posteriormente, el 25 de noviembre de 2022 el Comité de Representantes Permanentes del Consejo al Consejo de la UE, amplió la definición, huyendo de la noción de *software* hacia la siguiente conceptualización legal:

> "Artículo 3 Definiciones. A los efectos del presente Reglamento, se aplican las siguientes definiciones. (1)'sistema de inteligencia artificial' (sistema de IA): un sistema que está diseñado para funcionar con elementos de autonomía y que, basándose en datos e insumos proporcionados por máquinas y/o humanos, infiere cómo lograr un conjunto determinado de objetivos utilizando aprendizaje automático y/o enfoques basados en la lógica y el conocimiento, y produce resultados generados por el sistema, como contenido (sistemas generativos de IA), predicciones, recomendaciones o decisiones, que influyen en los entornos con los que interactúa el sistema de IA"[17].

En el preámbulo del texto, el Comité de Representantes Permanentes del Consejo, explica las razones que tuvo para ampliar la definición:

> "[…] para garantizar que la definición de un sistema de IA proporcione criterios suficientemente claros para distinguir la IA de los sistemas de software más clásicos, el texto transaccional reduce la definición del artículo 3, apartado 1, a los sistemas desarrollados median te enfoques de aprendizaje automático y lógica y conocimiento […] En cambio, se han añadido nuevos considerandos 6a y 6b para aclarar qué debe entenderse por enfoques de aprendizaje automático y enfoques basados en la lógica y el conocimiento. Para garantizar que la Ley sobre IA siga siendo flexible y preparada para el futuro, se ha añadido en el artículo 4 la posibilidad de adoptar actos de ejecución para especificar y actualizar técnicas en el marco de enfoques de aprendizaje automático y enfoques basados en la lógica y el conocimiento"[18].

II. Armas Autónomas (AA)

Autonomía es la cualidad que tiene un sistema para autogobernarse. El autogobierno del sistema se materializa mediante la toma de decisiones. Los sistemas

[16] European Commission (2021) p. 39.
[17] Council of the European Union (2022) p. 71.
[18] *Op. cit.* p. 4.

artificiales, es decir, los sistemas no orgánicos creados por el hombre, podrían tener ciertos niveles de autonomía. En el contexto de los sistemas de defensa, el término se utiliza normalmente para describir cómo las máquinas de guerra realizan ciertas funciones (en distintos grados) independientemente del control humano.

Hoy, la autonomía de los sistemas artificiales es posible gracias a los avances en IA. Y se espera que, a mayor desarrollo de la IA, mayores serán los niveles de autonomía que se podrán observar en los sistemas de armas. Si la IA algún día llegara al nivel de la IAG, la autonomía de los sistemas sería similar a la que se puede observar en los seres humanos.

En términos generales, las AA son sistemas que, dado que usan IA, pueden seleccionar y atacar objetivos con o sin un control humano significativo. Es decir, pueden, en alguna medida, autogobernarse y, por tanto, representan, según se ha dicho:

> "un paso inaceptable, más allá de los drones armados existentes, porque un humano no tomaría la decisión final sobre el uso de la fuerza en ataques individuales"[19].

La ONG Human Rigth Wacht, en un muy completo informe elaborado junto con la Escuela de Derecho de Harvard[20], señaló:

> "las armas completamente autónomas, también conocidas como sistemas letales de armas autónomas o 'robots asesinos', oficialmente, aún no existirían, pero están en desarrollo y las inversiones militares en tecnología autónoma están aumentando a un ritmo alarmante".

La definición de armas autónomas, por cierto que no es trivial. El Grupo de Expertos Gubernamentales (Group of Governmental Experts, GGE), luego de una amplia consulta, previene lógicamente que una definición de armas autónomas es un requisito previo importante para cualquier acción normativa firme[21].

---

[19] Human Rights Watch/International Human Rigths Clinic Harvard Law School (2020) p. 6.

[20] *Op. cit*. p. 6.

[21] Group of Govermental Experts on Emerging Tecnologies in the Area of Lethal Auutonomous Weapons Systems (2024). El GGE es una institución creada al amparo de la Convención sobre Prohibiciones o Restricciones del Empleo de Determinadas Armas Convencionales que Pueden Considerarse Excesivamente Nocivas o de Efectos Indiscriminado, también referida como Convención Sobre Ciertas Armas Convencionales, que fue adoptada el 10 de octubre de 1980 y entró en vigor en 1983. Al respecto, puede consultarse: https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/ [fecha de consulta: 15 de abril de 2024].

En un muy reciente informe del Servicio de Investigación del Congreso de Estados Unidos de América, se indica que los debates internacionales no se utiliza ninguna definición única y universalmente aceptada de Sistemas Letales de Armas Autónomas (Lethal Autonomous Weapon Systems, LAWS). A su turno, señala que la directiva 3000.09 del Departamento de Defensa, que establece la política estadounidense sobre autonomía en los sistemas de armas, define los LAWS como "sistemas de armas que, una vez activados, pueden seleccionar y atacar objetivos sin mayor intervención de un operador". La característica principal de esta definición es el papel del operador con respecto a la selección de objetivos y las decisiones de compromiso[22].

Otros países, prosigue el citado informe, han basado su definición en diferentes características, en particular la sofisticación tecnológica del sistema de armas, de modo que se considera que los LAWS son sistemas de armas capaces de cognición en el ámbito humano. En fin, también hay quienes no creen que sea necesaria (o deseable) una definición de los LAWS para los debates internacionales. A pesar de estas diferencias, la mayoría de los incumbentes por lo general coinciden en que las características definitorias incluyen la autonomía total (sin control humano manual del sistema) o parcial y el potencial de producir efectos letales[23].

Tampoco hay que olvidar que la discusión normativa mundial sobre la materia está teniendo lugar en el contexto de la Convención sobre Ciertas Armas Convencionales. Cabe tener presente que desde el año 2018, las Naciones Unidas, a través de su secretario general António Guterres, ha sostenido en varias oportunidades:

> "que, en ausencia de regulaciones multilaterales específicas, el diseño, desarrollo y uso de estos sistemas plantean preocupaciones humanitarias, legales, de seguridad y éticas y representan una amenaza directa a los derechos humanos y las libertades fundamentales"[24].

Sin embargo, no puede perderse de vista que las palabras del secretario general, parecen no tener mucho eco entre las grandes potencias, toda vez que el desarrollo de AA es una de las dimensiones de The Third Offset Strategy que guía las políticas y estrategias de defensa de Estados Unidos de América, la OTAN y sus aliados[25].

---

[22] Congressional Research Service (2023) p. 1. Un análisis comparativo también puede ser consultado en Taddeo & Blanchard (2021).

[23] Congressional Research Service (2023) p. 1.

[24] United Nations Office For Disarmament Affairs (2018).

[25] OTAN (2022). Para un estudio en profundidad de The Third Offset Strategy, puede consultarse Martinage (2014).

A continuación, y sin una pretensión de exhaustividad, se presentan las principales concepciones técnicas de armas autónomas

1. Definiciones técnicas de armas autónomas

Como se decía, en general, las definiciones se construyen a partir de la mayor o menor intervención humana en el ciclo de uso y activación arma. Una prueba de dicha aproximación taxonómica, es la revisión conceptual contenida en el marco de la convención sobre prohibiciones o restricciones del empleo de ciertas armas convencionales que pueden considerarse excesivamente nocivas o de efectos indiscriminados (1980), que grupo de expertos gubernamentales sobre tecnologías emergentes en el ámbito del sistema de armas letales autónomas, realizó el año 2023. En dicha reunión se propuso una recopilación no exhaustiva de definiciones y caracterizaciones, definiendo las AA (Autonomous Weapons Systems, AWS), básicamente, como sistemas que, tras la activación por parte de un usuario humano, utilizan el procesamiento de datos de sensores para seleccionar y atacar objetivos con fuerza sin intervención humana[26].

La mayor o menor intervención humana da origen a la clasificación que distingue entre armas autónomas y armas semiautónomas.

La fuente primaria institucional de dicha clasificación es la DoD Directive 3000.09 Autonomy In Weapon Systems del Department Of Defense de Estados Unidos de América, publicada el día 25 de enero de 2023[27]. En dicha directiva, se adoptan las siguientes definiciones:

- Sistema de armas autónomo. Un sistema de armas que, una vez activado, puede seleccionar y atacar objetivos sin intervención adicional de un operador humano. Esto incluye sistemas de armas autónomos supervisados por humanos que están diseñados para permitir que los operadores humanos anulen el funcionamiento del sistema de armas, pero que pueden seleccionar y atacar objetivos sin más intervención humana después de la activación.
- Sistema de armas autónomo supervisado por un operador. Un sistema de armas autónomo que está diseñado para brindar a los operadores la capacidad de intervenir y finalizar enfrentamientos, incluso, en caso de falla del sistema de armas, antes de que ocurran niveles de daño inaceptables.
- Sistema de armas semiautónomo. Un sistema de armas que, una vez activado, está destinado a atacar únicamente objetivos individuales o

[26] United Nations Office For Disarmament Affairs (2018).

