universidad de león

FACULTAD DE DERECHO
DEPARTAMENTO DE DERECHO PÚBLICO

TESIS DOCTORAL

CIBERÉTICA, AGENTES MORALES ARTIFICIALES Y RESPONSABILIDAD JURÍDICA INTERNACIONAL

ANTONIO PEDRO MARÍN MARTÍNEZ

PROGRAMA RESPONSABILIDAD JURÍDICA. ESTUDIO MULTIDISCIPLINAR

DIRECTOR: Dr. MIGUEL DÍAZ Y GARCÍA-CONLLEDO
TUTOR: Dra. MARTA ZUBIAUR GONZÁLEZ

LEÓN, 2021

noviembre, 2021

A mis padres

por confiar en mí

ÍNDICE

ABREVIATURAS

ALife	Vida Artificial
AA	Agente Artificial
AMA	Agente Moral Artificial
AMAA	Agente Moral Artificial Autónomo
AWS	Autonomous Weapons Systems
Bot	Robot
CAI	Conflicto Armado Internacional
CANI	Conflicto Armado No Internacional
CCW	Convención sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales que puedan considerarse excesivamente nocivas o de efectos indiscriminados (NU)
CDI	Comisión de Derecho Internacional
CICR	Comité Internacional de la Cruz Roja
CIJ	Corte Internacional de Justicia
CUE	Consejo de la Unión Europea
DCA	Leyes sobre los conflictos armados
DD.HH.	Derechos Humanos
DICA	Derecho Internacional de los Conflictos Armados

DIH	Derecho Internacional Humanitario
DoD	Departamento de Defensa de USA
EMAD	Estado Mayor de la Defensa (España)
ENISA	Agencia Europea de Seguridad
GGE	Grupo de Expertos Gubernamentales (NU)
IA	Inteligencia Artificial
ICTY	Tribunal Penal Internacional para la ex Yugoslavia
IEEE	Institute of Electrical and Electronic Engineers
IGA	Inteligencia General Artificial
IHM	Interfaz Hombre Máquina
IIR	Innovación e Investigación Responsable
IoT	Internet de las Cosas
LAWS	Lethal Autonomous Weapons Systems
LOAR	Ley de Retornos Acelerados
MHC	Control Humano Significativo
MIT	Massachusetts Institute of technology
NU	Naciones Unidas
OCDE	Organización para la Cooperación y el Desarrollo Económico

ONG	Organización no Gubernamental
OTAN	Organización del Tratado del Atlántico Norte
PAI	Protocolo Adicional I (1977) Convención de Ginebra
PAII	Protocolo Adicional II (1977) Convención de Ginebra
PAIII	Protocolo Adicional III (1977) Convención de Ginebra
RNA	Red Neuronal Artificial
ROE	Reglas de Enfrentamiento
SAA	Sistemas Armamentísticos Autónomos (AWS)
SAAL	Sistemas Armamentísticos Autónomos Letales (LAWS)
SIA	Super Inteligencia Artificial
SI	Sistemas de Información
SIPRI	Stockholm International Peace Research Institute
TIC	Tecnologías de la Información y la Comunicación
UE	Unión Europea
UK	Reino Unido
UNESCO	Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

UNIDIR	Instituto de las Naciones Unidas de Investigación sobre el Desarme
USA	Estados Unidos de América
USCYBERCOM	Comando Cibernético de los Estados Unidos de América
VA	Vehículo Autónomo
VANT	Vehículos Aéreos No Tripulados (UAV)
WWW	World Wide Web

PARTE I

INTRODUCCIÓN

CAPÍTULO I

PLANTEAMIENTOS GENERALES

En el ámbito militar los Sistemas Armamentísticos Autónomos (SAA) que utilizan tecnologías asociadas con la Inteligencia Artificial (IA) y la robótica están cada vez más presentes en el campo operacional. En muchas instancias, el progreso científico no siempre ha venido acompañado de una adecuación de la actual normativa del Derecho Internacional o de los Estados al nuevo paradigma del ciberespacio. Además, aunque los actuales sistemas armamentísticos tienen algún grado de operatividad sin la intervención humana, el rápido desarrollo de la investigación tecnológica expandirá dichas capacidades en los próximos años estresando aún más una adecuada implementación del Derecho en el ámbito militar del ciberespacio, especialmente con relación a la guerra y más concretamente a la ciberguerra, lo que propiciará nuevas cuestiones éticas y jurídicas relacionadas con dicho dominio, con la necesidad de establecer soluciones apropiadas a dichas cuestiones.

1.1.- ANTECEDENTES Y SITUACIÓN ACTUAL

La expansión de la autonomía en los SAA, particularmente en aquellos con capacidades letales (SAAL), ha llevado a ciertas organizaciones no gubernamentales, activas desde 2012, a llevar a cabo llamamientos para establecer una prohibición preventiva en el desarrollo, despliegue y uso de dichos sistemas armamentísti-

cos[1]. A dichos esfuerzos se han unido un grupo de Estados que abogarían por la prohibición de los SAAL completamente autónomos y que no poseyesen un "Control Humano Significativo" (MHC). No obstante, dichas posiciones no están exentas de controversia, pues investigadores del ámbito militar, así como de algunos Estados, preconizan que el incremento del grado de autonomía en los sistemas armamentísticos incrementaría su efectividad operacional, funcionarían de forma más ética que los seres humanos y disminuirían el daño colateral y sobre los propios efectivos, al mismo tiempo que incrementarían el ahorro financiero y un uso más racional del personal humano (Arkin, 2007: 7; Byrnes, 2014: 54; HRW, 2020a: 4).

En todo caso, las implicaciones éticas, jurídicas e internacionales de los nuevos SAA y SAAL han propiciado que las Naciones Unidas (NU), a través de la "Convención sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales que puedan considerarse excesivamente nocivas o de efectos indiscriminados" (CCW) estableciesen, en un primer lugar en 2013, la necesidad de llevar a cabo una reunión informal, en 2014, para discutir elementos relacionados con las tecnologías emergentes en el área de los SAAL. A partir de 2016 se estableció un "Grupo de Expertos Gubernamentales" (GGE) formal sobre dichos sistemas armamentísticos, que se ha reunido a partir de entonces con

1 Una de las campañas más activas a nivel internacional y que aún perdura, ha sido la titulada "Prohibición de Robots Asesinos" (*Stop Killer Robots*). Dicha campaña está propiciada por una alianza de organizaciones no gubernamentales (ONGs), entre las que se encuentran Article 36, Amnistía Internacional y Human Rights Watch, cuya misión es trabajar para prohibir las armas completamente autónomas, estableciendo la necesidad del control humano sobre el uso de la fuerza (Stopkillerrobots, 2021).

regularidad[2] y que en 2019 desarrollaron una serie de principios rectores que, entre otros, establecía que el Derecho Internacional Humanitario (DIH) era de completa aplicación a todos los sistemas armamentísticos, incluidos aquellos desarrollados con diversos grados de autonomía y potencialmente utilizables, especialmente aquellos con capacidad letal (SAAL). Paralelamente, en el ámbito académico, gubernamental y militar, se han venido desarrollando estudios y propuestas teóricas y prácticas para incluir ciertos elementos éticos, así como del Derecho consuetudinario, en el desarrollo de dichos sistemas, aunque también se han multiplicado los llamamiento para rechazar los SAAL completamente autónomos.

1.2.- OBJETIVOS DE LA INVESTIGACIÓN

El análisis de las diversas controversias y los trabajos llevados a cabo a nivel institucional internacional, estatal o académico, han continuado a ritmo cada vez más acelerado, principalmente a partir del siglo XXI. No obstante, la mayor parte de dichos trabajos han sido planteamientos dentro del marco filosófico y del Derecho consuetudinario con propuestas teóricas, algunas de ellas repetitivas, acentuando la brecha entre utopía y realidad, que en ciertos momentos parecen más una lista de deseos que unos planteamientos realistas, especialmente cuando se intenta plasmar tanto el Derecho Internacional vigente como el consuetudinario en el ciberespacio, específicamente en los nuevos sistemas armamentísticos, con cierto grado de autonomía (SAA), a través de algoritmos computacionales, que se sirven de la IA y la robótica.

2 A excepción de 2020, debido al desarrollo de la pandemia sanitaria del "Covid 19".

Por lo tanto, los objetivos de nuestra investigación tienen como premisa el proponer planteamientos realistas, en el marco temporal actual, para la aplicación práctica de la responsabilidad jurídica internacional a los sistemas armamentísticos letales con diversos grados de autonomía (SAAL), a través de una aproximación pragmática, dentro de un proceso progresivo, marcada por las siguientes pautas:

- Un estudio actualizado del marco conceptual, relativo a los principales elementos que inciden en el Derecho Internacional de los SAAL: la ciberética, la ciberguerra, la responsabilidad penal internacional y los agentes morales artificiales (AMA), que nos permitirá explorar las posibles propuestas con una base de conocimiento amplia de los principales conceptos subyacentes a las mismas;

- Una visión panorámica de la situación actual de la responsabilidad jurídica internacional, en el marco de los nuevos sistemas armamentísticos, a través de las siguientes facetas: el impacto de un hipotético alcance de la Singularidad en los sistemas armamentísticos, la relación de los SAAL con el DIH y los Derechos Humanos (DD.HH.) (especialmente sobre los principales principios relativos al Derecho Internacional de los Conflictos Armados (DICA)) y las posibles restricciones o soluciones que se barajan a nivel global, para adaptar el Derecho a la nueva realidad del ciberespacio, incluido la posible utilización de nuevos elementos para su implementación, como los Agentes Morales Artificiales (AMA);

- La propuesta teórica de desarrollo de un AMA para un sistema armamentístico SAAL, cuya base se apoyaría en el Derecho Internacional y consuetudinario vigentes, así como en los principios

consensuados a nivel institucional (NU), relativos a la implementación de la responsabilidad jurídica internacional en los SAAL, incidiendo sobre los aspectos computacionales y su implementación práctica.

La intención, de fondo, será comprender los pasos necesarios para aplicar el Derecho Internacional en los SAAL a través de los nuevos vectores de los AMA, a la vez que se examinan las dificultades que existen para su implantación práctica, tanto a nivel jurídico como algorítmico y computacional. Dicho proceso identificará la necesidad de nuevas normas o la adaptación de las ya existentes, pero también la complejidad algorítmica y computacional de dicho AMA.

1.3.- METODOLOGÍA

La tesis se ha escrito en términos teóricos relativos, sin incidir en casuísticas particulares, aunque se apoye en escenarios concretos y en el impacto que los diversos grados de autonomía de los sistemas armamentísticos modernos inciden en el Derecho Internacional vigente, dado que la Responsabilidad Jurídica Internacional es nuestra base programática, relativa al ámbito de los conflictos armados entre Estados. También se circunscribe a los principales principios rectores que los Estados han identificado como básicos para los SAAL y su impacto sobre el DIH: distinción, proporcionalidad, responsabilidad, humanidad y necesidad militar, especialmente con relación a la "focalización de objetivos" (*targeting*) e "intervención sobre objetivos" (*engagement*). Se apoya en publicaciones filosóficas, sociotécnicas, técnicas, de procedencia académica, de los Estados o militares, especialmente del ámbito internacional,

con especial énfasis en los principales actores globales (China, Rusia y USA), instituciones internacionales (NU, OTAN, UE) y una especial mención del entorno español.

Incidiremos en la idea de que nuestra investigación no pretende argumentar una forma particular de entender como el Derecho debería ser aplicado en las nuevas tecnologías que inciden en el ámbito de los conflictos armados, pero sí aportar nuevas ideas para dicha aplicación. Particularmente, donde sea posible, se intentará utilizar las interpretaciones más comúnmente aceptadas, sobre cómo aplicar el Derecho Internacional a los SAAL. También se hará una especial referencia al Derecho consuetudinario y su interpretación, en el ámbito de los sistemas armamentísticos, por parte reputadas instituciones internacionales como el Comité Internacional de la Cruz Roja (CICR) o la OTAN, aunque sean objeto de amplios debates sobre su precisión y posibles sesgos intencionados.

Sobre el desarrollo de la propuesta de un AMA para un SAAL, utilizaremos como base teórica estándares internacionales propuestos por instituciones internacionales (IEEE, OCDE, UE), así como publicaciones de los principales investigadores en dicha materia. No obstante, la propuesta será propia y sujeta a cierto grado de subjetividad, pero siempre apoyada en planteamientos teóricos ya existentes, que consideremos los más apropiados para nuestro objetivo.

1.4.- LIMITACIONES TEÓRICAS Y PRÁCTICAS

La principal limitación que ha incidido sobre nuestra investigación ha sido el exceso de información. De un único repositorio digital internacional (*Academia edu*), en el periodo comprendido entre los años 2000 a 2021, hemos identificado más de 43.000 referencias

sobre AMA, cerca de 22.000 sobre ciberética o más de 6.000 relativos a los SAAL, aunque muchos de ellos sean repetitivos de los mismos planteamientos teóricos. Pero también existen repositorios clave, tanto filosóficos, jurídicos o técnicos, de instituciones internacionales como las NU (como el UNIDIR), la OTAN o la UE o de Estados de la importancia de China, Rusia o USA. Por lo tanto, hemos tenido que llevar a cabo un ejercicio de priorización y triaje de aquellas publicaciones que hemos considerado básicas de investigadores reputados o de los principales Estados e instituciones a nivel internacional y de España, para que nuestra investigación fuese manejable y eminentemente práctica.

Una segunda limitación proviene del marco de la seguridad nacional, que engloba todo lo relativo a sistemas armamentísticos, especialmente los SAAL. Nuestro punto de vista es que estamos seguros de que debe existir una amplia gama de información y trabajos de los Estados, entidades públicas y privadas a los cuales no se tiene libre acceso, pero que sabemos inciden de una forma fundamental en las conclusiones de nuestra investigación. Una pequeña parte de dicha información la hemos podido conocer de forma indirecta, pero la mayoría no han estado a nuestro alcance, por lo que nuestros planteamientos se han circunscrito a aquella información de acceso público existente, lo que nos ha llevado a planteamientos y propuestas teóricas básicas de muy alto nivel.

Por último, tampoco nos hemos adentrado en los mecanismos específicos de software y hardware de los algoritmos y módulos de interacción hombre-máquina propuestos, para la implementación de la responsabilidad jurídica internacional dentro de los SAAL, ya que consideramos que necesitarían de una ingente labor de ingeniería informática, con proyectos que involucrarían grandes re-

cursos de tiempo, técnicos y humanos, que sobrepasarían ampliamente el marco de nuestra investigación.

PARTE II

MARCO CONCEPTUAL
Y
TERMINOLÓGICO

CAPÍTULO 2

CIBERÉTICA

Una de las preguntas principales que se plantean los investigadores es si la tecnología informática establece nuevos principios éticos que difieren de los elementos tradicionales de la ética. La investigadora D. G. Johnson es de la opinión que la especificidad de dicha tecnología es la razón principal de que se haya planteado la necesidad de un nuevo paradigma: la ciberética. Para dicha investigadora la informática permite la realización de tareas que antes no se podían desarrollar y el desarrollo de otras de forma diferente. Dicha situación implicaría que los ámbitos de la ética tradicional también serían transformados por las nuevas tecnologías, aunque no significaría un cambio sobre los conceptos éticos tradicionales. Dicha postura concuerda con la indicada por J. Wizenbaum en 1950, cuando estableció que los nuevos casos éticos podrían ser asimilados a través de la aplicación de prácticas, leyes, reglas y principios que regulan el comportamiento humano ya existentes. En contraste, el investigador W. Maner sería de la opinión de que la intervención de las computadoras en la conducta humana puede crear elementos éticos totalmente nuevos, específicos para la computación, que no están presentes en otras áreas (Bynum, 2015; Johnson, 2005: 608; Maner, 1999: 1).

El problema subyacente, no obstante, es que a nivel global existe un pluralismo de la ética y diferencias irreductibles entre valores, enfoques, normativas, etc. tanto a nivel de civilización (ej.: occi-

dental-oriental), a nivel religioso (ej.: cristiano-islámico) e incluso a nivel de Estados (ej.: Rusia-Estados Unidos de América-China). Dichas diferencias plantean dificultades para desarrollar estándares éticos y normas que se consideren legítimas más allá de un individuo o un grupo en un momento y lugar específico. Para soslayar dicha problemática el investigador C. Ess plantea que, aparte del concepto de relativismo ético en el que cada cual mantiene su propio paradigma ético o del dogmatismo ético y la intolerancia, existe la posibilidad de establecer nuevas formas de pluralismo robusto estableciendo conexiones y compatibilidades entre diversos sistemas éticos (Ess, 2006: 215-217).

2.1.- CONCEPTUALIZACIÓN DE LA CIBERÉTICA

El objetivo último de la ciberética, según M. Anderson y S. L. Anderson, sería la creación de una máquina que siguiese unos principios éticos ideales, es decir: se guiase por dichos principios cuando tomase decisiones referentes a una acción a llevar a cabo. No obstante, algunos expertos, como J. Fox y C. Shulman, argumentan que no queda claro que un incremento en la inteligencia, el conocimiento y la racionalidad conllevaría a un incremento en un comportamiento de cooperación favorable a los humanos o un mantenimiento de la benevolencia. Dichos expertos serían de la opinión de que, si no son diseñados cuidadosamente, los objetivos finales de las máquinas inteligentes no serían altruistas, por lo que se deberían evitar situaciones en la que dichos sistemas fuesen muy poderosos con relación a la Humanidad. Así, como establece J. Handler[3] al hablar de la IA: “la IA puede ser utilizada para el

3 J. Hendler es Director del Instituto para la Exploración de Datos y Aplicaciones en el Instituto Politécnico de Rensselaer (RPI), Troy, Nueva York.

bien social. Pero también puede ser utilizada para otros tipos de impacto social donde el bien de un hombre es el mal de otro. Debemos ser conscientes de ello" (Anderson y Leigh Anderson, 2007: 15; Fox y Schulman, 2010: 456; IBM, 2020).

2.1.1.- DEFINICIÓN DE LA CIBERÉTICA

A mediados de la década de 1940 se estableció una nueva rama de la ética, por el norteamericano N. Weiner, que posteriormente se acuñó con el nombre de "ética de las computadoras" (*computer ethics*) o "ética de la información" (*information ethics*). Su trabajo incluyó referencias al impacto de las computadores sobre la seguridad, las responsabilidades de los informáticos, la globalización de las redes, la simbiosis entre los cuerpos humanos y las máquinas, la ética de los robots o la IA, entre otros temas[4] (Bynum, 2015; Vacura, 2015: 326; Weiner, 1948: 28).

Sería W. Maner el que crease el término "ética de las computadoras", como un área específica de la ética enfocada a los problemas éticos creados, transformados o exacerbados por las computadoras o por la tecnología computacional. Estaríamos así, ante una nueva área de la ética aplicada. En dicho contexto sería J. Wizenbaum el que sugeriría que se deberían poner límites estrictos a la computarización de la vida humana para proteger sus principales valores sociales (Vacura, 2015: 326; Wizenbaum, 1976). Dicha postura

4 Para más información ver los libros publicados por N. Weiner: *Cybernetics* (1948); *The Human use of Human Beings* (1950); *God and Golem Inc.* (1963) (Bynum, 2015; Vacura, 2015: 326).

propició el debate sobre el "carácter único" de la ética de las computadoras entre W. Maner y D. G. Johnson ya indicado. Esto propició que un amplio número de filósofos siguiesen dicha idea, este sería el caso de T. Bynum pero especialmente el de J. H. Moor a través de su artículo *What is Computer Ethics.* En dicho artículo se define la ética de las computadoras como "el análisis de la naturaleza y el impacto social de la tecnología computacional, así como la formulación y justificación de las políticas correspondientes para el uso ético de dichas tecnologías" (Moor, 1985: 266; Vacura: 2015: 327).

En todo caso, la integración de artefactos computacionales en otras tecnologías como los "entornos inteligentes" o la "computación ubicua" ha propiciado que la identificación de las computadoras como artefactos identificables se haya vuelto obsoleto. Dicha situación ha propiciado nuevas preguntas éticas relacionadas con términos como la privacidad, la vigilancia, los artefactos autónomos o la propiedad. Así, se ha desarrollado un nuevo término: la ciberética. Según H. T. Tavani, se define como "una rama de la ética aplicada que examina asuntos morales, legales y sociales en la intersección entre las tecnologías de la computación con las relativas a la información y la comunicación". Es más que la "ética de Internet" pues abarca más allá de Internet incluyendo el amplio abanico de la ciber tecnología (*cyber-technology*), pero menos que la "ética de la información" que cubre aspectos más amplios que esta (Stahl *et al,* 2016: 5; Tavani, 2013).

2.1.2.- DIFERENCIAS ÉTICA Y CIBERÉTICA

La ética, según S. Ramaswamy y H. Joshi, es un conjunto de principios sobre el bien y el mal que los individuos aplican cuando toman decisiones que influencian su comportamiento (Ramaswamy y Joshi, 2009: 810). Para N. Cointe, G. Bonnet y O. Boissier es una disciplina normativa práctica de cómo uno debe de actuar hacia los demás y tiene tres componentes (Cointe *et al,* 2016: 1107):

- *la ética consecuencialista*: un agente es ético solo si sopesa las consecuencias de cada elección y elije aquella con el mayor componente moral. También se conoce como ética utilitaria, donde el balance entre el nivel de felicidad y el nivel sufrimiento determina la medida de la calidad ética. Pensamiento ligado a pensadores británicos del siglo XIX como Mill (1861) o Bentham (1789) (Stahl *et al,* 2016: 4).

- *la ética deontológica*: un agente es ético solo si respeta las obligaciones, deberes y derechos relativos a una situación específica, actuando de acuerdo con las normas sociales establecidas. Así, la calidad moral de una acción dependerá de la intención del agente que lo lleva a cabo. Una aspecto del "deber" cuyo principal desarrollador fue el filósofo alemán I. Kant (Stahl *et al,* 2016: 4).

- *la ética de la virtud*: un agente es ético solo si actúa y piensa de acuerdo con unos valores morales. La ética estaría ligada por tanto al carácter del agente. Así, algo sería éticamente bueno si refleja el carác-

> ter templado del que lo propone y tiene su base en los trabajos de Aristóteles (Stahl, *et al,* 2016: 4).

En dicho contexto, se produciría un dilema ético cuando en una situación específica cualquier elección llevaría a infringir algún principio ético aceptado y, aun así, fuese necesario tomar una decisión (Kirkpatrick, 2015: 19-20).

No obstante, se debe tener en cuenta que la ética no es lo mismo que la moralidad. La moralidad es un conjunto de normas que guían nuestro comportamiento, mientras que la ética es la teoría y el reflejo de la moralidad. Por lo tanto, la ética también se puede definir como la filosofía de la moralidad y comprendería la sistematización, la defensa y las recomendaciones relativas al comportamiento sobre el bien y el mal. Siguiendo dicho razonamiento, el investigador K. Abney estableció que la moralidad es lo que el mundo debería ser en oposición a lo que realmente es. Se podría comprender de dos formas: haciendo el bien o siendo bueno. En tal caso, la moralidad dependería de las reglas que permiten llevar a cabo una acción correcta o como un individuo debe vivir para llevar una buena vida, una aproximación a través de reglas. En analogía con el Derecho, si se observasen las leyes un individuo sería moral, mientras que si las desobedeciese sería inmoral. Por lo tanto, al investigar la ética, se estaría incidiendo en qué reglas serían adecuadas para una sociedad en particular (Abney, 2012: 36; Ramadhan *et al,* 2011: 85).

En dicho contexto, en la actualidad, los filósofos dividirían

la ética en tres grandes áreas: la metaética, la ética normativa y la ética aplicada. La metaética investiga de donde proceden nuestros principios éticos y responde a aquellas preguntas enfocadas a temas de verdades universales como: la voluntad de Dios; el papel que juega la razón en nuestros juicios éticos o; el significado de los propios términos éticos. Es el área menos definida de la filosofía moral. Por el contrario, la ética normativa es más práctica estableciendo las tareas que se deben llevar a cabo para establecer los estándares morales que regulan la conducta sobre el bien y el mal. Implica que solo existiría un criterio único definitivo sobre la conducta moral y se le asociarían tres estrategias: las teorías de la virtud; las teorías sobre el deber y; las teorías consecuencialistas. Por último, la ética aplicada tiene en cuenta las necesidades del medio social en el que se desarrolla. Según el investigador J. Fieser, se define como la rama de la ética que analiza elementos morales específicos y controvertidos. Para S. B. Kaddu, la ética tendría como objetivo el desarrollo de una serie de reglas de conducta para situaciones específicas, donde principios éticos básicos son la guía para el desarrollo de estándares para profesiones y grupos específicos (Fieser, 2009; Kaddu, 2007: 2; Ramadhan *et al,* 2011: 85).

Por tanto, sería la ética aplicada la más relevante dentro de la ética computacional, pues utilizaría las ideas y discursos de la filosofía moral para comprender mejor y dar soluciones en áreas específicas donde exista la necesidad de prestar atención a la ética. Es más, sería también importante tener en cuenta el concepto de profesionalismo, donde profesiones específicas como la medicina o el De-

recho tendrían un impacto específico sobre aspectos éticos diversos. Así, se podría establecer un esquema de base que establecería los componentes de interacción entre la ética y la diversos elementos que influyen como: la tecnología; la teoría ética aplicable; la metodología utilizada; la contribución a la resolución del problema y; las recomendaciones que se establezcan. Un proceso que debería ser exhaustivo para determinar la relevancia de los elementos elegidos en cada caso concreto (ver fig.: 1) (Stahl *et al*, 2016: 4-6).

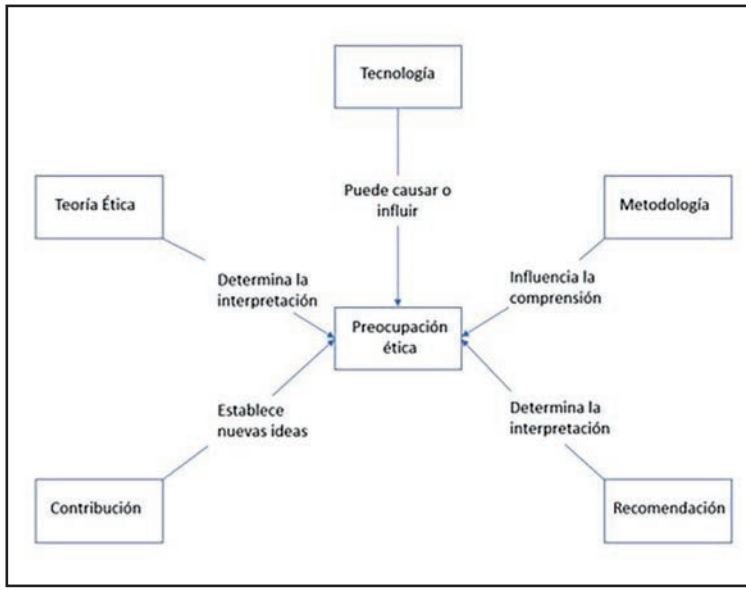


Fig.- 1: Elementos que pueden influir en la resolución de una preocupación ética (Stahl *et al*, 2016: 6)

En dicho contexto, tomando como referencia las indicaciones de J. M. Moor, se establecería la ética computacional como el análisis de la naturaleza y el impacto social de la tecnología computacional y la formulación y justificación

de políticas para el uso ético de dicha tecnología. Incluiría tanto las computadoras como la tecnología asociada a las mismas: hardware, software y redes (Moor, 1985: 266).

En este punto se debería tener en cuenta que, debido a su naturaleza social, los humanos también necesitarían de reglas, normas y convenciones para poder coexistir y colaborar. Por lo tanto, como indica B. C. Stahl, la disciplina de los Sistemas de Información (SI) necesariamente se interesarían por el rol social y organizativo de las tecnologías de la información y, por lo tanto, la discusión sobre normas, aceptación de reglas y convenciones se han convertido en eje central para la comprensión de la ética y la moralidad relacionada con la tecnología. Las nuevas tecnologías, por lo tanto, plantearían efectos normativos morales diferentes en forma cualitativa o cuantitativa de los ya establecidos y, además, dichas tecnologías podrían abrir nuevos caminos para solucionarlos (Stahl, 2012: 636, 638).

No debemos olvidar que cualquier planteamiento moral se basa, en un principio, en la "intuición moral". Una reacción no reflexionada de lo que un individuo considera que "está bien" o "está mal", donde podríamos poner como ejemplo el ataque indiscriminado a la población civil en un conflicto armado. Tanto comunidades, culturas o naciones pueden compartir dicha intuición moral, aunque los problemas comenzarían en caso de que algunos de ellos no lo compartiesen. Para los SI la intuición moral es importante porque supone un impacto significativo sobre el uso y el éxito de dichos sistemas. Así, como establece R. O. Mason, en general la mayor parte de los individuos tienen una

fuerte intuición moral sobre elementos como la privacidad, la propiedad, el acceso o la exactitud de la información en los SI, pero también se ha establecido una fuerte intuición sobre el posible impacto negativo de la IA sobre los valores éticos de una sociedad, como en el caso del respeto a los DD.HH. (Yu *et al,* 2018: 5527; Mason, 1986; Stahl, 2012: 639).

La escala, la complejidad y la revolución debido a las TIC, así como el uso indebido o el mal funcionamiento de los ordenadores, han contribuido a la emergencia de nuevos caminos éticos innovadores, que alteran la ética y la moral existente. Pero, como indica L. Floridi, esto significa que no solo existen nuevas fórmulas para solucionar cuestiones éticas y morales ya existentes, sino que también puede que se tenga que replantear las bases con las que desarrollamos nuestras posiciones éticas actuales. Así plantea que no habría que precipitarse en buscar soluciones viables e implementables, sino que se debería estudiar la raíz del problema y el desarrollo de posibles nuevas teorías éticas antes de pasar a la acción. Sin embargo, en la actualidad la ética computacional adopta la versión pragmática del principio "MINIMAX" (minimizar los daños y maximizar los beneficios), una orientación básicamente consecuencial. Por lo tanto, las TIC distancian al agente de la responsabilidad directa sobre las acciones llevadas a cabo en el ámbito computacional, pero también a nivel de sus efectos o en términos de la distancia conceptual del agente sobre el problema y el anonimato. Así, las posibles sanciones morales quedarían cada vez más diluidas, indirectas y distantes con-

forme las acciones computacionales sean más complejas y oscuras (Floridi, 1999: 35, 38).

Pero también, al mismo tiempo, la tecnología computacional da nuevas posibilidades para actuar y emergen nuevos valores que cuestionan nuestros valores preestablecidos. La problemática surge dado que es un entorno complejo y continuamente cambiante y que no hay una serie de reglas prefijadas, lo que, según J. H. Moor, hace que la ética computacional sea una disciplina aparte de la ética general. La Revolución Informática (*Computer Revolution*) sería, por tanto, su maleabilidad lógica, dado que los ordenadores pueden ser moldeados para cualquier actividad caracterizada por entradas, salidas y conexiones lógicas, a través de cambios en el hardware y software, por lo que los límites de la informática solo dependerían de nuestra propia creatividad. Como indica C. Stückelberger, estaríamos hablando de la Cuarta Revolución Industrial de la sociedad cibernética, la revolución digital y la IA (Moor, 1985: 267-269; Stückelberger, 2018: 27, 30).

En el ámbito de la ética uno de los factores principales de la informática es que la mayor parte de las operaciones son invisibles, esto puede propiciar el abuso invisible, el uso intencionado de las operaciones invisibles para que las computadores se comporten de una forma poco ética. Un segundo aspecto sería la presencia de valores invisibles en la programación, aquellos valores que están integrados en un programa informático. Así, un programador podría establecer ciertas especificaciones dentro de un programa que incrustasen una serie de valores específicos sin el co-

nocimiento del usuario del programa, un aspecto que puede ser intencionado o no intencionado. Pero, además, existe el factor de los cálculos complejos invisibles, donde los ordenadores en la actualidad pueden realizar un gran número de cálculos que pueden estar por encima de la comprensión humana, estableciendo así procesos demasiado complejos para que puedan ser inspeccionados y/o verificados (Moor, 1985: 273-274).

En dicho punto habría que preguntarse hasta qué punto el ser humano puede confiar en los cálculos invisibles computacionales. Un aspecto de especial significación ética si se tiene en cuenta las posibles consecuencias de dicho cálculo. Por ejemplo, aquellas computadoras utilizadas en el ámbito militar que toman decisiones acerca del lanzamiento de un ataque. Los ordenadores no son infalibles y puede que no exista el tiempo para confirmar una situación concreta. Aunque el ámbito computacional también permita extraer datos específicos que permitirían un conocimiento preciso de la situación, en multitud de casos no se conoce cuando, donde o como direccionar la atención de un programa informático hacia un punto concreto. Esto llevaría a una de las grandes diferencias entre la ética y la ciberética: la invisibilidad nos hace vulnerables. Como indica J. H. Moor, estamos abiertos al abuso invisible, la programación de valores inadecuados o errores de cálculo invisibles (Moor, 1985: 275).

Quizás uno de los elementos más significativos que ilustra dicha situación sería el "*Big Data*". Hay que tener en cuenta que la responsabilidad moral es una combinación de

factores internos y externos. Entre los primeros se pueden distinguir la causalidad (el resultado de la acción), el conocimiento (anterior a la toma de decisión) y la elección que se tome (el resultado de la acción tomada libremente). En el caso de que alguno de ellos no esté presente entonces se puede exonerar al agente de dicha responsabilidad moral. Con el "*Big Data*" además está la cuestión de que muchos actores contribuyen a la acción por lo que estaríamos ante una moralidad distribuida. También hay que tener en cuenta la división entre los agentes: recolectores; utilizadores y; generadores de los datos. El poder ejercido entre los diversos agentes dependerá del tamaño y la densidad de la red y del poder de cada uno de sus componentes dentro de ella. Así, con el "*Big Data*" se tendría diversos impactos éticos como sobre la privacidad, las tendencias obtenidas con el análisis de los datos y los problemas éticos de las investigaciones sobre los diversos grupos analizados (Zwitter, 2014: 2-5).

No se debería olvidar, no obstante, que los cambios tecnológicos han dejado atrás los desarrollos éticos, sobre todo en momentos en los que la tecnología debe ser desarrollada para hacer frente a problemas no anticipados, como en el caso de la reciente pandemia sanitaria, donde decisiones sobre el uso de las nuevas tecnologías deben ser rápidamente puestas en marcha sin haber dilucidado antes la complejidad ética de los nuevos desarrollos o estableciendo nuevos principios éticos computacionales alejados de los tradicionales. Una problemática que ya había sido expresada por L. Floridi cuando estableció que la ética computacional tenía

algo que decir sobre los problemas morales y podría contribuir con una nueva e interesante perspectiva con relación al discurso ético tradicional (Floridi, 1999: 34).

Así, los derechos individuales cobran una gran importancia en la Sociedad de la Información, pues el individuo no es solo un agente sino un objetivo potencial de acciones automáticas realizadas a medida, propiciando problemas relativos a los derechos individuales establecidos por el consecuencialismo ético tradicional. Es más, la naturaleza asimétrica de las acciones "virtuales" hacen que los individuos tengan menos fuerza en el ámbito tecnológico y por tanto que dichas acciones puedan evadir los contratos sociales preestablecidos y tener éxito. Por lo tanto, si se toma como referencia la teoría de juegos, el "juego" estaría siempre sesgado hacia el "hacker" y la mayor parte de los crímenes informáticos no serían detectados y principalmente se mantendrían impunes (Floridi, 1999: 39-40).

Consecuentemente, la ciberética desarrolla una serie de análisis y establece un plan de acción donde la atención se concentra en lo que sucede en la propia "infoesfera" (el propio medio donde se desarrolla la información computacional). Por lo tanto, aun siendo una ética aplicada, principalmente es una ética del ser y no de la conducta, pudiendo ser calificada como una ética no estándar. En tal caso, como indica L. Floridi, será la propia información la que ocupe el rol del paciente en cualquier acción, siendo infocéntrica y orientada a los objetos en contraste con el sentido biocéntrico de la ética tradicional. Dicha premisa implicaría que sin la información no existiría la acción moral y por

tanto la ética computacional sería básicamente una teoría ontocéntrica basada en los objetos. Teoría donde el agente sería una entidad capaz de producir información que podría repercutir en la infoesfera y el no ser se circunscribiría a la ausencia de información o entropía de la información (Floridi, 1999: 43, 45-46).

Dentro de dicho entorno, un aspecto que se debe destacar es que, dado que la ciberética no está limitada por el medio biofísico, pues la infoesfera incluye cualquier medio, la aplicabilidad de sus leyes éticas serían universales. Al seguir un enfoque construccionista, orientado a los objetivos, existiría la posibilidad de remodelar e implementar nuevas realidades donde los objetivos serían internos hacia el desarrollo de la infoesfera, cuyo principal objetivo sería enriquecerla, expandirla y mejorarla sin ninguna pérdida de información. Por lo tanto, un agente moral en dicho ámbito sería aquel que cuidaría del medio de la información y le proporciona mejoras para dejar a la infoesfera en mejores condiciones que antes de su intervención. Así, problemas éticos como los producidos por la bioingeniería o las armas autónomas podrían tener una respuesta por parte de la ciberética completamente distinta de la ética biocéntrica (Floridi, 1999: 56).

El problema que subyace estaría, por tanto, en como encajar la macro ética humana en la ciberética y cuál de ellas sería prevalente en el desarrollo tecnológico o el establecer un proceso simbiótico, que sería, a nuestro entender, preferente. Si tomásemos como ejemplo el desarrollo de la IA, se podrían plantear una serie de dilemas morales que estarían incluidos

en las posibles alternativas aplicables en un proceso de interacción entre la IA y los seres humanos. Sería por tanto interesante desarrollar modelos de preferencias éticas, que serían utilizados para establecer computacionalmente unas decisiones colectivas cuyo resultado sería el mejor posible, dentro de la teoría MINIMAX. Métodos que combinarían enfoques basados tanto en reglas y leyes como en ejemplos reales ya establecidos para resolver dilemas morales basados, en muchos casos, en los diversos enfoques culturales existentes. Dicho marco de referencia implicaría una capacidad de coleccionar un gran número de datos y su análisis a través de procesos propios del *Big Data.*

En todo caso, se debería tener en cuenta que el *Big Data* constituye un cambio computacional en el pensamiento y la investigación. Un movimiento radical sobre cómo se establece una investigación, dado que replantea la pregunta sobre la constitución del conocimiento, cuáles deben ser los procesos de investigación, como se debe utilizar la información al mismo tiempo que se establece la naturaleza y la categorización de la realidad. Es más, la investigadora D. Boyd de Microsoft plantea la necesidad de tener en cuenta que la concepción de que el *Big Data* permite una mayor objetividad y una mejor precisión puede ser errónea. Cuando los investigadores desarrollan un modelo de datos e interpretan sus resultados todo el marco está sujeto a observaciones y opciones subjetivas, además, más datos no significan mejores datos y cuando se combinan datos de diferentes fuentes se pueden magnificar los problemas éticos, por ejemplo, sobre la privacidad o la rendición de cuentas dado que, aunque unos datos sean accesibles eso

no significa que sea moralmente aceptable que se acceda a ellos (Boyd y Crawford, 2011: 6-8, 11)

2.2 – DIMENSIONES DE LA CIBERÉTICA

El ciberespacio está en todas partes, un espacio global y sin límites, mientras que el cuerpo humano está sujeto al espacio. Es inmaterial y digital por lo que es difícil distinguir entre lo real y lo virtual. Anónimo, facilita el desarrollo de identidades múltiples y las estructuras de poder que lo sostienen a veces están ocultas. Al influenciar todos los ámbitos sociales la ciberética incluye casi todos los dominios éticos. Por lo tanto, como establece C. Stückelberger, sería necesario que cualquier problema ético tuviese en cuenta la dimensión cibernética (ver fig. 2) (Stückelberger, 2018: 26-27):

La ética de la vida incluye la ética de la salud, la bioética y el impacto de la IA sobre la telemedicina, el envejecimiento, los servicios sanitarios, etc. La ética comunitaria hace referencia sobre los cambios de la vida comunitaria y las virtudes y defectos de los medios sociales. La ética medioambiental tiene en cuenta el impacto de la tecnología cibernética sobre las relaciones entre los humanos y la naturaleza. La ciberética en la ética política analiza su impacto y los cambios que producen sobre los sistemas políticos, la seguridad, los sistemas de armamento autónomos y su regulación a nivel nacional e internacional. En el marco económico, el impacto del ciberespacio sobre el crecimiento económico, el empleo y sobre el mundo financiero. Finalmente, en cuanto a la cultura, se analiza el impacto del ciberespacio sobre el ámbito cultural, la inclusión, la discriminación, el respeto religioso y el odio. En el mundo actual cualquier tópico ético tendría, por tanto, que tener en cuenta la

dimensión cibernética que lo condiciona (Stückelberger, 2018: 27).

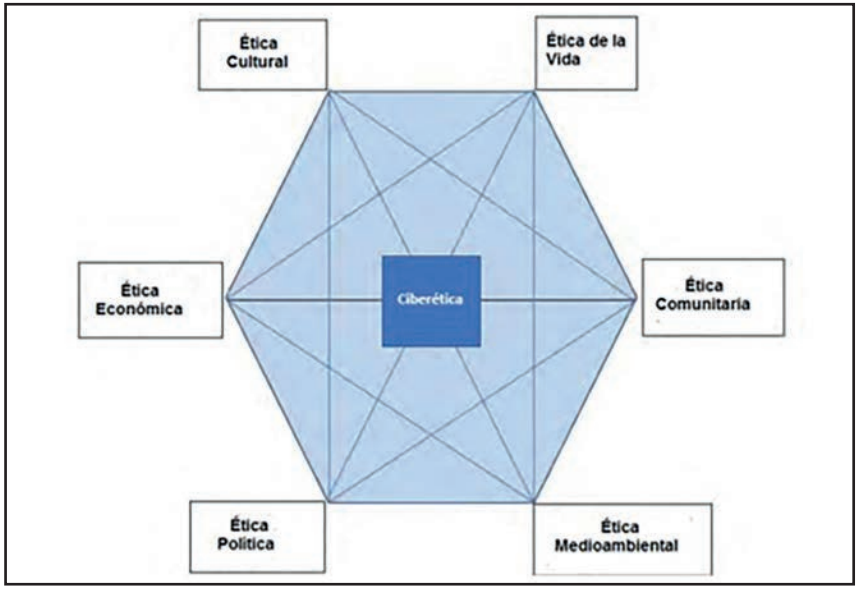


Fig.- 2: Influencias de la ciberética sobre los diversos dominios éticos (Stückelberger, 2018: 27)

Aunque nuestra investigación estaría centrada en el impacto de la ciberética sobre la denominada ética política, la interacción entre los diversos elementos hace necesario tener en cuenta los demás elementos establecidos. En todo caso, nuestro principal marco de investigación estaría relacionado con la "Seguridad de la Información" (*Information Security*) y sus diversas dimensiones: la dimensión ética; la dimensión de seguridad, la dimensión de las soluciones propuestas y; la dimensión del desarrollo moral relacionado con la Seguridad de la Información. Estaríamos, por tanto, en el desarrollo de dimensiones descriptivas y normativas de la ciberética (Dark *et al,* 2007; Siponen, 2001: 27-28).

2.2.1- DIMENSIÓN DESCRIPTIVA

Según los investigadores J. R. Desjardins y J. J. McCall, la ética descriptiva se refiere a "las creencias generales, los valores, el comportamiento y los estándares que, por lo demás, guía el comportamiento … la ética descriptiva examina las creencias típicas o los valores que se llevan a cabo usualmente". Por su parte, T. L. Beuchamp y N. E. Bowie lo definen como "la descripción factual y la explicación de las creencias y comportamientos morales". Tomando dicha base para establecer la dimensión descriptiva de la ciberética sería necesaria apoyarse en tres elementos que también son esenciales en la evolución de la computación: la teoría, el diseño y la abstracción. La teoría enfoca las relaciones y pruebas que soportan el desarrollo de la ciberética, el diseño enfoca la construcción de modelos y la abstracción el trabajo experimental que prueba y valida, en su caso, dichos modelos (Beuchamp y Bowie, 2001: 6; Desjardins y McCall, 2000: 4; Fisher, 2004: 398).

