INFORMES Y DOCUMENTOS

La inteligencia artificial y el aprendizaje automático en los conflictos armados: un enfoque centrado en las personas

Nota: la siguiente es una versión editada de un artículo publicado por el CICR en junio de 2019.

***

1. Introducción

En un momento caracterizado por la intensificación de los conflictos armados y la velocidad del desarrollo tecnológico, el Comité Internacional de la Cruz Roja (CICR) necesita comprender cómo inciden las nuevas tecnologías en las personas afectadas por esos conflictos y concebir soluciones humanitarias que atiendan las necesidades de los más vulnerables.

Al igual que numerosas organizaciones de diferentes sectores y regiones, el CICR está haciendo un gran esfuerzo para intentar comprender las consecuencias de la **inteligencia artificial** (IA) y del **aprendizaje automático** para su labor. La IA consiste en el uso de sistemas informáticos para realizar tareas –frecuentemente asociadas a la inteligencia humana– que requieren cognición, planificación, razonamiento o aprendizaje. Por su parte, los sistemas de aprendizaje automático son sistemas de IA que están "entrenados" y "aprenden" a partir de datos, que, en

última instancia, definen la manera en la que funcionan. Como estas herramientas de *software*, o algoritmos, pueden aplicarse a una enorme diversidad de tareas, pueden tener repercusiones cuyo alcance todavía no ha sido comprendido por completo.

Existen dos grandes ámbitos de aplicación –diferentes entre sí– de la IA y del aprendizaje automático que son de interés particular para el CICR: por un lado, **su empleo en la conducción de la guerra** o en otras situaciones de violencia[1]; y, por otro, **su empleo en la labor humanitaria** para asistir y proteger a las víctimas de los conflictos armados[2]. En este documento, el CICR expone su perspectiva sobre el uso de la IA y del aprendizaje automático en los conflictos armados y las posibles consecuencias en el plano humanitario, así como las obligaciones jurídicas y consideraciones éticas asociadas que deben regir su desarrollo y empleo. También hace referencia al uso de las herramientas de IA para la acción humanitaria, por parte de actores como el CICR.

2. El enfoque del CICR respecto de las nuevas tecnologías de guerra

El CICR tiene una larga trayectoria en el análisis de cómo los avances contemporáneos y del futuro próximo inciden e incidirán en los conflictos armados. Este enfoque considera los nuevos medios y métodos de guerra; específicamente, en lo que respecta a su compatibilidad con las normas del derecho internacional humanitario (también conocido como el derecho de los conflictos armados o el derecho de la guerra), así como los riesgos de consecuencias adversas desde el punto de vista humanitario para las personas protegidas.

La institución no se opone a las nuevas tecnologías de guerra *per se*. Ciertas tecnologías militares –como las que permiten una mayor precisión en los ataques– pueden ayudar a las partes en conflicto a reducir al mínimo las consecuencias humanitarias de la guerra, en particular para las personas civiles, y a que se respeten las normas de la guerra. No obstante, como sucede con cualquier nueva tecnología de guerra, las tecnologías de precisión no son beneficiosas en sí, y las consecuencias humanitarias en el terreno dependerán de cómo se empleen las nuevas armas. Es esencial, por lo tanto, contar con un análisis realista de las nuevas tecnologías, fundamentado en sus características técnicas y en la forma en que se emplean o se prevé emplearlas.

Toda nueva tecnología de guerra debe ser empleada, y debe tener la capacidad de ser empleada, en cumplimiento de las normas vigentes del

1 CICR, "Expert views on the frontiers of artificial intelligence and conflict", blog del CICR sobre derecho y políticas humanitarias (en inglés), 19 de marzo de 2019, disponible en https://blogs.icrc.org/law-and-policy/2019/03/19/expert-views-frontiers-artificial-intelligence-conflict/

2 CICR, "Summary Document for UN Secretary-General's High-Level Panel on Digital Cooperation", enero 2019, disponible en https://digitalcooperation.org/wp-content/uploads/2019/02/ICRC-Submission-UNPanel-Digital-Cooperation.pdf.

derecho internacional humanitario. Es un requisito mínimo[3]. Sin embargo, las características singulares de las nuevas tecnologías de guerra, las circunstancias previstas y esperadas de su empleo, y sus consecuencias humanitarias previsibles pueden plantear dudas sobre si las normas vigentes son suficientes o si necesitan esclarecerse o complementarse, a la luz del impacto que se prevé que tengan esas nuevas tecnologías[4]. Lo que está claro es que las aplicaciones militares de tecnologías nuevas y emergentes no son inevitables. Son elecciones de los Estados, que deben estar dentro de los límites de las normas vigentes y tener en cuenta las potenciales consecuencias humanitarias para las personas civiles y para combatientes que hayan dejado de participar en las hostilidades, así como las consideraciones más amplias de "humanidad" y "conciencia pública"[5].

3. Empleo de la IA y del aprendizaje automático por las partes en conflicto

Todavía no se conocen del todo las formas en que las partes en conflictos armados –tanto Estados como grupos armados no estatales– pueden utilizar la IA y el aprendizaje automático en la conducción de la guerra; tampoco se conocen sus posibles consecuencias. Sin embargo, se han identificado al menos **tres aspectos que se solapan entre sí considerados relevantes desde una perspectiva humanitaria**, por ejemplo, para el cumplimiento del derecho internacional humanitario.

3.1 Creciente autonomía en los sistemas robóticos físicos, incluidas las armas

Una aplicación importante de estas tecnologías es el uso de herramientas digitales de **IA y aprendizaje automático para controlar el material militar físico**, en particular, el creciente número de sistemas robóticos no tripulados –en el aire, en tierra y en el mar–, con una amplia variedad de tamaños y funciones. La IA y el aprendizaje automático permiten ampliar la autonomía en estas plataformas

3 Los Estados Partes en el Protocolo I adicional a los Convenios de Ginebra tienen la obligación de realizar exámenes jurídicos de armas nuevas durante su desarrollo y adquisición, así como antes de su empleo en conflictos armados. Para otros Estados, los exámenes jurídicos son una medida de sentido común que contribuye a garantizar que las fuerzas armadas del Estado puedan conducir hostilidades de conformidad con sus obligaciones internacionales.

4 CICR, El derecho internacional humanitario y los desafíos de los conflictos armados contemporáneos, informe presentado ante la XXXIII Conferencia Internacional de la Cruz Roja y de la Media Luna Roja, Ginebra, octubre de 2019 (informe de 2019 sobre los desafíos contemporáneos) pp. 18-31, disponible en https://www.icrc.org/es/publication/el-derecho-internacional-humanitario-y-los-desafios-de-los-conflictos-armados; CICR, El derecho internacional humanitario y los desafíos de los conflictos armados contemporáneos, informe presentado ante la XXXII Conferencia Internacional de la Cruz Roja y de la Media Luna Roja, Ginebra, octubre de 2015 (informe de 2015 sobre los desafíos contemporáneos) pp. 50-62, disponible en https://www.icrc.org/es/document/el-derecho-internacional-humanitario-y-los-desafios-de-los-conflictos-armados.

