Contextualización

En este apartado, la autora proporcionará una contextualización del debate en torno a la IA y los SAA, ofreciendo una guía conceptual de los términos y tecnologías que conforman el objeto de estudio en este artículo. Cabe precisar que este es un artículo *legal* que no pretende autoridad alguna en los aspectos científicos o meramente técnicos de dichos conceptos, ya que la mayoría de ellos ni siquiera están definidos de manera conclusiva por los principales expertos en sus campos, y porque la velocidad exponencial con la que evolucionan hace que sean de la naturaleza más fluida.

Comencemos con un examen semántico de las nociones relevantes para este estudio:

La palabra *artificial*, derivada del latín *artificialis*, es equivalente a los términos "facticio", "sintético", "falso", "antinatural"; una cosa que es artificial está hecha por el hombre o es producida por humanos, generalmente para que se parezca a algo que sí es natural.[3]

Por otro lado, la palabra *inteligencia* es más difícil de definir. La inteligencia se explica en términos generales como la "capacidad de comprender", "la capacidad de resolver

[3] Voz: "Artificial, adj.", *Diccionario de la lengua española* (*DLE*), Real Academia Española (RAE), disponible en *https://dle.rae.es/artificial?m=form*, ingresado el 2 de julio de 2020.

problemas" o simplemente como "conocimiento o el acto de entender".[4] Sin embargo, esta noción todavía es cuestionada entre los psicólogos, ya que algunos de ellos la relacionan con el intelecto humano y, por lo tanto, se limitan al cerebro cognitivo. Así pues, esta noción estaría tradicionalmente ligada a la condición *humana*.

Para Stephen Hawking, "la inteligencia es la capacidad de adaptarse al cambio".[5] Esta cita nos sirve para ligar el próximo concepto *compuesto*, la IA.

No hay consenso sobre una definición universal de este concepto, ya que los principales expertos en la materia la afinan y desafían constantemente.[6] El único aspecto definitivo y sobre el cual existe consenso es que la IA es una tecnología *muy* disruptiva.

Como campo de investigación, el nombre de IA se decidió durante un taller en la Universidad de Dartmouth en 1956, donde un grupo de científicos destacados se reunieron durante ocho semanas, llevando a cabo una lluvia de ideas sobre la concepción sobre el concepto de "máquinas que piensan".[7]

---

4 Voz: "Inteligencia", *DLE*, RAE, disponible en *https://dle.rae.es/inteligencia?m=form*, ingresado el 2 de julio de 2020.

5 Stephen Hawking declaró esto en su graduación de la Universidad de Oxford, "Professor Stephen Hawking: 13 of his most Inspirational Quotes", *The Telegraph*, Londres, 8 de enero de 2016, disponible en *www.telegraph.co.uk/news/science/stephen-hawking/12088816/Professor-Stephen-Hawking-13-of-hismostinspirational-quotes.html*, ingresado el 2 de julio de 2020.

6 Scherer, Matthew U., "Regulating Artificial Intelligence Systems, Risks, Challenges, Competencies and Strategies", *Harv. J. L. & Tech*, vol. 29, núm. 2, pp. 353, 359-362, 2016 (en adelante: Scherer, "Regulating AI").

7 Sileno, Giovanni, *History of AI, Current Trends, Prospective Trajectories,* Asser Institute, Winter Academy on Artificial Intelligence and International Law, 2021 (en adelante: Sileno, *History of AI*); Sileno menciona el grupo de 20 científicos e ingenieros notables, que estaban en el taller de Dartmouth en 1956, entre ellos: John McCarty (lenguaje LISP, cálculo de situaciones, lógicas no monótonas) Marvin Minsky (marcos, perceptrón, sociedad de mentes), Herbert Simon (teórico de la lógica, solucionador de problemas generales, racionalidad limitada), Allen Newell (teórico de la lógica, solucionador de problemas generales, el nivel de conocimiento), Ray Solomonoff