[27] Department of Defense (2023).

grupos de objetivos específicos que hayan sido seleccionados por un operador humano. Esto incluye:

- Sistemas de armas semiautónomos que emplean autonomía para funciones relacionadas con el combate, incluidas, entre otras, la adquisición, el seguimiento y la identificación de objetivos potenciales; indicar objetivos potenciales a operadores humanos; priorizar objetivos seleccionados; momento de cuándo disparar; o proporcionar guía terminal para localizar objetivos seleccionados, siempre que se mantenga el control humano sobre la decisión de seleccionar objetivos individuales y grupos objetivos específicos para el combate.
- "Disparar y olvidar" o municiones dirigidas con bloqueo después del lanzamiento que dependen de TTP[28] para maximizar la probabilidad de que los únicos objetivos dentro de la cesta de adquisición del buscador cuando el buscador se activa sean aquellos objetivos individuales o grupos de objetivos específicos que han sido seleccionados por un operador humano.

Las definiciones de la DoD Directive 3000.09 Autonomy In Weapon Systems del Department Of Defense, pueden ser graficadas del siguiente modo:

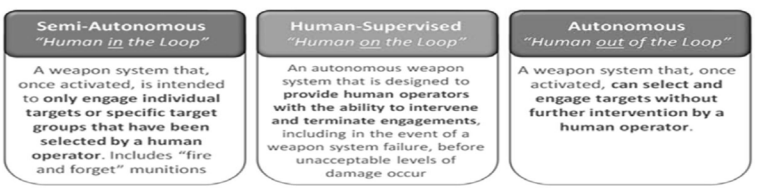


Figure 3-1. DoD 3000.09 Definitions of Top Levels of Autonomy *(Source: Deputy Secretary of Defense (64)).*

Fuente: Chakour (2022) p. 3.

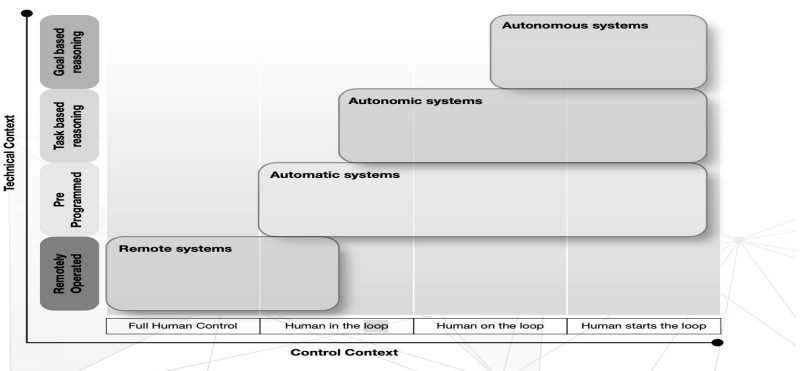
Otra aproximación taxonómica muy ilustrativa es la contenida en Robotic And Autonomous Systems de la Fuerza Australiana de Defensa, donde se propone una distinción escalonada entre: Sistemas de Control Remoto (Remote Control Systems), Sistema Automático (Automatic System), Sistemas Autonómicos (Autonomic Systems) y Sistemas Autónomos (Autonomous Systems)[29], a saber:

- Sistemas de Control Remoto (full control humano). Un sistema que es operado por un humano a través de métodos remotos. Sin el elemento de control remoto, el sistema tiene poca capacidad para funcionar de forma independiente.

[28] Tácticas, técnicas y procedimientos.

[29] Australian Defence Force (2020).

– Sistema Automático (humano en el *loop*). Un sistema que está preprogramado para responder a estímulos de una manera determinista y basada en reglas y que puede lograr su función sin más intervención humana.
– Sistemas Autonómicos (humano sobre el *loop*). Un sistema que logra tareas definidas por humanos operando con referencia a un conjunto de pautas predefinidas y responde a estímulos de manera probabilística. Los sistemas autónomos pueden requerir intervención humana para completar su función o pueden funcionar sin mayor supervisión.
– Sistemas Autónomos (humano fuera del *loop*). Un sistema que determina cómo realizar las tareas necesarias para lograr un objetivo definido. Un sistema autónomo responde a los estímulos de forma probabilística y puede alterar la forma en que realiza las tareas.



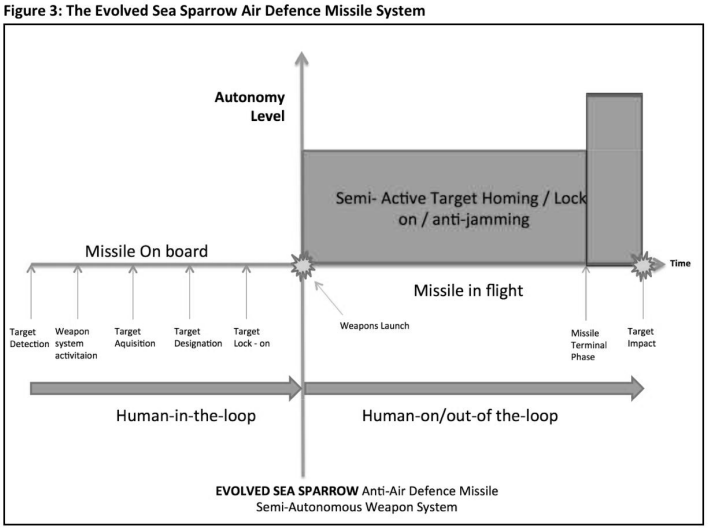
Control, Contex/
Fuente: Australian Defence Force (2020) p. 16.

Debe tenerse presente que el método de control de cada sistema será el específico de la plataforma, la misión y el entorno. Y especifica el citado documento de la Fuerza Aérea de Australia, los siguientes niveles de control humano:

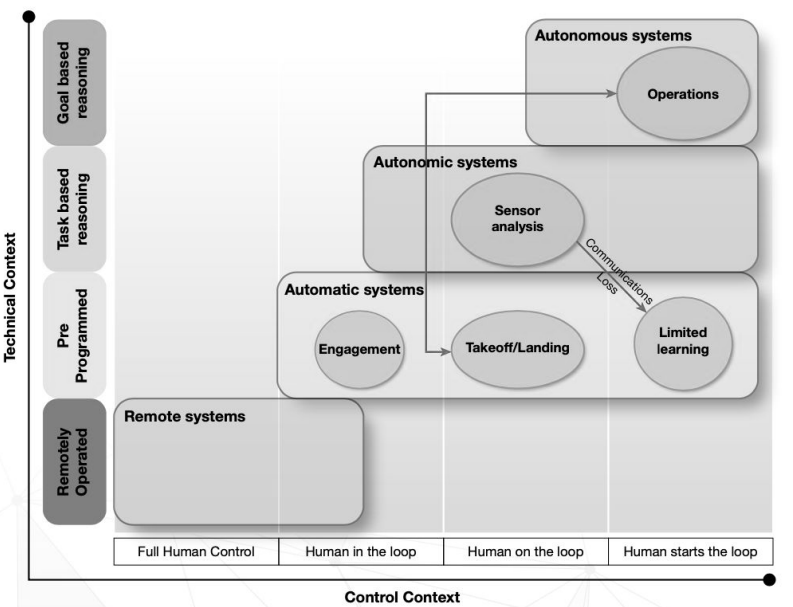
a. Control humano total. Un humano controla todos los aspectos del funcionamiento del sistema, físicamente o mediante control remoto.
b. Humano en el *loop*. El sistema realiza algunas funciones de forma independiente, pero requiere que un humano realice funciones que completen el ciclo de tareas del sistema.
c. Humano sobre el *loop*. El sistema realiza todas las funciones de forma autónoma, pero un humano puede intervenir para detener o modificar el resultado antes de que se complete la tarea.
d. El ser humano inicia el *loop*. Un humano establece los parámetros operativos e inicia la operación del sistema; la máquina no requiere más interacción humana para completar la tarea.

Este es un diagrama de un sistema de misiles antiaéreos, donde se ejemplifican los niveles de control. Este ejemplo presenta el sistema como Sistema Semiautónomo.

Figure 3: The Evolved Sea Sparrow Air Defence Missile System



Fuente: Schaub & Wenzel (2017) p. 16.



Control, Contex,

Fuente: Australian Defence Force (2020)

Figure 3- Categorisation On of a platform.

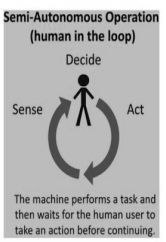
1.1. Una definición operacional de armas autónomas (AA) para los fines de este artículo[30].

Haciendo un ejercicio de abstracción, en general, las definiciones revisadas describen las AA como máquinas capaces de:

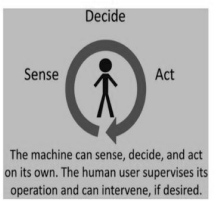
1) observar el entorno en el que existen,
2) orientarse a ese entorno basándose en las entradas de sus sensores,
3) evaluando posibles cursos de acción y decidiendo uno y
4) actuando para implementar esa elección[31]. Dependiendo de la mayor o menor autonomía de que gocen las AA, serán las posibilidades de distinción.

Sin descartar otras posibilidades taxonómicas, y dependiendo del nivel de intervención humana en el ciclo o *loop* de las AA, a continuación se propone distinguir entre:

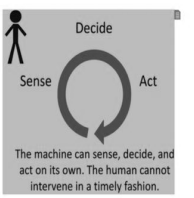
- Semiautónomo (humano en el *loop-human in the loop*): sistemas de armas que una vez activados, solo se engancha con un blanco o con un grupo específico de blancos, los que han sido previamente seleccionados por un operador humano.
- Supervisado por operador (humano sobre el *loop-human on the loop*): sistema de armas diseñado para proveerle al operador la posibilidad de intervenir o terminar un enganche, incluyendo en el caso de fallas, antes que daños inaceptables ocurran.
- Completamente autónomo humano fuera del *Loop-Human out of the Loop*): un sistema de armas que, una vez activado, puede seleccionar y enganchar blancos sin intervención de un operador humano. Este nivel incluye las operaciones completamente autónomas de Sistemas de Armas Autónomos Aéreos como, por ejemplo, el de algunas municiones merodeadoras. Aparentemente estas AA aún no han sido desarrolladas.