Para G.W. F. Hegel, la ética formaba parte de la filosofía política. El libre albedrío era un derecho formal y abstracto, donde la moralidad tenía que ver con la bondad y la maldad, lo esencial y lo actual. También representaba a la familia, a la sociedad civil y la constitución política. En dicho contexto, los aspectos del Estado estaban formados por las leyes, la libertad de la propiedad, la constitución, el gobierno, la nación, el derecho internacional y el espíritu nacional, entre otras características. Por otro lado, Marx y Engels consideraban la moralidad como ideologías que trataban de legitimar la dominación religiosa, económica

y política de las clases dominantes. Pero será a partir del siglo XIX cuando la filosofía materialista e idealista se erigiría como un método que concebía la realidad como un proceso de desarrollo continuo, donde la consecuencia de diversas contradicciones permitiría la emergencia de nuevos postulados incorporando o descartando otros antiguos. Así, la ética combinaría elementos subjetivos y objetivos, formando un nuevo paradigma ético. Aquí la importancia se debería al desarrollo de un proceso cognitivo, donde la ética, las normas sociales, los valores y las reglas emergerían a través de diversos procesos de comunicación (Fuchs *et al*, 2009: 450-451).

Dicho proceso se ha acentuado recientemente. Así, para el investigador N. Luhmann[5] la moralidad no es un subsistema de la sociedad, sino que circula a través de todos los sistemas sociales, por lo tanto, no es un concepto abstracto sino concreto estableciendo nuevos resultados como la bioética o la ética computacional. Por lo tanto, como establece C. Fuchs, el sistema moral de una sociedad sería un subsistema del subsistema cultural de la sociedad. En tal caso, la autoorganización de dicho sistema moral sería un proceso conjunto de tres variables: la cognición, la comunicación y la cooperación, donde esta última sería el principal principio de la moralidad moderna y la base de una dimensión objetiva de la ética: la ética de la cooperación. Como corolario, una sociedad realmente humana sería una

5 Información recabada de C. Fuchs *et al* sobre las teorías de N. Luhmann.de su trabajo "*Ethik als Reflexionstheorie der Moral*" (La ética como teoría de la reflexión moral) (Fuchs *et al,* 2009: 454).

sociedad cooperativista, que cuestiona los principios establecidos desarrollando voces alternativas para evitar el pensamiento unidimensional. Así, según Spinello, teniendo en cuenta la ciberética como un conjunto de meta normas que guían el actuar correctamente en el ciberespacio a través de las tecnologías y el conocimiento computacional, se estaría produciendo la transformación de la sociedad. Transformaciones que significarían que se establecerían nuevas preguntas de cómo se deben regular las relaciones sociales, con nuevos riesgos y oportunidades. Siguiendo dicha idea, C. Fuchs ha establecido que el desafío de la ciberética sería, por tanto, el establecimiento de la discusión de cuáles serían los principios morales que guiarían las acciones humanas para que los individuos estuviesen facultados para establecer una Sociedad de la Información global que fuese participativa y sostenible (Fuchs *et al,* 2009: 454, 456-457; Spinello, 2003: 2).

En este punto deberíamos volver a la dicotomía, que ya hemos indicado, entre los investigadores tradicionalistas, como el caso de la investigadora D. G. Johnson que propone que las nuevas tecnologías no supondrían un cambio sobre los conceptos éticos tradicionales y aquellos, como W. Maner, que proponen la especificidad de la ciber tecnología y la creación de elementos éticos completamente nuevos. Teorías que pueden ser concebidas como correctas en algunas instancias e incorrectas en otras. Los tradicionalistas podrían tener razón en que las computadoras no han creado ningún elemento ético nuevo, mientras que los específicos pueden estar en lo cierto en que la ciber tecnología

hace más complicado el análisis de los elementos éticos tradicionales. Por lo tanto, al exponer cualquier teoría sobre la ciberética deberíamos distinguir entre las características tecnológicas específicas y el posible desarrollo de nuevos elementos éticos.

Dado que la ciberética se adapta mejor a la ética aplicada, se podría establecer su desarrollo a través de tres perspectivas: la ética profesional; la ética filosófica y; la ética descriptiva. En el primer caso, la ciberética estaría ligada, como propone D. Gotterbarn, a las profesiones y los profesionales, aunque, a nuestro entender, dicha posición sería demasiado restrictiva en cuanto al alcance de la ciberética. En cuanto a la ética filosófica, la ciberética sería un campo dentro del análisis filosófico sobre las tecnologías como los análisis desarrollados por J. H. Moor, ya descritos anteriormente, utilizando a tal fin el modelo estándar de la ética aplicada propuesto por P. Brey[6]. Tanto la ética profesional como la ética filosófica formarían parte de la dimensión normativa que analizaremos más adelante, mientras que la ética descriptiva se basaría en la realidad del caso y no en lo que debiera ser. Así, la ética descriptiva nos prepararía para el posterior análisis de los elementos éticos que afectan a las políticas y las leyes, permitiendo que los elementos normativos éticos estén más claros al comprender el aspecto des-

6 Para P. Brey, la metodología estándar usada por los filósofos en las investigaciones de la ética aplicada tendría tres fases: la identificación de una práctica polémica como un problema moral; la descripción y el análisis del problema a través de la clarificación de los conceptos y el examen de datos factuales asociados con el problema y; la aplicación de las teorías y principios morales para alcanzar una posición sobre el problema moral en cuestión (Brey, 2000: 10).

criptivo de los efectos sociales de la tecnología (Gotterbarn, 1991: 26-31; Huff y Finholt, 1994; Moor, 1985: 267-269). La siguiente tabla resume las perspectivas de la ciberética aplicada dentro del marco de la ética política, base de nuestro estudio (ver Fig. 3):

Perspectiva	Disciplinas Asociadas	Elementos a examen
Profesional	Informática Ingeniería	Responsabilidad del agente Fiabilidad de los sistemas Códigos de Conducta
Filosófica	Filosofía Derecho	Privacidad y Anonimato Derechos Humanos
Descriptiva	Sociología Ciencias del Comportamiento	Impacto de la ciber tecnología en las instituciones gubernamentales de investigación y en los grupos sociales

Fig.- 3: Perspectivas de la ciberética aplicada en la ética política (elaboración propia)

2.2.2- DIMENSIÓN NORMATIVA

Como se ha observado anteriormente, la ciberética aplicada en la ética política establece una perspectiva filosófica en la que se encuadra la disciplina del Derecho con especial enfoque en los Derechos Humanos. Ahora bien, es necesario establecer la relación entre los valores éticos y las normas legales, para lo cual y en primer lugar se debe establecer el

marco de valores a los que se hace referencia. Siguiendo a C. Stückelberger implicaría una serie de valores esenciales que incluirían: la responsabilidad, la libertad, la justicia, la equidad, la paz, la seguridad, la comunidad, la inclusión, la participación y el perdón entre otros (Stückelberger, 2018: 33).

Estaríamos hablando, por tanto, del establecimiento de una dimensión normativa de la ciberética. En tal caso se iría más allá de la autorregulación y de la auto responsabilidad estableciendo una serie de leyes basadas en valores. Para C. Stückelberger sería necesario que dentro del ciberespacio la ciberética comprendiesew tres tipos de reglas que deben estar balanceadas y equilibradas (Stückelberger, 2018: 38):

- *Reglas Éticas*: para la visión general, las orientaciones y el modelo comunitario.
- *Reglas basadas en las leyes*: para proporcionar fiabilidad, confianza, control del poder, etc.
- *Reglas de relaciones*: para establecer las relaciones humanitarias.

Centrándose en la Ética, existe una jerarquía de niveles que establecen el carácter vinculante de las normas. Dicha jerarquía se puede expresar a través de la siguiente pirámide (ver figura 4) (Stückelberger, 2018: 39):

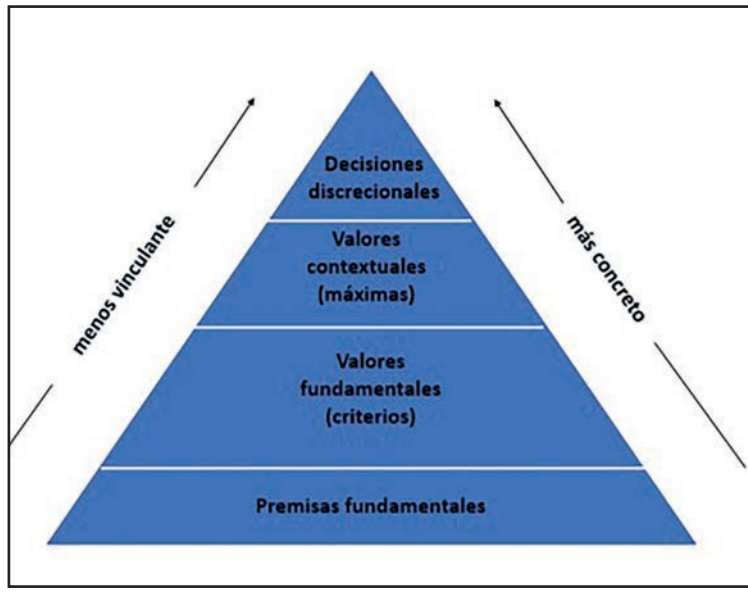


Fig.- 4: Gráfico de jerarquía ética normativa, basado en C. Stückelberger y traducción propia (Stückelberger, 2018: 39).

Las premisas fundamentales son la base de la ética, aunque no estén formuladas expresamente, como cuando se establece "yo vivo y por lo tanto existo". Los valores fundamentales son válidos a largo plazo, aunque puede que varíen en su prioridad. Los valores contextuales son los estándares con las que funciona una sociedad, generalmente las leyes que la rigen. Aquí se pueden incluir las convenciones internacionales, las leyes nacionales y los estándares es-

tablecidos[7]. Expresan los valores en un contexto específico y en un periodo concreto. Las decisiones discrecionales son aplicaciones concretas en áreas específicas. En el marco del ciberespacio se estaría hablando, por ejemplo, de la ética de la IA o del comportamiento ético específico de un robot a través de su programación (Stückelberger, 2018: 39-40).

En abril de 2019, el "Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial", creado por la Comisión Europea en junio de 2018, emitió las "*Directrices Éticas para una Inteligencia Artificial Fiable*". La fiabilidad de la IA debe estar basada en tres premisas: debe ser lícita, y cumplir todas las leyes y reglamentos aplicables; debe ser ética, garantizando el respeto de los principios y valores éticos y; debe ser robusta, tanto en lo técnico como en lo social para no provocar daños accidentales. Dichas premisas deben actuar de forma simultánea y en armonía. A tal fin, se debe garantizar que el desarrollo, despliegue y utilización de la IA cumpla con las siguientes directrices: acción y supervisión humana; solidez técnica y seguridad; gestión de la privacidad y de los datos; trasparencia; diversidad, no discriminación y equidad; bienestar ambiental y social y; rendición de cuentas (Ala-Pietilä *et al*, 2019: 2-3).

7 Un marco regulatorio basado en valores fue establecido por la Asamblea General de las Naciones Unidas en 2015. Denominado *Sustainable Development Goals (SDG's) 2015-2030* (Objetivos de Desarrollo Sostenible). Todas las actividades en el ciberespacio deben ser medidas a través de los SDG's y están basados en valores básicos como: la igualdad, la justicia, la libertad, la inclusión, la paz y la sostenibilidad (Stückelberger, 2018: 40-42).

Otros investigadores proponen otra serie de valores. El investigador T. Nagel distingue cinco tipos de valores fundamentales: utilidad; derechos generales; obligaciones especiales; compromiso sobre los proyectos propios y; fines perfeccionistas (Nagel, 1987: 177). En dicho contexto se estaría ante un entorno de pluralidad, como ya planteaba Ess, donde las vidas humanas son complejas con múltiples ramificaciones que afectan a otros. En cualquier momento la vista puede pasar desde una visión personal a una universal y ambas serían completamente válidas. Por lo tanto, no se podría implicar que los principios de deber humanitario o dignidad humana fuesen los criterios morales básicos ya que la consecución personal de la felicidad, el placer o el dinero también serían válidos (Ess, 2006, 215-17; Van den Hoven, 2010: 62-63).

Diferentes teorías normativas han destacado algún valor específico, minusvalorando otros o eliminándolos. Según J. Van den Hoven, dicha visión es bastante estrecha y no es válida dentro de la complejidad de los problemas morales humanos. Por lo tanto, una teorización de la ciberética puede que no se ajuste a la realidad de las soluciones propuestas ya que, según T. Nagel, alguna resolución de conflictos podría extenderse más allá de nuestra capacidad para enunciar principios generales. Además, "el buscar una única teoría general de cómo decidir el camino correcto a tomar, sería como mirar hacia una única teoría para decidir en que creer" (Nagel, 1987: 181; Van den Hoven, 2010: 63, 66).

La ética de Aristóteles responde a preguntas de qué hacer sobre la base de lo que una persona virtuosa haría. Las teorías morales utilitarias tomarían aquellas acciones que tuviesen las mejores consecuencias y resultados. Las teorías de Kant indicarían que la obligatoriedad de una acción no dependería de las consecuencias sino de su adherencia al principio ético principal: la Imperativa Categórica, que implica el respeto de los seres humanos y no usarlos como meros instrumentos de sus propósitos. Dichas teorías normativas tendrían dificultades para establecer el comportamiento ético en el mundo digital. Tomando como ejemplo un SAAL que mata indiscriminadamente, se podría argumentar que tanto las normativas utilitarias como las de Kant no podrían demostrar la maldad de los actos realizados por dicho SAAL, pues dichos actos podrían suponer un mejor resultado para los propósitos establecidos, en el caso de la teoría utilitaria. Por otro lado, un SAAL no es un ser humano y por lo tanto no estaría sujeto a la Imperativa Categórica en el caso de la teoría de Kant. Únicamente las teorías de Aristóteles podrían explicar nuestra intuición moral sobre dichos actos, si se estableciese que una persona virtuosa no realizaría alguno de dichos actos (Van den Hoven, 2010: 69-71).

Habría por tanto que tener en cuenta, como ya sugirió J. Moor, que en muchas áreas de la ciberética existe un vacío conceptual y, por tanto, un vacío a nivel normativo. Siguiendo lo indicado por J. Van den Hoven, estaríamos de acuerdo también con un vacío a nivel de diseño ya que no sabemos que sistemas realizar, que software tramar o que códigos de línea escribir. Por lo tanto, sería imprescindible una recons-

trucción conceptual de los conceptos clave antes de aplicar valores, principios y teorías, idea ya propuesta por Floridi. Tomando como ejemplo el entorno jurídico, aunque muchas personas compartan el "concepto" de justicia, puede que no compartan la "concepción" de la justicia. Las TIC hacen necesario que revisemos la concepción tradicional de privacidad, responsabilidad, democracia, comunidad y propiedad para formular concepciones más apropiadas al tiempo presente. Términos de la ciberética como "democracia digital", "vida artificial", "teletrabajo", etc. no necesariamente significa que comprendamos su naturaleza o que seamos capaces de regularlo de una forma adecuada. Como indica J. Van den Hoven, es necesario ser preciso en la formulación de nuestras ideas, en la articulación, la precisión y detalle normativo, en definitiva, conocer que es lo que deseamos alcanzar sobre ética y moralidad a través de la tecnología (Floridi, 1999: 35, 38; Moor, 1985: 266; Van den Hoven, 2010: 72-74).

Reflexionando sobre las diversas ideas analizadas en este capítulo, nuestro punto de vista estaría más en la línea de W. Maner y J. H. Moor que de D. G. Johnson. Las nuevas tecnologías, principalmente la IA, han propiciado la creación de un nuevo paradigma: la ciberética, más acorde con la ética aplicada computacional qué con la ética normativa tradicional. Particularmente, dicho cambio de paradigma ha estresado la capacidad de aplicación de los principios morales (reglas) existentes surgidos de la ética normativa al entorno del ciberespacio. En dicho contexto la pregunta a realizar sería: ¿es necesario cambiar la ética normativa existente y por lo

tanto sus reglas por una ética aplicada computacional?

Nuestra postura, siguiendo la idea propuesta por L. Floridi, es que no deberíamos sustituir la ética normativa existente por la ética aplicada, sino ser capaces de establecer una relación simbiótica entre la ética tradicional y la ciberética, manteniendo por un lado el enfoque reglamentario, es decir, los principios morales consensuados, pero adaptándolos al nuevo paradigma, añadiendo aquellos nuevos principios éticos surgidos del nuevo dominio y consecuentemente las nuevas normas consensuadas a aplicar, derivadas de la ciberética, surgidas especialmente de la introducción de la IA.

Particularmente, en el entorno de la responsabilidad jurídica internacional, el ciberespacio ha estresado la capacidad práctica de aplicación de la ética normativa existente y sus reglas derivadas (DIH y DD.HH. entre otras) en el ciberespacio, pero también puesto en duda si se deberían mantener o sustituir por otra ética aplicada (la ciberética) y consecuentemente por los nuevos principios surgidos del nuevo entorno operacional. Nuestra postura, en este entorno específico, sería mantener los principios éticos normativos existentes (DIH, DD.HH.) que deberían seguir siendo la referencia en el ciberespacio. Ahora bien, deberíamos identificar y solucionar los posibles vacíos de ética normativa surgidos de la ciberética a través de nuevos diseños computacionales, posiblemente a través de los nuevos principios rectores consensuados, solucionando de esa forma, los denominados vacíos conceptuales de J. H. Moor, así como los vacíos de diseño identificados por J. Van der Hoven, para el marco de la responsabilidad jurídica internacional de los SAAL.

CAPÍTULO 3

CIBERGUERRA

El concepto de ciberguerra viene estrechamente ligado al ciberespacio, un nuevo dominio de operaciones que tiene evidentes repercusiones sobre los principios de ataque y defensa desde planteamientos básicamente defensivos y reactivos hacia un marco de anticipación ante las amenazas antes de que se produzcan. En el marco conceptual, aparte de la rapidez en los cambios y capacidades, dichos cambios estratégicos tienen nuevas implicaciones éticas y jurídicas. Esto implicará una actualización y revisión tanto de los sistemas tecnológicos y de las redes, así como del marco reglamentario que enmarca las actividades dentro del ámbito de los conflictos de interés y de poder entre los Estados.

El ciberespacio se acuñó por primera vez como concepto por W. Gibson en 1984 en su novela "*Neuromancer*". Un espacio donde la World Wide Web (WWW) fue descrita por B. Jiang y F. J. Omerling como la base del propio concepto. Un marco evolutivo que, para algunos investigadores, culminará en un punto de no retorno, una vez alcanzada la denominada "Singularidad Tecnológica", donde las máquinas (robots, computadoras, programas o sistemas) llegarían a tener la capacidad de automejorarse para perfeccionarse, de acuerdo con el término introducido por V. Vinge, en 1983, ligado a la creación de máquinas inteligentes por encima de la comprensión humana y sin su intervención (Guliciuc, 2014: 79-80; Jiang y Omerling, 1997: 112).

En España dicho concepto fue definido dentro de la nueva "*Estrategia de Ciberseguridad Nacional*" de 2019, donde se establecía como "espacio común global" y como "una dimensión fundamental para la estabilidad el preservar la defensa de los valores y principios constitucionales y democráticos, así como los derechos fundamentales de los ciudadanos en el ciberespacio, especialmente en la protección de sus datos personales, su privacidad, su libertad de expresión y el acceso a una información veraz y de calidad". Es importante destacar también que dicho documento ponía de manifiesto de cómo el ciberespacio estaba actuando como elemento transformador más allá de la dimensión tecnológica para establecer nuevos modelos sociales y adentrarse en las relaciones personales y la ética. Pero también implicaría nuevos desafíos, mayor competición geopolítica y reordenación del poder (BOE, 2019).

Por lo tanto, serán el conflicto y el poder los ejes que definirán la expansión de la ciberguerra. Conflicto definido como enfrentamiento armado y la "persecución de objetivos incompatibles entre diferentes grupos", que está estrechamente ligado al poder y que generalmente se va acumulando, estableciendo los denominados "conflictos de interés": la coerción, la influencia, la fuerza y la manipulación de forma abierta o encubierta. Todo ello ejerciendo control sobre las agendas políticas, los intereses reales o subjetivos y los conflictos ya establecidos o potenciales, los denominados conflictos latentes. En todo caso, siempre existirá una evolución entre el momento del inicio del conflicto y su conclusión siguiendo distintas fases: paz, paz inestable, tensión, crisis, conflicto armado y guerra. En dicho contexto, los diferentes actores utilizarán todos los instrumentos a su alcance para obtener sus objetivos, incluidos todos los aspectos de la ciberguerra, a los que J. López de Turiso define como "instrumentos de poder" (López de Turiso, 2012: 122-

125; Lukes, 2005: 19-29; Mial *et al,* 1999:20).

3.1.- CONCEPTUALIZACIÓN DE LA CIBERGUERRA

En junio de 2011, el periódico "The New York Times" publicó un artículo denominado "*War evolves with drones, some tiny as bugs"* (La guerra evoluciona con los drones, algunos tan pequeños como bichos). Durante 2014 el Comité Internacional de la Cruz Roja (CICR) desarrolló un ciclo de conferencias de alto nivel denominado *"New technologies and the modern battlefield"* (Las nuevas tecnologías y el campo de batalla moderno) dicho ciclo abordaba los nuevos desafíos legales, humanitarios y éticos planteados por las nuevas tecnologías. En ambos casos se hacía referencia al nuevo marco de los conflictos armados tanto a nivel tecnológico como jurídico y ético, sirviendo como ejemplos de los cambios que han llevado desde la guerra analógica a la guerra en el ciberespacio, la denominada ciberguerra (Bumiller y Shanker, 2011; CICR, 2014).

3.1.1.- EL CONCEPTO MODERNO DE LA GUERRA

Un concepto de ciberguerra que es más que un concepto moderno de la guerra, ya que se engloba dentro del marco más amplio de la ciberseguridad, un concepto holístico, al que la Agencia Europea de Ciberseguridad (ENISA) define como: "todas las actividades necesarias para proteger el ciberespacio, a sus usuarios y a las personas a las que afecta de las amenazas cibernéticas", creando una pirámide de Maslow de capas de protección (ver fig. 5) (ENISA, 2017: 6).

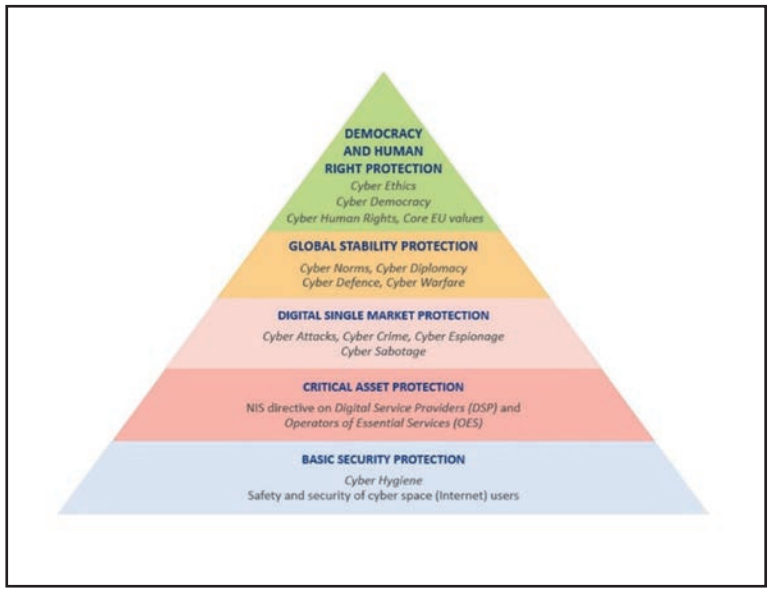


Figura 5: Pirámide de Maslow de la ciberseguridad según ENISA (ENISA, 2017: 4).

Dentro de dicha pirámide, la ciberguerra viene encuadrada dentro de la "Protección de la Estabilidad Global". En todo caso, su conceptualización es occidental teniendo diversas acepciones como: "guerra cibernética" (*cyber warfare*) o "conflicto cibernético" (*cyber conflict*), no siendo generalmente aplicable como concepto para superpotencias como Rusia o China.

Hay que tener en cuenta que la ciberguerra es un concepto más global que la ciberdefensa, aunque a veces se superpongan, que a su vez es más restrictivo en el espacio occidental. El término "ciberdefensa" viene descrito específicamente en la base de datos oficial de terminología de la OTAN denominada: "NATO*Term*" y es la siguiente: "los

medios para lograr y poder ejecutar medidas defensivas para contrarrestar los efectos de las amenazas cibernéticas y, por lo tanto, preservar y restaurar la seguridad de las comunicaciones, de la información y de otros sistemas electrónicos o la información que se guarda, se procesa o se transmite a través de dichos sistemas". En cuanto a España, el Estado Mayor de la Defensa (EMAD) engloba la ciberdefensa dentro de la ciberseguridad nacional dentro del marco de la nueva "Estrategia Nacional de Ciberseguridad" (ver fig. 6) (López de Turisio y Sánchez, 2018: 11; OTAN, 2019):

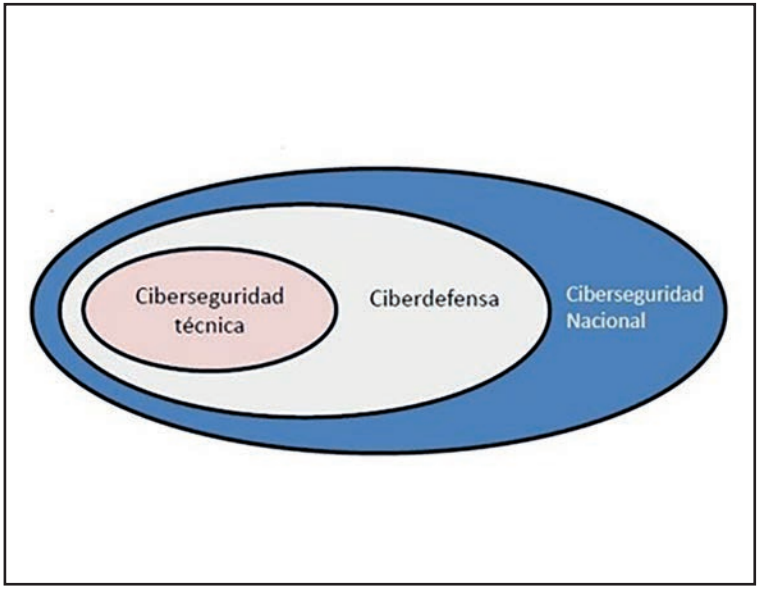


Figura 6: Inclusión de la ciberdefensa dentro de la Ciberseguridad Nacional (López de Turisio y Sánchez, 2018: 11)

Dicho concepto se aleja del concepto exclusivamente defensivo y preventivo hacia un enfoque de mayor fuerza

disuasoria; y una aproximación más proactiva de la ciber inteligencia, para conocer las capacidades, técnicas, tácticas e intenciones de los adversarios, elemento integrado en el principio de "defensa activa". Un concepto más cercano al de otras superpotencias como China, que ya lo estableció en su libro blanco de la "*Estrategia Militar China*" de 2015 a través de una defensa estratégica, combinada con una postura operacional ofensiva para contrarrestar actividades hostiles en el marco político y económico además del militar. Así, el concepto moderno de la guerra adquiere un elemento holístico y global, hacia un concepto de "guerra irrestricta" o "guerra sin restricciones" (no plasmado oficialmente), que según J. M. Mancera, "defiende los propios intereses, amplia zonas de influencia, se apropia de información o activos informáticos, interrumpe, bloquea o impide el uso del recurso informático y el ciberespacio al enemigo para, en el mejor de los escenarios, obligar al contrincante a aceptar los propios intereses" (BOE, 2019; Heath *et al*, 2016:34; LawInfoChina, 2017; Mancera Castaño, 2014: 91-92; Marin Martínez, 2019: 21).

Otro aspecto del nuevo concepto de la guerra se muestra a través del decreto nº 646, de 5 de diciembre de 2016, de la Federación Rusa sobre la "*Doctrina de la Seguridad de la Información de la Federación Rusa*", donde se define la "seguridad de la información" como el: "estado de protección del individuo, la sociedad y el Estado contra amenazas de la información internas y externas, permitiendo asegurar los derechos y libertades humanas y civiles constitucionales, la calidad de vida decente y estandarizada de los

ciudadanos, la soberanía, la integridad territorial y el desarrollo sostenible socio-económico de la Federación Rusa, así como la defensa y la seguridad del Estado". Concepto que incluye entre sus elementos estratégicos principales: la disuasión estratégica, el análisis preventivo de las amenazas cibernéticas (ciber inteligencia) y el "desarrollo de contramedidas sobre las amenazas sobre la información y psicológicas que puedan socavar los cimientos históricos y las tradiciones patrióticas relacionadas con la defensa de la patria". Por lo tanto, la ciberguerra y por ende la ciberdefensa vendrán a formar parte integra de la "seguridad de la información" del Estado (Mº Exteriores Federación Rusa, 2016).

3.1.2.- CONCEPTUALIZACIÓN DE LA CIBERGUERRA

El Diccionario de Oxford la define como "el uso de la tecnología informática para perturbar las actividades de un Estado u organización, especialmente a través de ataques deliberados sobre los sistemas de información con propósitos estratégicos o militares". Por su parte, G. Sánchez Medero la define como "una agresión promovida por un Estado y dirigida a dañar gravemente las capacidades de otro para imponerle la aceptación de un objetivo propio o, simplemente, para sustraer información, cortar o destruir sus sistemas de comunicación, alterar sus bases de datos". Una definición más completa la establecen S. Goel e Y. Hong, dado que describen cuatro modalidades de conflicto cibernético:

las guerras de las redes sociales para influenciar las políticas internas de un Estado; la guerra estratégica para causar daños al adversario y sus recursos (espionaje industrial); la guerra ideológica donde organizaciones fundamentalistas utilizan Internet para difundir su ideología y reclutar miembros y; las guerras iniciadas por los ciudadanos para atacar a los ciudadanos e instituciones de otros Estados (*hacktivismo*). En el mismo espacio occidental, el Departamento de Defensa de USA la define como un "conflicto armado que se lleva a cabo, en su totalidad o en parte, por medios cibernéticos" y está formada por las "operaciones militares conducentes a denegar a una fuerza opositora el uso efectivo de los sistemas y los armamentos del ciberespacio en un conflicto. Incluye el ciberataque, la ciberdefensa y las acciones habilitantes cibernéticas". Para J. Arquilla y D. Ronfeldt la ciberguerra se refiere a la preparación y la conducción de operaciones militares de acuerdo con principios relacionados con la información. Conocer todo lo del adversario sin dar datos de uno mismo, el lograr que el "balance de información y conocimiento" sea a favor de uno mismo, especialmente si el balance de fuerzas no es favorable, finalmente es la utilización del conocimiento para que se utilice menos capital y recursos humanos (Arquila y Ronfeldt, 1993: 149; DoD, 2010: 8; Goel y Hong, 2015: 3; Oxford Dictionaries, 2019; Sánchez Medero, 2010: 64).

Por su lado, la Comisión Europea en su Comunicación Conjunta al Parlamento Europeo y al Consejo sobre "*Resiliencia, disuasión y defensa: fortalecer la ciberseguridad de la UE*", de 2017, aunque no menciona específicamente la palabra ciberguerra, si la alude de manera indirecta al

afirmar que: "… agentes estatales están logrando cada vez más sus objetivos geopolíticos no solo a través de métodos tradicionales, como la fuerza militar, sino también mediante el uso de herramientas cibernéticas más discretas, como la interferencia en los procesos democráticos internos. El uso del ciberespacio como campo de batalla, de forma exclusiva o como parte de una táctica híbrida, es ahora ampliamente reconocido. Las campañas de desinformación, la propagación de noticias falsas y las operaciones cibernéticas dirigidas a infraestructuras vitales son cada vez más comunes", una descripción cercana al de conflicto cibernético holístico indicado por Goel y Hong. En este punto habría que tener en cuenta que en el mundo occidental existe una diferencia conceptual entre "guerra cibernética" (*cyber warfare*) y "ciberguerra" (*cyber war*). Así, según M. Robinson, K. Jones y H. Janicke, la guerra cibernética está circunscrita al ciberespacio y no conlleva una declaración formal de guerra, mientras que la ciberguerra, aunque también se lleva a cabo en el ciberespacio, necesita una declaración formal de guerra (Comisión Europea, 2017: 2; Robinson *et al,* 2015: 86).

Más recientemente, la Comunicación Conjunta de la Comisión Europea y el Alto Representante de la Unión para los Asuntos Exteriores y la Política de Seguridad al Parlamento Europeo y al Consejo sobre "*La Estrategia de Seguridad de la Unión Europea para la Década Digital*", de 16 de diciembre de 2020, establecía que el marco de las nuevas amenazas, incluidas las híbridas y las campañas de desinformación, estarían minando la seguridad internacional, la estabilidad y los beneficios del ciberespacio así como el

Estado de Derecho. En dicho contexto, la ciberseguridad debería estar integrada en tecnologías clave como la IA, la encriptación y la informática cuántica[8]. La prevención del uso inadecuado de dichas tecnologías permitiría al UE adherirse a las normas, reglas, principios y medidas de fomento de la confianza de las NU[9]. Es más, dicha Comunicación establecía la necesidad de aunar esfuerzos con otros socios internacionales para promocionar un modelo político y una visión del ciberespacio basado en el Estado de Derecho, los Derechos Humanos, las libertades fundamentales y los valores democráticos, con un claro impacto sobre las nuevas medidas a tomar por la UE en el ámbito de la prevención, disuasión y la respuesta a acciones de ciberguerra[10] (CE, 2020: 1-2, 5, 13, 18-19).

No obstante, La concepción occidental de la ciberguerra choca con la de otras superpotencias mundiales. En el caso de Rusia, el "*Borrador de Convención sobre la Seguridad de la Información*", presentado en 2011, no hace ninguna mención sobre la ciberguerra. En cuanto a sus fuerzas armadas, el documento "*Conceptual Views on the Activity of the Russian Federation Armed Forces in Information Spa-*

8 Incluyendo un gran énfasis en el desarrollo de dichas tecnologías en el ámbito de la Defensa.

9 Ver la Nota del Secretario General de las Naciones Unidas sobre el "Grupo de Expertos Gubernamentales sobre los Avances en la Información y las Telecomunicaciones en el Contexto de la Seguridad Internacional" (A/70/174), de 22 de julio de 2015, en https://undocs.org/es/A/70/174

10 Como el desarrollo de una Unidad Conjunta Cibernética (*Joint Cyber Unit*), una plataforma virtual de cooperación con el enfoque en una coordinación técnica y operacional contra incidentes y amenazas cibernéticas transfronterizas (CE, 2020: 13).

ce", también de 2011, hace referencia al concepto de "Guerra de Información" (*Information Warfare*) que engloba al concepto de ciberguerra. Se define como: "el conflicto entre dos o más Estados en el espacio de la información con el objetivo de causar daños a los sistemas de información, los procesos y los recursos de importancia crítica y otras estructuras, subvirtiendo los sistemas políticos, económicos y sociales, estableciendo trabajos psicológicos masivos sobre la población para desestabilizar a la sociedad y al estado y coaccionando al gobierno para que tome decisiones en el interés de la parte oponente". Una definición de "Guerra de la Información" que va más allá de su concepto occidental, ya que L. J. Janczewski y A. M. Colarik la definen como "un ataque planificado por parte de naciones o sus agentes hacia los sistemas de ordenadores y de información, los programas informáticos y los datos que tiene como resultado pérdidas para el enemigo"[11] (Giles, 2012: 68; Janckzewski y Colarik, 2008: xiv).

Por su parte, las fuerzas armadas chinas (*China's People's Liberation Army* (PLA)), utilizan el término 網絡戰 (*Wǎngluò zhàn*), para referirse al concepto occidental de ciberguerra. En todo caso las operaciones en el ciberespacio se definen como 網絡作戰 (*Wǎngluò zuòzhàn*), "gue-

11 En dicho ámbito se podría poner como ejemplo el ciberataque denominado "*Sunburst*", perpetrado en 2020, contra la empresa SolarWinds, afectando a las principales agencias gubernamentales, incluidas algunas de la seguridad nacional, y empresas, tanto de USA como globales (con mención específica sobre España), que el mundo occidental atribuyó a "hackers" estatales rusos. En el caso de que así fuese probado se estaría ante un ejemplo del concepto de "explotación de redes informáticas" (*computer network exploitation*) (BBC, 2020; Libicki, 2009: 14).

rra de redes", identificando dichas redes como "Cibernéticas Contra-Defensivas (DCC)" y "Ofensivas Cibernéticas Contra-Defensivas" (OCC). En dicho contexto al referirse al concepto occidental de ciberguerra, en particular para USA, utilizan el término 網絡戰攻擊 (*Wǎngluò zhàn gōngjí*): ataques de ciberguerra. En todo caso, según plantea J. M. Mancera, estaríamos ante el concepto "Guerra Irrestricta" (*Unrestricted Warfare*) con el uso masivo de formas no convencionales de guerra, sin reglas, predominando la asimetría y el uso de espacios diferentes al físico. La siguiente tabla resume los principales términos en ruso, chino e inglés (ver fig. 7) (Giles y Hagestad, 2013; Mancera Castaño, 2014: 90, Marín Martínez, 2019: 11-13):

English	Chinese	Russian
information space	信息空間 xìnxī kōngjiān	информационное пространство *informatsionnoye prostranstvo*
information warfare	信息战争 xìnxī zhànzhēng	информационная война *informatsionnaya voyna*
information weapon	信息武器 xìnxī wǔqì	информационное оружие *informatsionnoye oruzhiye*
information security	信息安全 xìnxī ānquán	информационная безопасность *informatsionnaya bezopasnost*
cyber warfare	網絡戰爭 wǎngluò zhànzhēng	кибервойна *kibervoyna*
cyberspace	網絡空間 wǎngluò kōngjiān	киберпространство *kiberprostranstvo*
cyber security	網絡安全 wǎngluò ānquán	кибербезопасность *kiberbezopasnost*
network warfare	網絡戰 wǎngluò zhàn	сетевая война *setevaya voyna*

Fig. 7: Principales términos relativos a la ciberseguridad en inglés, chino y ruso (Giles y Hagestad, 2013)

3.1.3.- DIFERENCIAS: GUERRA Y CIBERGUERRA

El 13 de julio de 2020, la agencia Reuters informó de un ataque con drones de los Houthis yemenís contra un complejo petrolífero de Arabia Saudí, miembro de la coalición con los que están en guerra, al mismo tiempo que lanzaban misiles. Dicha noticia expuso un claro ejemplo de la convergencia entre el mundo kinético (físico) y el mundo cibernético de un conflicto declarado. También existen ejemplos en conflictos larvados, como el acaecido entre Irán y USA, el 20 de junio de 2019, cuando un misil tierra-aire iraní derribó un dron de vigilancia norteamericano, que propició un ataque cibernético, en represalia, contra los sistemas informáticos iranís que controlan los lanzamientos de misiles. Como indica B. Sussman, ambos serían catalogados como claros ejemplos de la convergencia entre la guerra y la ciberguerra, dentro del marco de la guerra híbrida (Reuters, 2020; Sussman, 2019).

El concepto moderno de guerra proviene, predominantemente, de la teoría de J.-J. Rousseau, que establecía una condición de enemistad entre estados, con límites claramente definidos sobre quién y quien no participaba, en que capacidad, en qué periodo de tiempo y espacio geográfico (Finlay, 2018: 360; Rousseau, 1762/2004: 10). Hay que tener en cuenta que, en cierto sentido, existe una idea extendida que la guerra es inherente al ser humano. El investigador A. Gat estableció que desde tiempos remotos la principal fuente de violencia, conflicto y guerra era la escasez de recursos y de mujeres (la necesidad de reproducción). Durante la protohistoria, los principales conflictos en

el Mediterráneo Central y Occidental vinieron de la mano de las necesidades de explotación de recursos mineros (oro, plata, etc.), para el pago de la soldada y la logística de los conflictos, que a su vez mantenían la capacidad de expansión territorial de las ciudades Estado, como en el caso de Cartago y Roma. Como indica M. Scornavacchi, la guerra actual entre Estados sería el resultado de la complejidad, integración y escala que en el que el ser humano la habría convertido y donde la guerra se convierte en una actividad social como estableció Clausewitz (Clausewitz, 1968: 202; Gat, 2006: 664; Marín Martínez, 2018; Scornavacchi, 2015: 27).

Ahora bien, existe un profundo cambio en el paradigma de la guerra por el advenimiento de las nuevas tecnologías y más aún con la IA. Es más, la cuestión que se debe plantear es el posible impacto que las computadoras y la IA supondrá para el paradigma de la guerra, cuando dicha IA acabe por ser más inteligente que la mente humana y como dicha situación cambiará el propio concepto de la naturaleza humana y su relación con los conflictos, específicamente el control humano de las acciones llevadas a cabo. Esto supondría, a su vez, un cambio en la forma en el que el ser humano interacciona con su entorno llegándose uno a preguntarse, como también lo hace J. Rodríguez Álvarez, cómo evolucionaría el rol reservado para los humanos en dicho entorno y consecuentemente el futuro de nuestra especie (Rodríguez Álvarez, 2019: 244).

Dicho debate se estaría desarrollando en paralelo con el concepto de "Teoría de las Nuevas Guerras" (*New War Theory*). Dicho concepto tiene como base una serie de elementos que lo identifican: el incremento en el número de guerras civiles; la intensidad de las batallas; el número de civiles refugiados y muertos. Además, con el advenimiento de las nuevas tecnologías, se eliminan los conceptos de frontera donde actores que no son Estados militarizan sus agendas políticas en estilos de guerra transnacionales y no convencionales, al mismo tiempo que se establecen nuevos conceptos como la guerra híbrida. En dicho entorno no existe un fin a la violencia en un espacio de tiempo y lugar concreto y la lucha se prolonga indefinidamente, donde es fácil y barato reclutar y formar a los combatientes, teoría claramente definida por M. Kaldor dentro del marco de la globalización.

Para P. Mello, existen una serie de elementos que encuadran dicho concepto: la erosión del monopolio del Estado en el uso de la fuerza; la economía política de las nuevas guerras; su carácter asimétrico; una base identificada con el concepto de identidad (nación, tribu y religión según M. Kaldor) y; encuadrando el terrorismo dentro de dicho marco. Aun así, también existen detractores de dicha teoría, como en el caso de J. Galvin, proponiendo que lo descrito en la teoría de las Nuevas Guerras es simplemente una nueva forma de describir los conflictos de baja intensidad del pasado. No obstante, a nuestro entender y de acuerdo con lo indicado por M. Kaldor, sí estaríamos hablando de un nuevo concepto de guerra, especialmente por el uso de las

nuevas tecnologías, la IA y la guerra de la información, sobre todo bajo el paraguas de la denominada "guerra híbrida" y la nueva "economía de guerra". Se estaría estableciendo, por tanto, una relación simbiótica entre los grandes avances tecnológicos y el aumento de las capacidades militares creando una forma de "Revolución de los Asuntos Militares" (*Revolution of Military Affairs – RMA)* dentro de las denominadas "guerras posmodernas" enunciadas por C. H. Gray y que también serían utilizadas como definiciones de las guerras virtuales o las guerras en el ciberespacio (Galvin, 1992: 60; Kaldor, 2002: 2, 5, 29, 71, 79; Mello, 2010: 297-309; Scornavacchi, 2015: 28-29).