A grandes rasgos, puede entenderse como el uso de sistemas informáticos para realizar tareas que antes requerían de la inteligencia, cognición o razonamiento humanos.[8] Es una categoría de investigación destinada a desarrollar sistemas que sean capaces de resolver problemas o alcanzar metas en diferentes grados de dificultad mediante el razonamiento. Es decir, imitando las habilidades humanas de resolución de problemas, en algunos casos incluyendo la capacidad de aprender de la experiencia, y por lo tanto, mejorar las capacidades de la máquina sin ninguna intervención humana,[9] y que está diseñada para actuar como un agente racional.[10]

En ese sentido, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) ha dictado cinco principios básicos para la regulación de la IA —*lato sensu*— en un documento de acuerdo general adoptado por 42 Estados miembros.[11]

– La IA debe estar al servicio de las personas y del planeta, impulsando un crecimiento inclusivo, el desarrollo sostenible y el bienestar.

(padre de la probabilidad algorítmica, teoría de la información algorítmica), Arthur Lee Samuel (primer algoritmo de aprendizaje automático para damas), W. Ross Ashby (pionero en cibernética, ley de la variedad requerida), Claude Shannon (padre de la teoría de la información) y John Nash (padre de la teoría de juegos).

8 Voz: "Inteligencia (artificial)", *DLE*, DRAE, disponible en *https://dle.rae.es/inteligencia#2DxmhCT*, ingresado el 2 de julio de 2020.

9 Kononenko, Igor y Kukar, Matjaz, *Machine Learning and Data Mining: Introduction to Principles and Algorithms*, 2007, p. 38; véase también Scherer, "Regulating AI", *supra* nota 6, p. 361.

10 Russell, Stuart y Norvig, Peter, *Artificial Intelligence - A Modern Approach*, 3a. ed., 2010, pp. 4 y 5 (en adelante: Russell y Norvig, *AI-A Modern Approach*).

11 Organización para la Cooperación y el Desarrollo Económicos, *Principios de la OCDE sobre IA*, disponible en *https://www.oecd.org/centrodemexico/medios/cuarentaydospaisesadoptanlosprincipiosdelaocdesobreinteligenciaartificial.htm*, ingresado el 5 de julio de 2020.

- Los sistemas de IA deben diseñarse de manera que respeten el Estado de derecho, los derechos humanos, los valores democráticos y la diversidad, e incorporar salvaguardas adecuadas que permitan la intervención humana.
- Los sistemas de IA deben estar presididos por la transparencia y una divulgación responsable, a fin de garantizar que las personas sepan cuándo están interactuando con ellos y puedan oponerse a los resultados de esa interacción.
- Los sistemas de IA han de funcionar con robustez, de manera fiable y segura durante toda su vida útil, y los potenciales riesgos deberán evaluarse y gestionarse en todo momento.
- Las organizaciones y las personas que desarrollen, desplieguen o gestionen sistemas de IA deberán responder de su correcto funcionamiento en consonancia con los principios precedentes.

No obstante, el potencial de esta tecnología tiene un alcance muy amplio, por lo que es necesario hacer una distinción entre los diferentes tipos de IA:[12]

- IA acotada, que tiene un rango limitado de habilidades y es la IA que más prevalece en nuestro mundo actual. Está programada para realizar una tarea específica extremadamente bien.
- IA general, que está a la par con las capacidades humanas. Esta tecnología no se ha logrado aún. El propósito de la IA general es pensar, comprender y actuar de una manera indistinguible de la de un ser humano en una situación determinada. Cabe destacar que el objetivo de muchos proyectos relacionados con la IA es que estos sistemas puedan adaptarse a diferentes situaciones y

[12] O'Carroll, Brodie, *What are the 3 types of AI? A Guide to Narrow, General, and Super Artificial Intelligence*, 2017, disponible en *https://codebots.com/artificial-intelligence/the-3-types-of-ai-is-the-third-even-possible*, ingresado el 5 de julio de 2020.

funcionar sin control humano. Lo que los científicos no han logrado aún, es una forma de hacer que las máquinas sean *conscientes*, mediante la programación de un conjunto completo de habilidades cognitivas que hasta ahora sólo son conocidas en los humanos.