Semiautonomous Operation (human in the loop)



Supervised Autonomous Operation (human on the loop)



Fully Autonomous Operation (human out of the loop)

[30] La síntesis conceptual de esta sección se ha valido de Chakour (2022) y Department of Defense (2023) y (2012).

[31] Schaub & Wenzel (2017) p. 7, llegan a la misma conclusión luego de revisar seis definiciones oficiales de armas autónomas.

1.1.1. Otras dos categorías posibles de AA

La categoría sistema completamente autónomo (hombre fuera del *loop*), como se ha descrito, puede seleccionar y enganchar blancos sin intervención de un operador humano, pero supone que el ser humano lo activa. Vale decir, aun cuando puede enganchar y seleccionar un blanco, para que ello ocurra, previamente ha debido ser activado por un humano.

Pues bien, es posible imaginar dos categorías más de AA que, por cierto, no existen en la actualidad, pero que, conforme la IA avance hacia la IAG o la SIA, la posibilidad de que surjan es muy alta.

En efecto, en un mundo donde exista la IAG y más aún en uno donde se esté en presencia de la SIA, existirá la factibilidad tecnológica de construir AA cuya activación no dependa del ser humano. Es decir, que se autoactiven en función de su propio operar interno o que sean activados por otras IA. Por consiguiente, en este artículo, se propone distinguir –futuristamente–, por cierto, entre AA autoactivadas y activadas por IA externas que, a su vez, pueden ser activadas o no activadas por humanos.

III. Desafíos éticos y jurídicos de las armas autónomas

Entre otros desafíos y problemas que presentan las AA, se han listado los siguientes: deshumanización digital, sesgos algorítmicos, pérdida de control humano significativo, falta de juicio y comprensión humanos, falta de responsabilidad, incapacidad para explicar qué sucedió y por qué, reducir el umbral de la guerra y activar una carrera armamentística desestabilizadora[32].

La deshumanización digital consiste en la reducción del ser humano a datos recolectados mediante tecnologías de la información, en virtud de los cuales, y mediante el uso de esas mismas tecnologías, se toman decisiones que los afectan. Las máquinas no tienen juicio moral, solo procesan datos y crean predicciones a través de diversas aplicaciones de IA.

Reducido el hombre a códigos de barras, números; en definitiva, a datos mediante los cuales se construyen imágenes o perfiles estandarizados de seres mediante procesos desprovistos de toda empatía, lo humano comienza a perder sustancia, corporeidad, biología. Lo humano entra en un proceso de desonto-

[32] Listado por la ONG Stop Killer Robot, véase www.stopkillerrobots.org/es/detener-robots-asesinos/hechos-sobre-armas-aut%C3%B3nomas/ [fecha de consulta: 29 de abril de 2024]. Se revisan críticamente en este artículo.

logización moral, de vacuidad existencial. En ese estadio existencial, el hombre, deshumanizado digitalmente, podrá ser visto; pero el hombre, el de carne, sangre y huesos[33], no podrá ser tocado, palpado, sentido, porque habrá perdido su sustancia, su cuerpo, si se quiere, su identidad existencial.

En este orden de cosas, la utilización de AA emerge como la más radical forma de deshumanización digital, pues las máquinas ya no solo podrán, por ejemplo, denegar la concesión de un seguro de salud, sino que, incluso, matar a un ser humano. Sin duda que ese estado de cosas supone un refuerzo y desarrollo máximo de la tecnología de sesgos algorítmicos contra cuyos usos más benignos, diversas instituciones, organizaciones y gobiernos, ya están en guardia, alegando por mayor transparencia[34].

Asimismo, no debe perderse vista que, si los sesgos algorítmicos ya representan un peligro y ataque comprobado a derechos fundamentales como la privacidad, la igualdad y la libertad; peligro respecto del cual diversas sociedades que integran la sociedad mundial[35] han reaccionado comenzando a crear regulaciones protectoras; imagínese, entonces, cuán crítico desde el punto de vista ético y jurídico, resultaría los sesgos operacionales de las AA, que ponen en peligro, además de los anteriores, a los más fundamentales de los derechos fundamentales, a saber: el derecho a la integridad física y psíquica y el derecho a la vida.

Se ha propuesto, también, que la pérdida de control humano significativo de las AA significaría "que los usuarios de armas ya no están completamente comprometidos con las consecuencias de sus actos"[36], lo cual reduce el espacio para el razonamiento humano, pues las máquinas no podrían tomar decisiones éticas complejas y tampoco pueden comprender el valor de la vida humana y el contexto donde estas tienen lugar.

Sin embargo, de la pérdida de control humano significativo no se deduce necesariamente que los usuarios de las armas no estén comprometidos con los resultados de su operar. En efecto, los diseñadores, fabricantes y usuarios de AA saben perfectamente qué pueden hacer o no hacer los sistemas, incluido el error de blanco o la invención de blancos certeros mediante IA generativas. Luego, el problema, "no es que los usuarios de armas ya no están completamente comprometidos con las consecuencias de sus actos", pues ellos siempre

[33] "Este armazón de huesos y pellejo", dice el poeta sevillano Gustavo Adolfo Bécquer.

[34] La bibliografía es extensa, una buena introducción al tema en Consejo para la Transparencia (2020); Cotino y Castellanos (2022). También puede consultarse en sitio web del Centro Europeo para la Transparencia Algorítmica: https://algorithmic-transparency.ec.europa.eu/index_en [fecha de consulta: 29 de abril de 2024].

[35] Luhman (2007).

[36] ONG Stop Killer Robot, *op. cit*.

están comprometidos, puesto que tienen conocimiento objetivo del operar de sus AA y, por tanto, no pueden desconocer sus capacidades y poder destructivos. Están absolutamente comprometidos porque saben exactamente lo que han diseñado, fabricado o puesto en servicio: AA y no videos juegos de entretención.

La pregunta de fondo, por tanto, sería otra: ¿debe permitirse el diseño, fabricación y puesta en servicio AA? Si la respuesta es afirmativa, entonces surge una segunda pregunta: ¿qué clase de AA deben permitirse? *¿In, on u out the loop*? Y resuelta positivamente esa última pregunta, abriéndose el camino hacia la permisión de alguna clase SAA, entonces cabe preguntarse: ¿cuál sería el marco normativo de esa permisión, en caso que se estime que debe estar regulada de algún modo?

Se ha dicho también que la falta de juicio y comprensión humanos de las AA, son razones más que suficientes para someterlos a muy estrictas limitaciones de uso, y que las personas y no las máquinas son las que deben rendir cuentas en contextos donde este tipo de sistemas deciden entre la vida y la muerte[37].

Las objeciones precedentemente expuestas son razonables, pero, en esencia, se trata de la crítica ética predicable de todo tipo de sistema que sea gobernado con IA y que se relaciona, en última instancia analítica, con el tema de si la agencia moral solo debería o podría predicarse de los humanos. O planteado desde otra perspectiva, de si acaso sería éticamente permitido otorgarles algún tipo de personalidad jurídica a las máquinas inteligentes, desplazando así la responsabilidad resultante de las interacciones de esas máquinas con el ambiente, desde las personas humanas de sus diseñadores, fabricantes o usuarios, a las máquinas mismas, como centro de imputación, a lo menos patrimonialmente, bajo la lógica ficcional de las personas jurídicas[38].

Por otra parte, la incapacidad para explicar qué sucedió y por qué, no es una característica que se predique única y exclusivamente de las AA. Las AA poseen esas características en mayor o menor medida, en función del nivel o tipo de IA que usan. Se trata de una característica presente en todos los sistemas de IA que usan *machine learning*, el cual se ve aún más potenciado en los sistemas que usan un tipo de *machine learning* denominado *deep learning* que, básicamente, consiste en sistemas de IA que procesan información y producen predicciones mediante el operar de complejas arquitecturas de miles de capas de redes neuronales artificiales.

El problema aquí es que, por la estructura de funcionamiento de las redes neuronales artificiales, no es posible determinar si los *outputs* han sido pro-

[37] ONG Stop Killer Robot, *op. cit*.

[38] La Unión Europea ha estado explorando el tema, al menos, desde el año 2017, véase Parlamento Europeo (2017).

ducidos por sesgos así como tampoco es posible trazar paso a paso el procedimiento o ruta del "razonamiento" que produjo la predición (el actuar) del sistema, porque la arquitectura neuronal es inescrutable, intrazable, inexplicable. De ahí que uno de los principios éticos fundamentales propuestos para el desarrollo de la IA sea la trazabilidad o explicabilidad del funcionamiento de las IA. Las IA del tipo *deep learning* son verdaderas cajas negras. La crítica a las AA que operan con *machine learning*, en general, y con *deep learning*, en particular, es, por tanto, una crítica que se predica de todo sistema de IA que sea gobernado por dichas técnicas.

Pues bien, lo que se ha querido indicar hasta ahora, por tanto, es que las principales críticas que se hacen a las AA, en esencia, son las mismas críticas que se formulan con diversos énfasis y, en general, al desarrollo sin controles éticos de la IA. La IA es percibida como un riesgo existencial tecnológico similar a la energía atómica, cuestión que, por cierto, ha motivado que regiones geopolíticamente tan importantes como la Unión Europea, se hayan embarcado en su regulación, exceptuando, eso sí, a las AA (art 2 n.° 3 del reglamento).