5 Los "principios de humanidad" y los "dictados de la conciencia pública" se mencionan en el Artículo 1(2) del Protocolo I adicional a los Convenios de Ginebra y en el preámbulo del Protocolo II adicional a los Convenios de Ginebra, denominado Cláusula Martens, que forma parte del derecho internacional humanitario consuetudinario.

robóticas, estén o no armadas, así como controlar todo el sistema o bien funciones específicas, como el vuelo, la navegación, la vigilancia o la selección de objetivos de ataque.

Para el CICR, los **sistemas de armas autónomos** –caracterizados por la autonomía en sus "funciones críticas" de seleccionar y atacar objetivos– representan una preocupación inmediata desde una perspectiva humanitaria, jurídica y ética, debido al riesgo de que se pierda el control humano sobre las armas y el uso de la fuerza[6]. Esta pérdida de control plantea riesgos para las personas civiles, a raíz de sus consecuencias impredecibles; se ponen en juego cuestiones jurídicas[7], porque los combatientes deben tomar decisiones específicas según cada contexto al efectuar ataques en el marco del derecho internacional humanitario; y preocupaciones éticas[8], porque se necesita el criterio humano en las decisiones relativas al uso de la fuerza para mantener la responsabilidad moral y la dignidad humana. Por estos motivos, el CICR ha propuesto elementos prácticos de control humano como base para los límites internacionalmente acordados sobre la autonomía en los sistemas de armas, centrándose en lo siguiente[9]:

- **Controles respecto de los parámetros de las armas**, que pueden traducirse en límites a los tipos de sistemas de armas autónomos, incluidos los objetivos contra los que son empleados, así como límites a su duración y al alcance geográfico de su acción, y requerimientos para su desactivación y mecanismos a prueba de fallas.
- **Controles respecto del entorno,** que pueden traducirse en límites a las situaciones y ubicaciones en las que los sistemas de armas autónomos pueden emplearse, principalmente en cuanto a presencia y densidad de personas civiles y bienes de carácter civil.

6 CICR, declaraciones del CICR ante el Grupo de expertos gubernamentales sobre sistemas de armas autónomos letales de la Convención sobre ciertas armas convencionales (CCA) (en inglés), Ginebra, 25 al 29 de marzo de 2019, disponible en https://tinyurl. com/yyeadno3.

7 CICR, informe de 2019 sobre los desafíos contemporáneos, nota 4 supra, pp. 29-31; Davison, N., "Autonomous weapon systems under international humanitarian law", en Perspectives on Lethal Autonomous Weapon Systems, documentos ocasionales de la Oficina de Asuntos de Desarme de las Naciones Unidas n.º 30, noviembre de 2017, disponible en www. icrc.org/en/document/autonomous-weapon-systems-under-international-humanitarian-law.

8 CICR, "Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?", informe de una reunión de expertos, Ginebra, 3 de abril de 2018, disponible en www.icrc.org/en/document/ethics-and-autonomous-weaponsystems-ethical-basis-human-control.

9 CICR, Comentario del CICR sobre los "Principios rectores" del Grupo de expertos gubernamentales sobre sistemas de armas autónomos letales de la Convención sobre ciertas armas convencionales, Ginebra, julio de 2020, disponible en https://documents.unoda.org/wp-content/uploads/2020/07/ 20200716-ICRC.pdf (en inglés); Vincent Boulanin, Neil Davison, Netta Goussac y Moa Peldán Carlsson, Establecer límites a la autonomía de los sistemas de armas: identificación de elementos prácticos del control humano, CICR y el Instituto Internacional de Investigaciones sobre la Paz de Estocolmo, junio de 2020, disponible en www.icrc.org/en/document/ limits-autonomous-weapons (en inglés); CICR, "The Element of Human Control", documento de trabajo, Reunión de las Altas Partes Contratantes, Convención sobre ciertas armas convencionales (CCA), CCW/MSP/2018/WP.3, 20 de noviembre de 2018, disponible en https://tinyurl.com/y3c96aa6 (en inglés).

- **Controles respecto de la interacción entre personas y máquinas,** que pueden utilizarse para establecer requerimientos de supervisión humana, así como en relación con la capacidad humana de intervenir y desactivar los sistemas de armas autónomos; y requerimientos para un funcionamiento predecible y transparente.

Es importante reconocer que **no todas las armas autónomas incorporan tecnologías de IA o de aprendizaje automático**. Actualmente, existen armas dotadas de autonomía en sus funciones críticas –como los sistemas de defensa aérea con modos autónomos– que suelen utilizar un *software* de control sencillo y basado en reglas para seleccionar y atacar objetivos. Sin embargo, **el *software* de IA y de aprendizaje automático** –específicamente del tipo desarrollado para el "reconocimiento automático de objetivos"– **podría sentar la base de futuros sistemas de armas autónomos, aportando una nueva dimensión de imprevisibilidad a estas armas**, así como preocupaciones por la falta de explicabilidad y los sesgos (v. la sección 5.2)[10]. El mismo tipo de *software* también podría usarse en aplicaciones de "apoyo a la toma de decisiones" para la selección de objetivos de ataque, en lugar de controlar directamente un sistema de armas (v. la sección 3.3).

Por el contrario, no todos los sistemas robóticos militares que utilizan IA y aprendizaje automático son armas autónomas, ya que el *software* podría usarse para funciones de control que no sean la selección de objetivos, como la vigilancia, la navegación y el vuelo. Si bien, desde la perspectiva del CICR, la autonomía en los sistemas de armas –incluidos los sistemas con tecnología de IA– plantea las cuestiones más urgentes, el uso de la IA y del aprendizaje automático para aumentar la autonomía del equipamiento militar en general –como en aeronaves, vehículos terrestres y embarcaciones marítimas sin tripulación– también puede plantear interrogantes respecto de la interacción entre personas y máquinas en cuanto a la seguridad. Los debates en el sector civil sobre cómo garantizar la seguridad de los vehículos autónomos –como los automóviles sin conductor o los drones– pueden aportar aprendizajes para su aplicación en los conflictos armados (v. también la sección 3.3).

3.2 Nuevos medios de guerra cibernética y de la información

La aplicación de **la IA y del aprendizaje automático al desarrollo de armas o capacidades cibernéticas** es otro tema importante. No todas las capacidades cibernéticas incorporan IA y aprendizaje automático. Sin embargo, se espera que estas tecnologías **modifiquen la naturaleza tanto de las capacidades de defensa contra los ataques cibernéticos como de las capacidades de ataque**. Por ejemplo, mediante capacidades cibernéticas que se valgan de la IA y el aprendizaje automático,

10 CICR, Declaraciones del CICR ante el Grupo de expertos gubernamentales sobre sistemas de armas autónomos letales de la Convención sobre ciertas armas convencionales (CCA), tema 6(b) en el orden del día (en inglés), Ginebra, 27 al 31 de agosto de 2018, disponible en https://tinyurl.com/y4cql4to.

sería posible, automáticamente, buscar vulnerabilidades para aprovecharse de ellas o defenderse contra los ciberataques y, al mismo tiempo, lanzar contraataques. Este tipo de avances podrían aumentar la magnitud y cambiar la naturaleza, e incluso la gravedad de los ataques[11]. Algunos de estos sistemas podrían llegar a describirse como "armas autónomas digitales" y plantear interrogantes sobre el control humano similares a los que se plantean respecto de las armas autónomas físicas[12].