– Súper IA, que es incluso más capaz que un humano. Está destinada a superar nuestras capacidades y así superar las limitaciones de nuestra especie. Es superpoderosa y consciente de sí misma, más allá del sentido humano. Esto simboliza la máxima evolución de este campo, y es en lo que se basa la teoría de la *singularidad*, la cual dice que las máquinas algún día serán lo suficientemente inteligentes para programarse y mejorarse a sí mismas hasta independizarse de sus creadores humanos.

Por un lado, algunos tecnoescépticos creen que este escenario es improbable, por el otro, el inventor de Google, Ray Kurzweil, predice que esta "singularidad" ocurrirá en 2045 con la creación de una IA consciente de sí misma, que será millones de veces más poderosa que todos los cerebros humanos.[13]

La vía para lograr esto es a través de una tecnología llamada aprendizaje automático (en inglés *machine learning*) un concepto que fue inicialmente introducido en la década de 1940 por el matemático Alan Turing y desarrollada a través de su *juego de imitación,* lo que definió un estándar operativo para la inteligencia, conocido como el "Test de Turing",[14] el cual se convirtió en la base de la IA que conocemos actualmente.[15]

---

[13] Business, AI, *Ray Kurzweil Predicts that the Singularity will take Place in 2045*, (2017), disponible en *https://aibusiness.com/document.asp?doc_id=760200*, ingresado el 5 de julio de 2020.

[14] Russell y Norvig, *AI-A Modern Approach*, *supra* nota 10, pp. 16 y 17.

[15] *Ibidem*, pp. 16 y 17.

En esencia, el aprendizaje automático es un proceso que permite que los sistemas artificiales mejoren a partir de la experiencia,[16] permite que las máquinas se adapten a nuevos entornos y actúen de una manera que les permita lograr el objetivo asignado independientemente de obstáculos imprevistos y sin una dirección explícita de su programador.[17]

Idealmente, el aprendizaje automático se convertiría en una solución para abordar de manera más eficiente, efectiva y precisa la imprevisibilidad, sin recibir órdenes del programador.[18]

Por otro lado, el aprendizaje profundo (en inglés *deep learning*) es una subárea del aprendizaje automático que se ocupa de algoritmos inspirados en la estructura y función del cerebro, llamadas redes neuronales artificiales.[19] Se basa en una jerarquía de aprendizaje de representación, que produce diferentes niveles de abstracción.[20] Básicamente, es una expansión del aprendizaje automático en capas multiplicadas, asimilando así una cantidad exponencial de datos.

Estas tecnologías tienen una ventaja asimétrica notable sobre los humanos, dado que pueden acumular conocimiento de bases de datos potencialmente infinitas y, a su vez, nos dejan con un conocimiento muy limitado de sus capacidades.

---

[16] Sileno, *History of AI*, *supra* nota 7.

[17] Russell y Norvig, *AI-A Modern Approach*, *supra* nota 10, p. 693; Coglianese, Carry y Lehr, David, "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era", *Georgetown Law Journal*, SSRN, 2017, p. 1156 (en adelante: Coglianese y Lehr, "Regulating by Robot"); Rich, Michael L., "Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment", *U. PA. L. Rev.*, vol. 164, 2015-2016, pp. 871, 875 (en adelante: Rich, "Machine Learning").

[18] Russell y Norvig, *AI-A Modern Approach*, *supra* nota 10, p. 693; Coglianese y Lehr, "Regulating by Robot", *supra* nota 17, p. 1156; Rich, "Machine Learning", *supra* nota 17, p. 875.

[19] Brownlee, Jason, *What is Deep Learning?*, 2019, disponible en *https://machinelearningmastery.com/what-is-deep-learning/*, ingresado el 4 de febrero de 2021.

[20] Goodfellow, Ian *et al.*, *Deep Learning*, MIT Press, 2016.