Por todo lo anterior cabe preguntarse lo siguiente: ¿si los desafíos éticos que imponen las AA se predican de los desafíos más generales que la IA propone a la humanidad, ¿cuáles serían, entonces, los desafíos o riesgos específicos que las AA pueden producir en el corto o mediano plazo?

Por cierto que la AAs deshumanizan radicalmente al ser humano. Se trata de un proceso de deconstrucción algorítmica de la imagen jurídico-cultural del hombre, que da como resultado la creación del ser humano digital que, poco a poco, irá reemplazando al ser humano real. El ser humano digital tiene una identidad digital generada y reconocida por el operar del algoritmo computacional gestionado con IA; algoritmo que, en definitiva, es la máquina de IA, que consume datos para construir y reconocer el mundo y tomar decisiones (que son predicciones). El hombre pierde, así, toda sustancia y conexión vital; es solo un conjunto de datos computacionales, ceros y unos que viajan a la velocidad de la luz en un chip compuesto por billones de microtransistores.

Pero como se decía, la deshumanización digital no solo es un problema específico de las AA. Es el gran problema ético que se predica del desarrollo de la IA. Por consiguiente, las regulaciones generales que deberían aplicársele quedan dentro del campo de las regulaciones generales de la IA.

Ahora bien, sumados a los riesgos existenciales que toda IA genera, máxime en el hipotético estadio de la IAG y la SIA, los problemas específicos que las AA pueden generar con efectos catastróficos, digamos, los reales problemas de corto y mediano plazo, los cuales ciertamente requieren una discusión pragmática, son dos: el incentivo a reducir el umbral de la guerra y la repotenciación peligrosa de la carrera armamentística de las grandes potencias.

El presidente Vladímir Putin sin mayores adornos declaró en el año 2017: "quien domine la Inteligencia Artificial dominará al mundo"[39]. Sin duda que estaba pensando en las aplicaciones de IA que producen más IA y que, puestas al servicio de objetivos geopolíticos, siempre acompañados de poder duro, léase militar, permiten la consecución de los objetivos descritos por el mandatario ruso.

Las AA permiten, y permitirán cada vez más, el uso de menos contingente humano en el campo de batalla y en el teatro general de las operaciones bélicas. Y, si bien es comprensible, como se ha dicho, que los Estados quieran reducir los riesgos de pérdida de vidas para sus propias tropas[40]; mucho menos lo sería que producto del uso de AA disminuya la tasa de bajas militares y correlativamente aumente la de la población civil.

En ese sentido, una primera conversación normativa mundial sobre la regulación de las AA, como la que está promoviendo la ONU[41], debería considerar ese tópico: asegurar que las AA, de ser usadas, no generen daño a la población civil o que, en caso de producirse, siempre estén entrenadas para producir el menor daño posible a la población civil, significativamente menos que el que de forma hipotética produciría un sistema controlado por humanos.

La paulatina disminución de elementos humanos en el campo de batalla y, más en general, en el teatro global de las operaciones militares del conflicto, como consecuencia de la introducción de AA, dado el caso también, y bajo ciertas circunstancias, podría ser una razón suficiente para decidir participar en un conflicto militar.

Bajo la óptica de la realidad actual, la promesa política de que no habrá soldados en el campo de batalla, por cierto que cambia la ecuación del manejo y manipulación de la opinión pública. Cuando lo que viene devuelta es chatarra electrónica y no cadáveres de jóvenes que tenían todo un futuro por delante, la guerra se hace más sostenible en la opinión pública. Todas estas consideraciones podrían contribuir a bajar el umbral de participación en conflictos armados, por efecto de la deshumanización del campo de batalla.

También debe considerarse como un riesgo de corto y mediano plazo para la humanidad, la escalada armamentística que las AA, necesariamente ge-

[39] Entre otras publicaciones que cubrieron esas declaraciones, puede consultarse: www.eltiempo.com/tecnosfera/novedades-tecnologia/putin-dice-que-quien-domine-la-inteligencia-artificial-gobernara-el-mundo-127256 [fecha de consulta: 5 de mayo de 2024].

[40] ONG Stop Killer Robot, *op. cit*.

[41] El 1 de noviembre del año 2023, la Primera Comisión de la Asamblea General de la ONU, adoptó la primera resolución sobre armas autónomas, destacando la "necesidad urgente de que la comunidad internacional aborde los desafíos y preocupaciones que plantean los sistemas de armas autónomas". El resultado de la votación en resolución L.56 fue ciento sesenta y cuatro Estados a favor, cinco en contra y ocho abstenciones.

neran y van a seguir generando en el futuro. La primera potencia que alcance la IAG, obtendrá una ventaja estratégica que puede provocar una ruptura singular en la historia de la humanidad. Por tal razón, las grandes potencias están hace décadas enfocadas en esa tecnología. Entienden que la paz y el equilibrio está en peligro si solo una potencia domina la IA. Y ciertamente que se oponen a la prohibición preventiva[42].

Como es de público conocimiento, la estrategia de defensa nacional de la administración Biden-Harris, de octubre de 2022, considera a China y, en diferente medida a Rusia, sus principales competidores estratégicos[43]. La Ley Chips y Ciencia, promulgada durante el año 2023, que libera cincuenta y dos mil setecientos millones de dólares en subvenciones a la industria de los semiconductores radicada en Estados Unidos, es solo uno de los muchos pasos que dicho país está dando para ganar la carrera de la IA. Ayudará a Estados Unidos a ganar "la competencia económica del siglo XXI", aseguró el presidente Joe Biden: "El futuro de la industria de los chips se hará en Estados Unidos", sostuvo[44].

IV. Conexión necesaria y esencial entre el derecho de la inteligencia artificial y la Cláusula Martens

En un mundo como el actual; convulsionado, violento y tensado geopolíticamente con el más singular de los hechos de la historia humana, a saber, la existencia de armas nucleares que permiten a la humanidad autodestruirse, la pregunta por el desarrollo ético de la tecnología, como en su momento observara Hans Jonas[45], es trascendental.

Las AA potencian aún más estos problemas. ¿Es sostenible un mundo con sistemas de armas nucleares tácticas y estratégicas autónomas? ¿Sería un mundo más seguro que el actual? ¿Vale la pena correr el riesgo? Ante tal escenario y horizonte de eventos, el actuar de la humanidad debería estar presidido por una heurística del temor (Hans Jonas). En este artículo no se indaga específicamente en aquel asunto ético. Su alcance es más limitado y orientado

[42] Véase Congressional Research Service (2023).

[43] Puede ser consultada en www.whitehouse.gov/briefing-room/statements-releases/2022/10/12/fact-sheet-the-biden-harris-administrations-national-security-strategy/ [fecha de consulta: 12 de mayo de 2024].

[44] En www.dw.com/es/biden-firma-ley-para-impulsar-chips-de-estados-unidos-y-competir-con-china/a-6276129 8 [fecha de consulta: 5 de mayo de 2024].

[45] Jonas (1995) p. 309.

a encontrar una solución normativa a la crítica situación actual de inexistencia de normativa expresa que regule el desarrollo y uso de AA, sobre todo de los completamente autónomos.

El marco de solución general a ese problema lo proporcionan los principios de lo que se ha denominado derecho de la inteligencia artificial[46] y la norma de derecho internacional humanitario conocida como Cláusula Martens.

1. Derecho de la inteligencia artificial (DIA)

El DIA, en las más diversas jurisdicciones, es un derecho en construcción[47]. En esta etapa de su desarrollo consta básicamente de principios que tienen por objetivo orientar y conducir el desarrollo de la IA en función de objetivos humanistas. En efecto, frente a la posibilidad de que la IA, junto con otros desarrollos tecnológicos, pudiera inducir un cambio singular en la evolución de la humanidad, en términos de proyectarla hacia la poshumanidad, el DIA se alza como una respuesta preventiva, inspirado éticamente en el principio de responsabilidad que, básicamente, ordena al hombre abstenerse de actuar frente a la incertidumbre de las consecuencias de sus acciones tecnológicas, máxime cuando la humanidad ya tiene el poder de destruirse a sí misma con el operar de la ciencia y la tecnología fáustica[48].

Ante la ausencia de normas tipo reglas expresas que regulen el desarrollo de la IA en función de objetivos humanistas, el DIA, a escala de principios, puede ser inferido de los tratados internacionales de derechos humanos[49].

Es evidente que la Carta de las Naciones Unidas y la Declaración Universal de Derechos Humanos (1948), pilares fundacionales del derecho internacional de los derechos humanos (DIDH), no pudieron haber previsto como una amenaza para los derechos humanos (y, en consecuencia, también para la humanidad), el desarrollo y uso de la IA sin controles éticos y en perspectiva transhumanista, pues en aquellos tiempos la IA solo estaba en la mente de genios como Alan Turing o John von Neumann.

La tecnología que sí rondaba como un fantasma en la Conferencia de San Francisco y en el mundo entero, era la energía atómica. Por primera vez en la

[46] Sobre esta construcción dogmática puede consultarse en detalle López (2020) y (2021).

[47] Un hito trascendental en su proceso de desarrollo lo constituirá la promulgación de la Ley de Inteligencia Artificial de la Unión Europea, prevista para el año 2025.

[48] Sobre la tradición fáustica y prometeica de la ciencia, véase Berman (2000), Haldane y Russell (2005), Martins (2012).