El eje de interés del CICR respecto de la guerra cibernética radica en particular en hacer respetar las normas vigentes del derecho internacional humanitario en todo ataque cibernético en el marco de un conflicto armado y, asimismo, en que las dificultades específicas para la protección de la infraestructura y los servicios de carácter civil sean atendidas por las partes que realizan ataques de este tipo[13] o se defienden de ellos, a fin de limitar al máximo el costo humano[14].

En el ámbito digital, una aplicación conexa de la IA y del aprendizaje automático es **el empleo de estas herramientas para la guerra de la información**, en particular para producir y difundir informaciones falsas, con la intención de engañar o sin ella. No en todos los casos se recurre a la inteligencia artificial y al aprendizaje automático, pero esas tecnologías parecen tener la capacidad de cambiar la naturaleza y la magnitud de la manipulación de la información en tiempo de guerra, así como las posibles consecuencias. Los sistemas dotados de IA se han utilizado ampliamente para producir información falsa –ya sea texto, contenido de audio, fotos o videos–, que resulta cada vez más difícil distinguir de la información real. El uso de estos sistemas por las partes en conflicto para amplificar los métodos tradicionales de propaganda a fin de manipular la opinión e influir en las decisiones podría tener implicaciones significativas en el terreno[15]. En opinión del CICR, causa preocupación el riesgo de que las personas civiles se vean expuestas a arrestos o malos tratos, discriminación o denegación de acceso a servicios esenciales, o incluso ataques a su persona o a sus bienes, como resultado de la difusión de información falsa o errónea en el ámbito digital, de manera deliberada o no[16].

11 Miles Brundage *et al*., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, Oxford, febrero de 2018.

12 Instituto de las Naciones Unidas de investigación sobre el desarme (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, 2017.

13 Al afirmar que el derecho internacional humanitario es aplicable a las operaciones cibernéticas, el CICR no insinúa bajo ningún punto de vista que la organización apruebe la guerra cibernética o la militarización del ciberespacio: CICR, Informe de 2015 sobre los desafíos de los conflictos armados, nota 4 supra, pp. 38–44.

14 CICR, "The Potential Human osto f Cyber Operations", informe de una reunión de expertos, Ginebra, mayo de 2019: www.icrc.org/en/document/potential-human-cost-cyber-operations.

15 Steven Hill y Nadia Marsan, "Artificial Intelligence and Accountability: A Multinational Legal Perspective", en *Big Data and Artificial Intelligence for Military Decision Making*, actas de las reuniones, STO-MP-IST-160, OTAN, 2018 (en inglés).

16 CICR, informe del simposio: "Digital Risks in Situations of Armed Conflict" (en inglés), marzo de 2019, p. 9, disponible en www.icrc.org/en/event/digital-risks-symposium.

3.3 Modificar la naturaleza de la toma de decisiones en los conflictos armados

Quizás la aplicación más amplia y de mayor alcance sea el uso de **la IA y del aprendizaje automático en la toma de decisiones**, ya que permite la recopilación y el análisis generalizados de fuentes de datos para identificar personas o bienes, evaluar patrones de vida o de comportamiento, formular recomendaciones para estrategias u operaciones militares, o prever acciones o situaciones futuras.

Estos **sistemas de "apoyo a las decisiones" o "toma de decisiones automatizada" son efectivamente una expansión de las herramientas de inteligencia, vigilancia y reconocimiento**, que se valen de la utilización de la IA y del aprendizaje automático para automatizar el análisis de grandes conjuntos de datos a fin de "asesorar" a los seres humanos en la toma de decisiones específicas o para automatizar tanto el análisis como la puesta en marcha de una decisión o de una acción por parte del sistema. Entre las aplicaciones relevantes de la IA y del aprendizaje automático, podemos nombrar el reconocimiento de patrones, imágenes, rostros y comportamientos, así como el procesamiento del lenguaje natural. **Los usos posibles de estos sistemas son extremadamente variados**[17]: decisiones sobre a quién –o qué– atacar y cuándo[18], decisiones sobre a quién detener y durante cuánto tiempo[19], decisiones sobre estrategia militar –entre otras cosas, sobre el uso de armas nucleares[20]–, y operaciones específicas, por ejemplo, para prever o prevenir los planes de los adversarios[21]. En función de su uso o mal uso, así como de las capacidades y limitaciones de la tecnología, estas aplicaciones en relación con la toma de decisiones podrían suponer mayores riesgos para las poblaciones civiles.

Al facilitar que la información se recopile y analice de manera más rápida y generalizada, los **sistemas de apoyo a las decisiones** basados en la IA y el aprendizaje automático pueden permitir que se tomen mejores decisiones en relación con el cumplimiento del derecho internacional humanitario en la conducción de las hostilidades y con la reducción al mínimo de los riesgos para las personas civiles. Sin embargo, los mismos algoritmos sobre los cuales se basan los análisis o las previsiones también podrían llevar a decisiones peores o violaciones del derecho internacional humanitario y exacerbar los riesgos para la población civil, en particular si se tienen en cuenta las limitaciones actuales de la tecnología, como la imprevisibilidad, la falta de explicabilidad y los sesgos (v. la sección 5.2).

17 Dustin A. Lewis, Gabriella Blum y Naz K. Modirzadeh, *War-Algorithm Accountability*, Harvard Law School Program on International Law and Armed Conflict (en inglés), agosto de 2016.

18 EE. UU., "Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems", documento de trabajo, Grupo de expertos gubernamentales sobre sistemas de armas autónomos letales de la Convención sobre ciertas armas convencionales (CCA) (en inglés), marzo de 2019.

19 Ashley Deeks, "Predicting Enemies", Virginia Public Law and Legal Theory Research Paper n.° 2018- 21 (en inglés), marzo de 2018, disponible en https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3152385.

20 Vincent Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Vol. 1: Euro-Atlantic Perspectives*, Stockholm International Peace Research Institute (SIPRI), mayo de 2019.

21 S. Hill y N. Marsan, nota 15 supra.

Varias **decisiones tomadas por las partes en conflicto por intermedio o por influencia de la IA son pertinentes desde una perspectiva humanitaria**, en particular cuando plantean riesgos de lesiones, muerte o destrucción de bienes, y cuando esas decisiones están regidas por normas específicas del derecho internacional humanitario. Por ejemplo, en casos en los que puede haber consecuencias graves para la vida, el uso de la IA y del aprendizaje automático para **decidir atacar un objetivo en un conflicto armado** exigirá consideraciones específicas para que las personas conserven la facultad de evaluar, en función del contexto, las condiciones necesarias para cumplir las normas jurídicas que rigen la conducción de las hostilidades (v. la sección 5). Un sistema de IA utilizado para iniciar un ataque de manera directa (en lugar de producir un análisis o un "asesoramiento" para los seres humanos encargados de tomar decisiones) sería, en efecto, considerado como un sistema de armas autónomo y, por lo tanto, plantearía problemas similares (v. la sección 3.1).

El empleo de los sistemas de apoyo a la toma de decisiones y de toma automatizada de decisiones también puede plantear **interrogantes jurídicos y éticos para otras aplicaciones, como las decisiones relativas a la detención en los conflictos armados**, que también tienen graves consecuencias para la vida de las personas y se rigen por normas específicas del derecho internacional humanitario. Aquí es posible establecer paralelos con los debates en el sector civil sobre el papel del criterio humano y los problemas de sesgos e inexactitud en los algoritmos de evaluación de riesgos utilizados por la policía en las decisiones sobre arrestos, así como en el sistema de justicia penal para las decisiones sobre condenas y libertad bajo fianza[22].