Por consiguiente, existe una circunstancia que se debe tener en cuenta, un gran *caveat* derivado tanto del aprendizaje automático como del aprendizaje profundo, al cual podríamos referirnos como una "incógnita conocida". Esta incógnita comprendería el proceso de *toma de decisiones*, también conocido como "la caja negra". La razón de esto es que, a diferencia de los procesos de toma de decisiones de los humanos que son lineales, los procesos de toma de decisiones artificiales son demasiado complejos para que los entendamos nosotros, porque la propia máquina crea sus algoritmos, y es así como estos procesos conforman la caja negra.[21]

Esto significa que, independientemente de su "configuración" original, es el propio programa el que decide el valor adecuado que se le asigna a cada elemento que percibe.[22]

Además, el programador no sabe qué regla o incluso qué características específicas utilizó la máquina para llegar a una determinada conclusión, ni tampoco puede deconstruir las inferencias o rastrear los procesos de decisión que se aplicaron.[23]

A manera de ejemplo, una función de utilidad que esté programada para mitigar o evitar el sufrimiento humano, podría decidir matar en lugar de herir a una persona, ya que las personas no sufren cuando están muertas.[24]

---

[21] Nicholson Price II, W., "Black-Box Medicine", *Harvard Journal of Law & Technology*, vol. 28, 2014-2015, pp. 419, 432-434 (en adelante: Nicholson, "Black-Box Medicine"); Rich, "Machine Learning", *supra* nota 17, p. 886.

[22] Coglianese y Lehr, "Regulating by Robot", *supra* nota 17, p. 1156. "No podemos saber realmente en qué características precisas se basa cualquier algoritmo de aprendizaje automático".

[23] *Idem.*

[24] Para más ejemplos, véase Calo, Ryan, "Robotics and the Lessons of Cyberlaw", *California Law Review*, vol. 103, 2015, pp. 542 y 543 (en adelante: Calo, "Robotics and the Lessons of Cyberlaw").

En otras palabras, el programador controla los datos introducidos al programa en su fase de aprendizaje, proporciona pautas de optimización para la interpretación de estos datos (lo que se conoce como función de utilidad) y está al tanto del resultado que extrapola el programa, pero para todos los demás efectos, la entidad artificial se considera una caja negra que no ofrece una explicación intuitiva ni causal de sus acciones.[25]

En resumen, el uso de programas de aprendizaje automático debe hacerse siempre con la plena conciencia que conlleva el riesgo inherente de que no hay forma de predecir, comprender o auditar una decisión específica realizada por la IA en términos comprensibles para los humanos.[26]

Para asimilar estos conceptos, viene a la mente la siguiente cita: "Como sabemos, hay conocimientos conocidos; hay cosas que sabemos que sabemos. También sabemos que existen incógnitas conocidas; es decir, sabemos que hay algunas cosas que no sabemos. Pero también hay incógnitas desconocidas, las que no conocemos, no las conocemos".[27]

---

[25] Shilo, Liron, *When Turing Met Grotius AI, Indeterminism, and Responsibility,* SSRN, 2018, p. 14 (en adelante: Shilo, *When Turing Met Grotius*).

[26] *Ibidem*, pp. 11 y 12, 18 y 19. Para más información sobre la caja negra, véase Nicholson, "Black-Box Medicine", *supra* nota 21, pp. 432-437; 442-467. Él explica los pros y contras de las cajas negras en el contexto médico; Rich, "Machine Learning", *supra* nota 17, pp. 886, 923 y 924. Describe el intercambio entre la utilización de algoritmos que tienen las cajas negras, pero que son de gran precisión en sus predicciones, y concluye que, a pesar de la exactitud favorable en la predicción, las cajas negras deberían ser más transparentes en aras de analizar la implicación de la cuarta enmienda por el uso de tales algoritmos.

[27] Respuesta de Donald Rumsfeld a una pregunta del Departamento de Defensa de Estados Unidos, durante una conferencia de prensa en febrero de 2002, disponible en *https://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636*, ingresado el 7 de julio de 2020.