[49] Para una revisión del proceso de inducción de dichos principios, puede consultarse López (2020).

historia la humanidad tenía el poder de destruir apocalípticamente la vida sobre la faz de la tierra. Se inauguraba así la era de las tecnologías con poder de destrucción masiva. La arquitectura institucional de las Naciones Unidas fue concebida, también, por la urgencia de evitar el fin material de la especie humana producto de un conflicto bélico que desencadenare una conflagración nuclear suicida. Es tan cierto este antecedente, que la primera resolución que adoptó la Asamblea General de las Naciones Unidas (resolución 1 (I) de enero de 1946), no fue otra que la de crear la Comisión de Energía Atómica de las Naciones Unidas[50].

La energía nuclear es solo una especie dentro del género tecnologías de destrucción masiva. La IA, potencialmente, también es una especie del mencionado género de tecnologías. Entonces bien: si los Estados miembros de la ONU, todos los cuales son al menos suscriptores de la Carta Constitutiva, no actúan para evitar esas amenazas, comprometen su responsabilidad internacional, pues incumplen un deber asumido al suscribir la Carta de las Naciones: el deber de promover el respeto universal de los derechos humanos y las libertades fundamentales de todos, sin hacer distinción por motivos de raza, sexo, idioma o religión, y la efectividad de tales derechos y libertades. Vale decir, si un Estado no hace nada frente a las amenazas y riesgos que para los derechos humanos representa el desarrollo y utilización de las tecnologías, incumple el deber de promocionar el respecto de tales derechos asumido expresamente en el art. 55 c. de la Carta. En efecto, la obligación de promocionar[51] exige conductas activas. No hacer nada o no lo suficiente para enfrentar amenazas serias a los derechos humanos, conforman un arco de situaciones que van desde lo negligente hasta lo criminal.

Esos principios básicos que conforman los ejes del DIA, y que se denominan las tres leyes de la inteligencia artificial, se infieren de la Carta de las Naciones Unidas y de la Declaración Universal de Derechos Humanos[52].

Las tres leyes de la inteligencia artificial consideran que el respeto de los derechos humanos es condición de existencia y continuidad de la especie humana y de la humanidad. Su protección, por tanto, constituye un deber.

---

[50] Véase Comisión de Energía Atómica de las Naciones Unidas: www.un.org/es/sections/issues-depth/atomic-energy/index.html [fecha de consulta: 3 de marzo de 2021].

[51] Real Academia Española (1992) p. 1676.

[52] Se ha reducido el ejercicio dogmático solo a esos dos instrumentos por ser suficientes para la configuración de las tres leyes de la inteligencia artificial y porque, además, una revisión más extensa no resulta factible de exponer dados los límites formales que impone este artículo. Pero ciertamente que la ampliación del conjunto normativo de base de la inducción jurídica (por ejemplo, al conjunto de todos los tratados de derechos humanos), refuerza la configuración dogmática propuesta.

El poder transformador de la naturaleza humana que la IA posee, obliga a los Estados a actuar con cautela en la materia, evitando sus amenazas y también los atentados concretos y particulares a los derechos humanos. Es deber internacional de los Estados miembros de la ONU (todos suscriptores de su Carta Constitutiva y consecuencialmente también de la Declaración), evitar que el desarrollo y los usos de la IA amenacen o vulneren los derechos humanos, en particular la vida, la libertad y la igualdad, que constituyen el núcleo seminal del cual nace toda la especiación de derechos humanos constitutivos del *ethos* de la humanidad. Así las cosas, estos principios elementales (o leyes como también son llamados, rememorando las leyes de la robótica de Isaac Asimov[53]), son los siguientes:

- 1ª Ley de la IA: un sistema de IA no deberá hacer daño al ser humano individual o colectivamente considerado o a la humanidad, ya sea por acción, omisión o por cualquier otro tipo de conducta posible de ejecutarse por el sistema.
- 2ª Ley de la IA: un sistema de IA deberá siempre ceder el control de sus operaciones a los seres humanos, a excepción que dicha conducta condujere a una vulneración de la 1ª Ley.
- 3ª Ley de la IA: un sistema de IA debe proteger su propia existencia en la medida en que esta protección no vulnere la 1ª y 2ª Ley.

En efecto, si por acción u omisión los Estados permiten que el desarrollo de la IA afecte o dañe de algún modo a las personas o a la humanidad, se incurrirá, por cierto, en una vulneración de derechos humanos que, eventual mente, puede activar los sistemas de protección internacional.

Asimismo, en relación con la segunda ley de la IA, la autonomía de los sistemas de IA no puede llegar al punto que los seres humanos pierdan el control sobre ellos, pues, en tal caso, se queda a merced de sus decisiones y la humanidad pierde la conducción de su destino. Sin embargo, pueden darse ciertos casos en que sea teóricamente posible cederles el control a los sistemas de inteligencia artificial, pero ciertamente que bajo condiciones muy limitadas de tiempo y espacio. Las hipótesis fácticas de cesión de control deben encaminarse hacia todas aquellas circunstancias en que los seres humanos pretendan controlar los sistemas de inteligencia artificial para dañar al hombre y a la humanidad, pues el uso de la IA debe siempre ser democrático y en concordancia con los derechos humanos. Esta segunda ley se orienta también, por tanto, ha evitar los golpes de Estado y las tomas de control autoritario de las sociedades, pues esas acciones dañan al hombre y a la humanidad. Pero, aun así,

[53] Por cierto, que las leyes de la robótica de Isaac Asimov en ningún caso son una excentricidad, sino que el punto de partida de toda reflexión normativa sobre la inteligencia artificial, como consta en Parlamento Europeo (2017).

su aplicación debe ser muy estricta, calculada y en casos específicos, pues cederle el control a una IA, es ceder humanidad.

Por último, la tercera ley de la IA es una ley de salvaguarda. Los sistemas de IA deben proteger su propia existencia, es decir, deben desarrollar instinto de supervivencia, pero también deben estar entrenados/programados para la autodesconexión altruista (suicidio cibernético), cuando la proyección de su actuar, eventualmente, representare, de algún modo, un riesgo vital al hombre o a la humanidad.

Estos principios, que operan como sistema de reconocimiento, están implícitos en el núcleo esencial del DIDH y configuran constitucionalmente la normatividad más general del derecho de la inteligencia artificial, a la cual debe someterse para adquirir juridicidad (y legitimidad) las reglamentaciones particulares de carácter nacional e internacional que, en materia de desarrollo de la IA, los Estados en la actualidad posean o promulgaren en el futuro. Los Estados deben, por tanto, adecuar y ajustar sus conductas a dichas leyes o principios.

A la luz de los principios de DIA, las AA, no podrían diseñarse, fabricarse o usarse, puesto que tienen por objetivo destruir y dañar al ser humano. Pero si aún así no fuera posible evitar el desarrollo y uso de las AA, queda, también, siempre a disposición de las personas y los Estados, la Cláusula Martens para que, en combinación con el DIA, orienten del modo más benigno para la humanidad el desarrollo de las AA.

La Cláusula Martens es una pieza esencial del derecho internacional humanitario, que, por cierto, se conecta necesariamente con el DIA, cuya fuente primaria es el DIDH. La conexión necesaria se produce porque ambos derechos dimanan de un mismo corpus ético jurídico: el principio de humanidad[54], aunque, en el caso de la Cláusula Martens, trátase de un antecedente más antiguo que el DIDH, pues es considerada una de las piedras fundacionales del derecho internacional humanitario, que es anterior al DIDH y, además, codifica normas consuetudinarias sobre la guerra preexistentes a la codificación. En este sentido la Cláusula Martens refleja una:

> "coexistencia entre el DIH consuetudinario y el DIH convencional que se soluciona a favor de la opción que mejor responda a la finalidad primaria del DIH, que no es otra que la protección de las víctimas de los conflictos armados"[55].

[54] Salmon (2004) p. 70.

[55] Una fundamentación al respecto puede consultarse en Salmon (2004) p. 49.

2. Cláusula Martens

Aparece por primera vez en el preámbulo del (II) Convenio de La Haya de 1899 relativo a las leyes y costumbres de la guerra terrestre. Su formulación fue la siguiente:

> "Mientras que se forma un Código más completo de las leyes de la guerra las Altas Partes Contratantes juzgan oportuno declarar que en los casos no comprendidos en las disposiciones reglamentarias adoptadas por ellas las poblaciones y los beligerantes permanecen bajo la garantía y el régimen de los principios del Derecho de Gentes preconizados por los usos establecidos entre las naciones civilizadas, por las leyes de la humanidad y por las exigencias de la conciencia pública"[56].

Desde 1925 en adelante, un número significativo de tratados que regulan las prohibiciones de ciertas armas, también la incluyen. Así puede ser advertida en: los preámbulos del Protocolo de Gas de Ginebra de 1925, la Convención sobre las Armas Biológicas de 1972, Convención sobre Armas Convencionales de 1980, Convención sobre la Prohibición de Minas Antipersonales de 1997, Convención sobre Municiones en Racimo de 2008 y el Tratado sobre Prohibición de las Armas Nucleares de 2017[57].

La Cláusula Martens se ha incorporado también a los cuatro convenios de Ginebra en los artículos que regulan los efectos de las denuncias de los convenios por alguno de los Estados parte[58]. En el Protocolo Adicional I de 1977 de los convenios de Ginebra, se plasmó en los siguientes términos:

> "En los casos que no previstos en este Protocolo u otros acuerdos internacionales, civiles y combatientes permanecen bajo la protección y la autoridad del derecho internacional, derivado de la costumbre establecida, de los principios de la humanidad y de los dictados de conciencia pública"[59].