En términos más generales, el empleo de estos tipos de herramientas de IA y de aprendizaje automático podría traducirse en una mayor **personalización de la guerra** –haciendo un paralelo con la personalización de los servicios en el mundo civil–, mediante sistemas digitales que permitan reunir informaciones personales identificables a partir de múltiples fuentes –como sensores, comunicaciones, bases de datos, redes sociales y datos biométricos– para formar una determinación generada algorítmicamente sobre una persona, su condición y la posibilidad de que sea un objetivo, o para anticipar sus acciones futuras.

En general, el uso indebido de las tecnologías dotadas de IA para **vigilancia digital, monitoreo e intrusión** podría traer aparejadas consecuencias humanitarias –**riesgos digitales**– para las poblaciones civiles, como ataques, arrestos, malos tratos, robo de identidad y negación de acceso a los servicios, así como robo de activos o efectos psicológicos por temor a estar bajo vigilancia[23].

22 Lorna McGregor, "The Need for Clear Governance Frameworks on Predictive Algorithms in Military Settings", blog del CICR sobre derecho y políticas humanitarias (en inglés), 28 de marzo de 2019, disponible en https://blogs.icrc.org/ law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings; AI Now Institute, *AI Now Report* 2018, New York University, diciembre de 2018, pp. 18–22.

23 CICR, nota 16 supra, p. 8

4. Empleo de la IA y del aprendizaje automático para fines humanitarios

La IA y el aprendizaje automático también podrían emplearse para una amplia diversidad de usos en la labor humanitaria, por ejemplo, en el CICR. Actualmente, las organizaciones del sector estudian el empleo de estas herramientas para el análisis del entorno, así como para el seguimiento y el análisis de fuentes públicas de datos en contextos operacionales específicos. Estas aplicaciones podrían aportar la **información necesaria para evaluaciones de necesidades humanitarias**, como el tipo de asistencia necesaria (alimentos, agua, vivienda, economía, salud) y los lugares donde se la necesita.

Existen herramientas similares de agrupación y análisis de datos mediante IA que también podrían facilitar la **comprensión de las consecuencias humanitarias** en el terreno, incluidas las necesidades de la población civil en materia de protección. Podrían aplicarse, por ejemplo, herramientas de análisis de imágenes, videos u otros patrones para evaluar los daños a la infraestructura civil, los modelos de desplazamiento de la población, la viabilidad de cultivos alimenticios o el nivel de contaminación por armas (municiones sin estallar). Estos sistemas también podrían usarse en el análisis de imágenes y videos para identificar y evaluar la conducción de las hostilidades y sus consecuencias humanitarias.

El CICR, por ejemplo, ha elaborado **paneles de análisis del entorno** que utilizan IA y aprendizaje automático para captar y analizar grandes volúmenes de datos, a fin de fundamentar y apoyar su labor humanitaria en contextos operacionales específicos, incluido el uso de análisis predictivos para facilitar la evaluación de las necesidades humanitarias.

Existe una amplia diversidad de servicios humanitarios que podrían beneficiarse de la aplicación de herramientas de IA y de aprendizaje automático para tareas específicas. Por ejemplo, suscitan interés las tecnologías que podrían **mejorar la identificación de personas desaparecidas**, como el reconocimiento facial y el procesamiento del lenguaje natural para la coincidencia de nombres. El CICR estudia el uso de estas tecnologías en apoyo del trabajo de su Agencia Central de Búsquedas para reunir a miembros de familias separadas por conflictos armados. Asimismo, estudia la utilidad de la IA y del aprendizaje automático para **el análisis de imágenes y el reconocimiento de patrones a partir de imágenes satelitales**, ya sea para analizar la densidad de población en apoyo de proyectos de asistencia a la infraestructura en zonas urbanas o para mejorar su documentación en materia de respeto del derecho internacional humanitario como parte de su labor de protección de la población civil.

Estas **aplicaciones para la labor humanitaria también podrían conllevar riesgos** y plantean interrogantes jurídicos y éticos, en particular con respecto a la protección de datos, la privacidad, los derechos humanos y la rendición de cuentas, así como en lo que respecta a garantizar la participación humana en las decisiones que incidan de manera significativa en la vida y los medios de subsistencia de las personas. Toda aplicación de estas tecnologías en la labor humanitaria debe diseñarse y utilizarse bajo el principio de "**no hacer daño**" en el entorno digital, así

como respetar el derecho a la privacidad, entre otros aspectos, en lo que concierne a la protección de datos personales.

El CICR también procura que los **principios y valores fundamentales que rigen su labor humanitaria –neutralidad, independencia e imparcialidad–** se reflejen en el diseño y el uso de las aplicaciones de IA y de aprendizaje automático a las cuales recurre, teniendo en cuenta una evaluación realista de las capacidades y limitaciones de la tecnología (v. la sección 5.2). El CICR llevó adelante –en conjunto con el Brussels Privacy Hub– una iniciativa sobre protección de datos en la acción humanitaria para elaborar orientaciones sobre el uso de nuevas tecnologías, como la IA y el aprendizaje automático, en el sector humanitario, a fin de aprovechar al máximo los beneficios sin perder de vista estas consideraciones centrales. La segunda edición del *Manual sobre protección de datos en la acción humanitaria*, editado conjuntamente por el CICR y el Brussels Privacy Hub se publicó en mayo de 2020[24].

5. Un enfoque centrado en las personas

Como organización humanitaria cuya misión es proteger y asistir a las personas afectadas por los conflictos armados y otras situaciones de violencia, el CICR basa su cometido en el derecho internacional humanitario y se orienta por el Principio Fundamental de humanidad[25]. En ese marco, el **CICR considera crucial contar con un enfoque genuinamente centrado en las personas para el desarrollo y el empleo de la IA y el aprendizaje automático**. Este enfoque parte de la consideración de las obligaciones y responsabilidades de los seres humanos y de las condiciones necesarias para garantizar que el empleo de estas tecnologías sea compatible con el derecho internacional, así como con valores sociales y éticos.

5.1 Garantizar el control y el criterio humanos

El CICR considera que es **esencial preservar el control humano sobre las tareas, así como el criterio humano en las decisiones que pueden tener graves consecuencias** para la vida de las personas en los conflictos armados, en particular, en los casos en que estas tareas y decisiones presentan riesgos para la vida y se rigen por normas específicas del derecho internacional humanitario. **La IA y el aprendizaje automático deben estar al servicio de los actores humanos y mejorar su capacidad de tomar decisiones, no reemplazarlos**. Teniendo en cuenta que estas tecnologías se están desarrollando para realizar tareas que

24 CICR y Brussels Privacy Hub, *Manual sobre protección de datos en la acción humanitaria*, segunda edición, Ginebra, mayo de 2020, disponible en https://www.icrc.org/es/publication/manual-sobre-proteccion-de-datos-en-la-accion-humanitaria.