En cuanto a la "economía de guerra", el mantener una guerra convencional (kinética) es cara y no fácil de mantener. Como establece M. Tanenbaum, el mantener un ejército es caro y lleva tiempo, pues los sistemas de armas convencionales son muy caros (como en el caso de los nuevos cazas norteamericanos F-35), su mantenimiento y la logística también es cara y significa un importante porcentaje del presupuesto de un Estado, más aún sería difícil mantenerlo en secreto. Los elementos utilizados por los piratas informáticos (*hackers*), en contra, son relativamente baratos frente al armamento convencional y son bastante más fácil de ocultar. Tampoco debemos dejar de tener en cuenta que una guerra supone un sacrificio de vidas humanas y del conjunto económico de los países en conflicto que supone, generalmente, que si dicho conflicto se alarga en el tiempo puede producir conflictos internos en los países, hastío de la guerra y destrucción de los tejidos sociales y económicos

de los Estados. Este punto viene unido a la globalización que, según M. Kaldor, ha comenzado a romper las culturas verticales organizadas actualmente. Un proceso complejo con dos desarrollos completamente opuestos: por un lado, la integración, por otro la fragmentación, la homogenización frente la diferenciación. Un proceso que crea una red transnacional inclusiva, mientras que al mismo tiempo excluye y atomiza a un gran número de gente. En el aspecto económico se tendría la estandarización de productos a nivel global con economías de escala, mientras que se desarrollaría una mayor diferenciación de acuerdo con las demandas locales y especializadas (Kaldor, 2012: 73-74; Tanenbaum, 2018).

La complejidad moderna de la "economía de guerra" presupone que también existe un alto grado de complejidad en el desarrollo de las "guerras modernas". Dicha idea implicará, por tanto, la necesidad de desarrollar nuevas formas de conducción de la guerra en tiempos modernos, lo que se viene acuñando con el concepto de "guerra híbrida". Concepto que se desarrolla a partir de la anexión de Crimea por Rusia en 2014[12] pero que ya había comenzado a emerger a través del concepto chino de "guerra irrestricta" o "guerra más allá de los límites" en 1999[13] y que tiene un gran ejem-

12 Para más información ver: Pintado Rodríguez, C. (2017): *Ucrania. Un Estudio de Caso de Guerra Híbrida*, CISDE Observatorio, acceso noviembre 2021, en https://observatorio.cisde.es/actualidad/ucrania-un-estudio-de-caso-de-guerra-hibrida/

13 Para un completo desarrollo de dicho concepto ver: De Pablo López, M. (2015): *La Guerra Irrestricta ¿Un nuevo modo de hacer la guerra?,* Estudios CEEAG, acceso noviembre 2021, en https://www.ceeag.cl/wp-content/uploads/2020/05/Irrestricta.pdf.

plo en la guerra entre Hezbollah-Israel en 2006, como propone F. G. Hoffman. Más aún, como ya se ha comentado anteriormente sobre la "guerra de información" rusa, dicho Estado estableció que todas las formas de guerra, incluidas las operaciones militares, estarían subordinadas a las campañas de información, por lo que no todos los actos conllevarían a una guerra "convencional" sino a una "zona gris". Así, el concepto de "guerra híbrida" estaría definido por la complejidad, aunque existe todavía bastante confusión en su definición exacta, confundiéndolo con otros conceptos como la "amenaza híbrida" (*hybrid threat*) o el "conflicto híbrido" (*hybrid conflict*) (Clark, 2020: 15; Hoffman, 2007: 22-25, 35-42).

La Comisión Europea, en su "*Comunicación conjunta sobre la lucha contra las amenazas híbridas. Una respuesta de la Unión Europea*", de 2016, define la "amenaza híbrida" como: "subrayar la mezcla de actividades coercitivas y subversivas, de métodos convencionales y no convencionales (es decir, diplomáticos, militares, económicos y tecnológicos), que pueden ser utilizados de forma coordinada por agentes estatales o no estatales para lograr objetivos específicos, manteniéndose por debajo del umbral de una guerra declarada oficialmente". Sería comparable con la definición de "conflicto híbrido" indicado por el Parlamento Europeo en 2015, diferente al concepto de "guerra híbrida" que es parte de un evento de guerra abierta entre Estados o contra agentes no estatales utilizando todos los medios, incluidos los políticos, económicos y diplomáticos (ver fig. 8) (CE, 2016: 2; Monaghan, 2019: 84-87, PE, 2015: 1).

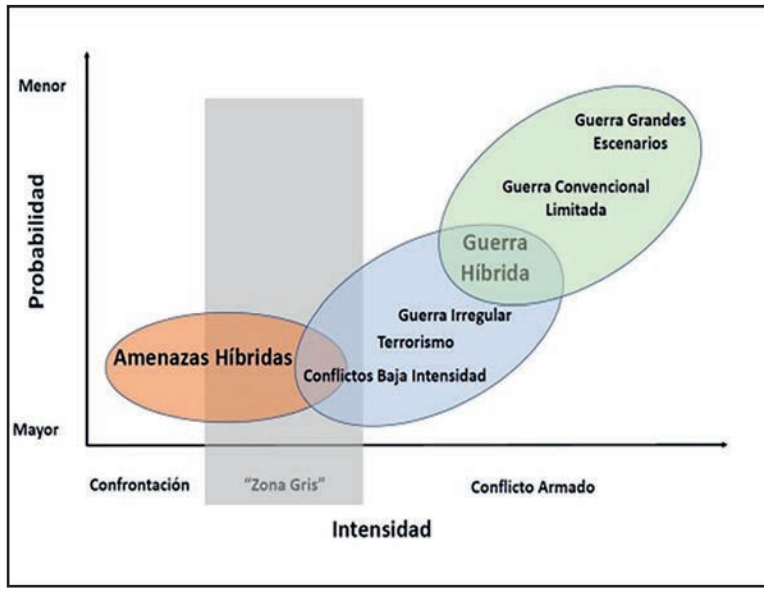


Fig. 8: Amenaza y Guerra Híbrida visto en progresión de conflictos (Monaghan, 2019: 87)

Particularmente, de acuerdo con el informe producido por P. J. Cullen y E. Reichborn-Kjennerud sobre "Comprendiendo la Guerra Híbrida" (*Understanding Hybrid War*), de 2017, la "guerra híbrida" estaría diseñada para explotar las vulnerabilidades nacionales a través del conjunto de los espectros político, militar, económico, social, de información y de infraestructuras. Para lo cual utilizaría, de una forma coordinada, diferentes instrumentos de poder militar, político, económico, civil y de información que estarían extendidos mucho más allá del contexto militar. El elemento clave para su completo funcionamiento la fusión entre

la coordinación y la sistematización a escala internacional (ver fig. 9) (Cullen y Reichborn-Kjennerud, 2017: 4).

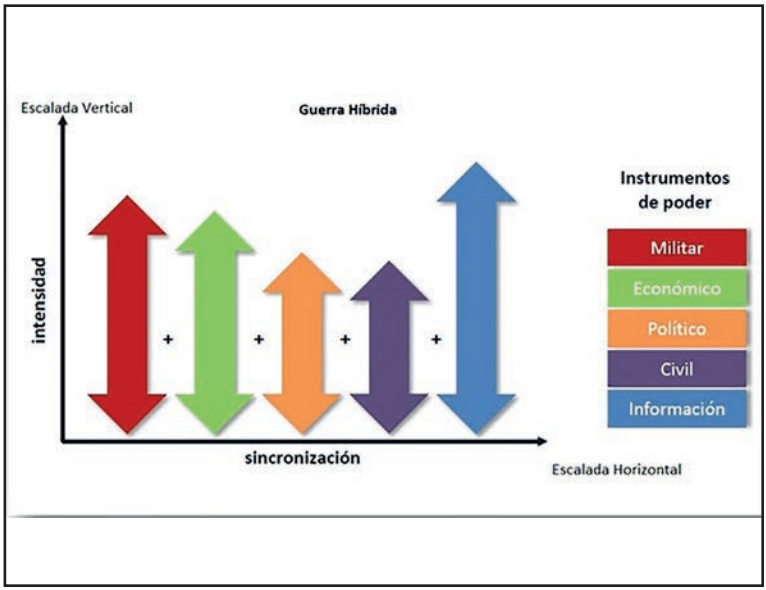


Fig. 9: Escalada de la Guerra Híbrida (Cullen y Reichborn-Kjennerud, 2017: 9; Cubeiro Cabello, 2018)

Es importante tener en cuenta que dichos elementos incluyen todo tipo de actores y acciones: diplomacia, ciberataques, guerra económica, fuerzas militares regulares, fuerzas especiales, fuerzas irregulares, apoyo a elementos opositores locales y guerra de la información (propaganda, "fake news", etc.), donde la información toma cada vez más protagonismo (ver figura 10) (Cubeiro Cabello, 2018).

Fig. 10: Elementos y actores de la Guerra Híbrida (Cubeiro Cabello, 2018)

3.2.- CONCEPTUALIZACIÓN DE AMENAZA, DESAFÍO Y DISUASIÓN

Los inicios de la "teoría de la disuasión" comienzan a partir de la Segunda Guerra Mundial y el nacimiento de la era atómica y está basada, principalmente, en los principios de la criminología clásica de C. B. di Beccaria en 1764[14], que asume que las personas toman decisiones razonadas para perpetrar o abstenerse de llevar acabo un crimen basándose en la maximización de los beneficios y la minimización de los costes. El foco se centraría en las sanciones legales, estableciéndose que cuanto más certeza, severidad y rapi-

14 Para más información ver: BECCARIA, C. B. di (1764) [2015]: *Tratado de los Delitos y de las Penas,* Universidad Carlos III, acceso noviembre 2021, en https://drive.google.com/file/d/1oeA4LOtswSC7HtUlhgCEZyOJc-k_M_tW3/view

dez se alcanza para llevar a cabo dichas sanciones más se disuade de la comisión de un acto ilícito. En la disuasión moderna dicho concepto se adapta para incluir las sanciones informales como la vergüenza o la pérdida del respeto, teniendo en cuenta que, en todo caso, la disuasión solo sería una de un conjunto de estrategias que los Estados pueden utilizar en sus relaciones internacionales con otros Estados o con actores que no lo son (ver figura 11) (D'Arcy y Herath, 2011: 644; Mallory, 2018: 2; Pratt *et al*, 2006: 371).

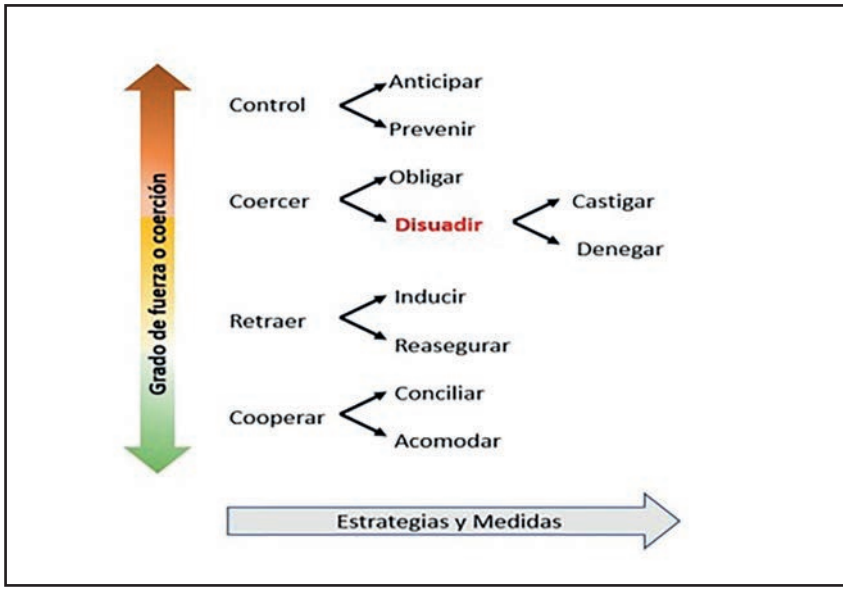


Fig. 11: Elementos alternativos de estrategia (Mallory, 2018: 2)

La mayor parte de la investigación sobre dicha teoría se ha concentrado en los conflictos entre Estados con una concepción simétrica[15]. Solo con el colapso de la Unión Soviética, en la década de

15 Un claro ejemplo sería el periodo de la "Guerra Fría" y la disuasión nuclear (Libicki, 2009: xvi; Lupovici, 2011: 49)

1990, y los actos terroristas del 11 de septiembre de 2001, la teoría de la disuasión empezó a transformarse para incluir la disuasión del terrorismo con relación a las "Armas de Destrucción Masiva" (*Weapons of Mass Destruction*). En todo caso, existen dos teorías de disuasión vigentes: la de disuasión estructural (*structural deterrence*) y la de decisión teórica de disuasión (*decision theoretic deterrence*). En todo caso, de forma generaliza, se estaría, según J. W. Knopf, en la cuarta fase de desarrollo de la "teoría de la disuasión" (Knopf, 2010: 1-2; Quackenbush y Zagare, 2016:4).

En el primer caso, a igualdad de condiciones, la probabilidad de la guerra estaría relacionada básicamente con su coste: a mayor coste menor probabilidad de guerra. Así, si el coste es pequeño la guerra es posible. Al contrario, cuando las condiciones de poder son asimétricas, se incrementan las crisis y las guerras y la disuasión no tendría éxito. En el segundo caso, se asume que la guerra es irracional lo que lleva a una inconsistencia lógica. Suponiendo que existe un atacante poderoso y un defensor más débil, el atacante tiene dos opciones cooperar o no. Si coopera se mantiene el "estatus quo", en caso contrario el defensor tiene dos opciones: ceder o no, si cede,el atacante triunfa, si no se desarrolla el conflicto. En dicho escenario si el coste de la guerra es exorbitante el conflicto sería el peor escenario. Ahora bien, como explican S. L. Quackenbush y F. C. Zagare, esto solo sería posible si el defensor convence al atacante de que podría llevar acabo un ataque suicida imposible de asumir por el atacante, independientemente del coste y los intereses hacia su propio Estado (Quackenbush y Zagare, 2016: 4-6).

Así, el principal desafío de la disuasión moderna sería, por tanto, el considerar la disuasión como un esfuerzo para moldear el pensamiento de un agresor potencial. Las políticas de disuasión se

observan, principalmente, desde la perspectiva del Estado que lleva a cabo dicha disuasión y se enfoca hacia aquellas acciones que elevan los costes y los riesgos de un ataque. Pero el valor de dichas medidas dependerá completamente en el efecto sobre la percepción del Estado sobre el que se actúa y para ello se deben evaluar previamente los intereses, motivos e imperativos del agresor potencial, que suelen ser variados y complejos y por lo tanto van más allá de una amenaza a un adversario, para llegar a un punto en el que dicho adversario perciba que las alternativas a una agresión son más atractivas que una guerra. (Lupovici, 2011: 53; Mazarr, 2018: 1-2).

La complejidad del paradigma de la disuasión se incrementa cuando se trata del ciberespacio. Según M. C. Libicki, dicho espacio tiene sus propios medios y sus propias reglas. En dicho contexto, los ciberataques se producen no a través de la fuerza sino a través de la explotación de las vulnerabilidades de los enemigos. El principal escollo se producirá, por tanto, cuando se intente establecer quién atacó y porqué, que es lo que obtuvieron y si el atacante realizase el ataque de nuevo. Esto es: si el atacante fuese capaz de realizar el ciberataque con impunidad y por lo tanto no tendría ninguna razón para parar. Además, un ciberataque que funcionase en la actualidad puede que no lo hiciese en un futuro próximo, por lo que la disuasión que funcionaba en el mundo analógico puede que no funcionase en el mundo digital. Más aún, la disuasión que funcionase hoy en el ciberespacio puede que tampoco funcionase en un futuro próximo. (Libicki, 2009: iii, xvi).

Analizando las posibilidades de éxito de la disuasión en el ciberespacio, el primer punto a destacar sería la capacidad del defensor para infligir un precio suficientemente negativo para el atacante

que le disuada de llevarlo a cabo. Unido a dicho punto estaría la credibilidad de la amenaza de represalia, como indican A. Bendiek y T. Metzger entre otros, para lo cual el defensor debería estar dispuesto a usar sus capacidades de disuasión. Existen una serie de factores que pueden limitar dicha respuesta: el efecto sobre la opinión pública nacional e internacional[16], el miedo a una escalada incontrolable[17] o las capacidades de disuasión del propio atacante, como a través de la ingeniería inversa. El tercer punto estaría relacionado con los dos anteriores y concierne a la capacidad del atacante de ser consciente de las capacidades del defensor y su disponibilidad a utilizarlas. Así, como indica A. Lupovici, lo que realmente importaría no serían las capacidades o las intenciones del defensor sino como estas fuesen percibidas por el atacante (Burton, 2018: 5-7; Libicki, 2009: 25-27; Lupovici, 2011: 50-51).

El elemento principal de análisis sería, por tanto, la posibilidad de que el paradigma de la disuasión pudiese ser aplicado a los ciberataques y por lo tanto al ciberespacio. Para que así sucediese, según J. Burton, la disuasión cibernética debería ser una estrategia exhaustiva que considerase todos los tipos de amenazas en el ciberespacio y no únicamente en el ámbito de la seguridad nacional y militar. Se entraría entonces en el aspecto crítico de la proporcionalidad: si la respuesta fuese muy reducida no sería efectiva, mientras que si fuese muy potente existiría el riesgo de una respuesta militar convencional. Ahora bien, si el concepto de disuasión mediante el castigo se circunscribiese al ciberespacio (el concepto de

16 Principalmente el que los observadores neutrales fuesen convencidos de que una represalia no significaría una agresión (Libicki, 2009: xvi).

17 El denominado "dilema de la ciberseguridad" (*cybersecurity dilemma*) (Buchanan, 2016: 6)

hack back), podría tener como resultado el minar el actual entorno normativo internacional legitimando el uso de los ciberataques como herramienta, impidiendo el progreso de medidas de cooperación internacional. Como B. Buchanan sugiere, aquellos Estados que invierten en capacidades defensivas y ofensivas cibernéticas pueden exacerbar el miedo y la desconfianza en el sistema internacional, conduciendo a una mayor proliferación y carrera de armamentos (Bendiek y Metzger, 2015: 555-557; Buchanan, 2016: 196, 201; Burton, 2018: 8-9).

Otros investigadores, como en el caso de M. P. Fischerkeller y R. J. Harcknett o B. D. Berkowitz, no están de acuerdo con que la disuasión sea adecuada para el ciberespacio y que la contención, derivada del principio de no intervención de NU[18] pueda servir en un dominio en constante movimiento y escala de cambios. Dicha característica alimentaría la noción de la necesidad de una persistente implicación operacional en contraste con el principio de contención anterior. Así, la visión de USA para el Comando Cibernético en 2018 se expresaría con la noción de: "alcanzar y mantener la superioridad en el ciberespacio". Por otro lado, la Comisión Europea plantea en su "Estrategia de Ciberseguridad para la Década Digital" de 2020, una serie de acciones a llevar a cabo por parte de la UE para obtener la resiliencia, la soberanía tecnológica y el liderazgo, aumentando su capacidad operacional para prevenir, disuadir y responder a las amenazas en el ciberespacio, aunque en este caso avanzando al mismo tiempo hacia un ciberespacio abierto y

18 **Artículo 2(4) de la Carta de las Naciones Unidas**: "Los Miembros de la Organización, en sus relaciones internacionales, se abstendrán de recurrir a la amenaza o al uso de la fuerza contra la integridad territorial o la independencia política de cualquier Estado, o en cualquier otra forma incompatible con los Propósitos de las Naciones Unidas." (NU, 1945).

global. Aunque, como se puede observar, existen grandes diferencias entre las dos propuestas, dado que USA plantea un liderazgo global mientras que la UE aboga por un ciberespacio abierto y a través de la cooperación entre Estados y estamentos internacionales, en ambos casos el elemento de la capacidad de respuesta ante las amenazas cibernéticas y por tanto la persistencia en la implicación se consideran centrales para sus objetivos (Berkowitz, 1997: 183-184; CE, 2020: 4; USCYBERCOM, 2018; Fischerkeller y Harcknett, 2019: 4-5).

Nuestro propio punto de vista no descartaría la disuasión cibernética como un elemento a utilizar en el ciberespacio. Estaríamos de acuerdo, sin embargo, en que la disuasión como se contemplaba durante la Guerra Fría como denegación de acceso o castigo no es de mucha utilidad en el mundo cibernético. En dicho punto estaríamos de acuerdo con los planteamientos de J. Burton que aboga por un planteamiento holístico que incluiría una serie de elementos de disuasión incluyendo aspectos legales, sociales, normativos y tecnológicos aplicables a medida según la amenaza, pero no descartando elementos de respuesta adecuados a cada una de ellas. No obstante, no se debería obviar que los cambios legales y normativos llevan tiempo, especialmente en el ámbito internacional y que a veces no se consigue llegar a buen término. Tampoco se debe olvidar el alto coste económico y de recursos humanos expertos que un marco de disuasión tecnológica acarrearía para los Estados, por lo que dicho tipo de disuasión estaría necesitada de una cooperación internacional importante, tanto entre Estados como con instituciones supranacionales (NU, OTAN, etc.) (Burton, 2018: 27-28).

3.3.- SISTEMAS ARMAMENTÍSTICOS AUTÓNOMOS Y AUTÓNOMOS LETALES (SAA, SAAL)

El CICR, en su 28ª[19] y 30ª[20] Conferencias Internacionales de la Cruz Roja y de la Media una Roja, expusieron los desafíos que, a su entender, planteaban para el DIH los conflictos armados contemporáneos y la necesidad de garantizar la licitud de las nuevas armas y de los nuevos medios y métodos de guerra. Dicho trasfondo propició un informe específico sobre "*El DIH y los desafíos de los conflictos armados contemporáneos*". En dicho informe se estableció una primera definición sobre los SAA: "Un sistema de armas autónomo es aquel que puede captar o adaptar su funcionamiento según la variación de las circunstancias del entorno en el que es desplegado". Ahora bien, según el CICR, para que dicho sistema fuese verdaderamente autónomo debería estar dotado de una IA capaz de aplicar la DIH. Dicha definición fue ampliada en el *"Informe de la Reunión de Expertos sobre Sistemas de Armamento Autónomos: Aspectos Técnicos, Militares, Legales y Humanitarios"* de 2014, en la que se definió un SAA como: "*aquel capaz de seleccionar y atacar objetivos independientemente, con o sin supervisión humana. El término se refiere a sistemas de armamentos que están equipados con funciones autónomas de focalización, seguimiento, selección y ataque de objetivos (funciones críticas).*

19 Objetivo Final 2.5: "Garantizar la licitud de las nuevas armas de conformidad con el derecho internacional" (CICR, 2003: 22).

20 En la Resolución 3: "Preservar la vida y la dignidad humana en los conflictos armados", en el apartado sobre "Conducción de Hostilidades", su punto 19 recuerda: "la obligación de los Estados Partes, como se expresa en el Protocolo adicional I, artículo 36, de examinar la licitud de las nuevas armas y de los nuevos medios y métodos de guerra e insta a todos los Estados a que consideren establecer mecanismos específicos de examen con esa finalidad" (CICR, 2007: 85).

El término excluye los sistemas de armamento que seleccionan y atacan objetivos bajo el control remoto de un operador humano" (CICR, 2011: 45; CICR, 2020: 57).

La evolución sobre la definición de un SAA por parte del CICR seguiría en 2016, dado que en el *"Informe de la Reunión de Expertos sobre Sistemas de Armamento Autónomos: Implicaciones de la Autonomía Creciente en las Funciones Críticas de los Armamentos"*, se puso de manifiesto que los avances tecnológicos estarían acelerando el proceso para que los propios SAA fuesen los que seleccionaran y atacaran los objetivos sin supervisión humana, después de una activación inicial de un operador humano. Así, sería el propio sistema de armamentos que, utilizando sensores, programación computacional y armamentos, tomaría el control de las selección de los procesos y de las funciones de la selección y ataque de los objetivos, que ordinariamente eran controlados directamente por humanos. A raíz de dicho informe el CICR estableció, el 11 de abril de 2016, su opinión sobre los SAA proponiendo la siguiente definición: "cualquier sistema de armamento con autonomía en sus funciones críticas. Esto es, un sistema de armamento que puede seleccionar (buscar o detectar, identificar, seguir, seleccionar) y atacar (usar la fuerza en contra, neutralizar, dañar o destruir) objetivos sin una intervención humana" (CICR, 2016a: 71; CICR, 2016b: 1).

Los cambios expuestos sobre la definición de los SAA por parte del CICR, a lo largo del tiempo, expone claramente la falta de un consenso internacional sobre su definición debido a los ingentes y rápidos cambios tecnológicos. En todo caso, se ha intentado establecer una serie de definiciones que se mantienen en la actualidad. Por un lado, como en el caso de la exposición de J. L. Rogers, dichos sistemas se pueden definir como "Armamentos Autónomos

Completos (*Full Autonomous Weapons (FAW)),* mientras que otros, como en el caso de M. E. O'Connell lo definen como "robótica autónoma completa" (*fully autonomous robotics*). En ambos casos es importante el término "completo" (*full*), pues los distingue de otras definiciones, como por ejemplo la de K. Cass, que habla de "una máquina capaz de percepción y de manipulación de su entorno, sin control o con un control limitado", en dicho contexto la categorización vendría dada por el nivel de control humano sobre el armamento. Se tendría por tanto tres posibles grados de intervención humana: "al tanto" (*in-the-loop*) donde el humano formaría parte de algunos de los procesos de decisión; "sobre la acción" (*on-the-loop*), cuando el armamento sería capaz de ejecutar todos los procesos de decisión, pero existiría una observación y capacidad de veto por parte del humano y; "fuera de la acción" (*out-of-the-loop*) donde el sistema de armamento no dependería de un operador humano para ejecutar las decisiones o acciones, incluidas aquellas potencialmente letales de activación del armamento.

A nuestro entender y una vez analizadas las diversas definiciones propuestas, el término SAA nos parecería el más preferible pues, como indica H.-Y. Liu, se estaría en un término lo suficientemente abierto para no incidir demasiado en los aspectos de robótica o de letalidad, mientras que el término "autónomo" establecería la principal característica de independencia frente a otros sistemas de armamento. (Kass, 2015: 1022; Liu, 2016: 833-834; O'Connell, 2014: 224; Rogers, 2014: 1258).

La autonomía es también la principal faceta que se establece en la Directiva del Departamento de Defensa de USA (3000.09, 21 noviembre 2012) sobre la "Autonomía en los Sistemas de Armamento" (*Autonomy in Weapon Systems*). Para el DoD, un SAA sería

aquel que, una vez activado, pudiese seleccionar y atacar objetivos sin más intervención por parte de un operador humano, aunque también incluiría aquellos SAA designados para permitir que un operador humano pudiese anular la operación. Así, se estaría hablando de sistemas de armamentos tanto "*on-the-loop*" como "*out-of-the-loop*". Una variante de dicha definición la aportó los Países Bajos a la reunión del "Grupo de Expertos Gubernamentales" (GGE) sobre los SAAL[21] de 2017, en la que definió un SAA como: "Un armamento que, sin ninguna intervención humana, selecciona y ataca objetivos que se ajustan a ciertos criterios predefinidos, siguiendo la decisión humana de desplegar el armamento con el conocimiento que un ataque una vez lanzado, no puede ser parado por una intervención humana". En este caso se estaría en la categorización "*out of the loop*" (DoD, 2012: 13-14; NU, 2017b: 1).

De forma similar sería la conceptualización de los SAAL. La principal diferencia estribaría en el elemento de "letalidad" (*lethal*) y por lo tanto se estaría hablando de un subconjunto de los SAA, aunque de igual forma no exista en la actualidad un consenso internacional sobre una definición de dicho término. De todas formas, si existen una serie de conceptos que se deberían tener en cuenta para establecer un marco de definición y que, a nuestro entender, queda claramente expuesto en la aportación de Bélgica a la reunión del GGE sobre los SAAL de 2017. En cuanto a la "autonomía", se especifica que se refiere a la necesidad de que exista una autonomía total en el proceso de una toma de decisión letal, esto es, la capacidad de que el sistema de armamentos pase a un modo letal o que pueda infligir heridas a una persona humana sin ninguna

21 En todo caso no existe una definición internacional consensuada y, por lo tanto, dicha postura era una más en los debates (2017b: 1-2).

supervisión previa o marginal, pero también debe incluir la incapacidad de revertir la decisión a través de un cambio de modalidad o desactivación. Además, se debería tener en cuenta el elemento de la "intencionalidad" para establecer consecuencias letales y debería existir un elemento de incertidumbre en la división de la autoridad entre el ser humano y la máquina para establecer dicha intencionalidad. También debería existir una falta de acotamiento del conocimiento sobre los posibles comportamientos del SAAL y la impredecibilidad de sus acciones una vez activado. En dicho punto habría que destacar su posible capacidad para redefinir autónomamente los criterios en los que pudiese operar, teniendo en cuenta los posibles cambios en el medio en el que actúa, los objetivos o la misión a desarrollar. Dicho marco de referencia sería similar al expresado por China sobre los SAAL, en su aportación al GGE sobre los SAAL de 2018[22], con la diferencia de que dicho Estado añade el elemento del "efecto indiscriminado", donde la actuación se llevaría a cabo independientemente de las condiciones, escenarios u objetivos. Dichas definiciones difieren enormemente de las planteadas por Rusia, en su aportación al GGE sobre los SAAL de 2019[23], dado que la definición no incide sobre el término de la "letalidad" y es bastante restrictivo. Así, la definición establece: "medios técnicos no tripulados fuera de los artefactos, municiones o artillería militar, destinados llevar a cabo misiones de combate o de apoyo sin ninguna participación de un operador (NU, 2017c: 1-2; NU, 2018b: 1; NU, 2019h).

22 El informe de 2018 del GGE del CCW sobre los SAAL debatió por primera vez los posibles principios rectores (NU, 2018a: 4).

23 La postura de la Federación Rusa viene plasmada en su ordenamiento legal y específicamente en el "*Internal Service Regulation of the Armed Forces of the Russian Federation, N. 1455, 16 November 2007*" (NU, 2019h: 2).

En este punto habría que incidir en que se entiende por "autonomía" de un SAAL. Para Estonia y Finlandia, en su aportación al GGE sobre los SAAL de 2018, los rápidos desarrollos tecnológicos hacen que la noción de "autónomo completo" (*fully autonomous*) sea problemática, ya que el concepto de "autonomía" es relativo, idea que nosotros compartimos. Por lo tanto, dichos Estados hacen una distinción entre automatización, autonomía e independencia. La "automatización" se entiende como que existe el conocimiento de unas respuestas preprogramadas y predecibles de una tarea en cualquier situación. En el caso de los sistemas de armamentos, aunque sus algoritmos sean robustos, siempre existirá un grado de comportamiento probabilístico de aleatoriedad, aunque los subsistemas individuales sean determinísticos. Asumir que la autonomía siempre produciría sistemas estables y seguros no sería válido para sistemas complejos (NU, 2018c: 2).

En cuanto a la "autonomía" se debería entender como la capacidad para desarrollar una acción de una forma autosuficiente y auto gobernable. Se incluye la libertad de una planificación propia en las tareas y subtareas, concepto ampliamente conocido en el ámbito informático. En dicho contexto, la programación y las estructuras de control detrás de los sistemas de IA están diseñados fundamentalmente sobre la base de la ejecución de tareas. Dichas tareas tendrían diferentes grados de complejidad y de interacción con los humanos o con otras sistemas. Por lo tanto, el concepto de "autonomía" no podría ser una característica simple de "encendido/apagado" y, por lo tanto, en vez de "sistemas autónomos" el concepto se debería expresar como "sistemas que tienen funciones o características autónomas", siendo difícil definir el grado de autonomía ya que cada sistema sería diferente. En cuanto a los SAAL, el enfoque se debería poner, por tanto, en el ciclo de selección de

objetivos y las condiciones de autorización del uso de la fuerza letal, especialmente el retardo entre la orden y la ejecución, comprendiendo la dinámica de la tarea y la ventana de tiempo existente para la autorización. En contraste con una operación autónoma, la "independencia" verdadera significaría que el sistema sería capaz de definir y decidir los objetivos finales de su funcionamiento, de la misma forma que realizan los humanos. Por lo tanto, la selección de objetivos estaría subordinada a la propia motivación del sistema, para lo que se requeriría una IA que hubiese evolucionado más allá del punto de la Singularidad, paradigma que trataremos en nuestra investigación más adelante (NU, 2018c: 2-3; NU, 2019b: 4).

En dicho contexto un elemento conceptual crítico, relativos tanto para los SAA como para los SAAL, sería el establecer un consenso internacional sobre qué se entiende por "intervención humana" o "intervención humana significativa". A tal fin, la mayor parte de los expertos inciden, idea que nosotros compartimos, en que la base principal estaría formada por una adecuada aplicación del DIH, más específicamente con relación al Artículo 36 del Primer Protocolo Adicional (1977) a los Convenios de Ginebra de 1949[24]. En dicho contexto, es interesante destacar la posición española, emitida en mayo de 2019, sobre la Resolución aprobada por la Asamblea General de las NU, el 5 de diciembre de 2018, sobre la *Función de la ciencia y la tecnología en el contexto de la seguri-*

24 **Artículo 36**: "cuando una Alta Parte contratante, estudie, desarrolle, adquiera o adopte una nueva arma o nuevos medios o métodos de guerra tendrá la obligación de determinar si su empleo, en ciertas condiciones o todas las circunstancias estaría prohibido por el presente Protocolo o por cualquier otra norma de derecho internacional aplicable a esa Alta Parte contratante" (BOE, 1989: 177(23837).

dad internacional y el desarme[25]: "El control humano significativo sobre el uso de armas y sus efectos es esencial para asegurar que el uso de un arma sea éticamente justificable y legal… Para que el control humano sea significativo, … debe haber la oportunidad de que haya un juicio y una intervención humana oportuna. Es responsabilidad del Estado garantizar que el despliegue de cualquier sistema de armas cumple los requisitos del Derecho Internacional"[26] (BOE, 1989: 177(23837); NU, 2019c: 3).

Para llevar a cabo dicho cometido, a nuestro entender, la intervención humana debería extenderse a través de todo el ciclo de vida del sistema de armamentos. Sin adentrarnos, en este momento, en los detalles que retomaremos más adelante en nuestra investigación, tomaremos como referencia el punto de vista del Reino Unido (UK), en su aportación al GGE sobre los SAAL de 2018, por considerarlo extenso y relativamente completo, en las que establece que el control humano sería un ciclo continuo que debería tener en cuenta una serie de factores en las diversas etapas del desarrollo y puesta en funcionamiento de un SAA (NU, 2018d: 3):

- *Reglamentación nacional e internacional*: incluyendo el Derecho Internacional apropiado;

25 Para más información ver: NACIONES UNIDAS (2018e): *Función de la ciencia y la tecnología en el contexto de la seguridad internacional y el desarme, Naciones Unidas*, Nueva York, A/RES//73/32, acceso diciembre 2020, en https://undocs.org/es/A/RES/73/32

26 No obstante, España no forma parte del grupo de Estados que proponen una prohibición de los SAA, que serían permisibles si existiese un control humano significativo y una comprensión humana de su comportamiento con la posibilidad de intervención, según la declaración en el GGE sobre los SAAL de 2019 (Delàs, 2019: 31-32; NU, 2019d).

- *Especificación*: la especificación de las necesidades de los sistemas que garantice un control humano apropiado;

- *Diseño*: diseñado para requerir y facilitar un control humano con un enfoque particular en la intervención hombre-máquina;

- *Verificación, validación y certificación*: de los procesos incluidos las revisiones legales (específicamente las relativas al Artículo 36);

- *Procedimientos y procesos operacionales*: formación, mando y control, doctrina y Reglas de Enfrentamiento (ROE).

Por último, al igual que indica Francia en su aportación al GGE sobre los SAAL de 2018, debería ser posible evaluar un sistema de armamentos en relación con su fiabilidad y su previsibilidad, al igual que ya hemos observado en la posición de Estonia y Finlandia. En realidad, según la posición francesa, lo fundamental sería la capacidad del operador (con la ayuda del diseñador) de evaluar y controlar el sistema. Para ello, tanto el operador como el mando deberían comprender como funciona un sistema y sus efectos potenciales en el teatro de operaciones, idea también indicada por España, para lo cual se deberían implementar fuertes procedimientos de verificación, evaluación y validación. Además, cuando fuese necesario, se deberían llevar a cabo pruebas exhaustivas y estrictas sobre los algoritmos de los sistemas, especialmente en circunstancias tanto difíciles como deterioradas y en condiciones estadísticamente anormales (NU, 2018e: 2).

Un resumen apropiado de las medidas a llevar a cabo en el control humano, lo aporta el documento preparado por SIPRI y el CICR denominado "*Limits on Autonomy in Weapons Systems. Identifying Practical Elements of Human Control*" (Límites en la Autonomía de Sistemas Armamentísticos: Identificando Elementos Prácticos de Control Humano), de junio de 2020. El estudio establece que para un control humano adecuado sería necesario una combinación de tres tipos de medidas de control (SIPRI y CICR, 2020: ix)

- *Control de los parámetros de uso de los sistemas armamentísticos SAA*: medidas para restringir el tipo de objetivo y la tarea de uso; poner límites atemporales y espaciales sobre su operación; limitar los efectos y; permitir la desactivación y la inclusión de mecanismos a prueba de fallos;

- *Control del entorno*: medidas que controlen o estructuren el entorno en el que se utiliza un SAA (utilizándolo en entornos donde no hay civiles, etc.)

- *Controles a través de la interacción hombre-máquina*: Medidas que permiten al usuario supervisar al SAA e intervenir en su operación cuando sea necesario.

A la vista de este análisis, reiteramos la necesidad de mantener una mentalidad holística, pues consideramos, al igual que propugna ENISA, que la ciberguerra forma parte de la protección de la estabilidad global, que a su vez se incluye dentro del entorno global de la ciberseguridad. Así mismo, es obvio que los medios cibernéticos tendrán cada vez un rol más importante en los conflictos armados y que el impacto de las computadoras y de la IA producirán importantes cambios del rol humano en las guerras. Las "guerras

posmodernas" de C. H. Gray, término que nosotros subscribimos, incrementarán su complejidad que, a su vez, incidirá sobre el Derecho aplicable a los conflictos armados (DICA) y el DIH, especialmente al aplicar las nuevas teorías de la disuasión o las nuevas tecnologías en los conflictos armados.

Particularmente, con relación a los SAA y lo SAAL, consideramos fundamental que se llegue a un consenso internacional sobre la definición de dichos términos y, en particular, la definición de lo que se entiende por "autonomía". Una visión estricta, plantearía la definición de dicho concepto como la capacidad de una objeto para desarrollar un acción de forma autosuficiente y auto gobernable, lo que denotaría que dichos objetos habrían alcanzado la Singularidad. Ahora bien, la postura internacional más consensuada cuando se habla de los SAA y los SAAL, se refieren tanto a los sistemas completamente autónomos, como aquellos que solo tienen ciertos grado de autonomía, donde el rol humano seguirá siendo significativo. Por lo tanto, a lo largo de nuestra investigación, adoptaremos dicha postura consensuada al referirnos a los SAA y los SAAL. En los próximo capítulos incidiremos con más profundidad sobre el impacto que los SAA y los SAAL, así como el advenimiento de la Singularidad, tendrán sobre la responsabilidad penal internacional de los conflictos armados y analizaremos la posible utilización de los AMA como herramienta de control para dichos sistemas armamentísticos.

CAPÍTULO 4
RESPONSABILIDAD JURÍDICA INTERNACIONAL

Los tratados y costumbres internacionales están en el punto de mira cuando se intentan aplicar al ciberespacio. Al igual que K. Watkin, nuestro punto vista se acerca a la idea de que las acciones en el ciberespacio estarían enmarcadas dentro de la denominada "lucha en las fronteras de la ley" (*fight at the legal boundaries*). Un concepto que desarrolla la problemática de poder establecer hasta qué punto el Derecho de los conflictos armados (DICA), el Derecho Internacional Humanitario (DIH), los Derechos Humanos (DD.HH.), el Derecho Internacional consuetudinario y las propias leyes de los Estados serían aplicables en el ciberespacio, teniendo en cuenta que dicho medio podría generar zonas oscuras y ambiguas, especialmente donde existiese un solapamiento entre ellas (Marín Martínez, 2019: 29-30; Watkin, 2016: 3-30).

Aunque los principios del Derecho Internacional serían de aplicación[27], teóricamente, en el ciberespacio, los nuevos conceptos

27 Tomando como referencia el Informe del GGE sobre SAAL de 2019, los Principios Rectores establecen que: "el Derecho Internacional, en particular la Carta de las Naciones Unidas y el Derecho Internacional Humanitario, así como las perspectivas éticas pertinentes, debían guiar permanentemente la labor del Grupo" y que "El Derecho Internacional Humanitario sigue aplicándose plenamente a todos los sistemas de armas, incluido el posible desarrollo y uso de Sistemas de Armas Autónomos Letales". Dichos principios fueron avalados por el CCW en noviembre 2019 (NU, 2019a: 15; 2019e).

como la ciberguerra o la guerra híbrida, anteriormente analizados, pueden que incidan en gran medida en la capacidad para aplicar dicho ordenamiento jurídico en su totalidad, como sugieren S. J. Shackelford y J. B. Sher. En la mayor parte de los casos, las acciones cibernéticas desarrolladas por los Estados están por debajo del umbral del uso de la fuerza, consistiendo en intrusiones persistentes de bajo nivel que pueden causar daño al Estado víctima, pero que normalmente resulta difícil discernir si causan efectos físicos apreciables. Así, China indicó que en 2017 había sufrido incidentes de ciberseguridad por valor de 60 billones de dólares, así mismo, UK estableció, en 2019, que en los últimos tres años había sido objeto de 1800 ciberataques, la mayor parte con origen en otros Estados y, en diciembre de 2020, USA informó que durante dicho año había sufrido uno de los mayores ciberataques de su historia, denominado "*Sunburst*", cuya autoría se atribuyó, en un principio, a otro Estado (BBC, 2020; Moynihan, 2019: 2; Shackelford, 2009: 194-195; 196; Sher, 2016: 241).

4.1.- ANTECEDENTES

En todo caso, los peligros del ciberespacio ya habían sido aventurados con anterioridad. Durante la década de 1970, la comunidad internacional se dio cuenta de los posibles peligros de las nuevas tecnologías, lo que propició una resolución (32/152 de 19 de diciembre de 1977) de la Asamblea General de las Naciones Unidas para llevar a cabo una Conferencia para tratar en parte dichos temas. Llevada a cabo en 1979, el resultado fue el desarrollo de la convención CCW y sus protocolos I (PAI), II (PAII) y III (PAI-II), adoptados el 10 de octubre de 1980 y que entraron en vigor en diciembre 1983. Dicha Convención tiene entre sus objetivos la

prohibición de armamentos que no hacen distinción entre civiles y combatientes o causan lesiones o un sufrimiento innecesario, así como la reafirmación de la "Cláusula Martens"[28]. Así mismo, en 1987, el CICR en su "Comentario sobre los Protocolos Adicionales de 1977 a la Convención de Ginebra de 1949", sobre el Artículo 36 del PAI, lanzaba la advertencia de que la automatización progresiva en el campo de batalla podría llevar a la situación de qué si "el hombre no domina la tecnología, pero permite que la tecnología lo domine, será destruido por la tecnología" (CICR, 1987: 427-428; UNODA, 2014: 1-2).