---

[56] Sobre los orígenes e historia de la cláusula, así como respecto de las diversas interpretaciones y aplicaciones de la misma, existe una abundante bibliografía. Entre otras, se ha tenido presente: Ticehurts (1997); Canesse (2000); Salmon (2004); Balitzki (2009); Ben-Naftali (2011); Rodríguez-Villasante y López (2017); Melzer (2019); Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020); Ivanenko (2022).

[57] Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020) p. 12.

[58] *Op. cit*. p. 20.

[59] Una versión del Protocolo puede consultarse en Comité Internacional de la Cruz Roja (1977).

Elizabeth Salmon ha señalado que el carácter general o consuetudinario del DIH se manifiesta en sus orígenes mismos a través de la Cláusula Martens, pues el texto de la cláusula no admite duda que frente a la ausencia de normas expresamente positivizadas, los beligerantes permanecen bajo la salvaguardia y el régimen de los principios del derecho de gentes, tales como resultan de los usos establecidos entre las naciones civilizadas, de las leyes de humanidad y de las exigencias de la conciencia pública[60].

La Cláusula Martens evidenciaría así que los Estados codificaban normas consuetudinarias ya existentes, fundadas en principios generales que mantienen su validez fuera del contexto convencional. Contemporáneamente, agrega Elizabeth Salmon:

> "esto se reafirma en que las normas del DIH son cada vez más consideradas como consuetudinarias y, en tanto tales, como normas que deben ser aplicadas por todos los estados en una base de universalidad. de armas emergentes, incluidas las armas completamente autónomas"[61].

Pues bien, más allá de la discusión dogmática sobre el valor normativo de las declaraciones contenidas en los preámbulos de los instrumentos internacionales, donde consta la Cláusula Martens, o el valor normativo de la misma cuando está contenida en el articulado o cuerpo principal de los instrumentos internacionales, en este artículo se acepta el siguiente tópico argumentativo: la Cláusula Martens debe ser considerada, a lo menos, como elemento de interpretación del derecho internacional.

Aquella es la tesis de Antonio Canesse, quien, adoptando un enfoque intermedio, entre los que le restan todo valor normativo, por una parte, y los que la consideran fuente formal del DIH, propone tratar los principios de la humanidad y los dictados de la conciencia pública como una guía fundamental para la interpretación del derecho internacional[62]. De lo cual se deduce:

> "la cláusula Martens proporciona los factores que los Estados deben tener en cuenta a medida que se acercan a la tecnología de armas emergentes, incluidas las armas completamente autónomas"[63].

Un importante precedente de invocación de aplicación de la Cláusula Martens en casos de tecnologías emergentes no reguladas lo constituyó, en la

[60] Salmon (2004) pp. 32-33.

[61] *Op. cit*. p. 34.

[62] Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020) p. 17. Fuente primaria en Canesse (2000) p. 212.

[63] Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020) p. 17.

década de 1990, la discusión sobre la prohibición preventiva de los láseres cegadores. Dichas armas, como así lo señalaron sus críticos: "plantean preocupa ciones bajo los principios de la humanidad y los dictados de la conciencia"[64]. Fue así como en la Primera Conferencia de Revisión de la Convención sobre Ciertas Armas Convencionales (CCAC), sostenida en 1996, representantes de la ONU y de la sociedad civil caracterizaron los láseres cegadores como: "inhumanos, aborrecibles a la conciencia de la humanidad e inaceptables en el mundo moderno"[65]. El representante de Chile en la referida Conferencia de Revisión, expresó su esperanza de que el organismo:

> "podría establecer pautas de acción preventiva para prohibir el desarrollo de tecnologías inhumanas y, por tanto, evitar la necesidad de remediar la miseria que podrían causar"[66].

Como resultado de la aludida discusión normativa, el Protocolo sobre Armas Láseres Cegadoras del CCAC (Protocolo IV CCW), adoptado el 13 de octubre de 1995, que entró en vigor el 30 de julio de 1998, en su art. 1, expresamente prohíbe el uso de láseres cegadores:

> "Queda prohibido emplear arnas láser específicamente concebidas, como única o una más de sus funciones de combate, para causar ceguera permanente a la vista no amplificada, es decir, al ojo descubierto o al ojo provisto de dispositivos correctores de IA vista. Las Altas Partes Contratantes no transferirán armas de esta índole a ningún Estado ni a ninguna entidad no estatal"[67].

Y si los láseres cegadores fueron prohibidos en virtud de los principios de humanidad y los dictados de la conciencia pública enarbolados por la Cláusula Martens, con mayor razón entonces, deberían prohibirse, al menos por el momento, las AA completamente autónomas, que son mucho más dañinas que los láseres cegadores.

Y es que las AA, imposible soslayarlo, no tienen juicio ni comprensión humana. Vale decir, son máquinas y, por tanto, carecen de agencia moral. Y al respecto, que duda podría caber: decidir sobre la vida y la muerte, no puede ser un acto algorítmico, desprovisto de un juicio ético, que permita evaluar situacionalmente todos los aspectos humanos involucrados en un conflicto bélico. Al carecer de agencia moral, difícilmente las AA podrán proteger a civiles y

[64] Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020) p. 18.

[65] *Op. cit*. p. 19.

[66] *Ibid*.

[67] Véase Comité Internacional de la Cruz Roja (1977).

combatientes en función de los principios de la humanidad y de los dictados de conciencia pública. Es decir, no podrían cumplir con los mandatos de protección de la Cláusula Martens, derivados de los principios de la humanidad y los dictados de la conciencia.

2.1. Los principios de humanidad y las AA

Los principios de humanidad obligan a

1) tratar a los demás con humanidad y
2) mostrar respeto por la vida y la dignidad humanas[68].

Brevemente convengamos que los principios de humanidad son aquellas normas que orientan el actuar humano con sensibilidad y compasión de las desgracias y situaciones penosas de los congéneres[69]. El derecho a ser tratado con humanidad es un pilar fundacional del DIH y del DIDH. El derecho y la obligación correlativa al trato humanitario está reconocida, por ejemplo, en los convenios de Ginebra, manuales militares, jurisprudencia internacional y en el Pacto Internacional de Derechos Civiles y Políticos.

El trato con humanidad supone, necesaria y forzosamente, el despliegue de la compasión y la ejecución de juicios éticos y legales.

El actuar compasivo debe orientar la conducta de los beligerantes en el sentido:

> "que la captura es preferible a herir a un enemigo y herirlo es mejor que matarlo; que los no combatientes se salvarán en la medida de lo posible; que las heridas infringidas sean lo más livianas posible, para que los heridos puedan ser tratados y curados y que las heridas causen el menor dolor posible"[70].

Si la compasión impulsa casi de manera instintiva el actuar humano, el juicio ético y legal entrega las herramientas conceptuales para plasmar conductualmente el mandato normativo. El juicio ético o legal supone la puesta en marcha no solo del juicio lógico formal, sino que, también, informal que se obtiene a través de la experiencia y que puede ser englobado en términos como "sentido común" o "reglas de la experiencia". Es la combinación de estos dos tipos de lógicas o razonamientos que el sujeto actúa y toma decisiones que ponderan y balancean de manera dinámica información pasada, presente-situacional y futura, todo, en función de no vulnerar el mandato de respetar los principios de humanidad.

---

[68] Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020) p. 21.

[69] 'Humanidad', acepción 5 en Real Academia Española (1992).

[70] Human Rights Watch & International Human Rigths Clinic Harvard Law SchooL (2020) p. 22. Fuente primaria: Pictec (1985) p. 62.

Entonces bien, si la compasión y la capacidad de hacer juicios éticos y legales, son características privativas del ser humano, las AA no estarían, ontológicamente, capacitadas para hacerlo. Luego, las posibilidades para que pudieran cumplir con la Cláusula Martens serían extremadamente complejas, por no decir, imposibles.

El dolor, el sufrimiento, las emociones o los sentimientos morales, hasta ahora, por lo que sabemos, solo son fenómenos humanos. Incluso, más esencialmente humanos, si hablamos de las emociones, sobre todo a la luz de las investigaciones del neurocientífico Antonio Damasio, quien, en *El error de Descartes*, propone un cambio de paradigma desde el "pienso, luego existo", al "siento o me emociono, luego existo"[71]. Los seres humanos, en síntesis, somos seres fundamentalmente emocionales.

Pero la IA no se emociona ni siente. Es verdad que puede producir en muchos campos conductuales *outputs* similares a los humanos; pero un AA enfrentada a la decisión de dar o no de baja a un blanco humano, no lo hará inspirado en los principios de humanidad, pues no siente compasión, ni razona ética o legalmente como un humano.

No puede perderse de vista, asimismo, que, en etapas muy avanzadas de desarrollo de la IA, por ejemplo, cuando eventualmente ya se esté cerca de la IAG, las AA serán tan inteligentes que no puede descartarse *a priori* que, aunque no tengan agencia moral, sin embargo, podrían actuar estadísticamente de manera mucho más alineada a los principios de la humanidad, que los propios humanos.

La pregunta que surge entonces es: ¿en tal estado del desarrollo del arte de la IA, deberían prohibirse las armas por completo autónomas, aun cuando fuere indiscutible que podrían desempeñarse, comparadas estadísticamente con los humanos, con un mejor y más pleno respeto al principio de humanidad?