Es destacable que la tecnología subyacente de la IA tiene el potencial de adaptarse tanto a usos civiles como militares. Este estudio se centra en el segundo dominio, cuyas aplicaciones podrían incluir inteligencia, vigilancia y reconocimiento, navegación, comando y control multidominio, defensa antimisiles, defensa cibernética, manipulación de información, reconocimiento de objetivos y desarrollo de armas.[28]

Los sistemas de armas más recientes son completamente *sui generis* y dan origen a una categoría propia *de jure* y *de facto*.

Estos sistemas son *entidades* altamente sofisticadas capaces de imitar las habilidades humanas de toma de decisiones para ejecutar una variedad de tareas *sin* intervención humana alguna.[29] También conocidos como sistemas de armas autónomas letales, estas entidades están diseñadas para *iniciar activamente* y *tomar decisiones letales* en lugar de actuar simplemente como sistemas defensivos y/o reactivos.[30]

Como tal, esta categoría de entidades no está definida actualmente en nuestro ordenamiento jurídico, dada la predominante y novedosa característica de contar con un carácter

---

[28] Nasu, Hitoshi, *Artificial Intelligence and the Obligation to Respect and to Ensure Respect for International Humanitarian Law*, Exeter Centre for International Law, 2019, p. 5.

[29] Véase Schmitt, Michael N., "Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics", *Harvard National Security Journal*, 2013, p. 4 (en adelante: Schmitt, "AWS and IHL"); Anderson, Kenneth *et al*., "Adapting the Law of Armed Conflict to Autonomous Weapon Systems", *International Legal Studies*, vol. 90, 2014, p. 386; véase Alston, Philip, *Interim report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, U.N. Human R. Comm., U.N. Doc. A/65/321, 23 de agosto de 2010, p. 18; Kerr, Ian y Szilagyi, Katie, *Asleep at the Switch? How Killer Robots Become a Force Multiplier of Military Necessity*, 2016, p. 333; Ben-Naftali, Orna y Triger, Zvi, "The Human Conditioning: International Law and Science Fiction", *Law, Culture and the Humanities*, vol. 14, 2016, p. 38.

[30] Shilo, *When Turing Met Grotius, supra* nota 25, p. 2.

autónomo a nivel cognitivo y de toma de decisiones. Por ello, es primordial recordar que, aunque pueden usarse de manera armamentista, no deben examinarse o definirse simplemente como armas, ya que son mucho más que eso.

Tras una observación minuciosa, uno se da cuenta de que estas entidades no son armas, plataformas convencionales ni agentes morales equivalentes a los humanos para efectos legales. Dicho esto, a menudo se les refiere como la primera, en ocasiones como lo segundo, y con frecuencia se les trata como el tercero.[31]

Naturalmente, este es otro concepto que carece de consenso en su definición, y cuyo margen de discrecionalidad varía enormemente. En un extremo del espectro, un SAA se considera un componente automatizado de un arma existente y, en el otro, como una plataforma que es capaz de detectar, aprender y lanzar ataques.[32]

De manera relacionada, y en un esfuerzo por simplificar —pero en efecto, ampliando— el concepto, académicos de Harvard han introducido el término "algoritmos de guerra", definido como cualquier algoritmo que se expresa en código de computadora, que se efectúa a través de un sistema construido y que es capaz de operar en el contexto de un conflicto armado. Estos sistemas incluyen arquitecturas de autoaprendizaje que están en el centro de los debates más importantes sobre el reemplazo del criterio humano con elecciones derivadas de algoritmos.[33]

Como se explicó anteriormente, el hecho de que estas elecciones puedan ser difíciles de anticipar o desentrañar para los humanos, *vis-à-vis* la posibilidad de que sean autó-

[31] *Ibidem*, p. 15.

[32] Lewis, Dustin A. *et al*., *War-Algorithm Accountability*, HLS PILAC, 2016, p. 5 (en adelante: Lewis *et al*., *War-Algorithm*).