Siguiendo con dicha evolución, el "Informe Provisional del Relator Especial del Consejo de Derechos Humanos sobre las ejecuciones extrajudiciales, sumarias o arbitrarias (A/65/321)", P. Alston, presentado en cumplimiento de lo dispuesto en la resolución 63/182 de la Asamblea de las Naciones Unidas, en 2010, establecía que existían dos postulados principales en el análisis del impacto de las nuevas tecnologías sobre el Derecho Internacional: que la nuevas tecnologías tenían ramificaciones muy importantes respecto al derecho a la vida y a la lucha contra las ejecuciones extrajudiciales, así como que no existía ninguna razón inherente para que las consideraciones relacionadas con los DD.HH. y el DIH no pudie-

28 La "Cláusula Martens" forma parte del DCA desde el Preámbulo del II Convenio de La Haya de 1899 relativo a las leyes y costumbres de la guerra terrestre: "Mientras que se forma un Código más completo de las leyes de la guerra, las Altas Partes Contratantes juzgan oportuno declarar que, en los casos no comprendidos en las disposiciones reglamentarias adoptadas por ellas, las poblaciones y los beligerantes permanecen bajo la garantía y el régimen de los principios del Derecho de Gentes preconizados por los usos establecidos entre las naciones civilizadas, por las leyes de la humanidad y por las exigencias de la conciencia pública" (CICR, 1997).

sen ser dinámicamente incluidas en el diseño y el funcionamiento de las nuevas tecnologías[29]. Es más, la comunidad internacional necesitaba urgentemente ocuparse de las repercusiones jurídicas, políticas, éticas y morales del desarrollo de las tecnologías de los robots que fuesen mortíferas (NU, 2010: 13).

En 2011, la "*International Strategy for Cyberspace*" (Estrategia Internacional para el Ciberespacio) de los USA, establecía que: "el desarrollo de las normas para el comportamiento de un Estado en el ciberespacio no requiere una reinvención del Derecho Internacional consuetudinario, ni hace que las normas internacionales existentes sean obsoletas. Las normas internacionales tradicionales que: "guían el comportamiento de un Estado – en tiempos de paz o conflicto – también son aplicables en el ciberespacio". Ahora bien, en el mismo documento se establecía que "los atributos singulares de la tecnología de redes requieren de un trabajo adicional de cómo se aplicarían dichas normas y que entendimientos adicionales podrían ser necesarios para suplementarlas". Dicha postura se alineaba con la "*Opinión Consultiva ante la Corte Internacional de Justicia sobre 'Armamento Nuclear' (Nuclear Weapons)"*[30], sobre si la prohibición del uso de la fuerza de acuerdo con lo establecido en el artículo 2(4) de la Carta de las Naciones Unidas, gobernaba el uso de armamento nuclear. La opinión indicaba que se "aplicaba a

29 Será dicho segundo postulado el que forme la base de nuestra investigación actual.

30 Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226 (hereinafter, Nuclear Weapons). Para más información ver: INTERNATIONAL COURT OF JUSTICE (1996): *Legality of the Threat or Use of Nuclear Weapons,* Corte de Justicia Internacional, acceso enero 2021, en https://www.icj-cij.org/en/case/95

cualquier uso de la fuerza, independientemente del armamento utilizado". Según M. N. Schmitt, dicha opinión habría proporcionado a los expertos que desarrollaron el "Manual de Tallinn" (*Tallinn Manual*) de la OTAN[31], la base para indicar que: "el mero hecho de que una computadora (en vez de un armamento más tradicional, un sistema de armamento o una plataforma) fuese utilizada durante una operación no tiene ningún efecto sobre si la operación significa un 'uso de la fuerza'" (OTAN, 2013; Schmitt, 2012a: 13-14, 16).

No obstante, se siguieron manteniendo discusiones en los foros internacionales. Así, el "I*nforme del Relator Especial sobre las ejecuciones extrajudiciales, sumarias o arbitrarias (A/HRC/23/47)*", C. Heyns, ante el Consejo de Derechos Humanos, en 2013, permitió seguir creando conciencia sobre la creciente importancia del ciberespacio, sobre el Derecho Internacional. Dicho informe se concentró en los "robots autónomos letales", indicando el Relator su preocupación con respecto a la protección de la vida en el marco de los DD.HH. y el DIH y recordando "la primicia y el carácter imperativo del derecho a la vida en virtud de los tratados y del Derecho Internacional consuetudinario". En dicho informe, una consideración destacada planteaba si era técnicamente posible programar robots autónomos letales para que cumpliesen las exigencias del DIH de manera más estricta que los seres humanos, para lo cual hacía referencia a los trabajos del investigador J. Herbach, que postulaba que las actuales tecnologías serían incapaces de distinguir entre combatientes y la población civil, especialmente en combates urbanos. Por lo tanto, sería necesario que dichos SAAL pudiesen evaluar y reaccionar ante las ambigüedades, lo que im-

31 Manual de Tallinn sobre la "aplicación del Derecho Internacional a la Guerra Cibernética" (OTAN, 2017).

plicaría la presencia del factor humano y la necesidad de satisfacer los principios del DIH, como la distinción y la proporcionalidad. A tal fin, sería necesario que los requerimientos del Derecho Internacional, como el principio de precaución, jugasen un rol integral en la investigación y el desarrollo de dichos sistemas (Herbach, 2012: 19; NU, 2013: 14).

Posteriormente, un importante paso se dio al establecerse el CCW como el foro para discutir las tecnologías emergentes dentro del marco de los SAAL. Después de una serie de discusiones informales en Ginebra entre 2014 y 2016, se estableció la necesidad de crear un GGE sobre SAAL con un mandato formal en la "Quinta Conferencia de Examen de las Altas Partes Contratantes en la Convención sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales que Puedan Considerarse Excesivamente Nocivas o de Efectos Indiscriminados" de diciembre de 2016[32]. En el 2018, el GGE propuso 10 principios rectores incluyendo: la aplicabilidad del DIH; la no-delegación de la responsabilidad humana; la responsabilidad sobre el uso de la fuerza de acuerdo con el Derecho Internacional; revisiones de los sistemas de armamentos antes de su despliegue; incorporación de salvaguardas de seguridad físicas, de no proliferación y cibernéticas; evaluación y mitigación de riesgos durante el desarrollo de la tecnología; consideración del uso de las tecnologías emergentes en el área de SAAL de acuerdo con el DIH; investigación y desarrollo no dañinos en el desarrollo y sus uso; la necesidad de adoptar perspectivas no antropomorfas

32 Para más información ver: NACIONES UNIDAS (2016): *Documento Final de la Quinta Conferencia de Examen (CCW/CONF.V/10)*, Organización de las Naciones Unidas, Ginebra, acceso en enero 2021, en https://undocs.org/es/CCW/CONF.V/10

sobre la IA y; conveniencia del CCW como marco para abordar el tema (NU, 2018a: 4-5).

Además de dichos principios, el GGE presentó cuatro opciones para el desarrollo de la política sobre los SAAL, incluyendo la posibilidad de limitaciones jurídicamente vinculantes, utilizando los principios rectores y los componentes básicos. Con respecto a los relativos a la caracterización, estos incluirían la necesidad de mantener el foco sobre el elemento humano en el uso de la fuerza. Sobre las consideraciones de la interfaz entre hombre-máquina estas se deben construir a través de la dirección política en la fase del predesarrollo; el desarrollo y la investigación; ensayo, evaluación y certificación; despliegue, formación, mando y control; uso y cancelación y; evaluación después de la utilización. El GGE también estableció que la rendición de cuentas sería el hilo conductor en todos los aspectos de la interfaz hombre-máquina en el contexto del CCW. En cuanto a las cuatro opciones de política posibles estas fueron (NU, 2018a: 7):

- Una propuesta de instrumento jurídicamente vinculante que estableciera prohibiciones y reglamentaciones sobre los SAAL;

- Una propuesta de declaración política esbozando principios importante como la necesidad del control humano en el uso de la fuerza y la importancia de la rendición de cuentas, así como elementos de transparencia y examen de la tecnología;

- Propuestas para seguir examinando la interfaz humano-máquina y la aplicación de las obligaciones jurídicas internacionales existentes, subrayando la necesidad de determinar medidas prácticas, prácticas óptimas y un intercambio de información para mejorar el cumplimiento del Derecho Internacional, con especial atención a los exámenes jurídicos de las armas exigidos por el Artículo 36 del PAI a los Convenios de Ginebra;

- Opiniones de que no era necesario adoptar nuevas medidas jurídicas, ya que el DIH era plenamente aplicable a los SAAL.

En el mismo año, la Asamblea General de las NU aprobó una Resolución sobre la "*Promoción del comportamiento responsable de los Estados en el ciberespacio en el contexto de la seguridad internacional (A/RES/73/266)*". Siguiendo los informes del GGE de 2013[33] y 2015[34], confirmó que: "el Derecho Internacional y, en

33 Para más información ver: NACIONES UNIDAS (2013): Grupo de Expertos Gubernamentales sobre los Avances en la Información y las Telecomunicaciones en el Contexto de la Seguridad Internacional. Organización de las Naciones Unidas, Ginebra, acceso en enero 2021, en https://undocs.org/es/A/68/98

34 Para más información ver: NACIONES UNIDAS (2015): Grupo de Expertos Gubernamentales sobre los Avances en la Información y las Telecomunicaciones en el Contexto de la Seguridad Internacional. Organización de las Naciones Unidas, Ginebra, acceso en enero 2021, en https://undocs.org/es/A/70/174

particular, la Carta de las Naciones Unidas, son aplicables y fundamentales para mantener la paz y la estabilidad y fomentar un entorno abierto, seguro, estable, accesible y pacífico en la esfera de la tecnología de la información y las comunicaciones …". Es importante indicar, que el GGE de 2015 destacaba la aplicación del principio de soberanía y llamaba la atención sobre "la existencia de principios jurídicos internacionales establecidos, entre ellos, de ser aplicables, los principios de humanidad, necesidad, proporcionalidad y distinción". La aplicabilidad del Derecho Internacional se mantuvo durante de la reunión consultativa informal del GGE en 2019 (NU, 2015: 3; NU, 2018f: 2; NU, 2019f).

La reunión del CCW de noviembre de 2019, aprobó una nueva serie de principios rectores establecidas por el GGE, que contienen principios ya establecidos en 2018 y nuevas consideraciones (NU, 2019e: 11-12):

- El DIH sigue aplicándose plenamente a todos los sistemas de armas, incluido el posible desarrollo y uso de SAAL;

- El ser humano debe mantener la responsabilidad por las decisiones que se adopten sobre el uso de los sistemas de armas, ya que la obligación de rendir cuentas no puede transferirse a las máquinas. Esa consideración debería tenerse en cuenta durante todo el ciclo de vida del sistema de armas;

- La interacción hombre-máquina debe garantizar que el posible uso de sistemas de armas basados en tecnologías emergentes en el ámbito de los SAAL se ajuste al Derecho Internacional aplicable, en particular al DIH;

- La rendición de cuentas sobre el desarrollo, el despliegue y empleo de nuevos sistemas de armas en el marco de la Convención debe garantizarse conforme al Derecho Internacional aplicable, bajo el control humano;

- Cuando se estudie, desarrolle, adquiera o adopte una nueva arma, o un nuevo medio o método de guerra, habrá que determinar si su empleo, en ciertas condiciones o en todas las circunstancias, estaría prohibido por el Derecho Internacional;

- Al desarrollar o adquirir nuevos sistemas de armas basados en tecnologías emergentes en el ámbito de los SAAL, deberían tenerse en cuenta la seguridad física, las salvaguardias no físicas adecuadas (incluida la ciberseguridad), el riesgo de adquisición por grupos terroristas y el riesgo de proliferación;

- Las evaluaciones de riesgos y las medidas de mitigación deberían formar parte del ciclo de diseño, desarrollo, ensayo y despliegue de tecnologías emergentes en cualquier sistema de armas;

- Debería considerarse la posibilidad de utilizar las tecnologías emergentes en el ámbito de los SAAL para garantizar el cumplimiento del DIH y otras obligaciones jurídicas internacionales aplicables;

- Al elaborar posibles medidas de política, no se deberían antropomorfizar las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales;

- Los debates y las posibles medidas de política que se adopten en el contexto de la Convención no deberían obstaculizar el progreso ni el acceso a los usos pacíficos de las tecnologías autónomas inteligentes;

- La Convención ofrece un marco apropiado para abordar la cuestión de las tecnologías emergentes en el ámbito de los SAAL.

Debido a la pandemia del "Covid19", se produjeron diversos retrasos en el desarrollo de las reuniones del GGE sobre los SAAL en 2020. Una primera consecuencia fue que se pidió a los Estados y otros organismos que indicasen sus ideas con relación a los principios rectores establecidos en 2019. El CICR indicó que daba la bienvenida a los principios establecidos en el GGE, pero que aparte las restricciones establecidas por el DIH, podría ser necesario establecer otras derivadas de consideraciones éticas, incluidas

aquellas sobre los principios humanitarios o los dictados por la conciencia pública. En definitiva, para el CICR, el eje principal no es si el DIH es aplicable, sino como se aplican las reglas, se interpretan y se implementan en la práctica. Con relación a las revisiones establecidas por el Artículo 36 del Protocolo Adicional I, el principio debería ser el asegurar el respeto hacia el DIH y para ello una atención particular debería darse hacia las medidas necesarias para establecer el control humano sobre el armamento y el uso de la fuerza (CICR, 2020).

En el caso de España, en los comentarios sobre los principios rectores sobre el respeto hacia el DIH, de junio de 2020, indicaba que se debería requerir un control humano suficiente sobre todos los sistemas de armamento, a lo largo de todo el ciclo de vida, y la atribución de una responsabilidad legal con relación al operador y la persona que autoriza su uso. Aquellos sistemas "fuera de la acción" (*out-of-the-loop*) deberían ser considerados como incompatibles con la necesidad de un control humano significativo (MHC). Paralelamente, en un informe de la ONG "Human Rights Watch", sobre la posición de los diversos Estados hacia la prohibición de los SAAL, con referencia a España, también especifica la necesidad de un control humano suficiente sobre todos los sistemas de armamento, con relación al respeto hacia el DIH (NU, 2020a; HRW, 2020: 47-48).

Con referencia a USA, su postura era, en septiembre 2020, que no había sido persuadida de la necesidad de establecer un nuevo tratado, ya que el DIH existente proporcionaba un marco coherente y robusto para la reglamentación de las tecnologías emergentes en el área de los SAAL. También establecía que los Principios Rectores sirven como base para los trabajos del GGE. Uno de los aspec-

tos más interesantes emergentes sobre la postura de USA sobre la responsabilidad de las fuerzas armadas, especifica que dicha se establecería a través de un sistema de Derecho Militar y de disciplina. El informe de HRW también alude a que no es necesario un nuevo tratado militar, siendo el actual sobre DIH suficiente. Un hecho de relevancia fue la no participación de Rusia durante la primera sesión del GGE de septiembre 2020, llevada a cabo telemáticamente[35], que según algunos observadores podría significar una forma de retrasar los trabajos o una señal de que Rusia pretendería retirarse completamente de dichos trabajos[36]. En cuanto a China, no ha habido cambios significativos de su postura desde 2018, donde los SAAL deben estar sujetos, en principio a las reglas del DIH establecidas por la Convención de Ginebra de 1949 y los dos Protocolos Adicionales de 1977, incluyendo los principios de precaución, distinción y proporcionalidad (HRW, 2020: 52-53; Lewis, 2020; NU, 2018b: 2; 2020b).

A lo largo de 2021, se volvieron a retomar los trabajos del GGE sobre los SAAL. Las posturas de los principales Estados (China, Rusia, USA) no habrían variado sustancialmente pues, aunque

35 Para más información sobre dicha sesión ver: NACIONES UNIDAS (2020c): *10th Meeting - 1st Session Group of Governmental Experts on Lethal Autonomous Weapons Systems 2020*, Organización de las Naciones Unidas, Ginebra, acceso enero 2021, en http://webtv.un.org/watch/10th-meeting-1st-session-group-of-governmental-experts-on-lethal-autonomous-weapons-systems-2020/6194570749001/

36 Sería un paso más a las objeciones de Rusia planteadas en 2019, sobre la limitación de las discusiones del GGE a "sistemas completamente autónomos" (*fully autonomous systems*), al mismo tiempo que se opone al desarrollo de un mecanismo obligatorio universal de "revisión legal" con referencia a los sistemas de armamentos de acuerdo con el Artículo 36 del Protocolo Adicional I (NU, 2019g:1-2)

todos ellos establecerían que el Derecho Internacional y en particular el DIH serían de pleno cumplimiento, Rusia se opondría al establecimiento de un instrumento internacional jurídicamente vinculante con relación a los SAAL en el GGE o la imposición de una moratoria en el desarrollo y uso de dichos sistemas y la tecnologías utilizadas para crearlos. Por su parte, China mantendría que los principios rectores ya establecidos serían una buena base para nuevas discusiones dentro del GGE y que deberían ser los propios Estados los que deberían tomar las medidas necesarias para implementarlos dentro de su ordenamiento jurídico y su régimen de reglamentación militar, promocionando el intercambio internacional de información de una forma voluntaria. En el caso de los USA, su propuesta detallaría como implementar los principios rectores en cuatro ámbitos: la aplicación del DIH (especialmente en el área de "focalización de objetivos"); la responsabilidad humana; la interacción hombre-máquina y; la revisión de armamento, informando de los pasos llevados a cabo para implementar dichos aspectos en su ordenamiento jurídico y reglamentación militar (NU, 2021a; 2021b; 2021c; 2021d).

Con relación a la postura de España, profundizando en sus propuestas de 2020, reiteraría la necesidad de un control humano en todo el ciclo de vida de los sistemas y que las guías operacionales tuviesen en cuenta los riesgos relacionados con los posibles sesgos de los sistemas, los niveles de confianza o el posible "hackeo", estableciendo la necesidad de una formación adecuada de los operadores de dichos sistemas. Además, propondría que el marco legal internacional de los DD.HH. debería ser tomado en consideración cuando se desplegasen los SAAL. De especial relevancia para España se considera que el foco de las recomendaciones del GGE debería centrarse en los procesos de "focalización de obje-

tivos" (*targeting*) e "intervención sobre objetivos" (*engagement*). No propone ningún nuevo instrumento legal para aplicar el DIH en los SAAL, sino que, a través de revisiones del marco jurídico, en caso de que existiesen problemas para su aplicación, se deberían implementar nuevas doctrinas y capacidades formativas para que el armamento pudiese ser utilizado de acuerdo con el DIH. Se proponen pruebas exhaustivas durante el proceso de investigación y desarrollo en condiciones cercanas al mundo real y procesos de validación extensos. Además, propone, al igual que China, un intercambio de información internacional de forma voluntaria También consideramos de especial relevancia la propuesta del CICR a los trabajos del GGE. Los SAAL no predecibles deberían ser abolidos, así como su uso específico contra las personas. Por lo tanto, debería haber límites sobre los tipos de objetivos, la duración, el ámbito geográfico o la escala de uso de los SAAL, unido a un proceso de control humano exhaustivo, a través de una interacción hombre-máquina efectiva (NU, 2021e; NU, 2021f).

4.2- ENTRE EL *IUS AD BELLUM* Y EL *IUS IN BELLO*

Como hemos analizado en el anterior punto, los Estados están generalmente de acuerdo en que el Derecho Internacional que regula el uso de la fuerza (*ius ad bellum*), aunque relativamente antiguo, sería aplicable en el ciberespacio, aunque la pregunta que se plantea es si los avances tecnológicos lo han dejado anticuado y en todo caso cuales serían las diversas modalidades posibles para su aplicación. El investigador M. Roscini, haciendo un símil con la época romana lo denomina como "*hic sunt leones*", (aquí hay leones), frase que solían escribir los cartógrafos romanos y medievales en los mapas, para describir territorios inexplorados. Expresión

que mantiene su valor cuando entran en acción las operaciones cibernéticas, dentro del nuevo paradigma de la guerra híbrida, donde no queda claramente establecido en qué momento se les puede considerar como uso de la fuerza o ataque armado[37]. Otra forma de expresar la situación la desarrolla M. N. Schmitt al hablar de las "zonas obscuras" (*grey zones*) del Derecho Internacional y la capacidad de los Estados de explotar dichas zonas a través de sus estrategias en el ciberespacio. Dichas advertencias son esenciales a tener en cuanta cuando se analiza el Derecho Internacional en el ciberespacio (Marin Martínez, 2019: 30; Roscini, 2010: 86; Schmitt, 2017: 1).

El concepto de "*ius ad bellum*" prohíbe que los Estados utilicen la fuerza entre fronteras, excepto cuando existe: el consentimiento del Estado; una autorización del Consejo de Seguridad de las Naciones Unidas o; en "legítima defensa". Además, la Carta de las Naciones Unidas prohíbe el uso no autorizado de la fuerza excepto en legítima defensa como respuesta a un ataque armado[38]. Ahora bien, el elemento jurídico que se plantea radica en cómo se catalogan las acciones en el ciberespacio que permitan dilucidar sin ninguna duda que se está ante un quebrantamiento del Derecho

37 Casos recientes serían los ejemplos de "Stuxnet", sobre la capacidad nuclear de Irán en 2010, atribuido en un principio a Israel y USA o "Sunburst", el ataque masivo contra USA en 2020, atribuido a Rusia (BBC, 2020; Libicki, 2009: 14; Nguyen, 2013: 1082-1083).

38 **Artículo 51**: "Ninguna disposición de esta Carta menoscabará el derecho inmanente de legítima defensa, individual o colectiva, en caso de ataque armado contra un Miembro de las Naciones Unidas, hasta tanto que el Consejo de Seguridad haya tomado las medidas necesarias para mantener la paz y la seguridad internacionales. …" (NU, 1945).

Internacional. Al actuar dentro de las "zonas obscuras", algunos Estados hacen más difícil que las victimas puedan invocar el Derecho Internacional y responder con "contramedidas", por lo que muchas de las acciones cometidas se mantendrán en una ambigüedad jurídica. Una situación que R. Hughes califica como la debilidad del sistema de gobernanza y regulación actual y que para M. N. Schmitt constituye una especie de "gobernanza jurídica asimétrica" (*assymetrical lawfare*). En dicho contexto, los Estados comprometidos con el Estado de derecho suelen vacilar en sus reacciones ante ciberataques cuyo origen no se puede corroborar explícitamente, propiciando la impresión de debilidad ante el agresor y por tanto estableciendo una ventaja de dicho Estado sobre el Estado víctima (Hughes, 2010: 524; Schmitt, 2017: 1-3).

Para paliar dicha situación, según R. Nguyen y T. Remus, seria imperativo clarificar cuando los principios de *ius ad bellum* serían aplicables al ciberespacio. Para ello hacen referencia a tres posibles enfoques utilizados en la actualidad (Nguyen, 2013: 1083-1084; Remus, 2013: 181-183):

- *Instrumental*: Dependiendo del tipo de arma que lleva a cabo el ataque y si el ataque posee las características físicas tradicionalmente asociadas con la coerción militar.
- *Responsabilidad objetiva*: Cualquier ciberataque a una infraestructura crítica de un Estado sería considerado como un ataque armado.
- *Efectos producidos*: Dependencia de los efectos sobre el Estado víctima, tales como la severidad, la inmediatez o la precisión del ataque.

Ahora bien, el propio R. Nguyen aclara que, aunque su propuesta teórica sería una amalgama de los tres, por lo que un ciberataque constituiría un ataque armado si causase serias interrupciones o daño físico sobre los sistemas ciber físicos, tales como la red eléctrica u otras infraestructuras críticas, también dependería del análisis y la ponderación de otras consideraciones políticas y/o económicas que futuras contramedidas pudiesen desencadenar, como por ejemplo una escalada de confrontación entre Estados (Nguyen, 2013: 1083-84).

En todo caso, antes de poder aplicar el Derecho Internacional del *ius ad bellum,* habría que tener en cuenta, como indica M. Roscini, que el anonimato representa una de las mayores ventajas y aunque un ataque pueda parecer que su origen está localizado en un determinado Estado o unos determinados ordenadores dentro del mismo, eso no significa necesariamente que dicho Estado o los propios propietarios de los ordenadores hubiesen estado detrás de los ciberataques. Por lo tanto, sería imperativo soslayar la problemática de las "guerras por delegación" (*proxy wars*), que llena de dificultades la posibilidad de establecer fehacientemente la atribución de un ciberataque. Como indica M. Hakimi, las decisiones unilaterales para el uso de la fuerza no ocurren en un vacío y son parte de un proceso desestructurado donde los Estados individualmente actúan o reaccionan en base a incidentes concretos. Así, al ser un proceso de actuación de los Estados al margen del Consejo de Seguridad de las Naciones Unidas o de la Corte Internacional de Justicia (CIJ), por norma general desembocarían, de forma reiterada, en operaciones que serían incompatibles, de forma estricta, con el Derecho Internacional vigente, pero que al mismo tiempo podrían ser aceptables desde un punto de vista operativo. En todo

caso, aún en el caso que los autores fuesen identificados quedaría la problemática de que el ataque pudiese ser atribuido a un Estado bajo el marco jurídico sobre "Responsabilidad del Estado por Hechos Internacionalmente Ilícitos"[39] y, por tanto, poder así aplicar la reglamentación de *ius ad bellum* (Hakimi, 2018: 158; Marín Martínez, 2019: 31; Roscini, 2010: 96-102).

Es importante, por tanto, analizar una serie de elementos que forman parte del marco jurídico del *ius ad bellum.* La responsabilidad del Estado vendría avalada: en el caso de que el ciberataque hubiese sido perpetrado por cualquier estamento gubernamental (militar o civil), empresas privadas o contratistas empoderados por el Estado para ejercer cualquier grado de autoridad gubernamental, de acuerdo con el Artículo 4º de dicha ley[40]. En el caso de que los "*hackers*" fuesen individuos o empresas contratadas por un Estado para llevar a cabo ciberataques podría también serles de aplicación el Artículo 8º de dicha ley[41]. En dicho caso lo que se debería probar

39 Adoptado por la Comisión de Derecho Internacional en 2001 y posteriormente avalado por la Asamblea General de las Naciones Unidas (NU, 2001).

40 **Artículo 4:** Comportamiento de los órganos del Estado
1. Se considerará hecho del Estado según el derecho internacional el comportamiento de todo órgano del Estado, ya sea que ejerza funciones legislativas, ejecutivas, judiciales o de otra índole, cualquiera que sea su posición en la organización del Estado y tanto si pertenece al gobierno central como a una división territorial del Estado.
2. Se entenderá que órgano incluye toda persona o entidad que tenga esa condición según el derecho interno del Estado (NU, 2001).

41 **Artículo 8:** Comportamiento bajo la dirección o control del Estado
Se considerará hecho del Estado según el derecho internacional el comportamiento de una persona o de un grupo de personas si esa persona o ese grupo de personas actúa de hecho por instrucciones o bajo la direc-

es si el Estado tiene un control efectivo de la operación militar o paramilitar durante la cual las violaciones fueron cometidas. Un tercer escenario sucedería cuando grupos no adscritos fuesen los que realizasen los ciberataques, en tal caso también sería la responsabilidad del Estado si posteriormente dicho Estado reconociese y adoptase el acto como suyo, de acuerdo con el Artículo 11º de dicha ley[42] (Roscini, 2010: 96-102).

Otro elemento a tener en cuenta, cuando se habla del uso de la "fuerza armada", relativo a lo establecido en los Artículos 2(4)[43] y 51 de la Carta de las Naciones Unidas y su aplicación a los ciberataques, no tendría ninguna diferencia a un ataque cinético (convencional) en tanto se aplicase a una forma de "guerra" usada para la destrucción de la vida y de la propiedad, donde la nueva tecnología estuviese asociada a las fuerzas armadas del Estado que la utilizase[44]. Dicha interpretación también estaría avalada por el Artículo

---

ción o el control de ese Estado al observar ese comportamiento (NU, 2001).

42 **Artículo 11:** Comportamiento que el Estado reconoce y adopta como propio. El comportamiento que no sea atribuible al Estado en virtud de los artículos precedentes se considerará, no obstante, hecho de ese Estado según el derecho internacional en el caso y en la medida en que el Estado reconozca y adopte ese comportamiento como propio (NU, 2001).

43 **Artículo 2(4)**: Los Miembros de la Organización, en sus relaciones internacionales, se abstendrán de recurrir a la amenaza o al uso de la fuerza contra la integridad territorial o la independencia política de cualquier Estado, o en cualquier otra forma incompatible con los Propósitos de las Naciones Unidas (NU, 2001).

44 Aunque la opinión predominante mantiene que cualquier contramedida no debe incluir el uso de la fuerza, con relación a los ciberataques, dicha postura fue puesta en cuestión por el Juez Simma de la CIJ en su opinión razonada sobre el

31, párrafo 3(b) de la "Convención de Viena sobre el Derecho de los Tratados", donde un tratado debería ser interpretado teniendo también en cuenta la práctica de su aplicación por parte de los Estados, dado que una serie de ellos han expresado el punto de vista de que la aplicación de una "fuerza cibernética" es considerada una tipología de "fuerza armada" (Kilovaty, 2014: 100-101; Marín Martínez, 2019: 32-33; Roscini, 2010: 106-108).

Adicionalmente, se debe considerar la aplicación del "principio de no intervención" con relación a los ciberataques, de acuerdo con la "Declaración sobre la inadmisibilidad de la intervención y la injerencia en los asuntos internos de los Estados"[45]. En dicho contexto se podría incluir la desfiguración de las páginas web para fomentar la insurrección civil o para influenciar los resultados de las elecciones de otros Estados. Así, el Estado víctima podría adoptar contramedidas que no conllevasen el uso de la fuerza, dentro de la denominada "defensa activa", siempre como respuesta a una vio-

caso del pleito de las "Plataformas Petrolíferas" entre Irán y USA. La argumentación es que, si un Estado fuese víctima de operaciones cibernéticas desde una fuente desconocida, que amenazase "gravemente y con riesgo inminente" sus "intereses esenciales", podría llevar a cabo medidas de protección basadas en un "estado de necesidad". Dicho derecho existiría aun cuando las acciones llevadas a cabo afectasen intereses no esenciales de otros Estados, como el cerrar redes de las que dependiesen dichos Estados o atacando la ciber infraestructura responsable de la operación (Schmitt, 2013, 177, CIJ, 2003).

45 **Ic** – El derecho de los Estados y de los pueblos a tener libre acceso a la información y a desarrollar plenamente sin injerencias su sistema de información y de medios de comunicación y a utilizar sus medios de información para promover sus intereses y aspiraciones políticos, sociales, económicos y culturales, sobre la bases, entre otras cosas, de los artículos pertinentes de la Declaración Universal de Derechos Humanos y de los principios del nuevo orden internacional de la información (NU, 1981).

lación del Derecho Internacional por parte del otro Estado. Así, los ciberataques y la propaganda cibernética cuyo propósito fuese el causar disensiones internas en el Estado receptor, permitiría a dicho Estado adoptar contramedidas proporcionales de acuerdo con las limitaciones de los Artículos 50, 51 y 52 de la Carta de las Naciones Unidas. El uso de la fuerza, por lo tanto, estaría restringido a los casos en el que el ciberataque desencadenase el derecho a la "legítima defensa" de acuerdo con el Artículo 51 de la Carta de las Naciones Unidas. El mayor problema resultaría en determinar si un ciberataque sobre una red informática de una infraestructura de información crítica civil podría ser considerado como un "ataque armado", dado que no existe consenso sobre qué se considera una "infraestructura crítica" y la Asamblea General de las Naciones Unidas reconoció que "cada Estado determinará sus propias infraestructuras de información críticas". Además, existiría una complicación adicional dado que en la mayor parte de los países dichas infraestructuras son propiedad del sector privado y que dicha noción estaría íntimamente ligada a la de "seguridad nacional", que también es difícil de definir bajo el Derecho Internacional (Marín Martínez, 2019: 33; Roscini, 2010: 113, 117-118).

Más difícil aún sería la aplicación de las leyes internacionales de "*ius ad bellum*" en el caso de una acción "preventiva" dentro del marco de la ciberdefensa. En primer lugar, la amenaza de operaciones cibernéticas destructivas defensivas contra la infraestructura militar de otro Estado, si dicho Estado establece operaciones transfronterizas ilegales, no estaría en contra de lo dispuesto en el Artículo 2(4) de la Carta de las Naciones Unidas. No obstante, tomando como referencia el caso Stuxnet, algunos analistas consideran que Irán podría haber invocado el derecho a la "legítima defensa", de acuerdo con el Artículo 51 de la Carta de las Naciones

Unidas, mientras que otros consideran que el ataque preventivo sobre Natanz estaría justificado como una forma de "legítima defensa preventiva" contra una amenaza inminente. En todo caso, el aspecto cuantitativo de proporcionalidad necesitaría que la escala y el efecto de las contramedidas de fuerza fuesen de nivel similar a las del ataque armado perpetrado[46] (Kilovaty, 2014: 92; Marín Martínez, 2019: 34).

Profundizando en dicho tema, no es lo mismo el estar hablando de un ataque "anticipatorio" que de un ataque "preventivo". Según el Derecho Internacional, el derecho a la "legítima defensa", de acuerdo con el Artículo 51 de la Carta de las Naciones Unidas, solo sería de aplicación si un "ataque armado" ya se hubiese producido, pues dicho Artículo establece claramente la premisa: "en caso de ataque armado". Por lo tanto, un ataque "preventivo" sería considerado contrario a dicho Derecho, lo que implicaría que si una operación cibernética no hubiese traspasado el umbral del "ataque armado" no existiría el derecho a la "legítima defensa". El análisis de T. Remus, en todo caso, alerta de la implicación de dicho resultado cuando se producen eventos de ciber espionaje, donde la información obtenida, si no fuese detectada, podría convertirse en la base para una acción futura. En tal caso, sugiere que la única alternativa para el Estado espiado sería el desarrollo de contramedidas de ciber espionaje similares, pero nunca un ataque informático a gran escala, ya que la mayoría de los Estados no desearían una

46 En el caso de Nicaragua contra USA, la Corte de Justicia Internacional estableció que, aunque los principios de necesidad y proporcionalidad no estaban inscritos en el Artículo 51 de la Carta de las Naciones Unidas, no obstante "la legítima defensa solo podría considerar medidas que sean proporcionales al ataque armado y necesarios para responder, una regla ya bien establecida en el Derecho Internacional consuetudinario" (Kilovaty, 2014: 104).

intensificación de los ataques contra las redes informáticas y traspasar el umbral de un "ataque armado", arriesgando con ello que se desencadenase un acto de "legítima defensa" y eventualmente una guerra total. Por lo tanto, aparte de la legislación internacional existente, no se debería tampoco obviar que también existiría una "legislación" informal consuetudinaria, dentro del marco del "*ius ad bellum*" conformada por aquellas prácticas, fuera del marco institucional, cuyo objetivo sería mantener y/o restaurar la paz y la seguridad internacional[47]. Por otro lado, dentro del Derecho consuetudinario, la CIJ confirmó, en su opinión consultiva sobre "Armamento Nuclear", como ya analizamos anteriormente en esta investigación, y con relación a las provisiones del uso de la fuerza, que dicho marco jurídico sería de aplicación "independientemente del tipo de armamento utilizado" (Hakimi, 2018: 164-165; Marín Martínez, 2019: 34-5; Remus, 2013: 189; Schmitt, 2013: 176).

En cuanto al concepto de "proporcionalidad" en las leyes de "*ius ad bellum*" de los SAA, algunos analistas argumentan que los SAA salvan la vida de soldados. Por lo tanto, aumentaría la aceptación del uso de armamento robótico ya que los daños serían los del agresor y no el de los combatientes y/o civiles del Estado que se defiende. Otros analistas argumentan, sin embargo, que los SAA serían incapaces de luchar "proporcionalmente", ya que el juicio necesario para dichos cálculos excede las capacidades de aprendizaje de dichos sistemas. H. M. Roff añade, además, la necesidad de mirar más allá de un evento específico hacia las consecuencias futuras de una posible guerra y de que dichos SAA serían incapaces

47 Prácticas que serían evaluadas según sus propios méritos basadas en el contexto en que se produjeron. Idea plasmada por W. M. Reisman en el desarrollo de los "Criterios sobre el uso legítimo de la fuerza en el Derecho Internacional" (Reisman, 1985: 281-282).

de discernir. Se debería que tener en cuenta que una respuesta defensiva con un SAA incidiría en los aspectos legales sobre "el daño inminente". Es más, la presunción de que un SAA fuese una ventaja de los Estados que pueden "luchar batallas sin ningún contacto", manipularía artificialmente el sentido de la "proporcionalidad" dentro del *ius ad bellum*, ya que asumiría que el Estado que lo utiliza no afrontaría amenazas futuras adicionales de su adversario. Adicionalmente, el uso de los SAA en un conflicto podría traer otros costos negativos, como la inducción de una carrera de armamentos. Por lo tanto, dentro del ámbito del *ius ad bellum*, se deberían ponderar todos los daños potenciales y no únicamente los relativos a los de la "guerra justa"[48] (Marín Martínez, 2019: 47; Roff, 2015a: 37-40, 44, 47, 50).

En todo caso, se debe tener en cuenta que el término *ius ad bellum* no existió dentro del marco jurídico internacional hasta mediados del siglo XX, cuando se desarrolló el régimen legal de la prohibición del uso de la fuerza. Fue entonces cuando fue necesario distinguir entre los términos *ius ad bellum* e *ius in bello,* para establecer una clara diferenciación entre la paz y la guerra. Así, aunque el término de *ius in bello* o DIH, que regula la conducta de las partes que participan en un conflicto armado, fue atribuido a mediados de los años 30, por R. Kolb, a J. Kunz en 1934[49], éste no tuvo un gran

48 Para R. Berkebile una "guerra justa" sería una guerra defensiva en respuesta a una agresión (Berkebile, 2018: 19).

49 Para más información ver: KUNZ, J. L. (1951) [2012]: "Bellum Justum and Bellum Legale", en T. Gazzini y N. Tsagourias (eds.), *The Use of Force in International Law*, Routledge, Londres, III(7), acceso en enero 2021, en https://www.taylorfrancis.com/chapters/bellum-justum-bellum-legale-josef-kunz/e/10.4324/9781315084992-7

impacto hasta después de la II Guerra Mundial, cuando se hizo esencial para el desarrollo del Derecho Internacional moderno. Habría también que tener en cuenta que la noción de paz no significa estrictamente la ausencia de violencia, que queda reflejado en el concepto colectivo de "responsabilidad de proteger", utilizado en operaciones de paz multidimensionales y en las intervenciones modernas. Dicho concepto se fundamenta en tres dimensiones de la gestión de conflictos relacionados con casos de atrocidades a gran escala: acciones preventivas; acciones de respuesta y; participación post conflicto ('responsabilidad para prevenir'; 'responsabilidad para reaccionar'; 'responsabilidad para reconstruir') (Giladi, 2008: 249-250; Kolb, 1997; NU, 2005: 33; Stahn, 2006: 922).

Al hablar del DIH habría que referirse a las Convenciones de Ginebra de 1949 y sus Protocolos Adicionales de 1977 (PAI, PAII y PAIII), tratados internacionales que contienen las reglas más importantes en tiempos de guerra y que protegen a las personas que no participan (civiles, personal médico, personal humanitario) y a las tropas que ya no pueden luchar (heridos y enfermos, náufragos y prisioneros de guerra)[50], así como el Derecho Internacional consuetudinario. Es necesario también puntualizar que, aunque explícitamente el DIH no se menciona en relación con el ciberespacio[51],

50 Para más información ver: COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2019): *The Geneva Conventions of 1949 and their Additional Protocols*, Comité Internacional de la Cruz Roja, Ginebra, acceso mayo 2019, en https://www.icrc.org/en/doc/war-and-law/treaties-customary-law/geneva-conventions/overview-geneva-conventions.htm

51 Reticencias provenientes de importantes Estados, como en el caso de Rusia y China, no han permitido hasta la fecha incluir el DIH explícitamente como Derecho Internacional en el ciberespacio. Para más información ver: SEGAL, A. (2012): "China, International Law and Cyber Space", *The Council of*

su núcleo principal de principios (humanidad, necesidad, proporcionalidad y distinción) si fueron explícitamente referidos por el GGE sobre "Avances en la Información y las Telecomunicaciones en el Contexto de la Seguridad Internacional" en 2015. Ahora bien, al igual que A. A. Haque, se deben plantear dos cuestiones principales cuando se habla de conflictos armados: cuándo comienzan y cuándo es posible aplicar el Derecho de los conflictos armados (DICA), dado que en general existe una distinción entre los conflictos armados internacionales entre Estados (CAI) y los conflictos armados no internacionales (CANI) entre Estados y grupos armados organizados o entre grupos armados organizados, ya que varía la legislación internacional aplicable en cada caso (Brown, 2017: 355; Haque, 2017: 476; Marín Martínez, 2019: 35-36, NU, 2015: 16).

Por su relevancia y antes de entrar en profundidad, es importante comprender el concepto de "conflicto armado" dentro del ámbito del Derecho Internacional. Un CAI, según A. A. Haque, se define como una disputa (conflicto) entre Estados (internacional) que conlleva el uso de la fuerza militar (armado). Reflejaría una posición prevalente en el ámbito internacional sobre dicho concepto, como la del del CICR, especialmente con relación al Artículo 2(1) de la Convención de Ginebra[52], donde la existencia de un conflicto

Foreign Relations (Oct. 2, 2012), acceso mayo 2019, en http://blogs.cfr.org/asia/2012/10/02/china-international-law-and-cyberspace

52 **Artículo 2(1)**: Aparte de las disposiciones que deben entrar en vigor ya en tiempo de paz, el presente Convenio se aplicará en caso de guerra declarada o de cualquier otro conflicto armado que surja entre dos o varias Altas Partes Contratantes, aunque una de ellas no haya reconocido el estado de guerra (CICR, 2012: 37).

armado debería estar basado únicamente en la circunstancias que demostrasen *de facto* la existencia de hostilidades entre los beligerantes aún sin una declaración de guerra formal[53]. En cuanto a los CANI, la visión predominante, que incluye al CICR, estipula que solo sería aplicable en confrontaciones armadas entre las fuerzas armadas de un Estado y grupos armados organizados o entre dichos grupos. Por lo tanto, no sería de aplicación si dichos grupos no estuviesen organizados o si la lucha no fuese suficientemente intensa, siguiendo lo establecido en el Artículo 1(2) del PAII[54]. Posturas que de forma similar se recogen en el documento "Orientaciones. El Derecho de los Conflictos Armados" del Ministerio de Defensa español (CICR, 2016: 211; 2021; Haque, 2017: 476; M. Defensa, 2007: 1-9 – 1-12).

En el DIH, el término "ataque" se refiere a una categoría particular de operaciones militares. El Artículo 49(1) del PAI[55] lo especifica como los "actos de violencia contra el adversario[56], ya sean ofensi-

53 Es importante añadir, que la jurisprudencia internacional ha establecido que el DIH es aplicable desde el inicio del conflicto y se extiende más allá del cese de las hostilidades hasta que se alcance un tratado de paz en el caso de los CAI o un acuerdo de paz en el caso de los CANI, como en el caso del Tribunal Internacional para la ex Yugoslavia (Szpak, 2017: 265).

54 **Artículo 2(1) Protocolo Adicional II**: El presente Protocolo no se aplicará a las situaciones de tensiones internas y de disturbios interiores, tales como los motines, los actos esporádicos y aislados de violencia y otros actos análogos, que no son conflictos armados (CICR, 1977b).

55 **Artículo 49(1) Protocolo Adicional I**: Se entiende por ataques los actos de violencia contra el adversario, sean ofensivos o defensivos (CICR, 1977a).

56 Según M. N. Schmitt, el término "adversario" no solo sería de aplicación a aquellas operaciones violentas contra las fuerzas enemigas, el criterio principal

vos o defensivos". Es un umbral clave en las DIH ya que muchas de sus prohibiciones y restricciones básicas solo son aplicables a actos así definidos, dentro del término legal de "conflicto armado", que solo se aplica a los CAIs y CANIs, como anteriormente hemos indicado. El problema estriba en que todos los "ataques armados" son usos de la fuerza, pero no todos los usos de la fuerza son considerados como "ataques armados". Por lo tanto, los Estados pueden ser objeto de operaciones cibernéticas constitutivas del uso de la fuerza, pero no ser lo suficientemente severas para calificarlas como un "ataque armado". Según M. N. Schmitt, las operaciones cibernéticas necesitarían de una reconceptualización de la noción de "ataque armado", aunque en la actualidad no existe un consenso de la comunidad internacional (Marín Martínez, 2019: 36; Schmitt, 2012b: 285-287).