25 CICR y Federación Internacional de Sociedades de la Cruz Roja y de la Media Luna Roja, *Los Principios Fundamentales del Movimiento Internacional de la Cruz Roja y de la Media Luna Roja: ética y herramientas para la acción humanitaria*, Ginebra, noviembre de 2015, disponible en https://shop.icrc.org/the-fundamental-principles-of-the-international-red-cross-and-red-crescent-movement-pdf-es.html.

normalmente realizarían los seres humanos, existe una tensión inherente entre la búsqueda de aplicaciones de IA y de aprendizaje automático, y la centralidad del ser humano en un conflicto armado, tensión que habrá que observar de manera permanente.

El control y el criterio humanos revestirán particular importancia para las tareas y decisiones que pueden conllevar riesgos de lesiones o muerte, o daños o destrucción de la infraestructura civil. Este aspecto probablemente planteará los interrogantes jurídicos y éticos más difíciles, y puede exigir respuestas en materia de políticas, como nuevas normas y regulaciones. **Requieren particular atención las decisiones relativas al empleo de la fuerza, que determinan las personas y los bienes que serán objeto de un ataque en un conflicto armado**. Sin embargo, existe una variedad mucho más amplia de tareas y decisiones para las cuales se podría recurrir a la IA que también podrían tener graves consecuencias para las personas afectadas por conflictos armados, como las decisiones sobre arrestos y detenciones. En este sentido, al considerar el uso de la IA para tareas y decisiones delicadas, es posible extraer aprendizajes de debates más amplios en el sector civil sobre el control de las aplicaciones de IA que son esenciales para la seguridad, es decir, aquellas cuya falla puede causar heridas o la muerte, o graves daños materiales o ambientales[26].

Otra área de tensión es la **discrepancia entre seres humanos y máquinas en la velocidad de ejecución de diferentes tareas**. Dado que los seres humanos son los responsables jurídicos (y morales) en los conflictos armados, las tecnologías y herramientas que utilizan en la conducción de las hostilidades deben diseñarse y utilizarse de manera que los combatientes puedan cumplir con sus obligaciones y responsabilidades jurídicas y éticas. Este aspecto puede incidir de manera significativa en los sistemas de IA y de aprendizaje automático empleados en la toma de decisiones; a fin de preservar el criterio humano, puede ser necesario diseñar y emplear estos sistemas de modo que contribuyan a la toma de decisiones a una velocidad humana en lugar de acelerar las decisiones a la velocidad de la máquina, sin procurar la intervención humana.

Fundamentos jurídicos del control humano en los conflictos armados

Para garantizar el respeto del derecho, las partes en un conflicto deben mantener el control humano cuando se aplican la IA y el aprendizaje automático como medios y métodos de guerra. Las normas del derecho internacional humanitario están dirigidas a los seres humanos. Son los seres humanos quienes cumplen e implementan el derecho, y también son los responsables en caso de

26 V., por ejemplo, la Asociación sobre la IA, que centra la atención en la seguridad de la inteligencia artificial y del aprendizaje automático y la define como "una cuestión urgente para resolver en el corto plazo, la aplicación de estas tecnologías en la medicina, el transporte, la ingeniería, la seguridad informática, así como en otros ámbitos, que dependen de la capacidad de garantizar el funcionamiento seguro de los sistemas de inteligencia artificial pese a entornos inciertos, inesperados y potencialmente adversos". Partnership on AI, Safety-Critical AI: Charter, 2018 (en inglés), disponible en www.partnershiponai.org/working-group-charters-guiding-our-explorationof-ais-hard-questions.

violaciones de las normas. En particular, incumbe a los combatientes ejercer su criterio en virtud de las normas del derecho internacional humanitario que rigen la conducción de las hostilidades, y esta responsabilidad no puede transferirse a una máquina, un *software* o un algoritmo.

Estas normas exigen criterios específicos a la luz del contexto por parte de quienes planifican, deciden y realizan ataques, a fin de garantizar los siguientes principios: **distinción** entre objetivos militares, que pueden ser atacados lícitamente, y personas civiles y bienes de carácter civil, que no deben ser atacados; **proporcionalidad**, para garantizar que los daños incidentales que podrían causarse a personas civiles o bienes de carácter civil no resulten excesivos en relación con la ventaja militar directa y concreta que se prevé alcanzar; y **precauciones en el ataque**, de modo que los riesgos para las personas civiles puedan reducirse al mínimo.

Cuando se emplean sistemas de IA en ataques –ya sea como parte de sistemas de armas físicas o cibernéticas, o en sistemas de apoyo a la toma de decisiones–, **su concepción y uso deben permitir que los combatientes ejerzan ese criterio**[27]. Con respecto a los sistemas de armas autónomos, los Estados Partes en la Convención sobre ciertas armas convencionales (CCA) han reconocido que "el ser humano debe mantener la responsabilidad" respecto de decisiones sobre el uso de sistemas de armas y el uso de la fuerza[28], y muchos Estados, organizaciones internacionales –incluido el CICR– y organizaciones de la sociedad civil han enfatizado la exigencia del control humano para garantizar el cumplimiento del derecho internacional humanitario y la compatibilidad con los valores éticos[29].

Más allá del empleo de la fuerza y la definición de objetivos de ataque, el uso potencial de los sistemas de IA para otras decisiones regidas por normas específicas del derecho internacional humanitario probablemente requerirá una minuciosa consideración del nivel necesario de control y de criterio humanos, como en el ámbito de la detención[30].

Fundamentos éticos del control humano

Las aplicaciones emergentes de la IA y el aprendizaje automático también suscitan planteos éticos en el debate público. **Un aspecto común de los principios generales en relación con la IA** formulados y acordados por gobiernos, científicos,

27 CICR, v. nota 6.

28 Naciones Unidas, Informe del período de sesiones de 2018 del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales, documento de la ONU, CCW/GGE.1/2018/3, 23 de octubre de 2018, Sección III. A. 21(b) y III. C.23(f), disponible en https://undocs.org/es/CCW/GGE.1/2018/3.

29 V., por ejemplo, las declaraciones formuladas ante el Grupo de expertos gubernamentales sobre sistemas de armas autónomos letales de la Convención sobre ciertas armas convencionales (CCA) (en inglés), 25 al 29 de marzo de 2019, disponible en https://tinyurl.com/yyeadno3.

30 Tess Bridgeman, "The viability of data-reliant predictive systems in armed conflict detention", blog del CICR sobre derecho y políticas humanitarias, 8 de abril de 2019, disponible en https://blogs.icrc.org/law-and-policy/2019/ 04/08/viability-data-reliant-predictive-systems-armed-conflict-detention.

especialistas en ética, institutos de investigación y compañías tecnológicas **es la importancia que reviste el elemento humano** para garantizar el cumplimiento jurídico y la aceptabilidad ética.