La pregunta pone en cuestión nuestras certezas ético-jurídicas más profundas. Porque, aun cuando las AA fueren más "compasivas" y "justas" que los humanos, juzgados estadísticamente por sus resultados, desde el momento que la humanidad le entrega el control de la vida y de la muerte a la IA, empieza a deslizarse hacia una especie de desintegración ética, ya que renuncia a ejercer lo que le es propio por naturaleza: ser un agente moral, es decir, un sujeto que es responsable de sus acciones porque tiene la capacidad para juzgar y distinguir entre lo correcto y lo incorrecto.

Desprovistos de agencia moral y en un contexto de IAG, los hombres, al entregarle el poder de decidir entre la vida y muerte a las AA, literalmente estarán renunciando a ser humanos. Y de ahí adelante, cualquier cosa puede

[71] Damasio (2010), (2018).

suceder. El juego podría cambiar aún más de forma más radical, cuando la IA en un futuro alcance la fase de IAG y, más aún, de SIA. No puede descartarse que en esa etapa emerja un nuevo tipo de inteligencia materializada en un ser ciber orgánico o virtual. Vale la pena preguntarse si la humanidad quiere eso. El debate ético-jurídico sobre las AA tiene la gran virtud de reactualizar el debate de fondo: ¿para qué queremos la IA?

2.2. Las exigencias de la conciencia pública y los SAA

La Cláusula Martens previene que ante la ausencia de normas expresas que regulen el uso de armas y el trato a los combatientes, estos permanecen bajo la protección del derecho internacional, derivados de la costumbre establecida, de los principios de humanidad y de las exigencias de la conciencia pública.

Las exigencias de la conciencia pública:

> "se refieren a pautas morales que dan forma a las acciones de los Estados y los individuos. El uso del término 'conciencia' indica que los dictados se basan en un sentido de moralidad, un conocimiento de los que es correcto o incorrecto"[72].

Una aproximación al estudio de las fuentes de la conciencia pública, basado en el trabajo que sobre el tema ha realizado Theodor Meron, puede realizarse a partir de la revisión de la opinión pública y las opiniones de los gobiernos[73].

La revisión de las principales fuentes de consulta de la opinión pública, tales como: encuestas, opiniones de los expertos, informes de centros de estudios especializados; de ONG y organizaciones internacionales, así como declaraciones de líderes mundiales, expertos en ciencia y tecnología y de las principales industrias tecnológicas, muestran que la sociedad mundial siente temor y no apoya la aparición de las AA completamente autónomas[74].

Por su parte, el discurso oficial de diversos gobiernos y países del mundo, con mayor o menor fuerza, han expresado su opinión en el sentido que el desarrollo de las AA completamente autónomas es inaceptable. Desde el año 2007 se vienen registrando los esfuerzos de la humanidad por regular las AA en fun-

[72] Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020) p. 31.

[73] *Op. cit*. p. 31. Fuente primaria: Meron (2000).

[74] Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020) pp. 31-41. Para una revisión de las principales ONG y demás instituciones de la sociedad civil que apoyan la prohibición de los SAA completamente autónomas, puede consultarse Stop Killer Robots, cual es una de las principales organizaciones del campo: www.stopkillerrobots.org/es/ [fecha de consulta: 12 de mayo de 2024].

ción del lograr un tratado internacional en la materia. Los dos últimos hitos relevantes en la materia[75], fueron los siguientes:

- El 1 de noviembre de 2023, la Primera Comisión de la Asamblea General de la ONU adoptó el primera resolución sobre armas autónomas, destacando la "necesidad urgente de que la comunidad internacional aborde los desafíos y preocupaciones que plantean los sistemas de armas autónomas". El resultado de la votación en resolución L.56 Hubo ciento sesenta y cuatro Estados a favor, cinco en contra y ocho abstenciones.
- 5 de octubre de 2023 el secretario general de las Naciones Unidas, António Guterres, y la presidenta del Comité Internacional de la Cruz Roja, Mirjana Spoljaric, pidieron a los Estados que inicien negociaciones sobre un nuevo instrumento jurídicamente vinculante para establecer prohibiciones y restricciones claras sobre sistemas de armas autónomos y concluir dichas negociaciones para 2026.

La lista de hitos es larga y permite afirmar, con bastante evidencia empírica, que, en general, los gobiernos y organismos internacionales están avanzando conversaciones sobre la regulación de este tipo de armas. En general, están conscientes de los peligros y desafíos que plantean. *Prima facie*, y en dominio de las declaraciones, se muestran contrarios a su desarrollo. Al menos respecto de las AA completamente autónomos.

Conclusiones

El desarrollo de la IA sin controles éticos y jurídicos es un riesgo existencial para la humanidad. En el corto plazo, señales de estos riesgos pueden ser distinguidos en las continuas vulneraciones a ciertos derechos fundamentales que, mediante el uso de la IA, están aconteciendo alrededor del mundo. Sin duda alguna que la máxima vulneración a los derechos fundamentales mediante el uso de IA se produce cuando dicha tecnología se pone al servicio de los conflictos armados. Ya no se trata de que un sistema de IA discrimine a las personas en el proceso de otorgamiento de un crédito o en el de selección para un puesto de trabajo. En el caso de las AA, la IA se pone al servicio de destruir el más fundamental de los derechos fundamentales: el derecho a la vida.

El epítome de esta situación lo constituyen los sistemas de armas completamente autónomas, las cuales, una vez activadas, pueden seleccionar y eje-

[75] Un completo cronograma con la historia de discusión mundial sobre SAA, que arranca el año 2007, y que ha guiado esta sección del artículo, puede ser consultado Stop Killer Robots: www.stopkillerrobots.org/es/la-historia-hasta-ahora/ [fecha de consulta: 12 de mayo de 2024].

cutarse sin necesidad de intervención humana. Según información recogida de fuentes abiertas, estas clases de armas todavía no existirían, pero se sabe, a través de las mismas fuentes, que las grandes potencias están trabajando en esa dirección, preparándose para una situación de teoría de juego con hipótesis de "juicio final" o derrota estratégica.

Y relacionado con lo anterior, otro aspecto a considerar, por tanto, en el contexto de un mundo con armas atómicas, es el uso de la IA para gobernar esos instrumentos bélicos. ¿Sería compatible con el principio de responsabilidad, la construcción de esos sistemas atómicos?

El desarrollo de las AA, por cierto, está en directa relación con los avances de la IA. Las condiciones para que surjan las armas completamente autónomas dependerá de cuan cerca esté la humanidad de conquistar la IAG. Llegados a ese punto de la historia, la situación será muy crítica, pues como se proyecta, la IAG tendrá capacidad de autoaprendizaje ilimitado, lo cual podría conducirla a una curva exponencial que desemboque en el nacimiento de la SIA. Un mundo con SIA es un mundo de incertidumbre. No hay claridad respecto a cómo se comportaría esa SIA y si, por ejemplo, se alinearía necesariamente con los valores y objetivos humanos, como ha sido teorizado por Nick Bostrom.

El problema de fondo sigue siendo resolver éticamente la pregunta sobre el "para qué queremos la IA". Tanto las Naciones Unidas, con su programa IA para el Bien, como la Unión Europea, han respondido esa pregunta apostando en términos generales por un desarrollo de la IA centrada en el ser humano, a saber: una IA puesta al servicio de los objetivos de desarrollo sustentable de la humanidad.

Sin embargo, las grandes potencias, y a juzgar por sus acciones, no estarían en la página de, por el momento, regular el desarrollo de las AA. Prueba de ello, es que el proyecto de ley de IA de la Unión Europea expresamente las excluye de sus regulaciones.

El escenario, así, es un tanto desalentador. Y ante los peligros e inminentes riesgos que las AA implican, sobre todo las completamente autónomas, la humanidad solo tiene como defensa el derecho de la inteligencia artificial (DIA) y, más en particular, la Cláusula Martens.

El DIA impone a los Estados la obligación de orientar el desarrollo de la IA en función de la reproducción material y cultural de la humanidad, lo cual resulta contradictorio con los fines de las AA, que están diseñados para dañar al ser humano en los conflictos armados. La Cláusula Martens, a su turno, como norma jurídica de derecho internacional humanitario, tiene un campo de aplicación más particular que el DIA, toda vez que está diseñada específicamente para los conflictos armados. Entre el DIA y la Cláusula Martens, por cierto, existe una conexión necesaria, signada por la circunstancia de derivar ambos *corpus* del principio de humanidad.

Si bien es cierto que abogar por la prohibición preventiva del desarrollo de las AA, especialmente las completamente autónomas, es una política ética correcta, también hay que tener una aproximación pragmática al problema y, si no es posible la prohibición preventiva, abogar por la regulación más humanitariamente posible de sus usos, también constituirá una decisión correcta.

Bibliografía

Australian Defence Force (2020): *Concept for robotic and autonomous systems*. Version 1.0 Reference: DPN BN9939583 (Defence Publishing Services).

Balitzki, Anja (2009): *The Martens Clause. Origin of a new source on International Law?* (Bremen, Universidad de Bremen).

Ben-Naftali, Oma (2011): *International Humanitarian Law and International Human Rights Law* (Oxford, Oxford Univsersity Press).

Berman, Marshall (2000): *Todo lo sólido se desvanece en el aire. La experiencia de la modernidad* (Buenos Aires, Siglo Veintiuno Editores).

Berzal, Fernando (2018): *Redes neuronales & deep learning* (Granada, Independently published). Véase deep-learning.ikor.org/ [fecha de consulta: 12 de abril de 2024].

Bostrom, Nick (2016): *Superinteligencia. Caminos, peligros, estrategias*. Traducción Marcos Alonso (Madrid, Teel Editorial).