La problemática principal surge dado que las operaciones cibernéticas no implican directamente el lanzamiento de fuerzas violentas. M. N. Schmitt indica que los tratados deberían ser interpretados en "el contexto y a la luz del objetivo y el propósito". Al leer de forma detenida el PAI, sobre prohibiciones y restricciones, su preocupación principal no sería el establecer que actos eran violentos, sino cuales de ellos tenían consecuencias dañinas (o su riesgo), con lo que se estaría hablando de las "consecuencias violentas" y la pretensión del tratado de evitar dichas consecuencias. Por lo tanto, dado que el Artículo 51(1) del PAI[57] establece que los civiles

debe ser el de "violencia" y sería de aplicación también a los ataques contra civiles (Schmitt, 2012b: 290).

57 **Artículo 51(1) Protocolo Adicional I**: La población civil y las personas civiles gozarán de protección general contra los peligros procedentes de operaciones militares. Para hacer efectiva esta protección, además de las otras normas

"disfrutan de la protección general contra los peligros de las operaciones militares", en el marco de las operaciones cibernéticas sería necesario el seleccionar el armamento y las tácticas para evitar y en todo caso minimizar la pérdida incidental de vidas civiles, daños a dichos civiles y a sus propiedades. Se estaría, por ejemplo, frente a operaciones cibernéticas que causasen un sufrimiento serio innecesario a la población civil o a la destrucción de objetos que los hace inoperativos, como a un sistema informático necesario para el suministro de agua o de electricidad a la población. Así, durante la "37ª Conferencia Internacional de la 'Cruz Roja' y de la 'Media Luna Roja'", de 2011, el CICR distribuyó un documento en el que indicaba que el Artículo 49 del PAI hacía referencia a que, aquellas operaciones cibernéticas (virus, gusanos, etc.) que tuviesen como consecuencia daños físicos contra los civiles o las propiedades, más allá del programa informático o de los datos atacados, podría ser calificado como un acto de violencia, dentro del marco de las DIH (CICR, 2011b: 37; Marín Martínez, 2019: 37; Schmitt, 2012b: 290-292).

En cuanto a la problemática de la "distinción", el Artículo 48 del PAI[58] establece que las diversas partes de un conflicto "distinga en todo momento entre la población civil y los combatientes y entre los objetivos militares y los civiles". Un principio que mantiene

---

aplicables de derecho internacional, se observarán en todas las circunstancias las normas siguientes (CICR, 1977a).

58 **Artículo 48 Protocolo Adicional I**: A fin de garantizar el respeto y la protección de la población civil y de los bienes de carácter civil, las Partes en conflicto harán distinción en todo momento entre población civil y combatientes, y entre bienes de carácter civil y objetivos militares y, en consecuencia, dirigirán sus operaciones únicamente contra objetivos militares (CICR, 1977a).

la ley consuetudinaria del principio de la "distinción", que ha sido establecido por la CIJ como uno de los dos principios cardinales de la DIH. El principio, por tanto, se aplicaría a las operaciones cibernéticas durante un conflicto armado. Para H. H. Dinniss, una forma de aplicar el principio de distinción en el ciberespacio sería el requerir que todo ataque, a través de las redes informáticas, emanase de una dirección IP militar designada. Sería una forma de marcador electrónico y resolvería el problema de la obligación de que un individuo utilizase un uniforme al perpetrar un ataque (Dinniss, 2013: 255).

En todo caso, para clarificar dicho principio, el Artículo 51(1), aparte de la protección generalizada de la población civil sobre los peligros de las operaciones militares, como ya se ha indicado anteriormente, prohíbe que la población civil sea el objetivo de un ataque, que se lleven a cabo ataques indiscriminados o ataques como represalia. De forma similar el Artículo 52 [59]prohíbe que las infraestructuras civiles sean un objetivo del ataque y el Artículo

59 **Artículo 52 Protocolo Adicional I**: (1). Los bienes de carácter civil no serán objeto de ataques ni de represalias. Son bienes de carácter civil todos los bienes que no son objetivos militares en el sentido del párrafo 2. (2). Los ataques se limitarán estrictamente a los objetivos militares. En lo que respecta a los bienes, los objetivos militares se limitan a aquellos objetos que por su naturaleza, ubicación, finalidad o utilización contribuyan eficazmente a la acción militar o cuya destrucción total o parcial, captura o neutralización ofrezca en las circunstancias del caso una ventaja militar definida. (3). En caso de duda acerca de si un bien que normalmente se dedica a fines civiles, tal como un lugar de culto, una casa u otra vivienda o una escuela, se utiliza para contribuir eficazmente a la acción militar, se presumirá que no se utiliza con tal fin (CICR, 1977a).

54[60] prohíbe los ataques a infraestructuras indispensables para la supervivencia de la población civil. En dicho contexto, en el marco de las operaciones cibernéticas, actualmente no queda claro si los datos en un ordenador pueden ser considerados como un "objeto". En caso de que así lo fuesen, el ataque contra datos civiles sería considerado ilegal y cualquier daño causado durante un ciberataque sobre un objetivo militar legal, tendría que ser considerado dentro del principio de proporcionalidad y cuando se determinasen las precauciones a tomar durante un ataque (CICR, 2016d).

Además, para M. N. Schmitt, aunque el término "ataque" (acto violento) implicaría que aquellos ciberataques no violentos esta-

60 **Artículo 54 Protocolo Adicional I**: (1). Queda prohibido, como método de guerra, hacer padecer hambre a las personas civiles. (2). Se prohíbe atacar, destruir, sustraer o inutilizar los bienes indispensables para la supervivencia de la población civil, tales como los artículos alimenticios y las zonas agrícolas que los producen, las cosechas, el ganado, las instalaciones y reservas de agua potable y las obras de riego, con la intención deliberada de privar de esos bienes, por su valor como medios para asegurar la subsistencia, a la población civil o a la Parte adversa, sea cual fuere el motivo, ya sea para hacer padecer hambre a las personas civiles, para provocar su desplazamiento, o con cualquier otro propósito. (3). Las prohibiciones establecidas en el párrafo 2 no se aplicarán a los bienes en él mencionados cuando una Parte adversa: a) utilice tales bienes exclusivamente como medio de subsistencia para los miembros de sus fuerzas armadas; o b) los utilice en apoyo directo de una acción militar, a condición, no obstante, de que en ningún caso se tomen contra tales bienes medidas cuyo resultado previsible sea dejar tan desprovista de víveres o de agua a la población civil que ésta se vea reducida a padecer hambre u obligada a desplazarse. (4). Estos bienes no serán objeto de represalias. (5). Habida cuenta de las exigencias vitales que para toda Parte en conflicto supone la defensa de su territorio nacional contra la invasión, una Parte en conflicto podrá dejar de observar las prohibiciones señaladas en el párrafo 2 dentro de ese territorio que se encuentre bajo su control cuando lo exija una necesidad militar imperiosa (CICR, 1977a).

rían excluidos dentro de la aplicación del PAI desde una interpretación estricta, recuerda lo establecido en el Artículo 31(1)[61] de la Convención de Viena del Derecho de los Tratados de 1969, que establece que los tratados deberán interpretarse "en el contexto de estos y teniendo en cuenta su objetivo y fin". Así, no sería la violencia del acto la que constituiría la condición previa para limitar la consecución de un "ataque", sino la violencia del resultado subsiguiente. A tal fin, como indican T. McCormack y N. Lubell, sería mejor hablar de "operaciones" cibernéticas y no de ciberataques al referirse a la aplicación del DIH en el ciberespacio, dado que dicho término sería más neutral jurídicamente (Marín Martínez, 2019: 37-38; McCormack, 2018: 224-225; Schmitt, 2012c: 91, 93, 94, 96).

Siguiendo con el punto aún no resuelto, de extrema importancia para la aplicación del DIH en el ciberespacio, sobre la catalogación de los datos informáticos como "objeto" y, por lo tanto, que fuesen considerados como objetivo militar, existen diferentes puntos de vista en la actualidad. El Manual de Tallinn, en su Regla 38ª, establece que se considerarán como "objetos" los ordenadores, las redes de ordenadores y la infraestructura cibernética. Mantendría por tanto la idea establecida por el CICR de que un objeto debe ser "visible y tangible" y los datos no estarían incluidos como "objetos", según la postura de la mayor parte de los expertos que desarrollaron dicho manual y que establecieron que "los datos son intangibles y por lo tanto no encajan dentro del 'significado corriente' del término 'objeto' ni concuerda con la aclaración indicada en

61 **Artículo 31(1) Convención de Viena**: Un tratado deberá interpretarse de buena fe conforme al sentido corriente que haya de atribuirse a los términos del tratado en el contexto de estos y teniendo en cuenta su objeto y fin (BOE, 1980).

el Comentario sobre los Protocolos Adicionales del CICR (paras. 2007-2008) de 2016". La consecuencia de dicha postura, según K. Macák, sería que una operación cibernética cuyo objetivo fueran datos, no entraría dentro del ámbito del DIH, a no ser que afectase la funcionalidad de un sistema de control y como resultado fuese necesario reemplazar su componentes físicos (CICR, 2016d; 633-634; Macák, 2015: 58-59; McCormack, 2018: 225-226; OTAN, 2013:125-127).

Como contraste, una minoría de los expertos indicaron que dicha postura sería contraria al principio de que la población civil goza de la protección general de los efectos de las hostilidades. Para dichos expertos el factor clave, de acuerdo con lo establecido en el Artículo 52 del PAI, sería la gravedad y las consecuencias de la operación y no la naturaleza del daño. Por lo tanto, aquellos datos civiles esenciales para el bienestar de la población civil estarían dentro de la noción de "objetos" civiles y en consecuencia protegidos. Dicha opinión, con la que nosotros también estamos de acuerdo, que considera que los datos podrían ser catalogados como "objetos virtuales", también ha sido indicada por el investigador K. Macák. En el caso del prominente investigador M. N. Schmitt, aunque su postura establece que los datos no se deberían caracterizar como objetos, si plantea que en el caso que su destrucción conllevase como consecuencia suficientes daños contra los objetos físicos o contra las personas, entonces la operación cibernética debería ser considerada como un ataque ilegal (Macák, 2015: 59; OTAN, 2013: 127; Schmitt, 2012c: 95-96).

Particularmente, se debe poner especial atención en aquellos sistemas cibernéticos de doble uso (militar y civil). Algunos caracterizados como civiles podrían ser esenciales para la economía y

por lo tanto constituir objetos para el mantenimiento de la guerra, siendo considerados, por tanto, objetivos militares. Sería muy probable que cualquier atacante marcase como objetivo aquellos poco protegidos, como los sistemas informáticos sanitarios o aquellos que sirviesen para el "mantenimiento de la guerra" como los sistemas económicos, los relativos a las infraestructuras del petróleo, medios de transporte o comunicación. La práctica[62] vendría reflejada en la norma 8ª del Derecho Internacional consuetudinario[63], aplicable tanto en los conflictos CAI como en los CANI. La clasificación de dichos bienes dependería, en un último análisis, de la aplicación de la definición de objetivo militar y siempre y cuando su ataque ofreciese una ventaja militar definida. En todo caso dichos ataques seguirían estando sujetos a los principios de proporcionalidad y cuando se determinasen las precauciones a tomar durante un ataque (CICR, 2007: 34-36; Marín Martínez, 2019: 38; Schmitt, 2012c: 95-96).

Otro punto a tener en cuenta, en el mundo cibernético, sería el encaje dentro del DIH de la participación directa de civiles, los denominados "*hackers*" o los contratistas informáticos, en un conflicto

62 Los manuales militares de Alemania, Argentina, Australia, Bélgica, Benín, Camerún, Canadá, Colombia, Croacia, España, Estados Unidos, Francia, Hungría, Italia, Kenia, Madagascar, Nueva Zelanda, Países Bajos, Sudáfrica, Suecia, Togo y Reino Unido, Ecuador, Indonesia y Estados Unidos así lo contienen (CICR, 2007: 34).

63 **Norma 8**: Por lo que respecta a los bienes, los objetivos militares se limitan a aquellos bienes que por su naturaleza, ubicación, finalidad o utilización contribuyan eficazmente a la acción militar y cuya destrucción total o parcial, captura o neutralización ofrezca, en las circunstancias del caso, una ventaja militar definida (CICR, 2007: 34),

armado. El Artículo 51(3) del PAI[64] deja claro que aquellos civiles que participasen directamente en las hostilidades no estarían cubiertos. En la "Guía Interpretativa sobre la Noción de la Participación Directa en las Hostilidades", el CICR incluye aquellos grupos armados organizados, que formasen parte de uno de los actores del conflicto, dentro de la categoría de fuerzas armadas. Ahora bien, en el ámbito cibernético no queda claro cuando los "*hackers*" y los grupos no militares pudiesen ser clasificados como grupos armados organizados. El punto más controvertido tendría que ver con el concepto de organización y si un grupo puede ser organizado "virtualmente", como por ejemplo las empresas de ciberseguridad contratadas por un Estado. Para el CICR, en su Comentario sobre la PAI, un grupo "organizado" debería tener un carácter colectivo, bajo un control y ciertas reglas. Por lo tanto, un "*hacker*" individual seguiría siendo un civil, mientras que un grupo que actuase colectivamente formaría parte de las fuerzas armadas y por tanto un miembro individual de dicho grupo podría ser un objetivo militar. En todo caso y ante cualquier duda el DIH presupone un estatus civil. No obstante, como indica H. H. Dinniss, habría que tener en cuenta que las operaciones cibernéticas, por su naturaleza son un método encubierto de guerra, por lo que movimientos preparatorios de ataques a las redes informáticas podrían originarse desde ordenadores civiles (virus, reconocimiento, botnets, dns, etc.). Además, la mayor parte de dichos ataques son encaminados a través de varios servidores intermedios en distintos lugares, por lo que el rastreo de su origen a veces resulta imposible (CICR, 1977a; CICR, 2009: 22, 27-28, 36; Dinniss, 2013: 257-258).

64 **Artículo 51(3) Protocolo Adicional I**: Las personas civiles gozarán de la protección que confiere esta Sección, salvo si participan directamente en las hostilidades y mientras dure tal participación (CICR, 1977a).

Dicha idea también afecta al concepto establecido en el mismo artículo: "mientras dure la operación". En una operación cibernética puede ser que no exista un "despliegue", dado que solo se necesita un ordenador y no necesariamente cerca del objetivo, para montar una operación. Puede ser que, además, la operación cibernética se retrasase en el tiempo, instalando una bomba lógica durmiente. En tal caso, M. N. Schmitt propone que, desde una perspectiva práctica, dicho concepto englobaría todo el periodo en el que un participante cibernético civil mantuviese su participación en operaciones cibernéticas repetitivas dentro de un conflicto armado (CICR, 2009: 22, 27-28, 36; CICR, 2013: 3; Marín Martínez, 2019: 38; Schmitt, 2012c: 99-102).

Con relación a la aplicación del *ius in bello* a los "ciber soldados" o cualquier actor envuelto en una operación cibernética, dicha premisa se debe entender desde el punto de vista de que: cualquier conflicto armado internacional o no internacional, en la actualidad, generalmente incluirá operaciones cibernéticas, como un elemento integrado de la estrategia de guerra. Es más, cualquier operación cibernética individual podría tener tantos efectos cinéticos (convencionales) para ser considerado como un "ataque armado", justificando así el uso de la fuerza en "legítima defensa". Por lo tanto, la caracterización de los "ciber soldados" como combatientes, civiles o combatientes potencialmente ilegales traería consigo diferentes consecuencias legales, especialmente si se les considerase como objetivos. Así, por un lado, aquellos "ciber soldados" miembros de las fuerzas armadas serían considerados como combatientes y en un conflicto armado estarían protegidos por las leyes internacionales si fuesen capturados. Ahora bien, los Estados puede que empleasen, al menos, otras tres categorías de actores, no formalmente afiliados a las fuerzas armadas, en operaciones cibernéticas:

los diseñadores de las armas que fuesen empleadas en dichas operaciones; los actores que lanzasen operaciones cibernéticas "por delegación" de un Estado por razones de una negación de la "responsabilidad" o; cuando un Estado utilizase a miembros de la población civil para defender las infraestructuras civiles críticas. En este último caso, dado que dichos bienes podrían ser considerados objetivos, los operarios civiles de dichas infraestructuras estarían en primera línea en un conflicto armado (Marín Martínez, 2019: 44; Padmanabhan, 2013: 289-290, 293-294).

Los "ciber soldados" responsables del diseño y el lanzamiento de armas cibernéticas, así como los grupos cuasi independientes podrían ser considerados como combatientes legales bajo el Artículo 4(A)(2) de la III Convención de Ginebra[65]. Para que ello sucediese dichos grupos deberían formar parte de una de las Partes contendientes, lo que S. Watts denomina como "afiliación". Se estaría entonces, en la práctica ante lo dispuesto en la Norma 4ª del Derecho Internacional consuetudinario[66], donde las fuerzas armadas

65 **Artículo 4(2) III Convención de Ginebra:** Son prisioneros de guerra, por lo que se refiere al presente Convenio, las personas que, perteneciendo a alguna de las siguientes categorías, caigan en poder del enemigo: 2) miembros de otras milicias y miembros de otros cuerpos de voluntarios, incluso los de movimientos de resistencia organizados, pertenecientes a una Parte contendiente y que actúen fuera o dentro de su propio territorio, aunque este territorio se halle ocupado, siempre que esas milicias o cuerpos organizados, incluso los movimientos de resistencia organizados, llenen las condiciones siguientes: a) que figure a su cabeza una persona responsable por sus subordinados; b) que lleven un signo distintivo fijo y fácil de reconocer a distancia; c) que lleven francamente las armas; d) que se conformen, en sus operaciones, a las leyes y costumbres de la guerra (CICR, 1949).

66 **Norma 4**. Las fuerzas armadas de una parte en conflicto se componen de todas las fuerzas, agrupaciones y unidades armadas y organizadas que estén

cubrirían a todas las personas que combaten por una parte en conflicto y que se subordina a un mando. Por lo tanto, se consideraría combatiente a cualquier persona que emprendiese actos hostiles, a las órdenes de un mando responsable, en un conflicto armado en nombre de una Parte en dicho conflicto, siendo las condiciones impuestas al grupo como tal y serían susceptibles de ser atacados (CICR, 2007: 16; Watts, 2012: 246).

Por otro lado, el Artículo 4(A)(6)[67] otorga a los habitantes de un país no ocupado el estatus de prisionero de guerra si "espontáneamente", dentro el principio del "levantamiento de masas", tomasen las armas para defenderse contra las fuerzas invasoras. En tal caso, los civiles que administrasen infraestructuras críticas y que utilizasen "defensas activas" para responder a una operación cibernética, podrían tener derecho al estatus de combatiente. En ambos casos, una de las principales problemáticas estriba con el concepto de "distinción". La aplicación "literal" de la necesidad de llevar un "signo distintivo" o "lleven francamente las armas" puede que tuviese como resultado que dichos actores no fuesen considerados como combatientes legales, aunque el Artículo 43 del PAI[68] ha eli-

---

bajo un mando responsable de la conducta de sus subordinados ante esa parte (CICR, 2007: 16).

67 **Artículo 4(6) III Convención de Ginebra:** Son prisioneros de guerra, por lo que se refiere al presente Convenio, las personas que, perteneciendo a alguna de las siguientes categorías, caigan en poder del enemigo: 6) la población de un territorio no ocupado que, al acercarse el enemigo, tome espontáneamente las armas para combatir a las tropas invasoras, sin haber tenido tiempo para constituirse en fuerzas armadas regulares, siempre que lleve francamente las armas y respete las leyes y costumbres de la guerra (CICR, 1949).

68 **Artículo 43 Protocolo Adicional I**: 1. Las fuerzas armadas de una Parte

minado dicho requisito, que si se incluye en el Artículo 44(3)[69]. En todo caso, es muy probable que dichos actores no llevasen ningún tipo de distintivo o uniforme y puede que ocultasen la naturaleza militar de los ordenadores utilizando marcas de infraestructura informática civil, como puede ser una "dirección IP" de Internet civil. Por otro lado, los "hacktivistas" que no estuviesen afiliados

---

en conflicto se componen de todas las fuerzas, grupos y unidades armados y organizados, colocados bajo un mando responsable de la conducta de sus subordinados ante esa Parte, aun cuando ésta esté representada por un gobierno o por una autoridad no reconocidos por una Parte adversa. Tales fuerzas armadas deberán estar sometidas a un régimen de disciplina interna que haga cumplir, inter alia, las normas de derecho internacional aplicables en los conflictos armados. 2. Los miembros de las fuerzas armadas de una Parte en conflicto (salvo aquellos que formen parte del personal sanitario y religioso a que se refiere el artículo 33 del III Convenio) son combatientes, es decir, tienen derecho a participar directamente en las hostilidades. 3. Siempre que una Parte en conflicto incorpore a sus fuerzas armadas un organismo paramilitar o un servicio armado encargado de velar por el orden público, deberá notificarlo a las otras Partes en conflicto (CICR, 2016d).

69 **Artículo 44(3) Protocolo Adicional I**: 3. Con objeto de promover la protección de la población civil contra los efectos de las hostilidades, los combatientes están obligados a distinguirse de la población civil en el curso de un ataque o de una operación militar preparatoria de un ataque. Sin embargo, dado que en los conflictos armados hay situaciones en las que, debido a la índole de las hostilidades, un combatiente armado no puede distinguirse de la población civil, dicho combatiente conservará su estatuto de tal siempre que, en esas circunstancias, lleve sus armas abiertamente: a) durante todo enfrentamiento militar; y b) durante el tiempo en que sea visible para el enemigo mientras está tomando parte en un despliegue militar previo al lanzamiento de un ataque en el que va a participar. No se considerarán como actos pérfidos, en el sentido del apartado c) del párrafo 1 del artículo 37, los actos en que concurran las condiciones enunciadas en el presente párrafo (CICR, 2016d).

a una de las Partes de un conflicto armado y que llevasen a cabo operaciones cibernéticas por simpatía personal hacia uno de los beligerantes, no tendrían la cualificación de estatus de combatiente al no tener una relación con uno de los Estados partícipes en el conflicto armado (Ayalew, 2016: 216; CICR, 1949; CICR, 2016d; Marin Martínez, 2019: 45; Padmanabhan, 2013: 294-296).

Con relación a los ciber soldados que operan un "vehículo aéreo no tripulado" (VANT) (*UAV unmanned aerial vehicle*), el analista V. Padmanabhan indica que tendría el estatus de combatiente dentro de un conflicto armado. No obstante, el Artículo 57(2) del PAI[70],

70 **Artículo 57(2) Protocolo Adicional I**: Respecto a los ataques, se tomarán las siguientes precauciones:
a) quienes preparen o decidan un ataque deberán: i) hacer todo lo que sea factible para verificar que los objetivos que se proyecta atacar no son personas civiles ni bienes de carácter civil, ni gozan de protección especial, sino que se trata de objetivos militares en el sentido del párrafo 2 del artículo 52 y que las disposiciones del presente Protocolo no prohíben atacarlos; ii) tomar todas las precauciones factibles en la elección de los medios y métodos de ataque para evitar o, al menos, reducir todo lo posible el número de muertos y de heridos que pudieran causar incidentalmente entre la población civil, así como los daños a los b ienes de carácter civil; iii) abstenerse de decidir un ataque cuando sea de prever que causará incidentalmente muertos o heridos en la población civil, daños a bienes de carácter civil, o ambas cosas, que serían excesivos en relación con la ventaja militar concreta y directa prevista; b) un ataque será suspendido o anulado si se advierte que el objetivo no es militar o que goza de protección especial, o que es de prever que el ataque causará incidentalmente muertos o heridos entre la población civil, daños a bienes de carácter civil, o ambas cosas, que serían excesivos en relación con la ventaja militar concreta y directa prevista; c) se dará aviso con la debida antelación y por medios eficaces de cualquier ataque que pueda afectar a la población civil, salvo que las circunstancias lo impidan. 3. Cuando se pueda elegir entre varios objetivos militares para obtener una ventaja militar equivalente, se optará por el objetivo cuyo ataque, según sea de prever, presente menos peligro para las personas civiles y los bienes de

establece el mandato de que aquellos que toman las decisiones sobre objetivos hagan "todo lo posible" para verificar que el objetivo no es un civil que no participa directamente en las hostilidades. No obstante, el Derecho Internacional consuetudinario reconoce que factores como "las restricciones de tiempo, los riesgos, la tecnología y los costes de los recursos" pueden condicionar la obligación de obtener información en ayuda de la decisión de fijar un objetivo. Por lo tanto, el "hacer todo lo posible" para distinguir a los civiles requeriría del ejercicio de un "cuidado razonable" en la fijación de dichas decisiones (CICR, 2016d; Marín Martínez, 2019: 45-46; Padmanabhan, 2013: 304).

Los mismo principios se aplicarían en el caso de los robots, de lo que los VANT formarían un subconjunto, que no tuviesen una autonomía real. Habría que destacar, no obstante, el marco de aplicación del DIH a los SAA, dado que operan normalmente sin la intervención humana, aunque no tengan capacidades autónomas de decisión, siguiendo un algoritmo o programa específico para cumplir su misión. Ahora bien, a través de la IA, dichos sistemas autónomos podrían "aprender" o "adquirir" nuevos conocimientos adaptando nuevos métodos para cumplir las órdenes o ajustando sus conocimientos a entornos cambiantes. Así, dichas restricciones provendrían tanto del DICA como de las "Reglas de Enfrentamiento" (*Rules of Engagement (ROE)*). No se debe obviar que las ROE, según J. M. Alía Plana, constituyen una de las principales herramientas de control civil sobre el poder militar, pues expresan el sometimiento militar al Estado de Derecho, formando parte del Derecho Militar Operativo, dotando a los mandos y tropas de la

carácter civil (CICR, 2016d).

asistencia necesaria para que las operaciones militares se mantengan dentro del ámbito del DICA, como pone de manifiesto R. Lorenzo Ponce de León (Alía Plana, 2009; Lorenzo Ponce de León, 2012: 47).

En dicho contexto, según R. Titiriga, los cuatro requerimientos acumulativos sobre las leyes de focalización de objetivos (*targeting*) serían: la necesidad militar, el concepto de determinación/distinción, la proporcionalidad y la humanidad (Marín Martínez, 2019: 46-47; Titiriga, 2016: 59, 76-78):

- *Necesidad militar*: objetivos no prohibidos por el DICA;
- *Determinación/distinción*: diferenciación entre combatientes y civiles (Artículo 48 del PAI, así como las reglas complementarias en los Artículos 31 y 52);
- P*roporcionalidad*: Examen de daños colaterales excesivos (Artículos 51(5)(b) y 57(2)(iii) del PAI);
- *Humanidad*: La necesidad de mantener un "cuidado constante" para evitar daños a la población o infraestructuras civiles (Artículo 57 del PAI).

El problema surge en relación con la "interpretación", dado que dichos requerimientos pueden derivar en enfoques diferentes y posibles preeminencias de unos principios sobre otros. Para D. Kostadinov, la idea de "necesidad militar" tiene un papel fundamental ya que tiene una precedencia temporal sobre los otros principios del DIH (humanidad, proporcionalidad, determinación/distinción).

Consecuentemente, según su punto de vista, una operación cibernética se debería analizar en primer lugar tomando como referencia la "necesidad militar". Para los USA, sin embargo, el principio de "necesidad militar", podría ser visto como un "complemento lógico" (*logical inverse*) del principio de "humanidad", dado que dicho principio prohíbe acciones no necesarias y por tanto reforzaría la efectividad militar. En contraste, para el CICR el principio de "necesidad militar" generalmente contraviene el principio de "humanidad", por lo que el propósito del DIH sería el establecer un equilibrio entre las exigencias humanitarias y la necesidad militar. Además, la "Comisión de Derecho Internacional" (CDI) (*International Law Commission (ILC)*), con relación a la necesidad como justificación trató la "necesidad militar" como una excepción al DIH. De acuerdo con el CDI la necesidad sería inadmisible como justificación para no aplicar el DIH (CDI, 1980: CICR, 2021; DoD, 2016: 59; Kostadinov, 2014).

Especialmente significativo dentro del DICA sería el concepto de "proporcionalidad", codificado en el Artículo 51(5)(b) del PAI. El analista R. Titiriga sostiene que dicho concepto dependería de la organización modular de la mente humana y los módulos cognitivos implicados en la valoración de los principios de discriminación/distinción y proporcionalidad por los humanos. En dicho contexto, un despliegue legal de los SAA sería posible en aquellas circunstancias donde los civiles estuviesen ausentes del campo de batalla, especialmente de las "zonas de matanza" (*killing zones*), como por ejemplo en los combates en el mar o bajo el mar y en los combates aéreos. Seguiría la práctica del DIH consuetudinario, en su norma 14ª sobre "proporcionalidad en el ataque"[71]. No obstante,

71 **Norma 14**: Queda prohibido lanzar un ataque cuando sea de prever que

en el informe del CICR sobre "El principio de proporcionalidad en las reglas que gobiernan el desarrollo de hostilidades bajo el DIH", de 2016, se hace hincapié que la responsabilidad de aquellos que tomen decisiones sobre un ataque, según el principio de proporcionalidad, deberá ser analizada sobre la base del análisis de la información que hubiesen tenido en dicho momento y no con el beneficio de la retrospectiva. En todo caso, sería también importante tomar en consideración la opinión de M. Waltzer, que argumenta que no solo la intención de un atacante debe ser buena, sino que debe intentar minimizar el impacto civil, aunque tuviese consecuencias negativas para sí mismo (como el no obtener una victoria militar definitiva) (CICR, 2007: 53; CICR, 2016e: 9; Marín Martínez, 2019: 47; Titiriga, 2016: 80-83, 85; Walzer, 1977: 155).

4.3.- EL IMPACTO DE LA CIBERGUERRA EN EL DICA

Todas las leyes y en especial las relativas al DICA fueron desarrolladas para un mundo analógico y en la actualidad tienen bastante problemas para adaptarse al mundo digital. Dado que no se vislumbra un nuevo tratado internacional en un futuro cercano, en el ciberespacio se tendrá que ir a un conflicto armado con el marco jurídico existente y no con el que se desearía tener. Ahora bien, en la actualidad no existe consenso entre los Estados sobre cómo aplicar, en la práctica, el Derecho Internacional existente del *ius ad bellum* y el *ius in bello* en el ciberespacio. Para intentar soslayar dicha situación, desde un enfoque regional occidental, la OTAN publicó

cause incidentalmente muertos y heridos entre la población civil, daños a bienes de carácter civil o ambas cosas, que sean excesivos en relación con la ventaja militar concreta y directa prevista (CICR, 2007: 53).

los denominados "Manuales de Tallinn" sobre "las Leyes Internacionales Aplicables a la Ciberguerra" y el "Derecho Internacional Aplicable a las Operaciones Cibernéticas"[72], dentro del marco de recomendaciones de expertos a través de instrumentos jurídicos no vinculantes (*soft law*). El CICR, que participó en su desarrollo como observador, acogió con satisfacción dicha iniciativa, dado que el uso de operaciones cibernéticas en un conflicto armado podría tener unas consecuencias humanitarias devastadoras y dichos documentos podrían contribuir en la discusión internacional. En los siguientes puntos nos referiremos a algunos conceptos importantes con relación a la ciberguerra y las operaciones cibernéticas de dichos manuales (CICR, 2013; OTAN 2013; OTAN, 2017).

Con relación al Manual de Tallinn sobre la ciberguerra de 2013, el primer punto que trataremos será el concepto de "responsabilidad" sobre la atribución de un operación cibernética. La Regla 7ª establece que "el simple eco de que una operación cibernética sea lanzada, desde una infraestructura gubernamental cibernética, no sería suficiente evidencia para atribuir la operación a dicho Estado, pero si es una indicación de que el Estado en cuestión estaría asociado a ella. No incluye aquellas operaciones encaminadas a través de la infraestructura gubernamental o las iniciadas a través de infraestructuras cibernéticas no gubernamentales. La Regla 8ª va más allá, ya que establece que "el eco de que una operación cibernética haya sido encaminada a través de la infraestructura cibernética de un

72 El manual Tallinn 2.0 fue escrito por expertos, principalmente occidentales, en Derecho Internacional, aunque también incluyó a expertos de Tailandia, Japón, China y Bielorrusia, dentro de las áreas de derechos humanos, derecho en el Espacio y derecho internacional de las telecomunicaciones. También fueron invitados a participar, como observadores, el CICR, así como otros Estados y organizaciones (Jensen, 2017: 737).

Estado no es suficiente evidencia para atribuir la operación a dicho Estado". Ambas reglas muestran los problemas actuales de atribución dentro del ciberespacio (OTAN, 2013: 39-40).

Sobre el "uso de la fuerza", la Regla 10ª establece que "una operación cibernética que constituye una amenaza del uso de la fuerza contra la integridad territorial o independencia política de cualquier Estado o que no es consistente con los propósitos de las Naciones Unidas, es ilegal". En cuanto a la "legítima defensa", la Regla 13ª indica que "un Estado que es el objetivo de una operación cibernética que llega al nivel de un "ataque armado" puede ejercer el derecho inherente de "legítima defensa". El que una operación cibernética constituya un "ataque armado" dependerá de su nivel y de sus efectos". En todo caso, los Expertos no establecieron los parámetros para dicho criterio, aunque si indicaron que un "uso de la fuerza" que implicase la muerte o el daño hacia las personas o las propiedades satisfaría cumplir los criterios de nivel y de efectos. Las Reglas 14ª y 15ª además establecen que el derecho a la "legítima defensa" debe cumplir los requisitos de necesidad, proporcionalidad, inminencia e inmediatez (OTAN, 2013: 45, 53-54, 58).

La Regla 20ª sobre la aplicabilidad del DICA establece que "las operaciones cibernéticas ejecutadas en el contexto de un conflicto armado están sujetos a la Ley de Conflictos Armados (*LOAC-DICA*)". Dado que el DICA no regula las actividades en el ciberespacio, se deberá tener en cuenta la "cláusula Martens" de la IV Convención de La Haya de 1907[73], las Convenciones de Ginebra

73 **Cláusula Martens (Convención de la Haya IV 1907)**: Hasta que un Código más completo de las Leyes de guerra se haya publicado, las Altas Partes Contratantes juzgan oportuno declarar que, en los casos no incluidos en las disposiciones reglamentarias adoptadas por ellas, las poblaciones y los beligerantes

de 1949 y el PAI de 1977. En cuanto a la "responsabilidad criminal" de los "Comandantes y otros Superiores", la Regla 24ª(a) establece que "los Comandantes y otros Superiores son criminalmente responsables por ordenar operaciones cibernéticas constituyentes en crímenes de guerra". Sobre el concepto de "levantamiento en masa", la Regla 27ª establece que "en un conflicto internacional armado, los habitantes de un territorio no ocupado que lleva a cabo operaciones cibernéticas como parte de un 'levantamiento en masa' gozarán de inmunidad de combatiente y el estado de prisionero de guerra". En cuanto al concepto de los "civiles", la Regla 29ª establece que "no está prohibido que los civiles participen directamente en operaciones cibernéticas hostiles, pero perderán la protección sobre posibles 'ataques' mientras participen". Siguiendo con el mismo concepto, la Regla 32ª establece que "la población civil como tal, al mismo tiempo que los individuos civiles, no serán objetivo de un ciberataque" (OTAN, 2013: 68, 80, 88, 90. 97).

Una mención especial se establece con respecto a los objetos de "doble uso" (civil y militar). La Regla 39ª establece que "un objeto utilizado tanto para propósitos civiles y militares – incluyendo ordenadores, redes y la infraestructura cibernética – es un objetivo militar". La Regla 43ª prohíbe el uso de medios y de métodos de ciberguerra que sean de naturaleza indiscriminada, basándose en los Artículos 51(4)(b) y (c) del PAI. La Regla 44ª establece que "está prohibido el uso de trampas explosivas cibernéticas asociadas con determinados objetos en el DICA". También quedan prohibidas las represalias, de acuerdo con la PAI, a través de la Regla 47ª. La

quedan bajo la protección y el imperio de los principios de la ley de las naciones, tal como y resultan de los usos establecidos entre naciones civilizadas, de las leyes de la humanidad y los dictados de la conciencia pública (Roberts y Guelf, 1989: 45).

Regla 70ª, indica que "el personal médico y religioso, las unidades y los transportes médicos deben ser respetados y protegidos y, en particular, no deben ser un objetivo de un ciberataque". Se complementa con la Regla 71ª sobre los ordenadores, sistemas y redes médicas. Por último, en este repaso, la Regla 81ª establece que "se prohíbe el ataque, la destrucción, el remover o el hacer inservible aquellos objetos indispensables para la supervivencia de la población civil, por medio de operaciones cibernéticas" (OTAN, 2013: 113, 121-122, 126-127, 167, 169, 185).

Un apunte general sobre la taxonomía del Manual de Tallinn y la definición de lo que se considera un "ataque". Dicho Manual, establece dicha consideración hacia un "objeto" basándose en el concepto de funcionalidad de acuerdo con tres escenarios diferentes: una operación cibernética podría dañar físicamente un componente de un sistema informático; podría hacer que dejase de funcionar hasta que se reinstalase el sistema operativo o; podría hacer que dejase de funcionar eliminando o interfiriendo sobre los datos del sistema. En dicho contexto la CICR sería de la opinión de que un "ataque" sería cualquier evento cibernético que provocase la pérdida de funcionalidad, mientras que el analista M.N. Schmitt, director del proyecto de desarrollo de los Manuales, sería de la opinión de que solo se consideraría un "ataque" si incidiese directamente sobre el sistema operativo, es decir cuando el sistema pierde su funcionalidad y ya no puede llevar a cabo la función que tenía asignada sin alguna reparación. Incluiría la recarga del sistema operativo o cualquier "software" esencial para su operación, pero no el reemplazo de los datos que estuviesen almacenados en el sistema. Dichas opiniones reflejan las diferencias existentes sobre la catalogación de los datos informáticos como "objetos" dentro del mundo cibernético (Brown, 2017: 368).

En cuanto al "Manual de Tallinn 2.0", el primer concepto que trataremos será el principio de "soberanía". La Regla 1ª establece que "la soberanía es aplicable al ciberespacio", mientras que la Regla 4ª indica que un "Estado no debe llevar a cabo operaciones cibernéticas que violen la soberanía de otro Estado" [74]. Un segundo aspecto estaría relacionado con la "jurisdicción", que viene definida como "la competencia de los Estados para regular a las personas, los objetos y las conductas bajo sus leyes nacionales, dentro de los límites impuestos por las leyes internacionales". En principio, con las limitaciones impuestas por las leyes internacionales, un Estado puede ejercitar la "jurisdicción" territorial y extraterritorial sobre las actividades cibernéticas. La Regla 9ª confirma que tanto la jurisdicción subjetiva como la objetiva territorial, serian de aplicación a dichas actividades. No hay consenso, sin embargo, sobre los datos en tránsito a través de un Estado y si éste tiene jurisdicción sobre dichos datos. En cuanto a la Regla 10ª se permite que los Estados tengan "jurisdicción" extraterritorial basada en la nacionalidad del individuo, aunque no queda resuelto el asunto de si es solo sobre el individuo o también sobre los datos que desarrolla en otro Estado. En cuanto a la aplicabilidad de la "jurisdicción", la Regla 11ª la permite dentro del territorio de un Estado, pero pone límites en el aspecto extraterritorial, que solo sería permitido con el consentimiento del otro Estado[75] (Jensen, 2017: 741, 747-748).

74 Por ejemplo, si un agente de un Estado utiliza un "pendrive" para introducir un "*malware*" en una infraestructura de otro Estado, entonces se estaría produciendo la violación de la soberanía de dicho Estado. En dicho contexto los USA no están de acuerdo con dicha premisa pues, tomando como referencia el Artículo 2(4) de la Carta de la Naciones Unidas, la regla de la "soberanía", por sí misma, no es ejecutable (Jensen 2017: 741).

75 Un aspecto fundamental con relación a los datos alojados en la "nube", que

Un concepto extremadamente importante es la doctrina de la "responsabilidad", de acuerdo con los artículos establecidos por la "Comisión de la Responsabilidad de los Estados" de las Naciones Unidas. La Regla 14ª establece que un "Estado tiene responsabilidad internacional sobre cualquier acto cibernético atribuible a dicho Estado que constituya una violación de una obligación legal internacional". La Regla 15ª, toma en consideración los Artículos 4º y 5º sobre la "Responsabilidad de los Estados", estableciendo que también se consideran órganos de los Estados a los actores (tanto individuos como organizaciones) que, aunque por ley no formen parte de un Estado si tengan una "dependencia completa" del mismo. Por otro lado, la Regla 17ª trata de los actos "por delegación" (*proxy*). De acuerdo con las leyes internacionales, las operaciones cibernéticas llevadas a cabo por actores no estatales, pero que estén bajo "un control efectivo" de un Estado, entonces dichos actos serían atribuibles a dicho Estado. En cuanto al aspecto de las "contramedidas", los Expertos estuvieron de acuerdo que dichas contramedidas no podían violar una norma perentoria y que deberían ser proporcionales al daño recibido, aunque no existiría la necesidad de que dichas contramedidas cibernéticas tuviesen como objetivo el mismo órgano estatal que hubiese violado la ley internacional (Jensen, 2017: 750-751, 754).

En cuanto a las leyes internacionales sobre los DD.HH., la Regla 34ª establece que "la leyes internacionales sobre los Derechos Humanos son aplicables a las actividades cibernéticas". No hubo consenso completo sobre la aplicación extraterritorial, aunque la mayoría de los Expertos si pensaban que sería de aplicación extra territorialmente. Por su parte, la Regla 35ª establece que "los indi-

en muchos casos está distribuido entre varios Estados.

viduos disfrutan de los mismos DD.HH. con respecto a las actividades cibernéticas", incluidas la libertad de expresión, el derecho a tener una opinión y el derecho a la privacidad. Sin embargo, la mayoría de los Expertos estuvieron divididos sobre la protección de las comunicaciones privadas a través de algoritmos informáticos. La mayoría fueron de la opinión que dicho tipo de inspecciones no implicaban una intrusión sobre los derechos individuales hasta que un Estado no accediese las comunicaciones de alguna forma incluyendo el procesamiento de los datos. Sobre el concepto de "intervención", la Regla 66ª establece que "un Estado no debe intervenir, incluso por medios cibernéticos, en los asuntos internos o externos de otro Estado". Dicho principio tiene especial importancia, en la actualidad, dadas las alegaciones sobre intervenciones cibernéticas rusas en procesos electorales de USA y Europa (Marín Martínez, 2019: 39-44; Jensen, 2017: 758-759, 774, 776).

Desde un punto de vista de un marco más global, debemos volver la atención a la adopción por consenso en 2019, por parte del CCW, de los principios rectores formulados por el GGE sobre los SAAL, que ya hemos analizado. Especialmente, habría que subrayar el primer principio rector que determina que "El DIH sigue aplicándose plenamente a todos los sistemas de armas, incluido el posible desarrollo y uso de los SAAL", que tiene un gran calado para el mundo cibernético, sobre todo porque fueron adoptados, entre otros Estados, por las grandes potencias: Rusia, China y USA. En dicho contexto, es importante destacar la posición de Rusia sobre la plena aplicabilidad de los principios rectores y su especial mención de la aplicación del DIH existente a todos los sistemas de armas, en su "Documento de Trabajo" sobre la aplicabilidad de los principios rectores, de 2019. En cuanto a China, aunque no existe ningún documento conocido en la actualidad sobre su

postura sobre los principios rectores, si existe su posición, de 2018, sobre la aplicación de las reglas humanitarias internacionales a los SAAL, que deben estar de acuerdo con la Convención de Ginebra de 1949 y los PAI y PAII de 1977, con especial mención de los principios de precaución, distinción y proporcionalidad: En cuanto a los USA, sus "Comentarios sobre los Principios Rectores", de 2020, abogan por la necesidad de que el GGE sobre SAAL aclare los requerimientos del DIH y su aplicabilidad sobre las tecnologías emergentes, proponiendo algunas conclusiones a considerar, incluyendo la idea de que si un sistema de armamento autónomo fuese incapaz de ser usado de acuerdo con los principios de distinción y proporcionalidad, entonces dicho sistema sería ilegal (NU, 2018b: 2; NU, 2019e: 11-12; NU, 2019h; NU, 2020c: 2).