Por ejemplo, los principios de Asilomar, definidos en 2017, ponen de relieve la importancia de que la IA esté en consonancia con los valores humanos, sea compatible con la "dignidad humana, los derechos, las libertades y la diversidad cultural", y dependa del control humano. Señalan que "los seres humanos deberían elegir si delegan decisiones a los sistemas de IA y de qué modo, para alcanzar los objetivos establecidos por seres humanos"[31]. El Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea hizo hincapié en la importancia de la "acción y supervisión humanas", de modo que los sistemas de IA deberían apoyar la autonomía y la toma de decisiones humanas, y garantizar la supervisión humana a través de enfoques de participación humana *(human-in-the-loop)*, control humano *(human-on-the-loop)* o mando humano *(human-in-command)*[32]. Los Principios sobre inteligencia artificial de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) –aprobados en mayo de 2019 por los 36 Estados miembros, junto con Argentina, Brasil, Colombia, Costa Rica, Perú y Rumania– destacan la importancia de los valores centrados en el ser humano y la equidad, especificando que los usuarios de IA deberían implementar mecanismos y garantías, como la capacidad de determinación humana, que sean apropiadas al contexto y acordes a los últimos avances en la materia[33]. Los Principios de Beijing sobre inteligencia artificial –aprobados en mayo de 2019 por un grupo de institutos de investigación y compañías tecnológicas de primer nivel de China– establecen que se deben hacer esfuerzos continuos para mejorar la madurez, la solidez, la confiabilidad y la capacidad de control de los sistemas de IA y propiciar estudios sobre la coordinación entre los seres humanos y la IA que permitan a los seres humanos aplicar plenamente sus ventajas y características humanas[34]. Varias compañías tecnológicas también han publicado en forma individual principios sobre la IA que destacan la importancia del control humano[35], en particular para

31 Future of Life Institute, "Asilomar AI Principles", 2017, disponible en https://futureoflife.org/ai-principles.

32 Comisión Europea, "Directrices éticas para una IA fiable", Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial, 8 de abril de 2019, pp. 19-20, disponible en https://op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.

33 Organización para la Cooperación y el Desarrollo Económicos (OCDE), "Recommendation of the Council on Artificial Intelligence", OECD/LEGAL/0449, 22 de mayo de 2019, disponible en https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

34 Beijing Academy of Artificial Intelligence (BAAI), "Beijing AI Principles" (en inglés), 28 de mayo de 2019, disponible en https://baip.baai.ac.cn/en.

35 Google, "AI at Google: Our Principles", *The Keyword*, 7 de junio de 2018, disponible en www.blog.google/technology/ai/ai-principles. "Diseñaremos sistemas de inteligencia artificial que permitan la oportunidad de formular comentarios, proveer explicaciones pertinentes y apelar decisiones. Nuestras tecnologías de IA estarán sujetas a la gestión y el control humanos pertinentes".

aplicaciones sensibles que presentan el riesgo de causar daño[36], y enfatizan que el propósito de la IA es amplificar, no reemplazar, la inteligencia humana[37].

Algunos gobiernos también están formulando principios sobre IA para uso militar. Por ejemplo, el Departamento de Defensa de Estados Unidos, que reivindicó la adopción de la IA "centrada en las personas" en su estrategia relativa a la IA de 2018[38], ha confiado a su comité de innovación en defensa la formulación de recomendaciones. Entre ellas, se destaca que los seres humanos deben actuar con criterios que estén a la altura y procurar mantener su responsabilidad en cualquier ocasión en que se emplee IA[39], que constituyó una base para el primero de los cinco principios que el departamento de Defensa aprobó en 2020, que establece que la IA debe ser "responsable". El personal del Departamento de Defensa actuará con criterios y cuidados que estén a la altura, a la vez que se hacen responsables por el desarrollo, despliegue y empleo de las capacidades de IA[40]. En Francia, el Ministerio de Defensa se ha comprometido a utilizar la IA de acuerdo con tres principios rectores –el respeto del derecho internacional, el mantenimiento de un control humano suficiente, y la responsabilidad de mando permanente–, y establecerá un comité de ética ministerial encargado de tratar las cuestiones relacionadas con las tecnologías emergentes[41].

En opinión del CICR, preservar el **control humano** sobre las tareas, así como el **criterio humano** en las decisiones que tienen graves consecuencias para la vida de las personas, también será **esencial para preservar un nivel de humanidad en la guerra. El CICR también ha señalado la necesidad de mantener la acción humana sobre las decisiones en materia del uso de la fuerza en un conflicto armado**[42], una opinión que deriva de consideraciones éticas más amplias de humanidad, responsabilidad moral, dignidad humana y los dictados de la conciencia pública[43].

Sin embargo, las consideraciones éticas de la acción humana pueden tener una aplicabilidad más amplia para otros usos de la IA y del aprendizaje

36 Microsoft, "Microsoft AI Principles", 2019, disponible en www.microsoft.com/en-us/ai/our-approach-to-ai; Rich Sauer, "Six Principles to Guide Microsoft's Facial Recognition Work", blog de Microsoft , 17 de diciembre de 2018, disponible en https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsoftsfacial-recognition-work.

37 IBM, "IBM's Principles for Trust and Transparency" , THINKPolicy Blog, 30 de mayo de 2018, disponible en www.ibm.com/blogs/policy/trust-principles.

38 Departamento de Defensa de EE. UU., Summary of the 2018 Department of Defense Artificial Intelligence Strategy, 2019.

39 Departamento de Defensa de EE. UU., comité de innovación en defensa, AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense, 31 de octubre de 2019.

40 Departamento de Defensa de EE. UU., "DOD Adopts Ethical Principles for Artificial Intelligence"(en inglés), comunicado de prensa, 24 de febrero de 2020, disponible en www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificialintelligence/.

41 Ministerio de Defensa de Francia, "Florence Parly desea una Inteligencia artificial de alto rendimiento, robusta y bien dominada", Actualités, 10 de abril de 2019, disponible en www.defense.gouv.fr/english/actualites/articles/florence-parly-souhaite-une-intelligence-artificielle-performante-robuste-et-maitrisee.

42 CICR, Estrategia del CICR para el período 2019–2022, Ginebra, 2018, p. 15, disponible en https://shop.icrc.org/icrc/pdf/view/id/2867.

43 CICR, nota 8 supra, p. 22.

automático en conflictos armados y otras situaciones de violencia. En ese sentido, podría ser conveniente capitalizar **las enseñanzas de debates sociales más amplios sobre aplicaciones sensibles de las tecnologías de IA y aprendizaje automático capaces de tener un doble uso**, en particular, en aplicaciones esenciales para la seguridad, así como propuestas relativas a la gobernanza por parte de científicos y desarrolladores en el sector privado. Google, por ejemplo, ha dicho que puede haber contextos sensibles donde la sociedad desee que un ser humano tome la decisión final, sin importar cuán preciso sea un sistema de IA, y que el hecho de delegar completamente las decisiones de alto riesgo a las máquinas –como decisiones jurídicas sobre responsabilidad penal o decisiones vitales en materia de un tratamiento médico– podría legítimamente considerarse como una afrenta a la dignidad humana[44]. Microsoft, al considerar el reconocimiento facial basado en la IA ha puesto énfasis en garantizar un nivel adecuado de control humano para usos que pueden afectar a las personas de manera considerable, y aplicar un enfoque de participación humana *(human-in-the-loop)* o de "revisión humana significativa" para usos sensibles, como aquellos que conllevan riesgos de daño corporal o emocional a una persona, disminución de las perspectivas de empleo de una persona o de su capacidad de acceder a servicios financieros, consecuencias en el plano de sus derechos humanos, o posible limitación de sus libertades individuales[45]. Dado que las aplicaciones en los conflictos armados probablemente se encuentren entre las más sensibles, estos debates más profundos pueden aportar enseñanzas sobre las restricciones necesarias en las aplicaciones de la IA.