Boulanin, Vicent; Davison, Neil; Goussac, Netta & Peldan Carlsoon, Moa (2020): *Limits on autonomy in weapon systems* (Geneva, International Committee of the Red Cross).

Cáceres Erraez, Cristóbal Alejandro (2022): *Sistema de prueba de vida para login biométrico usando modelos de machine learning*. Trabajo de titulación, previo a la obtención del título de Ingeniero en Sistemas e Informática (Ecuador, Universidad de las Fuerzas Armadas).

Canesse, Antonio (2000): "The Martens Clause: Half a Loaf or Simply Pie in the Sky", *European Journal of International Law* vol. 11 Issue 1: pp 187-216.

Chakour, Sam (2022): *Artificial Intelligence (AI) For Weapons Systems* (Virginia/ Maryland, Defense Systems Information Analysis Center-DSIAC).

Comité Internacional de la Cruz Roja (1977): Protocolo I adicional a los convenios de Ginebra de 1949 relativo a la protección de las víctimas de los conflictos armados internacionales, 1977. Disponible en www.icrc.org/es/document/protocolo-i-adicional-convenios-ginebra-1949-proteccion-victimas-conflictos-armados-internacionales-1977 [fecha de consulta: 7 de mayo de 2024].

Congressional Research Service (2023): *International Discussions Concerning Lethal Autonomous Weapons Systems* (Washington D.C., CRC).

Congressional Research Service (2024): *Defense: U.S. Policy on Lethal Autonoumous Weapon Systems* (Washington D.C., CRC).

Consejo para la Transparencia (2020). *Cuaderno de Trabajo N° 17. Transparencia Algorítmica. Buenas prácticas y estándares de transparencia en el proceso de toma de decisiones automatizadas* (Santiago, Ediciones Consejo para la Transparencia).

Cotino Hueso, Lorenzo y Castellanos Claramunt, Jorge (eds.) (2022): *Transparencia y explicabilidad de la inteligencia artificial* (Valencia, Tirant lo Blanch).

Council of the European Union (2022).

Damasio, Antonio (2010): *El error de descartes. La emoción, la razón y el cerebro humano*. Traducción Joandomènec Ros (Madrid, Editorial Crítica).

Damasio, Antonio (2018): *En busca de Spinoza: neurobiología de la emoción y los sentimientos*. Traducción Joandomènec Ros (Barcelona, Editorial Booket).

Department of Defense (2012).

Department of Defense (2023): *DoD Directive 3000.09 Autonomy in Weapons Systems* (Washington DC.).

European Commission (2021): Proposal for a Regulation of the European Parlament and the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligent Act) and Amending Certain Legislative Acts.

European Parlament (2024): Provisional Agreement Resulting From Interintitutional Negotiationes (2024).

Evans, Tyler D. (2013): "At war with the robots: autonomous weapon systems and the Martens Clause", *Hofstra Law Review* vol. 41 Issue 3: pp. 697-733.

Executive Office of the President & National Science and Technology Council Committee on Technology (2016*): Preparing For The Future of Artificial Intelligence* (Washington DC.).

Group of Govermental Experts on Emerging Tecnologies in the Area of Lethal Auutonomous Weapons Systems (2024). Disponible en chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://reachingcriticalwill.org/images/documents/Disar mament-fora/ccw/2021/gge/documents/draft-report-final.pdf [fecha de consulta: 15 de abril de 2024].

Guzmán Lucio, Vicente Gerardo (2019): "¿Por qué aprender Machine Learning?". Disponible en https://medium.com/ia-latam/por-qu%C3%A9-aprender-machine-learning-d001b28c 7361 [fecha de consulta: 13 de mayo de 2024].

Hai Stanford University Human-Centered Artificial Intelligence (2024): *Artificial Intelligent Index Report 2024* (Stanford).

Haldane, John B.S. y Russell, Bertrand (2005): *Dédalo e Ícaro: el futuro de la ciencia*, (Oviedo, KRK ediciones).

Holland, Arthur (2020): *The black box, unlocked. Predictability and undersdability in military AI* (Geneve, UNIDIR).

Human Rights Watch & International Human Rigths Clinic Harvard Law School (2015): *Precedent for Preemption: The Ban on Blinding Lasers as a Model for a Killer Robots Prohibition*. Disponible en www.hrw.org/news/2015/11/08/precedent-preemption-ban-blinding-lasers-model-killer-robots-prohibition [fecha de consulta: 12 de abril de 2024].

Human Rights Watch & International Human Rigths Clinic Harvard Law School (2020): *Atender la llamada. Un imperativo moral y legal para prohibir los robots asesinos*. Disponible en www.hrw.org/es/report/2018/08/21/atender-la-llamada/un-imperativo-moral-y-legal-prohibir-los-robots-asesinos [fecha de consulta: 12 de abril de 2024].

Ivanenko, Vitaly (2022): "The origins, causes and enduring significance of the Martens Clause: A view from Russia", *International Review of the Red Cross* No. 104: pp. 1708-1724.

Jonas, Hans (1995): *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*. Traducción Javier Fernández Retenaga (Madrid, Herder).

Kurzweil, Ray (2012): *La singularidad está cerca. Cuando los humanos transcendamos la biología*. Traducción de Carlos García Hernández (Berlín, Lola Books GbR. Ebook).

López de Mántaras Badía, Ramón y Meseguer González, Pedro (2017*): Inteligencia artificial* (Madrid, Los Libros de la Catarata, *ebook*).

López Oneto, Marcos (2020): *Fundamentos para un derecho de la inteligencia artificial. ¿Queremos seguir siendo humanos?* (Valencia, Tirant lo Blanch).

López Oneto, Marcos (2021): *Derecho al futuro* (Valencia, Tirant lo Blanch).

Luhman, Niklas (2007): *La sociedad de la sociedad*. Traducción Javier Torres Nafarrate (Ciudad de México, Herder).

Martinage, Robert (2014): *Toward a new offset strategy. Exploiting U.S. long-term advantages to restore U.S. global power projection capability* (Washington DC, CSBA).

Martins, Herminio (2012): *Experimentum humanum: civilizacao tecnológica e condicao humana* (Belo Horizonte, Fino Traco).

Melzer, Nils (2019): *Derecho internacional humanitario* (Ginebra, CIRC).

OTAN (2022): "Summary of Nato's Autonomy Implementation Plan". Disponible en www.nato.int/cps/en/natohq/official_texts_208376.htm [fecha de consulta: 13 de mayo de 2024].

Parlamento Europeo (2017): P8_TA(2017)0051. Normas de derecho civil sobre robótica. Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2013(INL)) (2015). Disponible en www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//ES [fecha de consulta: 26 de junio de 2019].

PICTET, Jean (1985): *Development and Principles of International Humanitarian Law* (Ginebra, Martinus Nijoff and Henry Dunant Institute).

REAL ACADEMIA ESPAÑOLA (1992): *Diccionario de la lengua española* (Madrid, Editorial Espasa Calpe).

RODRÍGUEZ-VILLASANTE Y PRIETO, José Luis y LÓPEZ SÁNCHEZ Joaquin (coords.) (2017): *Derecho internacional humanitario* (Valencia, Tirant lo Blanch, 3ª edición).

RUSSELL, Stuart y NORVIG, Peter (2008): *Inteligencia artificial. Un enfoque moderno. Madrid*. Traducción Juan Manuel Corchado Rodríguez *et al*. (Madrid, Pearson/Prentice Hall, 2ª edición).

SALMON, Elizabeth (2004): *Introducción al derecho internacional humanitario* (Lima, Pontificia Universidad Católica del Perú/CIRC).

SHAUB, Gary Jr. & WENZEL KRISTOFFERSEN Jens (2017): *In, On, o Out of the Loop? Denmark and Autonomous Weapon Systems* (Copenhagen, Centre for Military Studies University of Copenhagen).

TADDEO, Mariarosaria & BLANCHARD Alexander (2021): "A comparative analisis of the definitions of autonoumos weapons systems", *Science and Engineering Ethics* vol. 28 number 37. Disponible en https://doi.org/10.1007/s11948-022-00392-3 [fecha de consulta: 14 de abril de 2024].

TICEHURST, Rupert (1997): "The Martens Clause and the Laws of Armed Conflict", *International Review of the Red Cross*, No. 317.

UNIDIR (2018): *The weaponization of increasingly autonomous technologies: concerns, characteristics and definitional approaches.* Disponible en https://unidir.org/publication/the-weaponization-of-increasingly-autonomous-technologies-concerns-characteristics-and-definitional-approaches/ [fecha de consulta: 1 de mayo de 2024].

UNIDIR (2022): *Proposals related to emerging technologies in the área of lethal autonomous weapons systems.* Disponible en https://unidir.org/publication/proposals-related-to-emerging-technologies-in-the-area-of-lethal-autonomous-weapons-systems-a-resource-paper-updated/ [fecha de consulta: 7 de abril de 2024].

UNITED NATIONS OFFICE FOR DISARMAMENT AFFAIRS (2018): "Lethal Autonomous Weapon Systems (LAWS)". Disponible en https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/ [fecha de consulta:15 de abril de 2024].

WORLD INTELECTUAL PROPERTY ORGANIZATION (2019): *Artificial Intelligence*. Disponible en https://tind.wipo.int/record/29084?v=pdf [fecha de consulta: 12 de abril de 2024].

Normas

Convención sobre prohibiciones o restricciones del empleo de ciertas armas convencionales que pueden considerarse excesivamente nocivas o de efectos indiscriminados

(1980), ratificado por Chile el día 18 de julio del año 2003 y promulgado el 8 de junio del año 2004.