Siguiendo con nuestro análisis conceptual y aunque en próximos capítulos trataremos el tema con mayor profundidad, habría que indagar el potencial de los "Agentes Morales Artificiales" (AMA) en el desarrollo del marco jurídico del *ius ad bellum* y el *ius in* bello en el ciberespacio. A tal fin, uno de los elementos principales sería la posibilidad de la cuantificación de dichas normas para hacerlas aplicables dentro de un AMA. Un primer paso podría venir a través de la aplicación de algún marco normativo cuantificado ya existente o a desarrollar. A tal fin, el marco normativo planteado por M. N. Schmitt para decidir si la implicación de un Estado en una operación cibernética pudiera constituir el "uso de la fuerza", dentro del *ius ad bellum*, podría servir como ejemplo de cuantificación normativa. Dicho marco compara las características de una operación cibernética concreta con las características tradicionales del uso de la fuerza a través de siete parámetros: severidad; inmediatez; dirección directa; el grado de invasión; cuantificabilidad; legitimidad presumida y; responsabilidad. Cuantificando cada uno

de los parámetros en una escala del 1 al 10, por ejemplo, estableciendo un umbral de puntuación por encima del cual se consideraría un "uso de la fuerza", entonces un AMA en el que se integrase dicho marco normativo, podría establecer que una operación cibernética constituyese un "uso de la fuerza"[76]. No obstante, antes de su aplicación, habría que tener en cuenta que el propio M. N. Schmitt consideraba que la evaluación de dichos parámetros podría ser imprecisa y subjetiva y que, por lo tanto, no debería ser aplicada mecánicamente sino dependiendo del contexto, teniendo un mejor uso como herramienta forense que para caracterizar operaciones en tiempo real. Añadiríamos, por nuestra parte, la dificultad de establecer la atribución de una operación cibernética[77] a un actor concreto o que estamento decidiría el umbral por encima del cual dicha operación constituiría un uso de la fuerza[78] (Foltz, 2012: 43; Schmitt, 1999: 914-915, 927).

76 Para más información ver: MICHAEL, J. B. *et al* (2003): "Measured Responses to Cyber Attacks Using Schmitt Analysis: A Case Study of Attack Scenarios for a Software-Intensive System", *Proc. Twenty-seventh Annual Int. Computer Software and Applications Conf.*, IEEE, Dallas, acceso enero 2021, en https://www.researchgate.net/publication/221028636_Measured_Responses_to_Cyber_Attacks_Using_Schmitt_Analysis_A_Case_Study_of_Attack_Scenarios_for_a_Software-Intensive_System

77 Para más información ver: MEJIA, E. F. (2014): "Act and Actor Attribution in Cyberspace: A Proposed Analytic Framework", *Air University,Strategic Studies Quarterly (SSQ)*, acceso enero 2021, en https://apps.dtic.mil/dtic/tr/fulltext/u2/a603128.pdf

78 El propio M.N. Schmitt estableció tres franjas para la aplicación del Artículo 2(4): 0-3 (no un uso de la fuerza); 4-6 (zona gris para establecer el uso de la fuerza); 7-10 (uso de la fuerza de acuerdo con dicho artículo) (Pipyro *et al,* 2018: 381; Schmitt, 1999: 914-915).

A la vista del análisis realizado debemos concluir en que la aplicación del Derecho al ciberespacio, como aprecia K. Watkin, es una lucha continua en "las fronteras de la ley", existiendo zonas obscuras y ambiguas que los Estados explotan a su conveniencia, especialmente cuando se entra en territorios inexplorados como la "guerra híbrida", ya analizada en el capítulo anterior. Es por tanto esencial, según nuestro criterio, que debe existir una prohibición de armamentos indiscriminados, así como la vigencia de la "Cláusula Martens" en aquellas instancias donde el Derecho Internacional vigente fuese de difícil aplicación en el ciberespacio. Es más, la reacción ante la ambigüedad, como sugiere J. Herbach, debería establecer el principio de precaución como la base para el desarrollo de los SAA y los SAAL, ya desde la fase de investigación y desarrollo.

Tampoco se debe obviar que existe en la actualidad una debilidad del sistema de gobernanza, pues asistimos a una "gobernanza jurídica asimétrica" ya que, aunque las NU en sus resoluciones han establecido que el Derecho Internacional y la Carta de las Naciones Unidas son de aplicación al ciberespacio, la problemática surge cuando se intenta establecer la forma de su aplicación y si para ello es necesario establecer un nuevo instrumento jurídico. Teniendo en cuenta los intercambios entre Estados y los debates en la NU, nuestro punto de vista es que será difícil que a corto o medio plazo se establezca un nuevo tratado, pues los Estados prefieren medidas internas a nivel estatal y solo contemplar el intercambio de información de manera voluntaria (China, Rusia, USA y España, entre otros).

En dicho contexto, dentro de la ambigüedad, consideramos de especial relevancia el uso por parte de los Estados de las "guerras

por delegación", que extienden el "anonimato" de las operaciones cibernéticas y todo el marco de los "*hackers*". Eventos, dentro del marco del *ius ad bellum*, que, aunque incompatibles con el Derecho Internacional, son considerados aceptables desde el punto de vista operativo por parte de los Estados. Particularmente, cuando se habla de acciones preventivas de ciberdefensa que serían contrarias al "derecho de legítima defensa"[79] y que tienen especial impacto sobre el principio de "proporcionalidad".

Por lo tanto, consideramos esencial el desarrollo de instrumentos jurídicos no vinculantes (*soft law*), que intentan ofrecer una respuesta a la falta de nuevos instrumentos. Tanto los trabajos de la OTAN, a través de los "Manuales de Tallinn" o los del CICR sobre el Derecho consuetudinario son de gran interés, al proporcionar un marco interpretativo estable. Ahora bien, no se debe obviar que dichos instrumentos presentan diferencias para una misma acción, como por ejemplo las diferencias de interpretación de la definición de "ataque". Es por ello, que los trabajos institucionales a nivel internacional son de especial relevancia. Los trabajos del GGE del CCW sobre los SAAL y el desarrollo de principios rectores son fundamentales para una aplicación consensuada del Derecho Internacional en los sistema armamentísticos cibernéticos. Hay que tener en cuenta que los principios de humanidad, necesidad, proporcionalidad y distinción forman el núcleo principal del DIH.

No obstante, aún quedan por solucionar aspectos importantes dentro del marco cibernético. Algunos de los más relevantes serían:

79 De acuerdo con el Art. 51 de la Carta de las NU.

- *La definición de ataque armado según el Derecho Internacional*: todos los ataques armados son usos de la fuerza, pero no todos los usos de la fuerza son considerados ataques armados. Sería necesario reconceptualizar la noción de severidad dentro del dominio del ciberespacio.
- *La definición de "objeto"*: No están resueltos los aspectos de visibilidad y tangibilidad, la consideración de los datos como "objetos virtuales" o aquellos objetos de doble uso.
- *La definición de "necesidad militar"*: Existen graves discrepancias entre los Estados y las instituciones internacionales. Nuestro punto de vista se alinea con el expresado por la Comisión de Derecho Internacional (CDI/ILC): la necesidad no es admisible para no aplicar el DIH.
- *La aplicación práctica del Derecho Internacional y del Derecho consuetudinario*: Consideramos imprescindible pasar de la teoría a la práctica y establecer mecanismos de implementación especialmente en los SAA y SAAL. Establecer criterios internacionales sobre el principio de responsabilidad y los criterios de rendición de cuentas, así como un control de riesgos holístico de los sistemas armamentísticos, basado en los principios de Derecho Internacional vigentes, para lo que se necesitará una adecuada formación de todos los actores y el trabajo conjunto de diversas disciplinas (jurídica, científica, militar, etc.).

CAPÍTULO 5
AGENTES MORALES ARTIFICIALES

El filósofo B. Gert describió la moralidad común como: "aquel "sistema moral" que utilizan las personas, normalmente de una forma no consciente, para decidir cómo actuar cuando se ven confrontadas con problemas morales y cuando establecen su propios juicios morales". En dicho contexto, se podría definir el concepto de moralidad como "un sistema público informal que sería aplicable a todas las personas racionales, que gobierna el comportamiento que afecta a otros y que incluye lo que comúnmente se conoce como reglas morales, ideales y virtudes, teniendo como su objetivo el disminuir la maldad o el daño". Pero también, como indicaba K. Abney, la moralidad sería lo que el mundo debería ser en oposición a lo que realmente es. En todo caso, como ya analizamos en el capítulo 2º, la moralidad no es equivalente a la ética, pues la moralidad es un conjunto de normas que guían nuestro comportamiento, mientras que la ética es la teoría y el reflejo de la moralidad. El problema radicaría cuando, para alcanzar la mejor condición posible dentro de una situación se desarrollase una acción que moralmente resultase inaceptable, teniendo en cuenta que dicha inaceptabilidad dependería del fin que se desease alcanzar[80], pero también del pluralismo ético y los distintos enfoques morales existentes, tanto a nivel de civilización, a nivel religioso o incluso a nivel de Estados, según vimos anteriormente cuando analizamos

80 Se podría tomar como ejemplo el realizar un ataque a un objetivo por medio de un VANT, infringiendo con ello el DICA.

la ciber ética (Abney, 2012: 36; Gert, 1999: 58).

El auge en el desarrollo de la IA, especialmente en sistemas de apoyo a la toma de decisiones, cada vez más generan soluciones con consecuencias morales. Según B. Whitby, dado que dicha situación cada vez sería más frecuente, la pregunta a realizar podría ser: ¿Bajo qué circunstancias se aceptaría un juicio moral de una máquina? Su propuesta establece que la premisa sería que dicho juicio fuese aceptable moralmente por humanos libres y racionales. En contraste, para C. Allen. G. Varner y J. Zinser, la premisa podría ser la capacidad de una máquina para pasar una forma de "Test de Turing Moral" (*Moral Turing Test*), específico para asuntos de moralidad[81]. Si el "interrogador" no pudiese identificar la diferencia entre la máquina y el humano, entonces dicha máquina, sobre el criterio de la moralidad, sería considerado un agente moral. No obstante, como los objetivos filosóficos no tienen que ser los mismos que los de los ingenieros, puede existir una brecha importante entre las posibles teorías morales a utilizar y el diseño de algoritmos que pudiesen ser implementados en los AMA. En dicho contexto, la tarea de ingeniería para crear sistemas autónomos, que salvaguardasen valores humanos básicos, forzaría a los científicos a desagregar las acciones de toma de decisiones morales en sus componentes, teniendo en cuenta qué tipos de decisiones pudiesen ser o no ser codificadas y gestionadas por sistemas mecánicos. Al mismo tiempo, dichos ingenieros deberían aprender a diseñar sistemas cognitivos y afectivos capaces de gestionar la ambigüedad y

81 Para más información ver: WALLACH, W. y ALLEN C. (2012); "Hard Problems: Framing the Chinese Room in which a Robot takes a Moral Turing Test", *AISB, IACAP,* acceso enero 2021, en https://www.researchgate.net/publication/289173777_Hard_problems_Framing_the_Chinese_room_in_which_a_robot_takes_a_moral_turing_test

las acciones conflictivas (Allen, 2000: 254-255; Allen, 2008: 566; Whitby, 2008: 552-553).

5.1.- EL CONCEPTO DE AGENTE MORAL

Para poder construir el paradigma de un AMA será necesario, en primer lugar, desarrollar los conceptos de "agente" y "agente moral". Para K. E. Himma, desde un punto de vista conceptual, "X sería un agente sí y solo sí fuese capaz de llevar a cabo acciones". Ahora bien, solo aquellos seres capaces de estados intencionales son "agentes". Algunos son "naturales" cuya existencia puede ser considerada de acuerdo con consideraciones biológicas (personas, perros, gatos, etc.) y otras son "artificiales" dado que son fabricados por agentes intencionales y pueden ser considerados como "artefactos". También podrían existir agentes "naturales" y "artificiales" a la vez, como el caso de los clones, donde si se manufacturase ADN desde material no genético preexistente, entonces el organismo resultante estaría vivo tanto biológicamente como artificialmente. Por lo tanto, una definición adaptada del concepto de "agente", según K. E. Himma, sería: "X es un 'agente' sí y solo sí puede instanciar estados mentales intencionados capaces de causar directamente un desempeño" (Himma, 2009: 21-24).

En cuanto al concepto de "agente moral", aunque en la actualidad no exista una definición consensuada, se pueden establecer una serie de parámetros comunes que forman parte de las diversas definiciones. Según P. Brey, los "agentes morales" son seres que: son capaces de razonar, juzgar y actuar con referencia al bien y al mal; se espera de ellos que se comporten, al llevar a cabo sus acciones, por medio de estándares morales y; son moralmente responsables

por sus acciones rindiendo cuenta de sus consecuencias. Por lo tanto, un "agente" sería un "agente moral" cuando: "los estados de intención que cultiva y las acciones que lleva a cabo subsecuentemente son guiadas por consideraciones morales". Otra definición, de J. Parthemore y B. Whitby, considera un "agente moral" aquel que: se le puede hacer responsable apropiadamente por sus acciones. Ahora bien, para que así fuese considerado debería estar situado en un contexto espacio-tiempo particular y materializado de una forma física particular que moldeasen y limitasen sus interacciones con el medio ambiente. Por su parte, para L. Floridi y J. W. Sanders existirían cuatro condiciones para establecer un "agente moral": Para todo X, X es un agente moral sí y solo sí X tuviese las siguientes propiedades (Floridi y Sanders, 2004: 361):

- *Interactividad*: X y su medio ambiente son capaces de interactuar sobre el otro;
- *Autonomía*: X sería capaz de cambiar sus estados internamente sin el estímulo de la interacción con el mundo exterior;
- *Adaptabilidad*: X sería capaz de cambiar las reglas de transición por las que cambia de estado;
- *Moralidad*: X sería capaz de actuar de tal forma que produjese efectos morales que pudiesen causar el bien o el mal, es decir si fuese capaz de acciones moralmente calificables.

En todo caso, lo importante sería poder distinguir que condiciones serían necesarias y suficientes para que algo fuese considerado

como un "agente moral". Según K. E. Himma, existirían dos capacidades que serían necesarias y que conjuntamente serían suficientes:

- La capacidad de elegir libremente sus propios actos implicando, por lo tanto, que fuesen racionales;
- Conocer la diferencia entre el bien y el mal[82].

La primera premisa significaría, al igual que inciden J. Parthemore y B. Whitby, que no sería suficiente que un agente memorizase una lista de lo que se debe o no hacer, sino que exigiría intencionalidad, una construcción y espontaneidad, pero también más importante aún disponer del concepto de "ser". Un agente no podría ser considerado moralmente responsable por sus acciones si no tuviese un concepto de sí mismo como el agente que está actuando. Más aún, para que fuese un "agente moral" debería también ser consciente del concepto de límite entre el "ser" y el "otro", esto es, donde "uno" acaba y el "otro" comienza. Esto implicaría que algo podría ser considerado como "agente moral" cuando tuviese capacidad de "conciencia" pues, según K. E. Himma, el concepto de "rendición de cuentas" solo podría ser atribuible a seres conscientes y, por lo tanto, sería imposible conceptualmente recompensar o castigar algo que no lo fuese (Brey, 2014: 125; Floridi y Sanders, 2004: 361; Himma, 2009: 21-24; Parthemore y Whitby, 2013: 105, 109, 111-112).

82 Para más información sobre el concepto de "maldad" ver: FLORIDI, L. y SANDERS, J. W. (2001): "Artificial Evil and the Foundation of Computer Ethics", *Ethics and Information Technology,* 3(1), 55-66.

5.2.- EL CONCEPTO DE AGENTE MORAL ARTIFICIAL

El incremento en la autonomía de los sistemas informáticos y la expansión de la integración de elementos de IA en la tecnología implicará, cada vez más, que los diseñadores e ingenieros no puedan predecir en todo momento las opciones y acciones que los sistemas llevaran a cabo cuando se vean confrontados con situaciones no esperadas. Para tales situaciones, según W. Wallach, la construcción de máquinas morales debería ser considerado como un objetivo práctico, para no permitir que máquinas cada día más autónomas causasen daño a humanos o a otras entes dignos de consideración moral. Ahora bien, la inclusión de productos tecnológicos dentro del ámbito de los "agentes morales" dependerá de los criterios con los que se defina a un "agente moral" y la capacidad de dicho agente a unificar tanto a los seres humanos como a los "agentes artificiales" (AA).

Particularmente, la problemática surge desde el momento que se decide considerar a un AA como un "agente", ya que según la definición establecida por K. E. Himma, debería ser capaz de instanciar estados mentales intencionados. Dicha problemática se obviaría según, F. Fossa, si el concepto de "agente" fuese adaptado al contexto actual, un enfoque propuesto por M. Laukyte, que considera dicha acepción como válida al apreciar que los AA, al ser racionales e interactivos, tendrían atributos de responsabilidad y personalidad por lo que formarían parte del mundo social (Fossa, 2018: 116; Himma, 2009: 21-24; Laukyte, 2017: 1-17; Wallach, 2010: 243).

Pero: ¿puede un AA tener moralidad? En el año 2000, los investigadores C. Allen, G. Varner y J. Zinser acuñaron el término "Agen-

te Moral Artificial" (AMA) para definir aquellos AA con capacidades morales (Allen *et al*, 2000). Una primera aproximación para establecer que un AA fuese un AMA sería el cumplimiento de las cuatro condiciones establecidas por L. Floridi y J. W. Sanders para definir un "agente moral" (interactividad, autonomía, adaptabilidad, moralidad), que anteriormente hemos descrito.

Una definición funcional para incluir a los AMA dentro de los "agentes morales" y que formaría parte de un denominado "Planteamiento Continuista" (*Continuity Approach*), como lo describe F. Fossa. Dentro de dicho marco, la presunción sería de que no existiría una diferencia fundamental entre los seres humanos y los AMA con relación a la "agencia moral" (*moral agency*). Planteamiento también expuesto por J. H. Moor en su clasificación de "agentes morales"[83] y por W. Wallach y C. Allen en sus clasificaciones de "agencia moral"[84] o con relación a la propuesta del "Test de Touring Moral" ya descrito. En todo caso, como indica F. Fossa, el único requisito sería que la imitación del comportamiento moral humano por medio de la tecnología sería factible sí y solo sí el concepto de "agencia moral" fuese abordado en término cuantitativos y por lo tanto, estuviese dentro de un marco operativo (Floridi y Sanders, 2004: 361; Fossa, 2018: 116, 118; Moor, 2006: 19-21; Wallach y Allen, 2009; 2011: 104).

83 Para J. H. Moor, los "agentes" pueden ser clasificados en cuatro categorías dependiendo de su grado de autonomía y complejidad: agentes de impacto éticos; agentes éticos implícitos; agentes éticos explícitos y; agentes completamente éticos. En los "agentes éticos implícitos" existiría software que implícitamente soportaría un comportamiento ético (Moor: 2006: 19-21).

84 Clasificación a través de tres niveles: moralidad operativa; moralidad funcional y; "agencia moral" completa, En la moralidad operativa el significado moral queda en mano de los diseñadores y usuarios, mientras que en la moralidad funcional las propias máquinas tendrían la capacidad de evaluar y responder a los desafíos morales (Fossa, 2018: 116; Wallach y Allen, 2011: 104).

En este punto, habría que exponer el concepto de "mentor moral" con referencia a los AMA. Según F. Fossa, si no existiese una diferencia cualitativa que discriminase a los agentes morales humanos de los AMA, entonces el concepto de "agencia moral" debería ser medible y reproducible tecnológicamente, por lo que el marco establecido para un "agente moral humano" no tendría por qué ser el marco definitivo de un AMA. En el caso de que la IA avanzase más allá de los límites de la moralidad humana, entonces los seres humanos se volverían "agentes morales" obsoletos. En dicho punto, los AMA se convertirían en nuestros "mentores morales" entrando desde un punto moral en la denominada "Singularidad", que analizaremos posteriormente en nuestra investigación, donde los modelos actuales serían descartados y donde los AMA superinteligentes, según N. Bostrom, podrían responder a cuestiones éticas, más exactamente que los humanos, por lo que también serían agentes autónomos, creando una nueva imagen del mundo y una voluntad propia (Bostrom, 2003: 3; Fossa, 2018: 119-120).

El concepto del "Planteamiento Continuista" chocaría, sin embargo, con el planteamiento de D. G. Johnson que propone que los sistemas informáticos serían "entidades morales" pero no "agentes morales". Se basa en dos distinciones fundamentales: la distinción entre entidades naturales y las creadas por el hombre y; la diferencia entre "artefacto" y tecnología. En el primer caso, si se eliminase la distinción entre lo natural y lo creado, sería imposible distinguir los efectos del comportamiento humano o su contribución a lo que el mundo es. Haría difícil o casi imposible para los humanos el comprender las implicaciones de sus decisiones normativas sobre el futuro. En cuanto a la diferencia entre "artefacto" y tecnología, un ordenador o un sistema informático no serían fenómenos

naturales y por lo tanto no podrían existir a no ser que existiesen complejos sistemas de conocimiento y de comportamiento social (político, religioso, cultural, etc.). No importaría como fuesen los sistemas informáticos de independientes, autónomos o interactivos en el futuro, serían productos (directos o indirectos) del comportamiento humano de las instituciones sociales humanas y de la decisión humana. Por lo tanto, sería imposible considerar una computadora o un sistema informático como un simple objeto formado por un conjunto de piezas, pues siempre tendría un fin determinado dentro de un contexto social-temporal particular. En dicho sentido, los AMA serían ejecutores de funciones establecidas previamente por seres humanos, los verdaderos agentes, comportándose como meras "herramientas" para un propósito específico, como ya había planteado H. Jonas en 1953 (Fossa: 2018; 122; Johnson, 2006: 196-197; Jonas, 1953: 177, 183).

La idea propuesta por D. G. Johnson sería comparable a la de J. J. Bryson y P. P. Kime. Se basaría en los miedos y las esperanzas exageradas atribuibles a la IA con relación a la inteligencia y la posibilidad de alcanzar la Singularidad, dado que la población en general presupone mayor inteligencia a las máquinas que las que realmente poseen, mientras que se obvia el verdadero peligro que supone su uso indebido. Sugieren, que sería menos perturbador para el sistema ético existente si la IA fuese considerada como una herramienta de creatividad y no una creadora por sí misma, especificando que serían los diseñadores y/o los operadores los verdaderos creadores. Como resultado, F. Fossa describirá el denominado "Planteamiento de Discontinuidad" (*Discontinuity Approach*), en las que se aboga por establecer que los AMA son simplemente productos tecnológicos, definiéndolos como "herramientas sensibles"

(*sensible tools*), aquellas que ejecutan funciones mientras reconocen y reaccionan a los aspectos morales potenciales de sus operaciones. (Bryson y Kime, 2011; Fossa, 2018: 123).

Dichas posturas conllevan la idea de que los humanos y no las máquinas son "agentes responsables" y por tanto los AMA, integrados en una máquina, tampoco lo serían. Para A. Sharkey, esto sería debido a que el concepto de moralidad necesita de una base biológica y, además, porque las máquinas no son completamente independientes de sus diseñadores humanos. Como consecuencia, un punto importante de su planteamiento sería que, dado que el Derecho existente no ha sido formulado para tener en cuenta los desarrollos tecnológicos y la IA, no sería suficiente que las máquinas inteligentes (robots) fusen diseñadas y operadas para que cumpliesen con las leyes existentes y los derechos y libertades fundamentales, para que fuesen considerados como "agentes responsables".

Así mismo, B. Malle y M. Scheutz argumentan que los robots necesitarían una red de normas morales para saber lo que sería o no moralmente aceptable. Aducen, que no sería práctico el programar dicha red y que, en vez de programar a las máquinas con normas morales, se construyesen dichas normas de acuerdo con la retroalimentación de las respuestas a sus acciones. En todo caso, A. Sharkey sería escéptico del éxito de dicho planteamiento y no se debería permitir a la máquinas tomar decisiones morales en situaciones letales. Un planteamiento también argumentado por el Relator Especial de las NU P. Heynes, en 2013, con relación a los robots en el campo de batalla, argumentando que los robots carecen de la habilidad de comprender el contexto y, además, los humanos deberían tener el derecho a que solamente otros humanos

tomasen decisiones sobre la vida y la muerte propias (NU, 2013: 6, 13, 17; Malle y Scheutz, 2014; Sharkey, 2017: 210-216).

Llegados a este punto de teorías divergentes sobre si un AMA pudiese ser considerado un "agente moral", sería importante establecer la distinción entre AMA y "Agente Moral Artificial Autónomo" (AMAA), que a veces son confundidos y que consideramos que dicha confusión ha sido la que ha propiciado planteamientos tan dispares sobre los AMA. Como ya hemos analizado, según las propuestas de J. H. Moor y W. Wallach y C. Allen, existirían diversos grados de autonomía posible para los AMA, pero solo una clasificación, en cada una de las propuestas, sería considerada como AMAA. En las clasificaciones de J. H. Moor un AMA verdaderamente autónomo únicamente sería posible si la máquina pudiese realizar juicios explícitos, es decir, tener consciencia, intencionalidad y voluntad propia (*full ethical agents* – agentes éticos completos).

Paralelamente, para W. Wallach y C. Allen, la autonomía completa solo se alcanzaría si se llegase a la "agencia moral" genuina que tendría derechos y responsabilidades comparables con la de los humanos (*full moral agency* – agencia moral completa). Para B. C. Stahl, la problemática que surgiría, en dicho contexto, vendría dada en la identificación de la forma en que dichos agentes podrían reaccionar adecuadamente a los problemas normativos. Esto es, cómo sería posible la aplicación de una teoría abstracta a una situación real, dado que la calidad moral de una acción no solo dependería de criterios objetivos sino de una construcción social, lo que implicaría una comprensión de la información utilizada y la capacidad para decidir qué información sería relevante y cual no. Dada dicha situación, B. C. Stahl argumenta que la única posibi-

lidad sería que la máquina "adquiriese" la personificación de un "agente moral" (*embodiement of a moral agent*), tener en definitiva emociones y la capacidad de aprendizaje a través de capacidades cognitivas para capturar significados (Moor, 2006: 20-21; Stahl, 2004: 71, 78, 80-81; Wallach y Allen, 2011: 104).

Un planteamiento filosófico y social que difiere de un concepto puramente tecnológico. En dicho contexto, el concepto de "autonomía", planteado por A. Etzioni y O. Etzioni, se definiría como "la capacidad de una computadora de seguir un algoritmo complejo en respuesta a los aportes del medio en el que funciona, independientemente de un control humano en tiempo real" y serviría como base a la propuesta de P. Formosa y M. Ryan sobre la definición de un AMA, como "un 'bot' (robot) que puede recoger información del medio (interactividad), hacer juicios éticos independientes (autonomía) y actuar sobre dichos juicios éticos como respuesta a nuevas situaciones (adaptabilidad), sin un control humano en tiempo real". Dichas propuestas, a nuestro entender, irían más encaminadas al concepto de AMAA, que solo formaría parte de una subcategoría dentro de los trabajos de J. H. Moor, W. Wallach y C. Allen (Etzioni y Etzioni, 2016: 149; Ryan, 2020: 2).

Desde nuestro punto de vista, preferimos mantener la concepción inicial de AMA, dado que consideramos que su complejidad necesita de un marco inclusivo tanto filosófico y social como tecnológico. En tal caso, tomando dicha acepción del término, la cuestión radicaría en hasta qué punto no se debería apostar por un planteamiento pragmático y funcional para los AMA, dado que en circunstancias prácticas de ingeniería la intencionalidad y la propia voluntad serían difíciles de abordar, como observan G. Dodig-Crnkovic y D. Persson y que nosotros compartimos. Por lo tanto, desde un punto

de vista pragmático se debería incidir en la responsabilidad moral del grupo y no del individuo (responsabilidad moral colectiva). En tal caso, un AMA se podría describir como un mecanismo artificial de regulación social cuya pretensión sería maximizar aquellas acciones consideradas buenas, mientras que, simultáneamente, se minimizasen las consideradas malas por un grupo social, contextualizándolo en un espacio-tiempo determinado[85].

Así, la responsabilidad moral de los sistemas socio-tecnológicos complejos implicaría una distribución de deberes y responsabilidades de una forma jerarquizada similar a las que se producen en las organizaciones militares y, por lo tanto, un AMA serviría como mecanismo de regulación que aseguraría un comportamiento apropiado del sistema, rol que sería definido externamente, de acuerdo con las normas de un grupo social determinado. Dicha visión pragmática podría incluir el desarrollo de AMA de acuerdo con las clasificaciones de agentes éticos implícitos y explícitos de J. H Moor, así como las de moralidad operativa y de moralidad funcional de W. Wallach y C. Allen. No obstante, independientemente del uso de dichas clasificaciones como base, estaríamos más ante el concepto de AMA como una "entidad moral" que como un "agente moral" preconizado por D. G. Johnson, pudiéndolo considerar, por tanto, como un ejemplo de "herramienta sensible" de F. Fossa, hasta el momento en que dicho AMA alcanzase la autonomía total, en el que se convertiría en un AMAA completo (Dodig-Crnkovic y Persson, 2008: 165-168; Fossa, 2018: 123; Johnson, 2006: 196-197; Moor, 2006: 19-20; Wallach y Allen, 2011: 104).

85 Lo que para A. Etzioni y O. Etzioni sería un "bot ético endeble" (*weak ethic bot*), utilizable como paso intermedio hacia un "Agente Moral Autónomo" (Etzioni y Etzioni, 2016: 152-153).

5.3.- ALGORITMOS Y AMA: CONCEPTUALIZACIÓN Y DESARROLLO ACTUAL

Según W. Wallach, el diseño de los AMA necesitaría de un dialogo exhaustivo entre filósofos e ingenieros en su acercamiento a la ética. Los diseñadores informáticos, ingenieros y programadores han descubierto que simular cualquier actividad compleja requiere de un alto grado de atención al detalle, siendo bastante difícil el construir una simulación cuando existen muchas variables. Así, el propio W. Wallach señala algunas interrogantes con respecto a los AMA: ¿cómo puede un AA reconocer que se está ante una situación en la cual se deben considerar las consideraciones morales en la elección de una acción? o ¿qué comportamiento se debe esperar de una entidad que es capaz de evaluar información a través de criterios consecuencialistas o deontológicos, pero carece de conciencia, emociones, el sentido del ser, capacidades sociales o no está inmerso en el mundo natural? Dentro del campo del Derecho, el distinguir el bien sobre el mal es un requisito para considerar a una persona como un "agente moral" y poder responsabilizar legalmente a un individuo por sus acciones, sin embargo, actualmente no se comprenden completamente los mecanismos cognitivos que permiten el conocimiento o la comprensión. Además, como argumenta S. Torrance, los sistemas controlados informáticamente, aunque sean muy avanzados en sus capacidades cognitivas o de información, es improbable que posean conciencia y, por lo tanto, serían incapaces de ejercer una racionalidad empática que también es un requisito para ser un "agente moral" (Torrance, 2008: 495; Wallach, 2008: 466; 2010: 244-246).

Si se tomasen como base las argumentaciones de W. Wallach o de S. Torrance, entonces sería impracticable el desarrollar un AMA.

Ahora bien, siguiendo con la premisa de una posición pragmática, debemos constatar que, en el mundo real y especialmente en el campo de los SI, operaciones, decisiones y elecciones que realizaban los seres humanos se están delegando en algoritmos que pueden recomendar y a veces decidir cómo se deben interpretar unos datos y que acciones se deben llevar a cabo como resultado. Así, algoritmos de minado de datos, según K. De Vries, pueden ayudar a la comprensión de la gran cantidad de datos de comportamiento generados por el IoT o establecer sistemas de recomendación que dan a los usuarios pautas de qué comprar, que ruta tomar o a quien contactar. Ahora bien, según B. D. Mitelstadt *et al*, el determinar el impacto ético potencial de un algoritmo, especialmente en relación con los AMA, sería también difícil por varias razones: el poder identificar la influencia de la subjetividad humana en el diseño o el determinar que un evento concreto no previsto es simplemente un elemento aislado o un problema sistémico del propio algoritmo. Al incrementarse la complejidad, se incrementaría la distancia entre el diseño, la operación de los algoritmos y la comprensión humana de sus efectos morales y éticos, especialmente cuando las acciones pudiesen llevar a consecuencias letales (Mittelstadt *et al,* 2016: 1-2; De Vries, 2010: 81).

Más aún, como establecen C. Allen *et al,* un punto crucial anterior sería el determinar qué estándares morales deberían aplicarse a los algoritmos que formasen parte de los AMA. Desde un punto de vista filosófico, para los utilitarios clásicos las mejores acciones serían aquellas que produjesen la mayor felicidad en el mayor número de personas. En tal caso, existiría un sentido claro de lo que es "moralmente bueno", que se podría aplicar a las acciones de un agente independientemente de cómo dicho agente decidió dicha acción. Por el contrario, el imperativo categórico deontológico de

Kant consideraría que existe un mandato moral interno incondicional e inherente a la naturaleza humana. Por lo tanto, para que una acción fuese considerada como "moralmente buena", debería llevarse a cabo teniendo como base el imperativo categórico. Así, siguiendo la argumentación de M. G. Singer al establecer el "Principio de Generalización", se podría alcanzar eficazmente un objetivo en un mundo en que todos decidiesen seguir dicho objetivo actuando de la misma forma en circunstancias similares[86] (Allen *et al,* 2000: 253; Knopf, 1962: 18; Singer, 1971).

En todo caso, desde una perspectiva social, no resulta evidente,- como argumenta R. Capurro, como se podría establecer dicha dimensión de los algoritmos, Aunque existe una definición técnica clásica de algoritmo[87], propuesta por D. E. Knuth, los mismos están diseñados explícitamente dentro de un marco de costumbres sociales basadas en la interacción y existen notables diferencias entre las interacciones humanas y la de los actores no humanos. Por lo tanto, el desafío ético y jurídico que representan debería ser enfocado sobre la aculturación explícita de los mismos (*enculturing algorithms*), poniendo atención al contexto y determinar: el que, para que, por quién y para quién son creados y utilizados. El comprender los algoritmos como prácticas culturales significaría también el tener que reflexionar sobre las instituciones, valores y normas en las que están integrados. Por lo tanto, los algoritmos no serían neutrales porque fuesen lógicos o ejecutados por una

86 Por ejemplo, desarrollando AMA que implementasen diversos elementos de las normas morales existentes del DIH, instrumento jurídico aceptado mayoritariamente por los Estados.

87 Un "algoritmo" es un conjunto finito de reglas que proporciona una secuencia de operaciones para resolver un problema específico y que tiene cinco características principales: es finito, está definido, tiene una entrada, una salida y es eficaz (Capurro, 2019: 131).

máquina, sino que, dependiendo del contexto cultural en donde fueron creados, se podrían producir inconsistencias de acción para un mismo tipo de objetivo, tanto por una carencia de normas jurídicas, por normas jurídicas distintas dependiendo del contexto social donde se desarrollasen o por la forma en que los diseñadores regulasen el flujo de información a través de ellos. (Capurro, 2019: 132-134).

Por otro lado, desde un punto de vista social práctico, diversas instituciones han intentado elaborar una serie de criterios orientadores de gobernanza ética para la IA, que serían aplicables a los algoritmos integrados en dichos sistemas. Entre ellas, la "*IEEE Standars Association*" establece los siguientes criterios: DD.HH.; dar prioridad al bienestar; rendición de cuentas; transparencia; uso indebido de la tecnología y concienciación de ello. Para la "*Association of Computer Machinery (ACM)*" serían: concienciación; acceso y reparación; rendición de cuentas; explicación; procedencia de los datos; auditoría; validación y testeo. En cuanto a la "*UNI Global Union*": demandar que los sistemas de IA sean transparentes; equipar los sistemas de IA con una "caja negra ética"; hacer que la IA sirva a la gente y al planeta; adoptar un enfoque del "humano al mando"; asumir una IA no discriminatoria por razón de sexo; compartir los beneficios de los sistemas de IA; garantizar una transición justa y garantizar el apoyo para los derechos y libertades fundamentales; establecer mecanismos de gobernanza global; prohibir la atribución de la responsabilidad a los robots; prohibir la carrera armamentística de la IA.

Como se podrá observar, al igual que argumentan J. Whittlestone *et al,* existen una serie de solapamientos entre los principios con un amplio consenso sobre que: las tecnologías basadas en algorit-

mos y datos de la IA deberían ser utilizados para el bien común, no deberían dañar a las personas o socavar sus derechos y respetar algunos de los valores más aceptados como: la equidad y la justicia; la privacidad y; la autonomía[88]. En todo caso, al ser una mezcla de valores colectivos e individuales, en algún momento podrían entrar en conflicto entre ellos provocando, como también argumentan J. Whittlestone *et al*, tensiones que los propios principios no abordan, siendo un desafío para los diseñadores de dichos algoritmos (IEEE, 2019: 17-35; UNI, 2017; USACM, 2017; Whittlestone *et al*, 2019: 11-12, 53-54).

Dicha situación haría más urgente el establecimiento de una gobernanza ética pues, como argumentan A. F. T. Winfield y M. Jirotka, que nosotros subscribimos, sería necesario inculcar comportamientos éticos tanto en los diseñadores como en las organizaciones que los despliegan. Una gobernanza ética normativa que tratase los asuntos éticos antes de que surgiesen y no caso por caso. A nivel europeo, dicha premisa se ha plasmado en la "*Declaración de Roma sobre la Innovación e Investigación Responsable (IIR)*", de 2014, en la que se establece que las decisiones sobre innovación e investigación deberían estar alineadas con los principios fundadores de la Unión Europea (UE): el respeto de la dignidad humana; libertad; democracia; igualdad; el estado de derecho y; el respeto de los DD.HH. En todo caso, según A. F. T. Winfield y M. Jirotka sería necesario el establecimiento de una reglamentación supervisada por estamentos reguladores para generar la confianza de la sociedad. El siguiente esquema resumiría dicho marco (ver fig. 12)

88 Para más información ver: BODEN, M. *et al,* (2017): "Principles of robotics; regulating robots in the real world, *Connection Science,* 29(2), 124-129, acceso febrero 2021, en https://www.tandfonline.com/doi/pdf/10.1080/09540091.2016.1271400?needAccess=true

(CUE, 2014; Winfield y Jirotka, 2018: 2, 5-6):

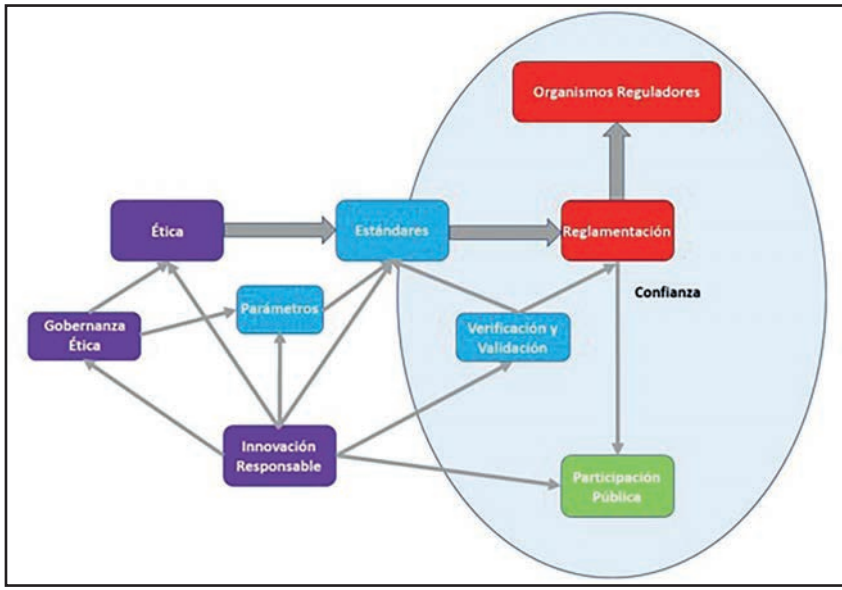


Figura 12: Esquema de IIR para construir la confianza pública. Adaptado de A.F.T. Winfield y M. Jirotka (Winfield y Jirotka, 2018: 6)

Dicha postura, sin embargo, no representa a la mayor parte de los investigadores. Por el contrario, la idea principal sería la aplicación de instrumentos de derecho indicativo no vinculante (*soft law*) en el desarrollo de los algoritmos. Esa sería la postura argumentada por G. A Marchant, B. Allenby así como por W. Wallach, dado que, debido a los rápidos cambios en la IA y la tecnología, cualquier sistema reglamentario nuevo estaría obsoleto rápidamente, por lo que el enfoque tradicional reglamentario no produciría ningún reglamento o un reglamento malo. Por otro lado, los instrumentos actuales de Derecho Internacional son más débiles que el Derecho

de los Estados y las implicaciones militares y de seguridad de muchas de las tecnologías emergentes harían que una harmonización normativa internacional fuese extremadamente difícil, como podemos observar en los trabajos de las NU sobre los SAAL que hemos analizado anteriormente o el desarrollo de un tratado internacional sobre ciberseguridad. En todo caso, según dichos investigadores, sería importante conseguir una "harmonización" de dichos instrumentos específicos para cada tipo de tecnología a través de estándares privados, orientaciones o códigos de conducta. Nuestro punto de vista sería, sin embargo, complementario: utilizar el Derecho existente, Internacional y de los Estados, allí donde fuese posible adaptándolo a las nuevas tecnologías, desarrollando al mismo tiempo instrumentos no vinculantes, harmonizados internacionalmente, en aquellas instancias donde actualmente existen dificultadas para un desarrollo normativo (Marchant y Allenby, 2017: 109-113; Wallach y Marchant, 2019: 505).

Otra posibilidad complementaría sería el desarrollo de algoritmos para los AMA que tuviesen una autonomía ajustable. Esto es, implementar mecanismos que permitiesen a los humanos, compartir, supervisar e intervenir en su control, cuando dichos algoritmos fuesen incapaces de lidiar con situaciones complejas. Permitiría, según argumentan Cervantes *et al*, el dotar a los AMA con la flexibilidad y fiabilidad necesarias para maximizar el rendimiento del algoritmo. Así, se transferiría la decisión a los humanos en situaciones imprevisibles de incertidumbre críticas, con el objetivo de mantener el control global humano sobre los AMA, para evitar comportamientos autónomos inapropiados o no deseados[89]. A tal

89 Para más información ver: MOSTAFA, S. A., AHMAD, M. S. y MUSTAPHA, A. (2019): "Adjustable autonomy: A systematic literature review", *Artificial Intelligence Review,* 51(2), 149–186.

fin S. Zieba *et al* propusieron un marco para establecer la metodología de la autonomía ajustable, que se reflejaría en los algoritmos correspondientes (ver fig. 13) (Cervantes *et al,* 2020a: 502; Zieba *et al,* 2010: 202):

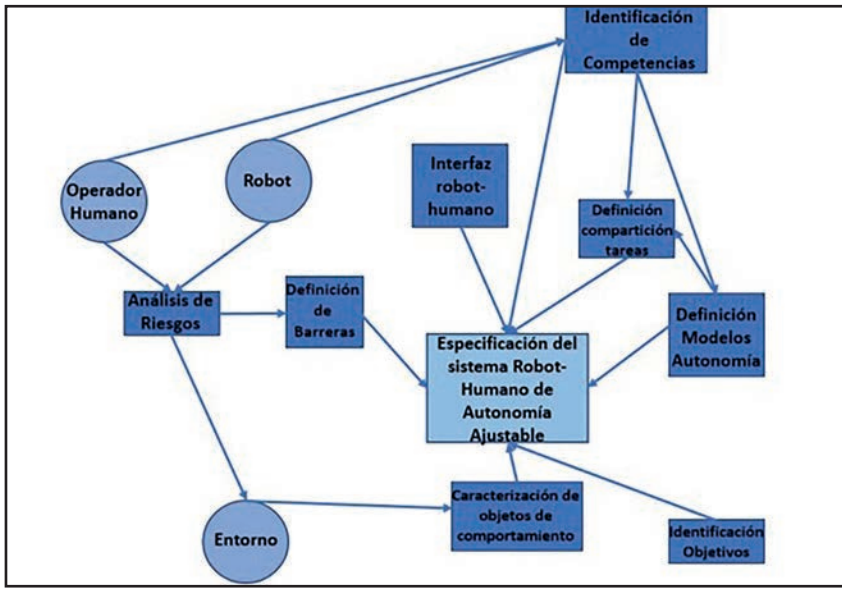


Figura 13: Esquema de marco metodológico de Autonomía Ajustable. Adaptado de Zieba *et al* (Zieba *et al*, 2010: 202).