[33] *Ibidem*, p. 10.

nomos y capaces de actuar físicamente en el mundo, definitivamente confronta las nociones jurídicas tal y como las conocemos hoy en día.[34]

En consecuencia, los SAA desafían las nociones fundamentales e interrelacionadas del derecho internacional público, el DIH, el derecho penal internacional (DPI) y los esquemas de responsabilidad relacionados. Estos conceptos incluyen atribución, control, previsibilidad y capacidad de reconstrucción.[35]

Es comprensible que el científico estadounidense Max Tegmark se refiera a la transición hacia la autonomía como "la tercera revolución de las armas", después de la invención de la pólvora en el siglo trece y la de las armas nucleares en el siglo veinte.[36]

Autonomía

Ahora bien, es importante abordar dos aspectos claves de esta nueva categoría: la autonomía y la toma de decisiones.

Dado que nuestras sociedades están cada día más interconectadas, es relativamente sencillo llevar a cabo no sólo los ciberataques tradicionales, sino los ciberataques cinéticos. Estos ataques son asaltos virtuales que tienen consecuencias tangibles en el mundo físico causando daños, lesiones o incluso la muerte, únicamente a través de la explotación de sistemas y procesos de información vulnerables. Esto es, utilizar el ciberespacio para infligir daños físicos en plantas de energía nuclear, instalaciones de agua, oleoductos, fábricas, hospitales,

---

[34] Calo, "Robotics and the Lessons of Cyberlaw", *supra* nota 24, p. 542.

[35] Lewis *et al*., *War-Algorithm*, *supra* nota 32, p. 77.

[36] Tegmark, Max, *Life 3.0: Being Human in the Age of Artificial Intelligence*, Nueva York, 2017.

bancos, sistemas de tránsito y estructuras de apartamentos.[37] El atacante puede estar ubicado en un lugar seguro y alejado, afectando de forma remota las vidas humanas y desestabilizando gobiernos nacionales o extranjeros al atacar su infraestructura crítica, pues todo lo que se requiere es una conexión a internet.

A pesar de lo peligrosa que es esta noción, a simple vista, es digno de mención que se trata de un tipo de guerra controlada, operada y dirigida por humanos… *en principio*.

Sin embargo, lo que actualmente concebimos como *espacio de batalla*[38] se está volviendo gradualmente, pero cada vez más, libre de humanos, tanto en la práctica como en la teoría. Es fácil entender la razón de esto, ya que las máquinas tienen una ventaja sobre los humanos, dado que pueden ejecutar algunas tareas de manera más rápida, precisa y económica que si las hiciéramos nosotros.

Teóricamente, los seres humanos todavía estamos a cargo de *cuándo* comenzar una guerra, *contra quién* se peleará (*ius ad bellum*), qué *armas*, *medios y métodos* se utilizarían y qué *objetivos* se deben alcanzar (*ius in bello*). Sin embargo, a nivel táctico, las máquinas se están convirtiendo en los "micro-gestores" y los ejecutores de cómo lograr estos objetivos. Decidirán cada vez más a quién, qué y cuándo atacar para hacerlo. Esto es lo que se conoce como automatización del espacio de batalla.[39]

La pregunta que debemos hacernos en este punto es, ¿podemos permitir que estas decisiones se deleguen de manera legal (y ética) a las máquinas?

---

[37] Goud, Naveen, *What is a Cyber Kinetic Attack?*, disponible en *https://www.cybersecurity-insiders.com/what-is-a-cyber-kinetic-attack/*, ingresado el 20 de julio de 2020.

[38] Previamente conocido como "campo de batalla", como se explicará *infra* en el epígrafe "Consideraciones del derecho internacional humanitario".

[39] Shilo, *When Turing Met Grotius*, *supra* nota 25, p. 1.

Además, es sólo cuestión de tiempo antes de que esta automatización del espacio de batalla llegue al siguiente nivel, el volverse autónoma.