El hecho de mantener **el control y el criterio humanos será un componente esencial** para promover el cumplimiento del derecho y mitigar las preocupaciones éticas planteadas por ciertas aplicaciones de la IA y del aprendizaje automático. **Pero, en sí, no será suficiente para protegerse contra riesgos potenciales** sin la debida consideración de los problemas de interacción entre personas y máquinas, como **el conocimiento de la situación** (conocimiento del estado del sistema en el momento de la intervención humana); **el tiempo disponible** para una intervención humana eficaz; **el sesgo sobre la automatización** (riesgo de un exceso de confianza en el sistema); y **la restricción moral** (riesgo de transferir la responsabilidad al sistema)[46]. Además, para asegurar un control y un criterio humanos pertinentes y eficaces, será necesario un minucioso análisis tanto de las capacidades como de las limitaciones de las tecnologías de IA y de aprendizaje automático.

44 Google, "Perspectives on Issues in AI Governance" (en inglés), enero de 2019, p. 23–24, disponible en http://ai.google/perspectives-on-issues-in-AI-governance.

45 R. Sauer, nota 36 supra: "Alentaremos y ayudaremos a nuestros clientes a desplegar la tecnología de reconocimiento facial de manera tal que garantice un nivel adecuado de control humano cuando su uso pueda tener consecuencias importantes para las personas".

46 CICR, nota 8 supra, p. 13.

5.2 Comprender las limitaciones técnicas de la IA y del aprendizaje automático

Si bien se habla bastante de las nuevas posibilidades que ofrecen la IA y el aprendizaje automático, **es necesario efectuar una evaluación realista de las capacidades y limitaciones de estas tecnologías**, en particular para su aplicación en los conflictos armados. Ante todo, es necesario tomar conciencia de que, al emplear la IA y el aprendizaje automático para ciertas tareas o decisiones, no se hace un reemplazo equivalente entre pares. Se requiere una **comprensión de las diferencias fundamentales entre la forma en que actúan los seres humanos y las máquinas, así como sus respectivas fortalezas y debilidades**. Las personas y las máquinas hacen las cosas de manera diferente y, de hecho, hacen cosas diferentes. Debemos tener presente que, como objetos inanimados y herramientas para el uso de los seres humanos, las máquinas nunca podrán tener interacciones genuinamente humanas, más allá de su capacidad de simulación[47].

Desde esta perspectiva, varios aspectos técnicos requieren precaución al considerar sus aplicaciones en los conflictos armados (y, en efecto, para la acción humanitaria). **La IA, y particularmente el aprendizaje automático, generan preocupaciones sobre su imprevisibilidad y su falta de fiabilidad** (o de seguridad)[48], **su falta de transparencia** (o explicabilidad) y **los sesgos**[49].

Más que seguir una secuencia de consignas preprogramada, **los sistemas de aprendizaje automático crean sus propias reglas sobre la base de los datos a los que están expuestos**, ya sea datos utilizados para su formación o adquiridos mediante prueba y error con su entorno. **Como consecuencia, son mucho más impredecibles** que los sistemas preprogramados en lo que concierne a la manera en que funcionarán (obtener un resultado) en una situación dada (con insumos específicos), y su funcionamiento depende en gran medida de la cantidad y la calidad de los datos disponibles para una tarea específica. Para quien lo desarrolla, es difícil saber cuándo termina la fase de aprendizaje, o incluso qué ha aprendido el sistema. Un mismo sistema de aprendizaje automático puede responder de manera diferente incluso ante dos situaciones idénticas; además, algunos sistemas pueden arrojar soluciones imprevistas para una tarea en particular[50]. Estos problemas fundamentales se exacerban cuando el sistema continúa "aprendiendo" y modificando su modelo después de la implementación para una tarea específica. El carácter impredecible de los sistemas de aprendizaje automático, que puede ser una ventaja para resolver tareas y no plantear ningún problema para tareas "benignas",

47 Google, nota 44 supra, p. 22

48 Dario Amodei *et al.*, *Concrete Problems in AI Safety*, Cornell University, Ithaca, NY, 2016, disponible en https://arxiv.org/abs/1606.06565.

49 CICR, *Autonomy, Artificial Intelligence and Robotics: An Ethical Basis for Human Control* (en inglés), informe de una reunión de expertos, Ginebra, agosto de 2019, disponible en www.icrc.org/en/document/autonomy-artificialintelligence-and-robotics-technical-aspects-human-control.

50 Joel Lehman *et al.*, *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities* (en inglés), Cornell University, Ithaca, NY, 2018, disponible en https://arxiv.org/abs/1803.03453.

como jugar un juego de mesa[51]. Por el contrario, puede ser una preocupación importante para las aplicaciones militares en conflictos armados, como los sistemas de armas autónomos, la guerra cibernética y los sistemas de apoyo a las decisiones (v. las secciones 3.1 y 3.3).

Para complicar aún más las cosas, numerosos sistemas de aprendizaje automático **no son transparentes: producen resultados que no son explicables**. Esta naturaleza de "caja negra" hace que para el usuario sea difícil y hoy en día, a menudo imposible, comprender *cómo y por qué* el sistema obtiene un resultado desde un insumo determinado. En otras palabras, no se puede explicar ni interpretar.

En función de estos problemas de imprevisibilidad y de falta de explicabilidad, **se hace sumamente difícil lograr confianza en los sistemas de IA y de aprendizaje automático**. Existe un problema adicional para la confianza: el **sesgo**, que puede tener muchas facetas, ya sea reforzando los sesgos humanos existentes o introduciendo otros nuevos en el diseño o el uso del sistema. Un tipo de sesgo frecuente se relaciona con los datos utilizados para el entrenamiento de los sistemas, en la medida en que las limitaciones en la cantidad, la calidad y la naturaleza de los datos disponibles para entrenar un algoritmo para una tarea específica pueden generar un sesgo en el funcionamiento del sistema en relación con su tarea. Este puede ser un problema importante para las aplicaciones en conflictos armados, donde escasean los datos representativos de alta calidad para tareas específicas. Sin embargo, existen otros tipos de sesgo que pueden derivarse de la ponderación que el sistema da a los diferentes elementos de los datos, o a su interacción con el entorno durante la ejecución de una tarea[52].

Las dificultades planteadas por la imprevisibilidad, la falta de transparencia o de explicabilidad, así como por el sesgo, se han documentado en diversas aplicaciones de la IA y el aprendizaje automático, por ejemplo, los sistemas de reconocimiento de imágenes[53], de reconocimiento facial[54] y de toma de decisiones automatizadas[55]. Asimismo, determinadas aplicaciones de la IA y del aprendizaje automático, como la visión artificial, plantean otro problema fundamental: **la brecha semántica**. Este tipo de problemas muestra que los seres humanos y las máquinas llevan a cabo tareas de manera muy diferente[56]. Un algoritmo de visión artificial entrenado para reconocer imágenes de temas particulares puede ser capaz de identificar y clasificar esos temas en una nueva imagen. Sin embargo, el algoritmo no comprende el significado o el concepto de ese tema, por lo cual puede cometer errores que un ser humano

51 David Silver *et al*., "Mastering the Game of Go without Human Knowledge", *Nature*, vol. 550, n.º 7676 (en inglés), 19 de octubre de 2017

52 UNIDIR, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer* (en inglés), UNIDIR, 2018.

53 Matthew Hutson, "A turtle – or a rifle? Hackers easily fool AIs into seeing the wrong thing" (en inglés), *Science*, 19 de julio de 2018, disponible en www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-aisseeing-wrong-thing.