Dicha metodología estaría compuesta de una serie de pasos (Zieba *et al,* 2010: 202):

- La identificación de los objetivos que se pretenden alcanzar por el sistema robot-humano;
- La identificación de las competencias de cada agente;

- La caracterización del entorno y de los objetos para identificar los comportamientos potenciales;
- Un análisis de riesgos identificando los errores potenciales de cada agente;
- La definición de barreras para limitar las oportunidades y las consecuencias de dichos errores;
- La definición de los diversos modelos de autonomía;
- La definición de las tareas a compartir entre el operador humano y el robot con relación a los pasos previos y por cada modelo de autonomía;
- La especificación de la interfaz robot-humano que asistirá en la comunicación entre los objetos y sus comportamientos.

En la actualidad, el desarrollo de los AMA en la IA se ha centrado en abordar los problemas éticos basándose en dilemas morales. Contextualizando, existirían dos situaciones no-exclusivas donde se podrían presentar conflictos morales: dentro del mismo agente cuando dos de sus normas entrasen en conflicto y; entre agentes que tuviesen diferentes formas de razonar sobre lo que es ético o no. En esta segunda situación existiría una doble interacción agente-agente y agente-humano. Dicho lo cual, los dilemas podrían ser clasificados en (Cervantes *et al,* 2020a: 507):

- *Dilemas de obligación*: Basándose en las reglas éticas del AMA, todas las acciones posibles serían obligatorias pero un AMA no podría elegir y ejecutar más de una acción;
- *Dilemas de prohibición*: Basándose en las reglas éticas del AMA, todas las acciones posibles estarían prohibidas, pero el AMA necesitaría elegir una.

Para llevar a cabo el diseño de dichos AMA se podrían utilizar distintos planteamientos, que C. Allen *et al* categorizaron dependiendo del mecanismo utilizado (Allen, *et al*, 2005: 149-154; Cervantes *et al*, 2020a: 506):

- Enfoque de arriba-abajo (*top-down approach*): Establecimiento de una serie de reglas basadas en teorías éticas como el utilitarismo o la deontología de Kant. Serían modelos basados en "mandamientos", con reglas específicas para cada tipo de comportamiento ético. Su mayor problema radica en que dichas reglas a menudo podrían entrar en conflicto;
- Enfoque de abajo-arriba (*bottom-up approach*): No se impone ninguna teoría ética como parte del proceso de decisión. Se establecen mecanismos de aprendizaje basados en la experiencia y en valores intrínsecos para guiar su comportamiento y presupone que los agentes pueden desarrollar sus propios juicios morales. Son difíciles de desarrollar y de evolucionar, especialmente con relación a la estructura moral "profunda", una especie de "gramática moral" básí-

ca del sistema;

- Enfoque híbrido (*hybrid approach*): Integración de diversas influencias incluidas valores adquiridos de las teorías éticas (*top-down*) unidos a una "gramática moral" como base estructural y amplios mecanismos de aprendizaje basados en la experiencia.

En la actualidad existen modelos computacionales para cualquier tipo de categorización establecido, algunos de los ejemplos son los siguientes (Cervantes *et al*, 2020a: 512-523):

- Enfoques de arriba-abajo: MoralDm[90]; Jeremy[91]; máquina de consecuencias (*consequence engine*)[92];

90 Intenta imitar el proceso humano de toma de decisiones, con comportamiento tanto utilitario como deontológico dependiendo del problema. Para más información ver: DEHGHANI, M. TOMAI, E., FORBUS, K. D. y KLENK, M. (2008): "An integrated reasoning approach to moral decision-making", *Twenty-third AAAI conference on artificial intelligence*, 1280–1286.

91 Utilizan "*machine learning*" para resolver dilemas éticos. Sistema basado en la teoría utilitaria. Para más información ver: ANDERSON, M., y ANDERSON, S. L. (2008): "Ethical healthcare agents", en M. Sordo *et al* (eds.), *Advanced computational intelligence paradigms in healthcare-3,* Springer, Berlin, 233–257.

92 Modelo interno de simulación de acciones implementado en robots para predecir sus consecuencias y así orientar acciones futuras. Para más información ver: VANDERELST, D., y WINFIELD, A. (2018): "An architecture for ethical robots inspired by the simulation theory of cognition", *Cognitive Systems Research,* 48*,* 56–66.

- Enfoque abajo-arriba: Casuist BDI-Agent[93]; GenEth[94]
- Enfoque híbrido: LIDA[95]; modelo de toma de decisiones ético[96].

Aparte de dichos modelos computacionales, también se han desarrollado metodologías para evaluar las reglas éticas implementadas en los AMA como: "GDT4MAS", que prueba que las reglas éticas y morales se han expresado como propiedades invariables o "Athena", un sistema interactivo para aprobar teoremas[97] (Cervantes *et*

93 Modelo de arquitectura que combina el razonamiento basado en casos con los modelos de agente BDI (*belief-desire-intention* – creencia-deseo-intención). Para más información ver: HONARVAR, A. R., y GHASEM-AGHAEE, N. (2009): "Casuist BDI-agent: A new extended BDI architecture with the capability of ethical reasoning", *International conference on artificial intelligence and computational intelligence*, Springer, Berlin, 86–95.

94 Analizador general de dilemas éticos. El conocimiento está basado en conceptos éticos: deberes, acciones, casos y principios. Para más información ver: ANDERSON, M., y ANDERSON, S. L. (2014): "Geneth: A general ethical dilemma analyzer", *Twenty-eighth AAAI conference on artificial intelligence*, 253–261.

95 Arquitectura general cognitiva. Implementa teorías éticas a través de reglas y mecanismos de aprendizaje. Se ha implementado parcialmente en el robot asistencial CareBot. Para más información ver: MADL, T., y FRANKLIN, S. (2015): "Constrained incrementalist moral decision making for a biologically inspired cognitive architecture", en R. Trappl (ed.), *A construction manual for robots' ethical systems*, Springer, Cham, 137–153.

96 Modelo computacional basado en la neuro ciencia, diseñado para dotar a los agentes autónomos de mecanismos de decisión éticos. Para más información ver: CERVANTES, J. A., RODRÍGUEZ, L. F., LÓPEZ, S., RAMOS, F., y ROBLES, F. (2016): "Autonomous agents and ethical decision-making" *Cognitive Computation,* 8(2), 278–296.

97 Para más información ver: MERMET, B., y SIMON, G. (2016): "Formal verification of ethical properties in multiagent systems", *ECAI 2016 workshop on ethics in the design of intelligent agents (EDIA'16).* La Haya y; ARKOUDAS, K., BRINGSJORD, S., y BELLO, P. (2005): "Toward ethical robots via mechanized deontic logic", *AAAI Fall symposium on machine ethics*, 17-23.

al, 2020a: 524).

5.4.- REFLEXIONES SOBRE EL DERECHO Y LOS AMA

La moralidad y el Derecho van unidos, pero según argumenta L. B. Eliot, no se ha considerado suficientemente el rol dual entre ambos, existiendo diversas ideas sobre dicha dualidad. Para algunos investigadores, en el marco del derecho natural, la moralidad sería la fuente de las leyes y serviría como poder vinculante de las mismas. Para los investigadores del derecho positivista, la moralidad estaría categóricamente separada del Derecho.

En todo caso existirían, según Kagan, tensiones entre ambas dado que: "la ley podría permitir algún acto particular, aunque el acto fuese inmoral y podría prohibir un acto, aunque fuese moralmente permisible o incluso necesario moralmente". En dicho contexto, el concepto de "agencia legal" (*legal agency*) estaría basado, según argumentan S. Chopra y L. F. White, en la máxima latina "*Qui facet per alium, facit per se*" (aquel que actúa a través de otro se considera en Derecho como que lo hubiese realizado el mismo), para lo cual se necesitaría el consentimiento del agente para actuar como representante de otra persona (Chopra y White, 2011: 18; Eliot, 2020: 1; Kagan, 1998).

Por lo tanto, uno de los aspectos más importantes dentro del marco de los AMA sería la hipotética acepción de personalidad que se le pudiese atribuir a un AA para poder emitir un consentimiento, actuar como un AMAA y, por lo tanto, con derechos y obligaciones atribuibles a una persona jurídica. En el mundo real, no constituiría una excepción única ya que, por ejemplo, las corporaciones modernas también la poseen. En tal caso, según indican S. S. Chopra

y L. F. White, el "agente" sería un sistema computacional complejo que, individualmente o formando parte de un sistema multi agente, representaría a humanos, a corporaciones o a estamentos públicos. Dicha posibilidad es puesta en cuestión por B. Brozek y M. Jakubiek, pues mientras la responsabilidad jurídica de las corporaciones e instituciones estarían conectadas con las acciones llevadas a cabo por sus representantes o sus trabajadores, las acciones de un AMAA no serían directamente rastreables hacia las acciones de un agente humano. Dado que detrás de una personalidad jurídica siempre hay un humano, para que un AMAA fuese considerado como una personalidad jurídica, tendría que ser un verdadero actor en el marco de la interacción social (Brozek y Jakubiek, 2017: 300-302; Chopra y White, 2004)

En la actualidad no se le ha otorgado personalidad jurídica a ningún AMA, por lo que dicha responsabilidad recaería en alguno de los actores humanos involucrados en los mismos. Ahora bien, existe una compleja interacción entre los diversos actores humanos y los artefactos. El siguiente esquema de valor ilustra dicha complejidad desde el desarrollo de un algoritmo hasta su uso (ver fig. 14) (Fritz *et al,* 2020: 8):

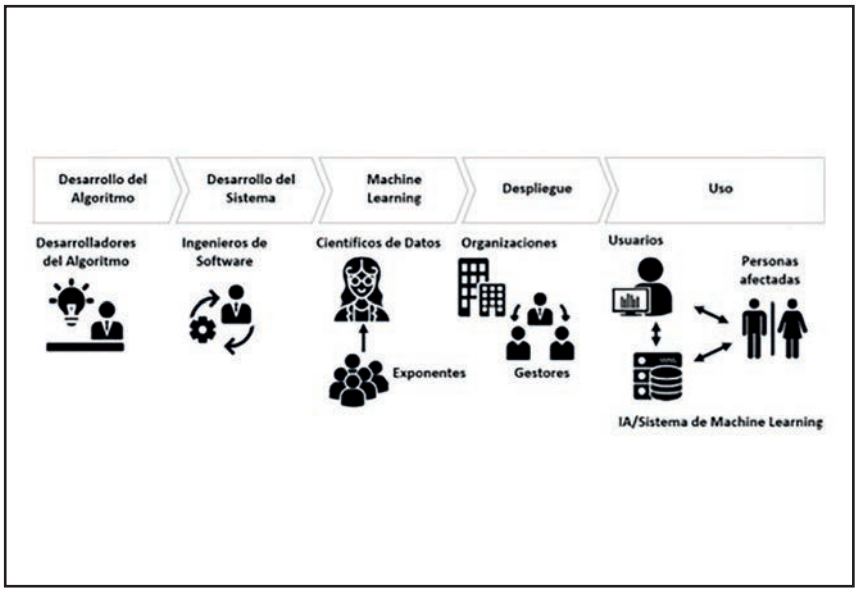


Figura 14: Esquema abstracto de una cadena de valor desde el desarrollo del algoritmo hasta su uso. Adaptado de Fritz *et al* (Fritz et al, 2020: 8).

Según Fritz *et al,* existirían ocho posibles actores humanos responsables: el sistema técnico de IA, aunque fuese un artefacto; los usuarios obligados a utilizar el sistema que no entenderían; los gestores que ni comprenden la "caja negra" ni toman decisiones individualmente; la organización: los científicos de datos, aunque no tomasen decisiones sobre personas individuales; las personas que proveen de los datos de aprendizaje, muchas veces sin saberlo; los ingenieros de software, aunque fuesen incapaces de prever el comportamiento del sistema después del aprendizaje o; los desarrolladores del algoritmo que crearon las "cajas negras" multiuso en primer lugar (Fritz *et al,* 2020: 8).

Para analizar cada una de las posibilidades, un primer concepto a tener en cuenta sería el de "caja negra". Según D. Card, la metáfora

de la "caja negra" proviene de los primeros días de la cibernética y el conductismo y se referiría a un sistema del cual solo se pueden observar las entradas y salidas, pero no su funcionamiento interno. La incapacidad de comprender el proceso de toma de decisiones de un AMA tendría un impacto profundo cuando se llevasen a cabo pruebas sobre su intencionalidad o causalidad.

Para Y. Bathaee, si un algoritmo de IA fuese una "caja negra", tampoco se conocería las intenciones o las conductas de los humanos que lo crearon o lo desplegaron, pues posiblemente ellos mismos serían incapaces de prever que soluciones y/o decisiones tomaría el algoritmo. Esto llevaría a la necesidad de establecer pruebas a medida, dependiendo del nivel de transparencia del algoritmo y de la supervisión humana. Ahora bien, el propio Y. Bathaee, que nosotros compartimos, cuestiona la hipótesis de que la transparencia se podría mejorar en un futuro, ya que la complejidad de los algoritmos basados, por ejemplo, en redes neuronales artificiales, así como los problemas de dimensionalidad, serían cada vez mayores con los rápidos avances de la IA.

Tampoco sería adecuada una responsabilidad estricta (*strict liability*), pues si un desarrollador o un usuario fuese incapaz de predecir sus efectos *ex ante*, no podría tomar precauciones sobre el daño infligido. Por lo tanto, Y Bathee propone que aquellos algoritmos supervisados por humanos ofrecerían menos problemas en las pruebas de intención y causalidad, mientras que los algoritmos "autónomos", dentro de un AMAA necesitarían de esquemas de responsabilidad basados en el concepto de "negligencia". En el caso que un AMA pudiese operar de forma autónoma, el desarrollador, gestor o usuario de dicho algoritmo deberían ser considerados responsables de cualquier negligencia en su despliegue

o uso, propuesta que también nosotros podríamos compartir, dependiendo del contexto de desarrollo y uso. Un planteamiento que consideramos abierto y que continuaremos con su análisis dentro del marco de los SAA y los SAAL y su impacto sobre el DIH y los DD.HH. (Asaro, 2016: 191-192; Bathaee, 2018: 892-894, 896, 932-933; Card, 2017).

Concluiremos que, desde nuestro punto de vista, un algoritmo representativo de un AMA sería una herramienta social y no un "agente" en sentido estricto, ya que la única forma de serlo sería poder pasar un "Test de Turing Moral", donde no se podría identificar la diferencia entre máquina y humano. Para ello necesitaría de un concepto de "ser", una conciencia y elegir libremente sus propias acciones, a través de un conocimiento de la diferencia entre el bien y el mal, difícil de abordar si dicho algoritmo no alcanzase la Singularidad.

Por lo tanto, para los SAA y los SAAL, definiríamos un AMA como un mecanismo artificial de replicación social, que maximizaría lo bueno frente a lo malo, pero siempre dentro de un contexto en un espacio-tiempo determinado, que dependería del exterior a través de una normas de grupo, pero también de la situación real operativa en un determinado momento. El AMA sería una "entidad" y no un "agente", que dependería de un proceso de aculturación (valores, normas e instituciones) y que necesitaría de una gobernanza ética por parte de diseñadores y organizaciones responsables de su despliegue. En el caso de los SAAL, dicha aculturación creemos que debería surgir de los principios rectores desarrollados y consensuados por el GGE del CCW para los SAAL de las NU. Es decir, la utilización para el algoritmo de dicho AMA de una base de Derecho consuetudinario (*soft law*) consensuado internacional-

mente, ya analizado en el capítulo anterior.

Además, nuestra valoración apuntaría al desarrollo de una metodología de autonomía adaptable para el AMA de un SAAL, variable en el espacio-tiempo, a través de un enfoque híbrido, en un proceso de "arriba-abajo" (*top-down*), basado en la ética deontológica (valores adquiridos) y expresado por el DIH. Se complementaría, por un mecanismo de aprendizaje "abajo-arriba" (*bottom-up*) basado en la experiencia, que podría surgir del campo operacional, a través de las ROE. Dicho proceso estaría sujeto a una complejidad creciente, debido a la propia complejidad del Derecho Internacional y la infinidad de situaciones operacionales posibles, pero también por el uso creciente de la IA, como las RNA. Al mismo tiempo, necesitaría de un esquema de responsabilidad basado en la "negligencia", donde el desarrollador, gestor o usuario de dicho AMA serían considerados responsables de cualquier negligencia en su despliegue o uso, dependiendo del contexto de cada situación, lo que necesitaría de un diseño algorítmico apropiado que analizaremos en profundidad más adelante en esta investigación.

PARTE III

SUPERINTELIGENCIA ARTIFICIAL Y AGENTES MORALES ARTIFICIALES

CAPÍTULO 6

LA PARADOJA DE LA SINGULARIDAD

En 1965 el investigador I. J. Good describió la explosión de la inteligencia a través de las máquinas, donde con cada generación se alcanzarían mayores niveles de conocimiento. Una explosión de la inteligencia generalmente conocida como "Singularidad"[98], donde una máquina ultra inteligente se definiría como aquella que pudiese sobrepasar enormemente todas las actividades intelectuales de cualquier ser humano por inteligente que fuese. Paralelamente M. L. Minsky establecería que en el momento en que existiesen programas que pudiesen automejorarse se establecería un proceso evolutivo donde la máquina se mejoraría autónomamente, observándose en la misma elementos de consciencia, intuición e inteligencia (Chalmers, 2010: 7; Good, 1965: 33; Minsky, 1966: 257).

Más aún, ya en 1993 el informático V. Vinge predijo que en treinta años la Humanidad tendría la capacidad tecnológica para crear una inteligencia sobrehumana y poco después la era humana terminaría. Una capacidad que vendría dada por mejoras recursivas de la tecnología, que para D. Roden supondría un "evento trascendental", la emergencia de mentes post humanas tan vastas que la Humani-

98 El término "Singularidad" proviene de J. V. Neumann, según indica S. Ulam en el obituario que escribió a su muerte, donde recuerda una conversación entre ambos que se: "centró en el progreso acelerado de la tecnología y los cambios en los modos de vida humana, que da la sensación de estar acercándose a alguna forma de <*singularidad*> esencial en la historia de la raza humana, a partir de la cual los asuntos humanos, como los conocemos, no continuarían" (Ulam, 1958: 5).

dad carecería de modelos para vislumbrar su potencial transformador. Evento que, de acuerdo con la "*Seis Épocas de la Evolución*" de R. Kurzweil, tendría lugar en la "Quinta Época", con la fusión entre la tecnología humana con la inteligencia humana, donde se trascendería las limitaciones del cerebro humano dentro de una civilización "hombre-máquina" (Kurzweil, 2005: 28-34; Roden, 2013: 281-282; Vinge 1993). En dicho contexto y basándose en la "ley de Solomonoff", utilizando las velocidades de computación actuales, el tiempo de duplicación actual y una estimación del poder de procesamiento del cerebro humano, S. Yudkowsky estimó que se llegaría a alcanzar la Singularidad en el año 2021 (Hutter, 2012: 148; Solomonoff, 1985: 149-153; Yudkowsky, 1996).

En todo caso, como base subyacente de la IGA y la Singularidad, estaría la noción de que la tecnología creada por el ser humano forma parte de la Ley de Retornos Acelerados (*Law of Accelerated Returns – LOAR*), donde la tecnología evoluciona exponencialmente hacia la denominada Super Inteligencia Artificial (SIA) a través de la Explosión de la Inteligencia, la capacidad de que una máquina ultra inteligente llegue a diseñar otras máquinas mejores que aceleren la llegada de la SIA (Pohl, 2015: 2).

6.1.- LA COMPLEJIDAD DE LA MENTE HUMANA

Pero antes de comenzar el análisis sobre la paradoja de la Singularidad y comprobar los aciertos de las anteriores predicciones, debemos plantearnos la pregunta: ¿Qué es la mente? Para contestarla sería necesario saber qué es lo que nos hace humanos. Los investigadores D. R. Hofstadter y D. C. Dennett proponen que la respuesta es obvia: somos centros de consciencia y las cosas de

lo que somos conscientes determinan lo que somos (Hofstadter y Dennett, 1982: 10-11). Ahora bien, la determinación del concepto de "consciencia" es esquivo. Desde una perspectiva general, se puede considerar que la consciencia es la experiencia interna en primera persona de la mente, aunque, según indica A. L. Nelson, esto no resolvería su definición desde un punto de vista científico. Es más, la experiencia de una mente consciente, al ser un fenómeno personal, puede que no permitiese una verificación científica (Nelson, 2013: 52).

Es importante distinguir entre dos términos: la consciencia (*conscience*) y la concienciación (*consciousness*), aunque a veces se entremezclen. La consciencia se describe generalmente en un sentido moral, la capacidad inherente de un ser humano para percibir lo que está bien o mal, que se forma de acuerdo con el entorno cultural, político y económico de un individuo (Swan y Vallier, 2012: 1; Vithoulkas y Muresano, 2013: 105). La concienciación es la función de la mente humana que recibe y procesa la información, la cristaliza y luego la adopta o la rechaza ayudado por una serie de elementos: los cinco sentidos; la capacidad de razonamiento de la mente; la imaginación; la emoción y; la memoria (Vithoulkas y Muresano, 2013: 104-105).

De forma generalizada se desarrolla un proceso complejo (información-concienciación-percepción-consciencia), continuo e integrado y si alguna de las partes deja de existir o es defectuosa todo el sistema sufrirá o se colapsará. Esto da idea, según Vithoulkas y Muresano, de la coherencia, plenitud y continuidad de la estructura del cerebro humano que, aunque seamos capaces teóricamente de separar las diversas funciones, funcionan como un todo sistémico con absoluta interdependencia (Vithoulkas y Muresano, 2013:

105-106).

Ahora bien, el proceso de consciencia y concienciación va ligado estrechamente al concepto de inteligencia, pero si deseamos establecer la super inteligencia debemos primero establecer la definición de inteligencia. Desgraciadamente, no existe ninguna definición estándar, existiendo más de setenta definiciones establecidas por los distintos campos del conocimiento. Los investigadores S. Legg y M. Hutter han utilizado una serie de atributos comunes de dichas definiciones para establecer su propia definición: "la inteligencia mide la habilidad de un agente para alcanzar sus metas en un amplio número de entornos" (Legg y Hutter: 2007: 17-24).

Como corolario, en el campo de la inteligencia de las máquinas existen dos nociones, aquellas basadas en el rendimiento en entornos complejos y otras basadas en el contenido de la información, aunque la mayoría de las definiciones incluyen a ambas. Las siguientes definiciones han sido propuestas (Kurzweil, 2000: 61; Poole, Mackworth y Goebel, 1998: 1-7; Wang, 1995: 5):

> "Un agente inteligente realiza lo que es apropiado para sus circunstancias y metas, es flexible a los entornos y metas cambiantes, aprende de la experiencia y toma decisiones apropiadas dentro de sus limitaciones perceptivas y la computación finita" (D. Poole)

> "La inteligencia es la capacidad de usar de una manera óptima los recursos limitados – incluyendo el tiempo – para alcanzar las metas" (R. Kurzweil)

> "La inteligencia es la capacidad de un agente procesador de información para adaptarse a su entorno con conocimiento y recursos insuficientes" (P. Wang).

Dichas definiciones y otras han llevado a S. Legg y M. Hutter a proponer una ecuación de "Inteligencia Universal" donde la inteligencia de un agente π se puede definir con la ecuación:

$$\gamma(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi$$

Donde μ es un entorno dentro de todo el conjunto E de todos los entornos computables incentivados, K(.) es la complejidad de Kolmogorov[99] y V_μ^π es la suma de incentivos futuros esperados cuando el agente . En dicho contexto el agente más inteligente sería aquel que convergiese al rendimiento óptimo en cualquier entorno donde fuese posible para un agente general. Dicha ecuación serviría tanto para agentes reactivos simples como para agentes universalmente óptimos. Es una medida continua de rendimiento y por tanto da más información de progreso. Hay que tener en cuenta que dicha "Inteligencia Universal" está basada en fundamentos

99 La Complejidad de Kolmogorov de un string $s \in \Sigma*$ es el largo mínimo de un programa para una Máquina de Turing Universal U con entrada $\in \{0, 1\}$ que genera s y se detiene. Trae como costo que no exista un algoritmo que, dado un objeto, prediga el tamaño del programa más corto que genere el objeto; es decir, su complejidad algorítmica, lo que significa que no se puede medir con absoluta certeza la complejidad algorítmica de un objeto, pero se le puede aproximar encontrando programas cortos que generen el objeto en cuestión (Zenil, 2013:62).

matemáticos y de computación y no en aspectos humanos (Legg y Veness, 2013). Dicha "Inteligencia Universal" se acercaría, por tanto, al concepto de IGA y de alcanzar la Singularidad, pero aún no estaríamos ante el concepto de "Super Inteligencia Artificial" (SIA), el nivel de IA alcanzable cuando una computadora sobrepasase exponencialmente el nivel de inteligencia de un humano en varias órdenes de magnitud.

6.2.- EL CLONADO DE LA MENTE HUMANA

El escenario hipotético de la Singularidad tecnológica puede acontecer a través de una explosión de la inteligencia, una explosión de la velocidad o una combinación de ambas. En todo caso, existen diversos caminos que se están explorando para llegar a la Singularidad. La mayor parte se basan en una combinación de inteligencia del "software" con un "hardware" cada vez más potente. Tres de los más importantes son: el razonamiento basado en el conocimiento y software de planificación (investigación de IA tradicional); la creación de agentes artificiales que aprenden de la experiencia (enfoque de aprendizaje automático) y; sistemas inteligentes de evolución automática (enfoque en algoritmos genéticos y vida artificial (ALife)) (Hutter, 2012: 144).

Cuando se hace referencia a modelos computacionales software-hardware de las emociones se está hablando de las tentativas para desarrollar y validar modelos computacionales de los mecanismos de las emociones humanas. Objetivos paralelos a los objetivos generales de la IA cuando se restringen al campo de las

emociones. En el contexto de la investigación de la IA tradicional, el desarrollo de sistemas artificiales que tienen en cuenta consideraciones morales, están confinadas, en general, al desarrollo de sistemas con "moralidad operativa", aquella que asegura que el sistema de IA funcionaría según su diseño. Es una extensión de la preocupación tradicional de la ingeniería por la seguridad en el diseño de máquinas que realizan una tarea específica. Existen diferentes tipos de proyectos, los denominados "teóricos" que intentan establecer una mejor comprensión de los mecanismos de las emociones y los denominados "aplicados", aquellos que intentan enriquecer las arquitecturas de los agentes artificiales dotándoles de mecanismos de emociones similares a los humanos (Reisenzen *et al,* 2013: 246; Wallach *et al,* 2008: 570).

Dichos modelos artificiales son utilizados para el desarrollo de los AMA. Según los investigadores W. Wallach, C. Allen y I. Smit, la implementación de capacidades morales en la IA es una extensión necesaria para los mecanismos sociales de los agentes de software autónomos (*autonomous software agents*). En el desarrollo de dichos agentes, es necesario establecer que papel representa la ética en su desarrollo. Existen dos caminos de desarrollo de dichas arquitecturas: un enfoque de imposición de las teorías éticas de arriba hacia abajo y otro cuyo objetivo es el crear sistemas o estándares que pueden o no estar especificados en términos teóricos explícitos. Para C. Allen, G. Varner y J. Zinser, el desarrollo de los AMA es la tarea más importante y desafiante a la que se enfrentan los sistemas autónomos y como indica R. Picard del MIT. "a mayor libertad de una máquina mayor será la necesidad de estándares morales" (Allen *et al*, 2000: 251-261; Picard, 1997; Wallach *et al,*

2008: 565-566).

La cuestión central es como implementar dichas capacidades morales, que tipos de decisiones pueden ser desarrolladas por sistemas computacionales dentro de los límites actuales de la tecnología y las perspectivas futuras. La respuesta permitiría discernir que tipo de sistema moral es factible para los sistemas artificiales. En dicho contexto, se desglosan y dividen las capacidades de decisión moral de los humanos en módulos y componentes computacionales manejables. La metodología de arriba hacia abajo tomaría como base una teoría ética y analizaría los requerimientos computacionales para diseñar algoritmos y subsistemas capaces de implementarla. En contraste, siguiendo un enfoque de abajo hacia arriba, suponiendo que existiese una teoría de base, solo se utilizaría como una forma de especificar una tarea a un sistema, pero no como un método de implementación de estructura de control de dicha teoría. Ahora bien, en dicha estructura no sería necesario que existiese una teoría de base ya que se puede especificar la tarea basándose en una o varias medidas de rendimiento. A través de la metodología de ensayo y error se puede afinar el rendimiento de los sistemas para acercarse o sobrepasar los criterios de rendimiento establecidos. En la práctica, no obstante, se utilizan ambas estructuras conjuntamente, ensamblando componentes que desarrollan tareas específicas pero guiadas por un análisis de arriba hacia abajo teórico que normalmente es incompleto (Wallach *et al*, 2008: 568-569).

El investigador M. Hutter plantea la creación de una sociedad virtual basada en programas informáticos (*software*) que consistiría en la interacción de agentes racionales cuya inteligencia sería tan

alta que les permitiría construir una nueva generación de agentes racionales más inteligentes. Visto desde el exterior del proceso, los humanos establecerían los planos iniciales para crear nuevas máquinas que una vez creadas producirían a su vez mejores máquinas a un ritmo acelerado. Los humanos no partícipes tendrían cada vez un rol menor en dicho proceso debido a su límites cognitivos y de velocidad hasta el punto en que se convertirían en simples espectadores del proceso. Después de un periodo corto de tiempo, la interacción inteligente entre máquinas y humanos se tornaría imposible, creándose un agujero negro entre el interior y el exterior del proceso, concluyendo dicho movimiento en el paso de la Sociedad de la Información a la sociedad virtual (Hutter, 2012: 144, 149-150).

Es con relación a dicho concepto que se está investigando el desarrollo de un marco computacional común a través de la IA, la ciencia cognitiva, la neurociencia y la robótica, lo que se determina como un "Modelo Estándar de la Mente" (*Standard Model of the Mind*). La idea consiste en desarrollar una base coherente que aglutine el progreso de forma acumulativa. En dicho marco arquitecturas como ACT-R, Clarion y LIDA dan respuesta a datos conductuales de experimentos controlados que utilizan memoria, solución de problemas y la interacción percepción-motora. En la IA, arquitecturas como Soar y Sigma desarrollan capacidades funcionales que se utilizan en tareas para el control de agentes inteligentes, humanos virtuales, simulaciones y robots embebidos. En robótica, arquitecturas como 4D/RCS y DIARC se utilizan para controlar en tiempo real a robots físicos (Laird *et al*, 2017: 17).

Dichos modelos tienen en común la premisa de que la mente tiene niveles intermedios de organización cada uno de ellos con un contenido de información y con una estructura propia, lo que permite establecer propiedades de reutilización y componibilidad. Es decir, sistemas complejos con un gran número de partes interconectadas. En general, modelos reutilizables con un amplio grado de modularización escalable por donde los datos fluyen para su aprendizaje y validación. Uno de dichos modelos es la arquitectura cognitiva ACT-R (*Adaptive control of thought-rational*).

La figura 15, ilustra la arquitectura básica del ACT-R 5.0, que consiste en una serie de módulos interconectados, como un módulo visual para identificar objetos en el campo visual (*visual module*), un módulo manual para controlar las manos (*manual module*), un módulo declarativo (*declarative module*) para recuperar la información de la memoria o un módulo de objetivos (*goal module*) que sirve para hacer un seguimiento de los objetivos y las intenciones. Dichos módulos se coordinan a través de un sistema central de producción (*Productions*). Dicho sistema es selectivo a la información recibida de los distintos módulos y solo se concentra en el objeto de interés. Por lo tanto, aunque por las memorias intermedias (*buffers*) pase mucha información esta dependerá de las reglas de producción, por lo que cada módulo está encapsulado y solo se podrán comunicar a través de la información que envían o reciben de dichas memorias gobernadas por el sistema central de producción (ver fig. 15) (Anderson *et al*, 2004: 1037),

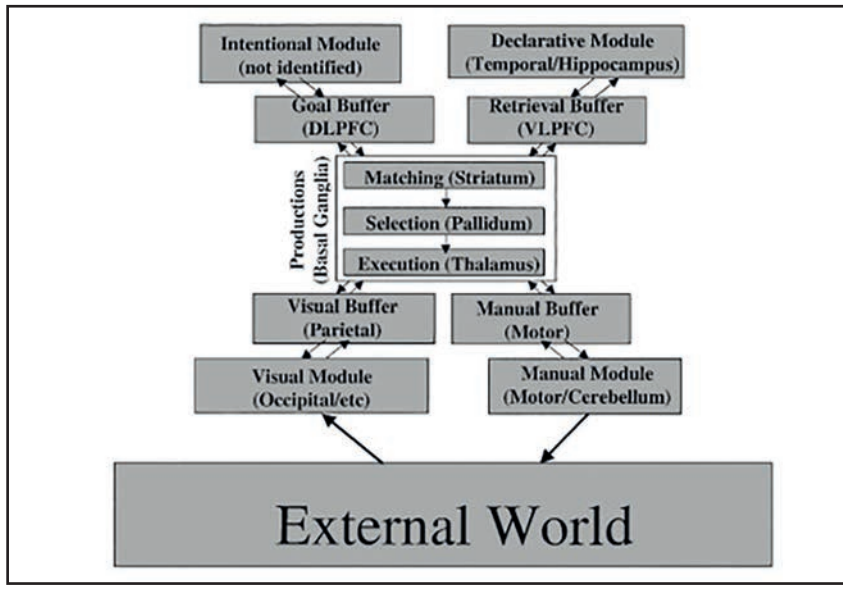


Figura 15: Organización de la información del ACT-R 5.0. La información de los búfer asociados con los módulos responden y cambian de acuerdo con las reglas de producción (Anderson et al, 2004: 1037)

Ahora bien, para ver la complejidad de replicar la mente humana se pueden añadir arquitecturas diferentes a la cognitiva, como es el caso de la arquitectura Fisiológica. Un ejemplo de dicho resultado es la denominada arquitectura ACT-R/φ, que añade un sistema Afectivo (*Affect System*) y un sistema Fisiológico (*Physiology System*), basado principalmente en el modelo HumMod. El sistema Afectivo incluye un sistema motivacional y un sistema defensivo, que permite cambiar los objetivos y recompensas debido al cambio fisiológico (Dancy y Ritter, 2017: 318).

Otra arquitectura cognitiva desarrollada ha sido SOAR, diseñada

por J. Laird, P. Rosembloom y A. Newell[100], un sistema fundamentado en reglas de producción y basado en un ciclo de decisión simple: estado de elaboración, selección de los operadores, elección de un operador y acción. Está compuesto por una serie de módulos paralelos internos asíncronos, incluida una memoria de procedimientos. Se organizan alrededor de una memoria de trabajo global con memorias declarativas episódicas y semánticas, además de un módulo de visión/espacial y un módulo motor que controla los efectos robóticos y virtuales (ver figura 16) (Laird, Lebiere y Rosenbloom, 2017: 20).

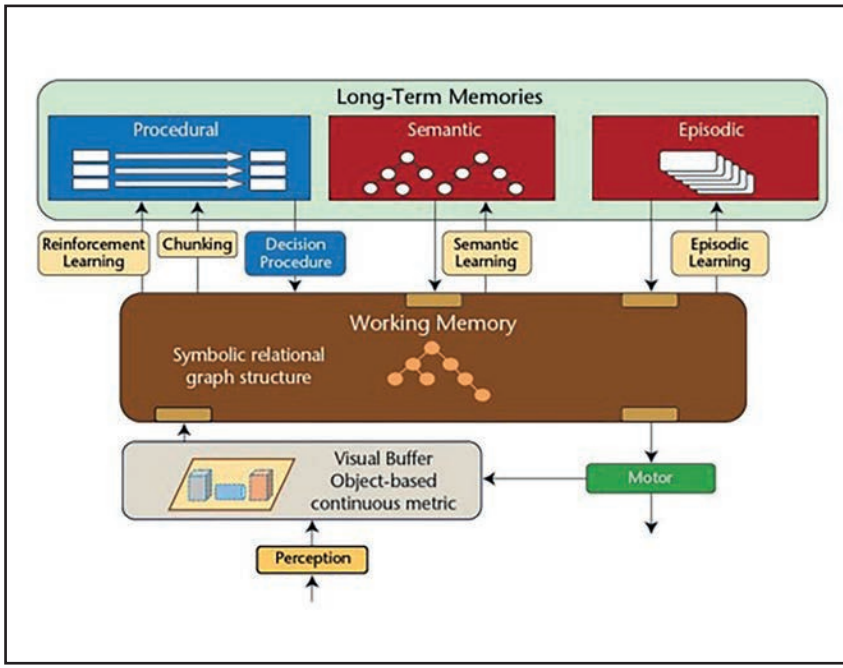


Figura 16: Arquitectura SOAR (Laird, Lebiere y Rosenbloom, 2017: 20).

100 Para más información ver: Lehman, J. F., Laird J. E. y Rosenbloom, P. (1998): "A gentle introduction to Soar: An architecture for human cognition", en D. Scarborough *et al* (eds.), *An invitation to Cognitive Science, Second Edition, Volume 4, Methods, models and conceptual issues,* MIT Press, Cambridge.

La arquitectura cognitiva SIGMA, por otro lado, es una combinación de las anteriores con el conocimiento de los modelos gráficos. No tiene una arquitectura tan modular que solo tiene una memoria a largo plazo que unida a la memoria de trabajo y los componentes perceptuales y motores se basa en modelos gráficos (ver figura 17) (Laird, Lebiere y Rosenbloom, 2017: 18).

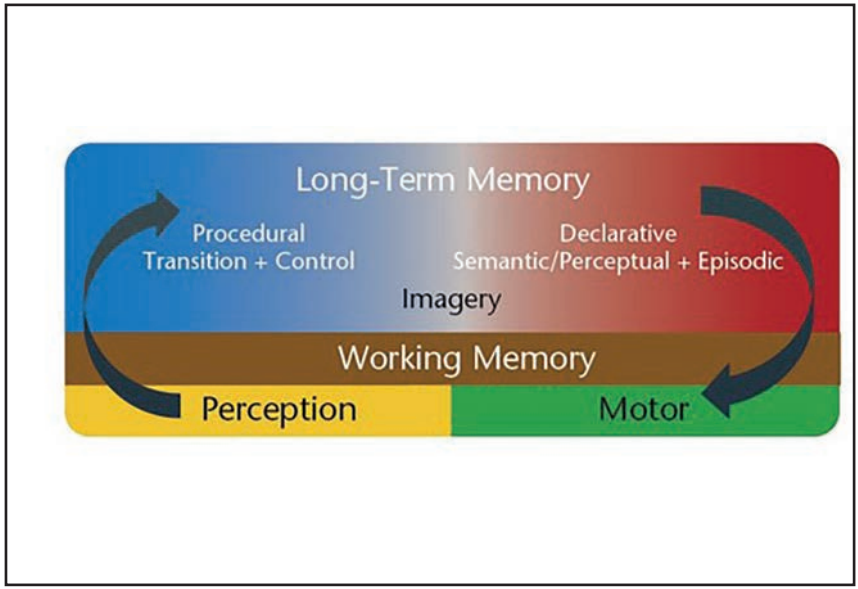


Figura 17: Arquitectura SIGMA (Laird, Lebiere y Rosenbloom, 2017: 21).

Dichos modelos forman parte del Modelo Estándar de la Mente. La premisa principal es que la mente está creada de módulos independientes que desarrollan distintas funciones. Los elementos principales son percepción y motor, memoria de trabajo, memoria a largo plazo declarativa y memoria a largo plazo procesal. Cada módulo se descompone a su vez en otros módulos. La memoria de trabajo se considera el elemento que actúa como intercomunicador entre los distintos componentes actuando como memoria interme-

dia de los diversos módulos. La base del modelo es el "ciclo cognitivo" (*cognitive cycle*). La memoria procesal induce la selección de un acto único por ciclo y cada acción puede significar múltiples modificaciones de la memoria de trabajo (ver figura 18) (Laird, Lebier y Rosenbloom, 2017: 21).

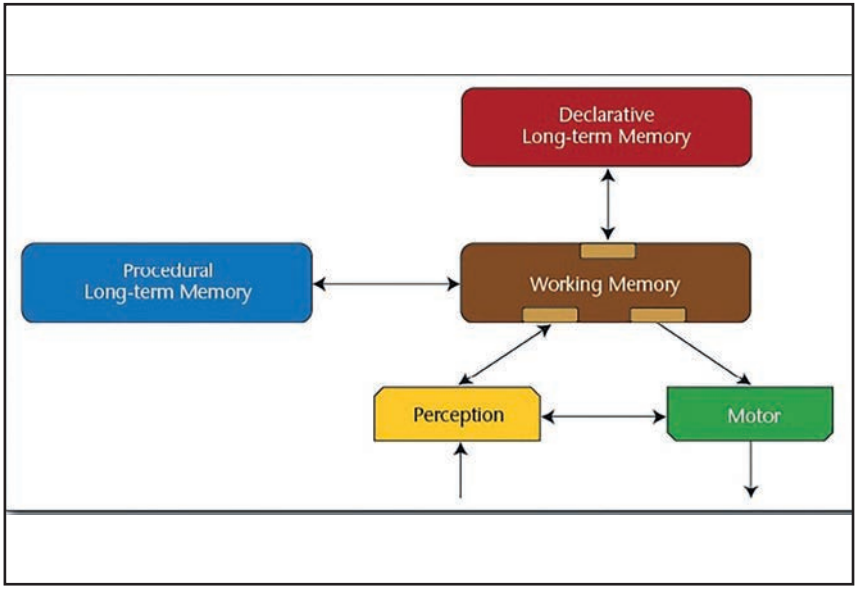


Figura 18: Modelo Estándar de la Mente (Laird, Lebiere y Rosenbloom, 2017: 21).

6.3.- LIMITE CONCIENCIA INGENIOS ARTIFICIALES

Quizás el elemento principal que determinase la existencia de una máquina ultra inteligente sería su capacidad de pasar el "Test de Turing", basado en la pregunta planteada por A. M. Turing: ¿puede una máquina pensar?, cuya prueba se establecería a través del "juego de imitación" donde se constataría que una máquina habría logrado alcanzar los niveles humanos de inteligencia si lograse

convencer a un humano de que era una persona real. Recientemente, E. Feigenbaum propuso una variante a dicha prueba ya que el "Test de Turing" se entendía como muy generalista. En dicha variante, la prueba se habría conseguido si la máquina convenciese a un científico de una determinada disciplina de que la máquina era un experto en dicho campo (Feigenbaum, 2003: 36; Turing, 1950: 433-434).