El Comité Internacional de la Cruz Roja (CICR) es la organización humanitaria encargada de salvaguardar los Convenios de Ginebra, y ha declarado que no se opone a las nuevas tecnologías de la guerra *per se*.[40]

En todos los casos, como requisito mínimo, cualquier tecnología de guerra nueva debe utilizarse, y debe poder utilizarse, de conformidad con las normas vigentes del DIH. Sin embargo, las características únicas de las nuevas tecnologías de guerra, las circunstancias previstas y esperadas de su uso y sus consecuencias humanitarias pronosticables, plantean dudas respecto a si las reglas existentes son suficientes o si deben aclararse o complementarse, a la luz de su presumible impacto.[41]

Por supuesto, ciertas tecnologías militares —como las que permiten una mayor precisión en los ataques— pueden ayudar a las partes en conflicto a minimizar las consecuencias humanitarias de la guerra. No obstante, como ocurre con cualquier nueva tecnología de guerra, su legalidad depende de la forma en que se utilicen en la práctica.

Para el CICR, una aplicación importante es el uso de herramientas de IA y aprendizaje automático para controlar *hardware* militar físico, en particular, el creciente número de sistemas robóticos no tripulados —en el aire, en tierra y en el mar— con una amplia gama de tamaños y funciones. La IA y el aprendizaje automático pueden permitir una mayor autonomía en estas

[40] CICR, Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centered Approach, Ginebra, 6 de junio de 2019, p. 1 (en adelante: CICR, AI and Machine Learning in Armed Conflict).

[41] CICR, "International Humanitarian Law and the Challenges of Contemporary Armed Conflicts", 32nd International Conference of the Red Cross and Red Crescent, Ginebra, octubre de 2015, pp. 38-47.

plataformas robóticas, ya sean armadas o desarmadas, y controlar todo el sistema o funciones específicas, como vuelo, navegación, vigilancia o selección de objetivos.[42]

La organización también está interesada en la aplicación de la IA y el aprendizaje automático para el desarrollo de armas cibernéticas como "armas digitales autónomas", ya que se espera que cambien la naturaleza tanto de las capacidades de ciberdefensa y ciberataques, aumentando así la escala y cambiando la naturaleza y gravedad de los ataques.[43]

Naturalmente, una preocupación recurrente entre los académicos es que su naturaleza autónoma crea una brecha de responsabilidad[44] —la caja negra— que niega cualquier culpabilidad moral de los actores humanos que participaron en su creación, uso o despliegue. Está claro que las cuestiones relativas a la atribución de responsabilidad son importantes, porque el hecho de que alguien pueda ser considerado responsable de desviarse de las reglas acordadas, es una condición *sine qua non* para librar una guerra justa.[45]

---

[42] CICR, AI and Machine Learning in Armed Conflict, *supra* nota 40, p. 3.

[43] Brundage, M. *et al*., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, 2018; United Nations Institute for Disarmament Research, *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, UNIDIR, 2017.

[44] Davison, N., "Autonomous Weapon Systems under International Humanitarian Law", *UNODA Occasional Papers. Perspectives on Lethal Autonomous Weapon Systems*, núm. 30, noviembre de 2017 (en adelante: Davison, "AWS under IHL"), disponible en *https://www.icrc.org/en/document/autonomous-weapon-systems-under-international-humanitarian-law*, ingresado el 2 de agosto de 2020.

[45] Véase Sparrow, Robert, "Killer Robots", *Journal of Applied Philosophy*, vol. 24, 2007, pp. 62, 67 (en adelante: Sparrow, "Killer Robots").

¿Tomadores de decisiones?

Quizás la aplicación más amplia y de mayor alcance es el uso de la IA y el aprendizaje automático para la toma de decisiones, permitiendo la recopilación y el análisis de datos para identificar personas u objetos, evaluar patrones de vida o comportamiento, hacer recomendaciones para estrategias militares u operaciones, o hacer predicciones sobre acciones o situaciones futuras.[46]