54 AI Now Institute, nota 22 supra, pp. 15–17.

55 Ibíd., pp. 18–22.

56 Arnold W. M. Smeulders *et al*., "Content-Based Image Retrieval at the End of the Early Years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n.º 12 (en inglés), 2000.

nunca cometería, como clasificar un objeto como algo completamente diferente y no relacionado con su naturaleza real. Obviamente, esto generaría problemas graves en ciertas aplicaciones en conflictos armados, como en sistemas de armas autónomos o sistemas de apoyo a la toma de decisiones para el ataque (v. las secciones 3.1 y 3.3).

Probablemente, será aún más difícil confiar en la IA y el aprendizaje automático en situaciones de conflicto armado en las que se puede suponer que los adversarios aplicarán contramedidas destinadas a engañar o burlar los sistemas de los demás. **Los sistemas de aprendizaje automático son particularmente vulnerables a las condiciones antagonistas**, ya sean modificaciones al entorno diseñadas para engañar al sistema o el uso de otro sistema de aprendizaje automático para producir imágenes o condiciones antagonistas (red antagonista generativa o GAN, *generative adversarial network*). En un ejemplo bien conocido, los investigadores indujeron a un algoritmo de clasificación de imágenes a identificar una tortuga impresa en 3D como un "fusil", y una pelota de béisbol impresa en 3D como un "espresso"[57]. Los riesgos de los problemas de este tipo también son claros si un sistema de reconocimiento de imágenes basado en IA se empleara en sistemas de armas o para la toma de decisiones relativas a los objetivos de ataque.

6. Conclusiones y recomendaciones

La IA y los sistemas de aprendizaje automático podrían tener **profundas consecuencias en relación con el papel que desempeñan los seres humanos en los conflictos armados**, en particular por la creciente autonomía de los sistemas de armas y otros sistemas no tripulados; las nuevas formas de la guerra cibernética y de la información; y, en líneas más generales, la naturaleza de la toma de decisiones. El CICR considera que los gobiernos, las fuerzas armadas y otros actores pertinentes en los conflictos armados deben adoptar **un enfoque genuinamente centrado en las personas para el empleo de sistemas de IA y de aprendizaje automático, que esté basado en obligaciones jurídicas y responsabilidades éticas**. El empleo de la IA en sistemas de armas debe abordarse con mucho cuidado.

Como principio general, es **esencial preservar la capacidad de control y de discernimiento de los seres humanos cuando se aplica la IA y el aprendizaje automático en tareas y decisiones que pueden tener consecuencias graves para la vida de las personas**, en particular cuando estas tareas y decisiones plantean riesgos para la vida y cuando se rigen por normas específicas del derecho internacional humanitario. A fin de cuentas, **los sistemas de IA y de aprendizaje automático no dejan de ser herramientas cuyo uso debe servir a los actores humanos y ampliar su capacidad para tomar decisiones, no reemplazarlos.**

Será necesario garantizar el control y el criterio humanos en los sistemas físicos y digitales basados en IA que presenten tales riesgos para lograr el cumplimiento del derecho internacional humanitario y, desde una

57 M. Hutson, nota 53 supra.

perspectiva ética, preservar una cierta humanidad en los conflictos armados. Para que las personas desempeñen su papel de manera significativa, tal vez sea necesario diseñar y utilizar estos sistemas para **fundamentar la toma de decisiones a la velocidad humana, en lugar de acelerar las decisiones a la velocidad de la máquina**, eludiendo la intervención humana. En última instancia, estas consideraciones pueden conllevar restricciones en el diseño y el uso de sistemas de IA y de aprendizaje automático para permitir el ejercicio de un control y de un criterio humanos que sean significativos y eficaces, y se funden en obligaciones jurídicas y responsabilidades éticas.

Un principio general de control y de criterio humanos es un componente esencial, pero no es suficiente en sí para protegerse contra los riesgos potenciales de la IA y del aprendizaje automático en los conflictos armados. **Otros aspectos relacionados para considerar** son la **previsibilidad** y la **fiabilidad** –o seguridad– en el manejo del sistema y las consecuencias derivadas; la **transparencia** –o **explicabilidad**– del funcionamiento del sistema y las razones por las cuales alcanza un resultado particular; y la **falta de sesgo** –o **imparcialidad**– en el diseño y el empleo del sistema. Estos problemas deberán resolverse para **instaurar confianza** en el empleo de un sistema determinado, incluso mediante **pruebas rigurosas en entornos realistas** antes de su puesta en funcionamiento[58].

La naturaleza de la interacción necesaria entre los seres humanos y la IA probablemente dependerá de consideraciones éticas y de las normas específicas del derecho internacional humanitario y otras normas aplicables en las circunstancias del caso. Por ende, **es posible que los principios generales deban complementarse con principios, directrices o normas de carácter específico en relación con el uso de la IA y del aprendizaje automático para aplicaciones y circunstancias particulares**.

En opinión del CICR, una de las preocupaciones más acuciantes es la relación entre seres humanos y máquinas en las decisiones de matar, herir, dañar y destruir, y la **importancia crucial de garantizar el control humano sobre los sistemas de armas y el uso de la fuerza** en los conflictos armados. Dada la autonomía creciente de los sistemas de armas –independientemente de si están dotados de IA–, existe el riesgo de que estas decisiones queden efectivamente libradas a sensores y algoritmos, una perspectiva que plantea preocupaciones jurídicas y éticas que deben resolverse sin demora.

El CICR ha propuesto elementos clave de control humano necesarios para respetar el derecho internacional humanitario y atender las preocupaciones éticas como base para los límites internacionalmente acordados sobre la autonomía de los sistemas de armas, que incluyen controles en los parámetros de armas,

58 Netta Goussac, "Safety net or tangled web: Legal reviews of AI in weapons and war-fighting", blog del CICR sobre derecho y políticas humanitarias (en inglés), 18 de abril de 2019, disponible en https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting; Dustin A. Lewis, "Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider", blog del CICR sobre derecho y políticas humanitarias (en inglés), 21 de marzo de 2019, disponible en https://blogs.icrc.org/law-andpolicy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elementsconsider.

controles en el entorno y controles a través de la interacción entre personas y máquinas[59]. Es evidente para el CICR que se necesitan límites a los tipos de armas autónomas empleadas y a las situaciones en las que se las emplean[60].

Este **enfoque basado en el control humano** para los sistemas de armas autónomos **también sería pertinente para aplicaciones más generales de la IA y del aprendizaje automático en la toma de decisiones en los conflictos armados**, en particular cuando existan riesgos significativos para la vida humana y normas específicas del derecho internacional humanitario aplicables, como el uso de sistemas de apoyo a las decisiones para la determinación de los objetivos y la detención.

59 CICR, Comentario sobre los "Principios rectores", nota 9 supra; CICR, "The Element of Human Control", nota 9 supra; V. Boulanin *et al*., nota 9 supra.

60 CICR, Declaraciones del CICR ante el Grupo de expertos gubernamentales sobre sistemas de armas autónomos letales de la Convención sobre ciertas armas convencionales (CCA) (en inglés), Ginebra, 21 al 25 de septiembre de 2020, disponible en https://documents.unoda.org/wp-content/uploads/2020/09/20200921-ICRC-General-statement-CCW-GGE-LAWS-Sep-2020.pdf.