Estos sistemas automatizados de toma de decisiones son efectivamente una expansión de las herramientas de inteligencia, vigilancia y reconocimiento, que utilizan IA y aprendizaje automático para automatizar el análisis de grandes conjuntos de datos y proporcionar "consejos" a los humanos en la toma de decisiones particulares y, cada vez más, para automatizar tanto el análisis y la posterior decisión y/o acción por parte del sistema.[47] Las aplicaciones relevantes de la IA y del aprendizaje automático incluyen el reconocimiento de patrones, procesamiento del lenguaje natural, reconocimiento de imágenes, reconocimiento facial y reconocimiento de comportamiento. La preocupación del CICR gira en torno al hecho de que el posible uso de estos sistemas es extremadamente amplio y puede incluir decisiones sobre a quién o qué atacar y cuándo,[48] decisiones sobre a quién detener y durante cuánto tiempo,[49] decisio-

[46] CICR, AI and Machine Learning in Armed Conflict, *supra* nota 40, p. 4.

[47] *Idem*.

[48] Estados Unidos, "Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems", Documento de Trabajo, Convención sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales que Puedan Considerarse Excesivamente Nocivas o de Efectos Indiscriminados (CAC) Grupo de Expertos Gubernamentales, marzo de 2019.

[49] Deeks, Ashley, "Predicting Enemies", *Virginia Public Law and Legal Theory Research Paper*, núm. 21, 2018, pp. 1549-54.

nes sobre estrategia militar —incluso sobre el uso de armas nucleares—[50] y las relacionadas con operaciones específicas, como los intentos de predecir o adelantarse a los adversarios.[51]

Es precisamente esta preocupación la que ha motivado discusiones agitadas sobre el papel de la toma de decisiones en la guerra, y sobre quién está mejor preparado para tomar decisiones de vida o muerte: los humanos o las máquinas. Sin embargo, esta es una cuestión casi teológica, ya que no está técnicamente claro que una máquina pueda cumplir en todo momento con el DIH o las reglas para las que fue programada, y por otro lado, un ser humano puede actuar deliberadamente en violación al DIH.[52]

En todo caso, también existe un desacuerdo significativo sobre el análisis costo-beneficio que podría resultar de distanciar a los combatientes humanos del campo de batalla y si los beneficios potenciales de salvar vidas de los SAA son superados por los riesgos inherentes al hecho de que la guerra también se vuelve, en un sentido práctico, más fácil de llevar a cabo,[53] lo cual tiene como resultado una mayor vulnerabilidad de las poblaciones civiles.

Los defensores de estas tecnologías argumentan que la IA y los sistemas de apoyo en la toma de decisiones, basados en el aprendizaje automático, pueden permitir a los humanos tomar mejores decisiones en la forma en que se conducen las hostilidades, y que éstas

---

[50] Boulanin, V. (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. 1: *Euro-Atlantic Perspectives*, Stockholm International Peace Research Institute, 2019.

[51] Hill, S. y Marsan, N., "Artificial Intelligence and Accountability: A Multinational Legal Perspective", *Big Data and Artificial Intelligence for Military Decision Making, Meeting Proceedings STO-MP-IST-160*, NATO, 2018.

[52] Sassóli, Marco, "Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified", *International Law Studies*, vol. 90, U.S. Naval War College, 2014, p. 310 (en adelante: Sassóli, "Autonomous Weapons and IHL").

[53] Lewis *et al*., *War-Algorithm*, *supra* nota 32, p. 8.

estarán apegadas al DIH. Afirman que esto minimizará los riesgos para los civiles, al facilitar una recopilación y un análisis de información más rápida y generalizada.

No obstante, el CICR correctamente reitera el problema de la caja negra, dado que esos mismos análisis algorítmicos también podrían facilitar peores decisiones, violaciones del DIH y exacerbar los riesgos para los civiles. Lo anterior, especialmente debido a las limitaciones actuales de la tecnología, tales como la imprevisibilidad, la falta de explicación y los prejuicios discriminatorios. Desde una perspectiva humanitaria, esta es una preocupación fundamental, ya que son riesgos de lesiones o muerte para las personas o de destrucción de objetos civiles, y además porque estas *decisiones* se rigen por la *lex specialis* del DIH.[54]

[54] CICR, AI and Machine Learning in Armed Conflict, *supra* nota 40, p. 7.