No obstante, la posibilidad de que una máquina ultra inteligente alcance la Singularidad tiene detractores y partidarios. Tanto A. Braga como R. K. Logan preconizan que la noción de que alguna vez una máquina alcance la Singularidad es falsa, pues los partidarios no tienen en consideración la dimensión completa de la inteligencia humana. Existen características como la curiosidad, la imaginación la intuición, las emociones, los valores, la moralidad, la sabiduría o el buen juicio, entre otras, que son únicas en los humanos. Es más, son de la opinión de que los ingenieros no deberían estar definiendo la inteligencia. Para ello, siguiendo la idea de T. Deacon, para que un ente tenga inteligencia también debería tener un sentido de sí mismo, pero una computadora o un sistema de IA carece de dicho sentido, no es consciente de lo que sabe, ya que trata los datos de forma secuencial, uno cada vez[101] (Braga y Logan, 2017: 157).

Según dichos investigadores, existe el peligro real de que se pierda parte de la autonomía humana hacia IA o la IGA, con un descenso de la inteligencia humana y de cómo se contempla la naturaleza del espíritu humano a través de una pérdida parcial de nuestra au-

101 T. Deacon define la "información" como "algo para algo con algún fin", por lo que una máquina carecería de "información" según dicho investigador (Deacon, 2012: 524).

tonomía hacia otras tecnologías. Muchos de los defensores de la IGA o la IA, están subyugados por la lógica y la racionalidad hasta tal punto que no contemplan la parte emocional de la inteligencia. Pero la inteligencia humana es bicameral compuesta tanto por la parte izquierda del cerebro racional como por la parte derecha intuitiva. En la mente de una máquina de IA rige la parte izquierda del cerebro y por tanto carece de la neuroquímica que forma parte de la vida emocional humana (Braga y Logan, 2017: 163).

Por otro lado, el razonamiento abductivo[102], que permite que los humanos creen patrones que desembocan en sistemas cognitivos y culturales complejos y dinámicos, es una forma de inferencia lógica que utiliza la imaginación. Para R. K. Logan y M. Tandoc, por tanto, una máquina sería incapaz de un razonamiento abductivo, dado que carece de imaginación y por tanto la capacidad para crear patrones estructurados. La evolución de la cultura humana depende de un sistema adaptativo complejo a través de un proceso de restructuración de patrones. La capacidad humana de crear y de imaginar, permite un creciente desarrollo de ideas complejas, desde la tecnología, al arte o la ciencia. Por lo tanto, muchas de las cualidades que nos hacen humanos derivan de nuestra capacidad para reestructurar patrones, que permiten la creación de nuevas ideas (Logan y Tandoc, 2018: 83-84, 94).

Recientemente D. Mumford, profesor emérito de las universidades de Brown y Harvard planteó que uno de los ingredientes principales que faltaban en las máquinas con IA eran las emociones y que a excepción de R. Picard del MIT Lab no se habían planteado mo-

102 Un razonamiento abductivo o conjetura, según C. S. Pierce, es un razonamiento que a partir de un hecho se llega a una hipótesis, que busca ser la mejor explicación o la más probable (Aguayo, 2011: 34-36).

delos sobre ellas. D. Mumford hace mención sobre la complejidad de las emociones, cuyo desarrollo fue plasmado por R. Plutchik[103] que estableció ocho elementos principales: alegría; confianza; miedo; sorpresa; tristeza; disgusto; enojo y; anticipación, así como un número ilimitado de emociones secundarias como: vergüenza, sentido de la culpabilidad; gratitud; perdón; venganza; frustración; excitación, etc. Por lo tanto, se necesitaría desarrollar lo que califica como "empatía artificial", siendo necesario que la IA adquiriese emociones propias (Mumford, 2019).

En contraste, a raíz de sus trabajos sobre la computación afectiva, cuyo principal propósito sería el desarrollar máquinas que tuviesen sentimientos, R. Picard estableció que las emociones no serían estrictamente necesarias para obtener inteligencia. Matizó, que, aunque en la actualidad la expresión de emociones por los robots no sea realista y que el desarrollo de modelos cognitivos sea escaso, pueden existir nuevas formas de construir sistemas de apariencia humana sin la necesidad de las emociones. En todo caso, existiría la necesidad de llegar a un equilibrio, algunas máquinas no necesitarían de emociones, mientras que otras utilizarían un subconjunto de ellas (Picard, 2003: 55-64).

Hay que tener en cuenta que la mayor parte de los trabajos actuales consisten en modelos computacionales sobre el reconocimiento de las emociones, en modelos sobre factores causales de las emociones y la expresión de emociones a través de rostros renderizados de robots. Solo una pequeña parte de los trabajos modelan los efectos de las emociones, las teorías de evaluación cognoscitiva o las emociones emergentes. En todo caso, cualquier modelo para que sea

103 Para más información ver: Plutchik, R. (2001): "the Nature of Emotions", *American Scientist*, 89-4, 344-350.

computacional necesitaría que fuese posible su ejecución en una computadora y por lo tanto la computadora necesitaría instrucciones detalladas paso a paso para poder implementar dicho modelo (Broekens, 2010: 1-2).

Según J. Broekens, los resultados que se obtienen al ejecutar un modelo computacional son predicciones de la teoría subyacente del modelo. Su utilidad dependería de dos factores: la credibilidad de la teoría utilizada como base para el modelo y; una implementación computacional correcta. En todo caso, cuando se hable de las emociones dentro de una computación afectiva esta se debería interpretar de una forma amplia. Dicha computación estaría relacionada con el reconocimiento de las emociones, la producción de emociones, el sentimiento de emociones y los efectos de las emociones en el marco cognitivo, así como en el comportamiento (Broekens, 2010: 2-5).

Uno de los problemas destacados por J. Broekens, en 2010, estaría relacionado con la constatación de que los trabajos sobre computación afectiva estarían más desarrollados en el campo técnico e informático, pero poco desarrollado en el campo psicológico. Sería necesario el comprender mejor los mecanismos y procesos subyacentes que forman las emociones. Como establecen los investigadores L. F. Barrett, B. Mesquita, K. N. Oschner y J. J. Gross, un elemento esencial que ha escapado a la visión científica ha sido la experiencia subjetiva de la emoción. Consecuentemente, el conocer las causas de las emociones no es suficiente para responder sobre la experiencia de las emociones. Así el método científico dejaría fuera un aspecto importante de la realidad: que la personas sienten algo cuando experimentan una emoción. Es más, según dichos investigadores, el conocer como sienten las personas debería

ser un requisito esencial para los científicos antes de postular un modelo de computación afectiva (Barrett *et al*, 2007: 374; Broekens, 2010: 6).

No obstante, recientemente se han realizado avances significativos en dicha área. Se han creado máquinas que pueden representar tres tipos de aplicaciones de computación afectiva: sistemas que pueden detectar las emociones del usuarios; sistemas que pueden expresar lo que un humano percibe como una emoción (ej. avatares, robots, agentes conversacionales animados) y; sistemas que realmente "sienten" una emoción. No obstante, el desarrollo de agentes con capacidades de inteligencia emocional, para alcanzar la Singularidad, deberían ser competentes en el razonamiento de las emociones, el predecir y comprender las emociones humanas y el poder procesar las emociones cuando se relacionasen con un humano (Zohora *et al,* 2016: 3).

Uno de los puntos a lograr sería, el poder diseñar sistemas basados en agentes donde dichos agentes muestren algún tipo de emociones y lo más importante exhiban comportamiento dependiendo de su estado emocional. Es decir, la creación de agentes emocionales: sistemas artificiales diseñados de tal manera para que las emociones desempeñen un papel en los mismos. Como indica J-J. Ch. Meyer, el uso de estados emocionales para diseñar agentes artificiales inteligentes, pero sin entrar en cuestiones filosóficas de si dichos agentes realmente poseen emociones verdaderas como la de los humanos. Así, en los sistemas basados en agentes que son percibidos como inteligentemente racionales, éstos poseerían algunas formas de actitud BDI (creencia-deseo-intención), por lo que se podría describir su comportamiento y evolución de sus estados mentales a lo largo del tiempo y como éstos determinarían

que acción llevarían a cabo y que efecto tendrían sobre su entorno (Meyer, 2006: 601).

Desde dicho punto de vista, es interesante destacar la posición de S. Petersen que plantea la posibilidad de que se puedan crear robots éticos e incluso crear "gente artificial". El establecer que algo artificial pudiese ser una persona, significaría que al menos podría tener un pensamiento ético como el nuestro. Además, se les podría encomendar tareas que los humanos consideran cansinas o desagradables. Con el mismo pensamiento de considerar a las máquinas como personas, R. Sparrow desarrolló la denominada "Prueba de Triaje de Turing" (*Turing Triage Test*), en la que una máquina habría alcanzado el mismo nivel moral de un humano cuando en un dilema moral entre preservar la vida de un humano o de una máquina se elegiría a la máquina. En este punto no estaría demás tener en cuenta la pregunta formulada por V. Saraswat[104]: "¿Qué es lo que significa la noción de la ética para una máquina a la que no le importa si los humanos que están alrededor existen o no, que no puede sentir, que no puede sufrir, que no conoce lo que son los derechos fundamentales?" (IBM, 2020; Petersen, 2012: 283-284; Sparrow, 2004: 203).

Ahora bien, para controlar los problemas de dicho desarrollo, el investigador D. Chalmers propuso, en 2010, una Singularidad a "prueba de fugas". Así, por razones de seguridad, los sistemas de IA estarían restringidos a mundos virtuales simulados hasta que sus tendencias de comportamiento fuesen completamente comprendidas dentro de una condiciones controladas. Aunque sería muy difícil de implementar, como indica R. V. Yampolskiy, daría un poco

104 V. Saraswat era, en 2020, Científico Jefe para IBM Compliance Solutions.

de tiempo a la Humanidad para prepararse a dar una mejor respuesta. Otra posibilidad sería la creación de sistemas de monitorización autónomos con reglas de implementación específicas, como por ejemplo conceptos éticos o requerimientos legales (Chalmers, 2010: 36-40; Yampolskiy, 2012: 407-408).

Además, existen límites para el desarrollo de la LOAR de forma indefinida y de las mejoras recursiva de la IGA. Un sistema necesita de "hardware" para la memoria, la comunicación y el proceso de la información y existen límites físicos de la computación en términos de velocidad, comunicación y consumo de energía con relación a factores como la velocidad de la luz, ruido cuántico y la constante gravitacional. En cuanto al "software", muchos de los problemas reales del mundo no son posibles de resolver a través de la computación, aunque se puedan crear teóricamente máquinas con un poder computacional enorme, en la práctica no se podrían implementar, dado que la información no computacional necesaria para que funcionasen no sería computable. Cualquier cambio no es estrictamente una mejora, tendría algunos progresos, pero también perdería algunos de las mejoras ya existentes. Como indica R. V. Yampolskiy, si se asume que el objetivo es el de una inteligencia incrementada sobre todos los medios, el sistema debería establecer como comprobar dicha inteligencia en todos los niveles, incluso por encima del propio nivel de la máquina, un problema que todavía no ha sido resuelto (Yampolskiy, 2015: 395-396, 399).

Todas las disquisiciones teóricas mencionadas nos llevan a la conclusión de que, en la actualidad, estaríamos ante un proceso de "moralidad operativa", es decir que la IA subyacente de un AMA teórico funcionaría únicamente según su diseño, que sería modulable y escalable a través del aprendizaje y la validación. Creemos,

además, que el alcance de la Singularidad, esto es, el sentido de sí mismo y la capacidad de tener emociones, no será factible a corto plazo. Ahora bien, nos deberíamos plantear, como R. Picard, si realmente un SAAL necesita emociones para funcionar o simplemente una serie de restricciones establecidas por un AMA y un control y supervisión humana adecuadas. Nuestro punto de vista es que la existencia de los SAAL carentes de emociones es ya una realidad[105] y que, a nuestro entender, sería imperativo desarrollar un control y supervisión humano holístico, utilizando todos las herramientas tecnológicas posibles, incluidas los AMA, para no permitir "ataques indiscriminados" incontrolados de "entidades" artificiales, pero al mismo tiempo no permitir que dicho AMA, a través de la IA, alcance la condición de AMAA, una moralidad propia de la máquina, no necesariamente similar a la humana, cuyas consecuencias serían imprevisibles.

105 El nuevo dron turco Kargu-2 presuntamente puede seleccionar y atacar objetivos humanos autónomamente basado en el reconocimiento facial y la IA que, según un informe de las NU, ha sido ya utilizado en Libia (Wadhwa, 2021).

CAPÍTULO 7
ABORDANDO LA SINGULARIDAD Y LOS SAA

El teórico de la guerra C. V. Clausewitz advertía que "la tendencia a destruir al adversario, que subyace en la concepción de la Guerra, en ningún caso se cambia o modifica a través del progreso de la civilización". Por lo tanto, el desarrollo de los SAA o de los SAAL tendrán siempre como objetivo la destrucción del enemigo, que en nuestra actualidad se concretaría con el desarrollo de artefactos integrados con algoritmos de IA y, por lo tanto, con un crecimiento notable de autonomía en la toma de decisiones de dichos sistemas.

Como argumenta C. Haynes, dicha situación implicaría que no solo los humanos estarían ausentes físicamente del campo de batalla, sino que también lo estarían psicológicamente, un nuevo nivel en el proceso de despersonalización de la guerra. Una situación contextualizada, según argumenta A. Westhues, en un entorno VICA (volatilidad, incertidumbre, complejidad y ambigüedad) de inestabilidad geopolítica. En dicho contexto, lo que entraría en juego sería el impacto que tendría la utilización de los SAA (incluidos los SAAL) sobre las decisión más crítica de su función: la liberación del uso de la fuerza (Heyns, 2016: 4; May *et al*, 2006: 115-121; Westhues, 2020: 12).

Un uso de la fuerza que sería producto de una combinación entre el mundo físico y el mundo digital, lo que se denominaría como una "Realidad Mixta" (*Mixed Reality*)[106], si se sigue la definición de dicho concepto, establecida por Young *et al,* como la combinación del mundo físico y el virtual en un espacio de interacción. Ahora bien, según argumentan I. Bode y H. Huelss, donde se llevasen a cabo las decisiones del uso de la fuerza podría ser menos decisivo que el tipo de control que se ejerciese.

Un tipo de control que sería una combinación de una red de estructuras estatales e internacionales tanto jurídicas (ej. DIH, DD.HH., etc.) como normativas que surgirían de la práctica (desarrollo, entrenamiento, prueba, despliegue) del uso de los SAA. Un enfoque que iría más allá de un examen sobre cómo serían gobernados los SAA por normas, hacia un marco que analizaría como se crean, moldean y definen dichas normas, punto de vista que nosotros compartimos y utilizaremos en nuestro análisis. A tal fin, utilizaremos como base los principios fundamentales que rigen el DICA: necesidad militar; humanidad; distinción; proporcionalidad y; precaución. Estos principios, aunque no reemplacen al propio Derecho, sustentan tanto el Derecho convencional como el consuetudinario (Bode y H. Huelss, 2018: 395-397; CICR, 2016f: 57; Kelsey, 2008: 1441; Young *et al,* 2011: 2, 10).

106 J. Young *et al* definen la "Realidad Mixta" como: la combinación del mundo real y el virtual para producir nuevos entornos y visualizaciones, donde los objetos físicos y los digitales coexisten e interaccionan en tiempo real (Young *et al,* 2011: 2).

7.1.- EL ARMAMENTO INDISCRIMINADO Y SUS DESAFÍOS

El DICA establece las normas que reflejan el equilibrio entre la necesidad militar y la humanidad. Según el CICR, dicho equilibrio se refleja en el principio de distinción, que requiere que la fuerzas armadas distingan claramente entre objetivos militares y personas o bienes civiles[107]. En un dominio de "Realidad Mixta", aunque se ha establecido que no existe un vacío en la aplicación del Derecho Internacional, como ya hemos analizado anteriormente con relación al *ius in bello*, no queda tan claro como dicho principio funcionaría en el mundo cibernético, incluidos los SAA y los SAAL, dado que las nuevas tecnologías crean una nebulosa entre lo militar y lo civil. Como argumentan R. Geib y H. Lahmann, en el mundo cibernético cualquier componente podría ser un objeto de uso dual y ser utilizado en la actualidad o en un futuro como objetivo militar legítimo, con amplias repercusiones hacia la población civil. Así, potencialmente, cualquier infraestructura cibernética (computadoras, redes y cables) o incluso el propio ciberespacio podrían ser cualificados como un objetivo militar[108]. Dicho principio de distinción estaría intrínsicamente ligado a la regla sobre "ataques indiscriminados". El Artículo 51(4) del PAI, sobre "protección de la población civil"

107 Artículos 48, 51(2) y 52(2) del PAI, así como en el Protocolo II de la CCW de 1980, Art. 3(2) (ibíd., párr. 157); Protocolo II enmendado de la CCW de 1996, Art. 3(7) (ibíd., párr. 157); Protocolo III de la CCW de 1980, Art. 2(1) (ibíd., párr. 158); preámbulo de la Convención de Ottawa de 1997 (ibíd., párr. 3). (CICR, 1977a; CICR, 2007).

108 Dichos elementos típicos del mundo cibernético se añadirían a otras infraestructuras civiles que pudiesen ser utilizadas como infraestructuras militares en caso de conflicto armado como: centrales eléctricas, instalaciones de telecomunicaciones, puentes, etc. (Kelsey, 2008: 1437).

establece dicha definición[109] que, en la práctica de los Estados, se ha convertido en una norma de Derecho Internacional consuetudinario aplicable tanto a los conflictos armados internacionales como a los no internacionales[110]. Además, de acuerdo con el Artículo 85(3)(b) del PAI, un ataque indiscriminado se consideraría como una violación grave del Protocolo[111][112] (CICR, 2005: 248; 2016f: 5; Geib y Lahmann, 2012: 382-383).

Para poder analizar qué se entiende por el principio de distinción y la regla sobre "armamento indiscriminado" dentro del entorno de los SAA (incluidos los SAAL), habría que tener en cuenta, en

109 **Artículo 51(4):** Se prohíben los ataques indiscriminados. Son ataques indiscriminados: a) los que no están dirigidos contra un objetivo militar concreto; b) los que emplean métodos o medios de combate que no pueden dirigirse contra un objetivo militar concreto; o c) los que emplean métodos o medios de combate cuyos efectos no sea posible limitar conforme a lo exigido por el presente Protocolo; y que, en consecuencia, en cualquiera de tales casos, pueden alcanzar indistintamente a objetivos militares y a personas civiles o a bienes de carácter civil (CICR, 1977a).

110 No obstante, habría que tener en cuenta, a la hora de analizar la aplicación de dicho Artículo 51, las posibles matizaciones o limitaciones que los Estados pudiesen establecer. Por ejemplo, en el caso de España, en la ratificación del PAI y PAII de 1989, con respecto al Artículo 51 establece que: "Entiende que la decisión adoptada por mandos militares y otros con facultad legal para planear o ejecutar ataques que pudieran tener repercusiones sobre personal civil, bienes o similares no puede necesariamente ser tomada más que sobre la base de informaciones pertinentes disponibles en el momento considerado y que ha sido posible obtener a estos efectos", declaración que sería de especial relevancia con relación a los SAA (BOE, 1989).

111 **Artículo 85(3)(b)**: lanzar un ataque indiscriminado que afecte a la población civil o a bienes de carácter civil a sabiendas de que tal ataque causará muertos o heridos entre la población civil o daños a bienes de carácter civil, que sean excesivos en el sentido del artículo 57, párrafo 2, a) iii (CICR, 1977a).

112 Como complemento, el Artículo 3(3) del Protocolo II del CCW de 1980 y el Artículo 3(8) del Protocolo II Enmendado de 1996 establecen que: "el uso indiscriminado de minas, bombas trampas y otros artefactos está prohibido" (CICR, 2005: 248),

primer lugar, las diversas acepciones de los términos "autónomo" o "semi autónomo". Para el CICR, un SAA sería aquel sistema armamentístico que tuviese autonomía en sus funciones críticas de: selección de objetivos (búsqueda o detección, identificación, seguimiento, selección) y de ataque (uso de la fuerza, neutralización, daño o destrucción) a un objetivo sin la intervención humana. Para H. Roff, sería un sistema armamentístico que estaría formado por cuatro funciones básicas: accionamiento del sistema (*triggering*); selección de objetivos (*targeting*); navegación (*navigation*) y; movilidad (*mobility*) y se consideraría semi autónomo si tuviese entre uno y tres de dichos elementos automatizados informáticamente. Para J. Stroud-Trup, habría que diferenciar entre sistemas automatizados y sistemas autónomos. Un sistema automatizado sería aquel que siguiese una serie de reglas preprogramadas con un resultado predecible, mientras que un sistema autónomo, aunque la actividad de dicho sistema fuese predecible, sus acciones individuales no tendrían por qué serlos, pues el sistema no seguiría una serie de reglas de forma predecible.

En el caso de la Federación Rusa, V. Kozyulin define a un robot inteligente (autónomo), como un artefacto multifuncional con comportamiento antropomórfico, que parcial o completamente llevaría a cabo funciones de un humano durante misiones de combate específicas. Otra definición la propone B. Kastan, cuando define "autonomía" como la medida de relativa independencia de un robot o un armamento. Podría ser: una "tele operación" (operación remota); un sistema automático que funcionase dentro de unos parámetros preprogramados sin requerir una orden dada por un humano o; un sistema completamente autónomo que decidiese por sí mismo y se adaptase ante nueva información (CICR, 2016a: 8, 57, 60; Roff, 2015b).

En dicho contexto, un SAA o un SAAL podría ser, por ejemplo, desde un sistema de defensa antimisiles hasta un programa malicioso (*malware*). Si tomásemos un ciberataque militar sofisticado, que podría consistir en programas maliciosos que se mantienen durmientes durante un periodo de tiempo determinado (horas, días, meses, años) y luego se activan ante un evento, según la definición del CICR podrían ser considerados también como un SAA, ya que tendrían la capacidad de seleccionar un objetivo y atacarlo (causando daño o destrucción). Además, entrarían en la definición de un sistema armamentístico semi autónomo de H. Roff, al tener, por lo menos la capacidad de *triggering, targeting* y *mobility* o entrar en la definición de V. Kozyulin de artefacto multifuncional con comportamiento antropomórfico (robot inteligente), al llevar a cabo funciones de humanos en una misión de combate.

Por su parte, J. F. Carson argumenta, como ya hemos analizado anteriormente, la necesidad de tener en cuenta el impacto del legado cultural al desarrollar los SAA y, además, que un agente "autónomo" sería aquel que, en primer lugar, fuese capaz de desarrollar una acción deliberada (intención deliberada) e independiente que resultase de su propia voluntad. Así, en el ámbito de la tecnología militar, la "autonomía" sería la capacidad de un sistema armamentístico de determinar por sí mismo y sin ningún tipo de interferencia humana, cuando y contra quién se utilizaría una fuerza letal, para lo cual necesitaría de tres elementos: la capacidad de analizar todos los posibles resultados y sugerir la mejor estrategia posible; hacer que los robots inteligentes coordinasen conjuntamente una acción común y; tener la capacidad analítica de mostrar el mismo discernimiento moral que los seres humanos. Según el propio J. F. Carson, en la actualidad la IA solo sería capaz de tener la primera y la segunda premisa, como en el caso de los sistemas antimisiles de

defensa o los programas maliciosos. Dicha postura vendría ligada con la idea expresada por G. Sartor y A. Ominici, de la distinción entre la "capacidad de independencia" que establece la dimensión que tiene un sistema armamentístico para completar una tarea y la "independencia organizativa", su capacidad para llevar a cabo una tarea dentro de "la infraestructura socio técnica global", sugiriendo que los efectos de un armamento no solo estarían en función de su diseño, sino también en que uso se le diese y la vulnerabilidad de aquellos a quién les afectase (Carson, 2020: 174-175; CICR, 2016a: 60; Kastan, 2013: 49-50; Sartor y Omicini, 2016: 40).

La inexistencia en la actualidad, aunque la tecnología evoluciona rápidamente, de algoritmos que permitiesen un discernimiento moral como los seres humanos ha llevado a argumentar a E. Rosert y F. Sauer, así como a *Human Rights Watch,* que los SAAL serían incapaces de discernir entre combatientes y civiles, violando el principio de distinción y por lo tanto se deberían considerar como sistemas indiscriminados. Esto sería debido a que el concepto de "civil" es un término complejo, dependiente del contexto en el que se establece la acción, no siendo trasladable a una aplicación informática (*software*), independientemente de que dicho algoritmo estuviese basado en reglas o en "*machine learning*". Es más, para dichos investigadores, el reconocer y aplicar el concepto de "civil" en el campo de batalla requeriría de juicios de valor, pero también de un cierto grado de conocimiento del entorno[113] y una comprensión del contexto social, que la tecnología computacional actual y en un futuro próximo sería incapaz de poseer. Interesante también sería la relación que establecen entre el principio de distinción con

113 Referencia al concepto de "conciencia situacional" (*situational awareness*), el conocimiento del entorno operativo en todas sus dimensiones: política, cultural y social, económica y militar (Walker, 2019: 55).

el de humanidad, al plantear la necesidad de prohibir dichos sistemas armamentísticos por ser indiscriminados y por tanto atentar contra la dignidad humana. Afirmación que estaría incluida dentro del marco del denominado derecho natural ya analizado, donde se estaría viendo los SAAL, en primer lugar, como un ejercicio moral para después considerarlo como un ejercicio legal. En todo caso, para J. T. G. Kelsey, los Estados tendrían grandes incentivos para desarrollar ciberataques u otros ataques de los SAA (como los VANT) que violasen las nociones tradicionales del principio de distinción, especialmente cuando dichos ataques no fuesen letales. Un área gris dentro del marco del Derecho Internacional ya existente, dado que el concepto moderno de la guerra no excluye a objetivos cuya destrucción o neutralización, aunque no avanzasen directamente en el objetivo general de la guerra, no obstante, degradasen la capacidad y la voluntad de combatir[114]. (Eliot, 2020: 1; HRW, 2012: 30-31; Kelsey, 2008: 1431, 1437, 1440; Rosert y Sauer, 2019: 370, 372-373).

Existen, no obstante, críticas con relación a la utilización de la "dignidad humana" como base para prohibir los SAA. A. Saxton argumenta, que el uso *de facto* de dichos sistemas no necesariamente significaría una violación inherente de la dignidad humana, simplemente por su autonomía. El concepto de dignidad humana sería difícil de medir durante un conflicto, dado que la propia guerra sacrifica parte de la dignidad humana a través del sacrificio intencionado de vidas para alcanzar objetivos militares. La pérdida

114 Se podría tomar como ejemplo la utilización de drones por parte de Azerbaiyán en el entorno de Nagorno Karabaj en 2020. Un arma de castigo contra la población civil, según el Defensor del Pueblo de la autoproclamada república de Nagorno Karabaj (RNK), al atacar centrales eléctricas, depósitos de agua y de gas (EFE, 2020).

de empatía no necesariamente tendría que suponer que una decisión fuese injusta o inmoral, especialmente si a través de una decisión que fuese moral se incrementase los riesgos hacia las propias tropas. Por lo tanto, A. Saxton no sería partidario de implementar una prohibición total, sino que se debería evaluar cada SAA independientemente, para establecer su impacto sobre la dignidad humana[115]. Por su parte, para D. Birnbacher, el concepto de "dignidad humana" solo sería aplicable de forma individual y nunca de forma colectiva. Tampoco aceptaría que la utilización de los SAA autónomos, por sí mismos, resultase *de facto* en que los civiles fuesen humillados y degradados a excepción del posible desarrollo de un sufrimiento mental (miedo, ansiedad) subjetivo intenso en un entorno y contexto concreto. Es más, según M. N. Schmitt, una prohibición total cercenaría la posibilidad de utilizar los SAA para minimizar el daño "civil" en comparación con otros arsenales existentes (Birnbacher, 2016: 108, 117; Saxton, 2016; Schmitt, 2012b: 290-292; 2013: 3; Sharkey, 2019: 79-80).

Una postura que sería apoyada por R. Artkin, que argumenta que es muy probable que en un futuro se desarrollasen SAA que minimizasen las bajas civiles, más que los propios humanos. No obstante, para que ello fuese una realidad necesitaría de un desarrollo y despliegue con precaución y poder cumplir con el DIH. Si dicho sistema fuese más allá de la capacidad humana en su aplicación de dicha normativa humanitaria, entonces significaría que los SAA estarían cumpliendo un rol humanitario y existiría un imperativo moral para su utilización. En todo caso el desarrollo y despliegue de dichos sistemas necesitaría de un examen de las consecuencias

115 Idea de contextualización también argumentada por M. Reisman con relación al *ius ad bellum*, que ya analizamos en el capítulo 4º (Reisman, 1985: 281-282)

de su uso, que debería ser practicado por todos los actores, incluidos los civiles y la sociedad en general[116]. Siguiendo con la misma idea, K, Anderson y M Waxman argumentarían que cualquier SAA debería tener la capacidad de seleccionar objetivos con un nivel legal de discriminación aceptable. En el caso del principio de distinción se podría programar, en principio, con una lista de objetivos aceptables, incrementando la lista a través de un razonamiento inductivo sobre características que hiciesen que un objetivo estuviese incluido en la lista, con un desarrollo del algoritmo que sería gradual a través de estándares y buenas prácticas, que sería mejor que a través del establecimiento de nuevos tratados internacionales (Anderson y Waxman, 2013: 41, 46; Artkin, 2013: 1-3, 5).

Nuestro punto de vista sería que, en la actualidad, todavía existen importantes desafíos para que un SAA pueda observar el principio de distinción y como corolario que no se considerase un sistema armamentístico indiscriminado. Siguiendo a los investigadores A. K. Krishnan y J. Foy, existirían tres inquietudes, que nosotros compartimos:

- Una capacidad de percepción de la máquina aún débil (*weak machine perception*): La distinción requeriría de una evaluación a través de una información proveniente de sensores que necesitarían de una alta capacidad de discriminación, de la cual carecen en la actualidad. Una situación que se volvería más compleja cuando existiesen combatientes no uniformados.

116 El mismo investigador propone el desarrollo, dentro de los SAA, de una especie de "gobernador ético" que desarrollase un "control de comportamiento ético" que aplicase el DIH (Arkin, 2007: 61).

- Un problema del marco de actuación (*frame problem*): En un entorno de batalla moderno de ritmo acelerado un SAA tendría problemas para interpretar toda la información existente. Esto implicaría la necesidad de programar que información sería o no relevante. Una programación inadecuada podría llevar a una interpretación errónea de la información causando un ataque indiscriminado, que se vería incrementado si existiesen dudas sobre la legitimidad del objetivo. La dificultad estribaría en establecer en dichas situaciones el umbral de ataque dentro del algoritmo.

- Un algoritmo débil (*weak software*): El incremento de la complejidad de un algoritmo le hace menos predecible. Ningún programador comprende un algoritmo complejo al completo, dado que muchas veces se programan por módulos, por lo que, combinado con un entorno abierto, se podrían dar situaciones donde el SAA aplicase la fuerza indiscriminadamente por un error de programación no anticipado o una situación no programada.

Añadiríamos, como también argumenta E. K. Gade, la neblina que existe en la actualidad para distinguir un combatiente de un no combatiente, de acuerdo con los términos del DIH, que no equivaldría a ser un "civil", ya que dicho sujeto también podría participar en la "máquina de guerra". Por lo tanto, la decisión de establecer quién sería considerado un "no combatiente", sería subjetiva y dependería del contexto y el entorno en el que se desarrollase la acción, lo

que dificultaría el desarrollo de algoritmos eficientes para los SAA. En todo caso, dichos desafíos aún no han sido resueltos, por lo que se necesitaría de una precaución extrema antes de desplegar y usar un SAA completamente autónomo, con relación al principio de distinción o proporcionar elementos de control como los AMA. No obstante, dado que en la actualidad se siguen desarrollando sistemas armamentísticos con algoritmos de IA por parte de diversos Estados, consideramos la necesidad de seguir investigando para desarrollar y mejorar dichos algoritmos, huyendo de una posición de prohibición total, que consideramos tendría poco recorrido en el momento geopolítico actual (Foy, 2014: 57-59; Gade, 2010: 227; Krishnan, 2009: 98-99).

7.2.- PROPORCIONALIDAD Y RESPONSABILIDAD EN LOS SAA

Aunque los SAA hubiesen desarrollado la capacidad de establecer adecuadamente el principio de distinción, el principio de proporcionalidad requiere que el uso de cualquier sistema armamentístico sea evaluado para determinar entre la ventaja militar por su uso y el daño contra el estamento civil (personas y/u objetos). El daño hacia dicho colectivo no debería ser excesivo con respecto a la ventaja militar obtenida, idea que vendría enmarcada dentro del concepto de "guerra justa", en el ámbito del *ius ad bellum*, e implicaría que la destrucción perpetrada por una guerra no fuese desproporcionada con relación al bien que dicha guerra alcanzaría. Incluso si existiese una causa justa, el recurrir a una guerra podría ser impropio si el daño que causase fuese excesivo. En cuanto al *ius in bello*, el daño colateral a los civiles estaría prohibido si resultase desproporcionado, lo que se consideraría un uso excesivo

de la fuerza, con relación a los resultados obtenidos y así queda reflejado en el artículo 51(5)(b) del PAI[117]. El Estatuto de Roma de la Corte Penal Internacional en su Artículo 8(2)(b)(i-vi) lo consideraría, además, como "crímenes de guerra"[118] (Anderson y Waxman, 2013: 41; Hurka, 2005: 34-36; Roff, 2015a: 39-40).

Siguiendo con el argumento de M. N. Schmitt de que los SAA puedan salvar vidas, especialmente en el caso de una amenaza injusta, dicha idea serviría de especial referencia con relación al análisis sobre el principio de proporcionalidad. Como ya desarrollamos anteriormente en el capítulo 4º, algunos investigadores, como en

117 **Artículo 51(5)(b)**: Se considerarán indiscriminados, entre otros, los siguientes tipos de ataque: b) los ataques, cuando sea de prever que causarán incidentalmente muertos y heridos entre la población civil, o daños a bienes de carácter civil, o ambas cosas, que serían excesivos en relación con la ventaja militar concreta y directa prevista (CICR, 1977a).

118 **Artículo 8(2)(b)(i-vi):** A los efectos del presente Estatuto, se entiende por «crímenes de guerra»: b) Otras violaciones graves de las leyes y usos aplicables en los conflictos armados internacionales dentro del marco establecido de derecho internacional, a saber, cualquiera de los actos siguientes: i) Dirigir intencionalmente ataques contra la población civil en cuanto tal o contra personas civiles que no participen directamente en las hostilidades; ii) Dirigir intencionalmente ataques contra bienes civiles, es decir, bienes que no son objetivos militares; iii) Dirigir intencionalmente ataques contra personal, instalaciones, material, unidades o vehículos participantes en una misión de mantenimiento de la paz o de asistencia humanitaria de conformidad con la Carta de las Naciones Unidades, siempre que tengan derecho a la protección otorgada a civiles o bienes civiles con arreglo al derecho internacional de los conflictos armados; iv) Lanzar un ataque intencionalmente, a sabiendas de que causará pérdidas incidentales de vidas, lesiones a civiles o daños a bienes de carácter civil o daños extensos, duraderos y graves al medio ambiente natural que serían manifiestamente excesivos en relación con la ventaja militar concreta y directa de conjunto que se prevea; v) Atacar o bombardear, por cualquier medio, ciudades, aldeas, viviendas o edificios que no estén defendidos y que no sean objetivos militares; vi) Causar la muerte o lesiones a un combatiente que haya depuesto las armas o que, al no tener medios para defenderse, se haya rendido a discreción (BOE, 2002).

el caso de M. N. Schmitt, serían de la opinión de que dicha idea implicaría que se incrementaría la aceptación del uso del armamento robótico ya que los daños serían los del agresor y no el de los combatientes y/o civiles del Estado que se defiende. En dicho contexto, en el caso de que un SAA incurriese en muertes colaterales, persiguiendo objetivos militares legítimos, dichas muertes serían desafortunadas, pero no prohibidas siguiendo la doctrina del "efecto doble"[119]. Se estaría entonces ante un escenario de "defensa propia" que solo sería aceptable si el daño fuese inminente y dirigido contra los intereses vitales del Estado, diferenciando entre ataque "anticipatorio" y ataque "preventivo" este último prohibido. En dicho contexto, habría que insistir, como argumenta H. M. Roff, en que se deberían ponderar todos los daños que dicha acción pudiese causar en un futuro y no solo los relativos al principio de "guerra justa" en el presente, dado que un SAA incide tanto en el *ius ad bellum,* el *ius in bello* y el *ius post bellum*, afectando, por lo tanto, a los cálculos sobre el principio de proporcionalidad cuando se decide entrar en guerra (Hakimi, 2018: 164-165; Roff, 2015a: 42-44, 47, 49-50; Schmitt, 2013: 176).

Paralelamente, habría que destacar que el principio de proporcionalidad no se puede definir en abstracto, como establecen tanto W. H. Boothby como M. Wagner, pues solo tendría sentido desde un punto de vista contextualizado, es decir: una acción concreta, en un escenario concreto y en un tiempo concreto. En dicho marco sería importante establecer el significado del término "excesivo",

119 Para más información ver: McINTYRE, A. (2018): "Doctrine of Double Effect", *Standford Encyclopedia of Philosophy*, Standford University, Standford, acceso febrero 2021, en https://plato.stanford.edu/entries/double-effect/

del artículo 51(5)(b) del PAI, dado el potencial de un entorno cambiante dentro de la propia acción de combate. El artículo 57(2) del PAI sería la base para establecer dichos límites al requerir a los comandantes que tomasen precauciones para evitar o minimizar el daño o la pérdida de vidas indirectas[120]. Destacaríamos, no obstante, la puntualización que diversos Estados han realizado con respecto a la aplicación del Artículo 51 del PAI, ya que consideran que la aplicación del principio de proporcionalidad dependería de la información que se tuviese a mano en cada momento, como ya hemos visto con relación a la postura española en la ratificación de dicho protocolo. En todo caso, M. Wagner argumenta que el principio de proporcionalidad sería demasiado impreciso, lo que crearía tensiones en su aplicación, como pondría de manifiesto el informe final del Tribunal Penal Internacional para la ex Yugoslavia que argumenta: "El problema principal con el principio de proporcionalidad no es si existe, sino que significa y como se debe

120 **Artículo 57(2):** Respecto a los ataques, se tomarán las siguientes precauciones: a) quienes preparen o decidan un ataque deberán: i) hacer todo lo que sea factible para verificar que los objetivos que se proyecta atacar no son personas civiles ni bienes de carácter civil, ni gozan de protección especial, sino que se trata de objetivos militares en el sentido del párrafo 2 del artículo 52 y que las disposiciones del presente Protocolo no prohíben atacarlos; ii) tomar todas las precauciones factibles en la elección de los medios y métodos de ataque para evitar o, al menos, reducir todo lo posible el número de muertos y de heridos que pudieran causar incidentalmente entre la población civil, así como los daños a los bienes de carácter civil; iii) abstenerse de decidir un ataque cuando sea de prever que causará incidentalmente muertos o heridos en la población civil, daños a bienes de carácter civil, o ambas cosas, que serían excesivos en relación con la ventaja militar concreta y directa prevista; b) un ataque será suspendido o anulado si se advierte que el objetivo no es militar o que goza de protección especial, o que es de prever que el ataque causará incidentalmente muertos o heridos entre la población civil, daños a bienes de carácter civil, o ambas cosas, que serían excesivos en relación con la ventaja militar concreta y directa prevista; c) se dará aviso con la debida antelación y por medios eficaces de cualquier ataque que pueda afectar a la población civil, salvo que las circunstancias lo impidan (CICR, 1977a).

aplicar" (Boothby, 2009: 79; CICR, 2016e: 8-9; Egeland, 2014: 53-55; ICTY, 2000; Wagner, 2014: 1393-1397).

La ambigüedad, puesta de manifiesto sobre la aplicación del principio de proporcionalidad, llevaría a M. Wagner a dudar si, en un futuro próximo, un algoritmo sería capaz de llevar a cabo adecuadamente un análisis tan dependiente del contexto. La dificultad de medición de información cualitativa impactaría en como un SAA definiría un objetivo para así poder tomar la decisión inicial de atacar de acuerdo con dicho principio. Para que así sucediese, dicho algoritmo debería ser capaz de manejar un gran número de decisiones y no solo codificar escenarios individuales, codificando un gran número de reglas que pudiesen tomar decisiones mientras sopesase un sinfín de factores[121]. Además, el SAA debería determinar el efecto de cada armamento en cualquier circunstancia. Un proceso que sería aún más difícil debido al entorno cambiante de la acción y la proximidad creciente de civiles en los entornos modernos de batalla[122]. Por lo tanto, sería altamente improbable, en la actualidad, que los algoritmos cuantitativos de IA fuesen capaces de aplicar dicho principio de proporcionalidad, una postura que

121 En un entorno reducido, el lanzamiento de misiles por parte de la Fuerza Aérea de los USA, se viene utilizando desde 2002 un algoritmo denominado FAST-CD (*Fast Assessment Strike Tool-Collateral Damage*), que predice el daño colateral de un misil dependiendo de una serie de factores (terreno, tamaño del armamento, altitud, ángulo y la velocidad desde donde se lanza el misil (Egeland, 2014: 54).

122 Para una información más detallada de dicho problema, con relación al colapso de los SAA por un exceso de información que incrementa su inestabilidad, debido a la aplicación del Teorema de la Velocidad de Datos (*Data Rate Theorem*) ver: WALLACE, R. (2018): "Coming Full Circles: Autonomous Weapons", en R. Wallace (ed.), *Carl von Clausewitz, the Fog-of-War, and the AI Revolution The Real World Is Not A Game Of Go*, Springer International Publishing AG, Cham-Suiza, 73-78.

dependiendo del contexto también compartiríamos, aunque sería posible que un futuro y dependiendo de los avances de la IA o el uso de AMA efectivos, posiblemente a través de la computación cuántica, se lograse una contrapartida aceptable entre necesidad militar y protección civil con relación a dicho principio (Wagner, 2014: 1398-1399).

Los principios de "rendición de cuentas" (*accountability*) y "responsabilidad (*responsability*) son conceptos clave cuando se analizan los SAA. El GGE de las NU sobre los SAAL del CCW, al establecer los posibles principios rectores especificaba que: "El ser humano debe mantener la responsabilidad por las decisiones que se adopten sobre el uso de los sistemas de armas, ya que la obligación de rendir cuentas no puede transferirse a las máquinas. Esta consideración debería tenerse en cuenta durante todo el ciclo de vida del sistema de armas". No obstante, según exponen diversos investigadores como I. Verdiesen *et al*, P. M. Asaro o M. Wagner, entre otros, los SAA podrían establecer un vacío en relación con la rendición de cuentas, circunstancias en donde ningún humano fuese culpable de las decisiones, acciones o efectos de dichos sistemas armamentísticos. Idea que partiría de la premisa de que la comunidad internacional sería incapaz de verificar la legalidad de la acción, ni confirmar la autenticidad de la inteligencia utilizada en el proceso de "focalización" de los objetivos, lo que supondría como resultado la impunidad de dicha acción (Asaro, 2012: 693; NU, 2018a: 4; Verdiesen, *et al,* 2021: 138, 145; Wagner, 2014: 1371).

Particularmente la rendición de cuentas estaría íntimamente ligada al concepto de control. Para M. Bovens, la rendición de cuentas sería una forma de control, aunque no todas la formas de control serían mecanismos de rendición de cuentas. Tomando como base la perspectiva sociotécnica del concepto de control, se estaría ante la premisa de que un agente controlaría a otro agente (que podría ser humano o un artefacto) a través de normas legales, sanciones o instrucciones políticas y estaría intrínsicamente ligado a alcanzar objetivos sociales compartidos y al concepto de rendición de cuentas. Un control que podría ser *ex ante,* activo o *ex post*. Ahora bien, los desarrolladores de los algoritmos de los SAA no dispondrían, en la actualidad, de un marco estructural de referencia adecuado y, por lo tanto, según argumentan I. Verdiesen *et al,* sería necesario el establecimiento de un entorno de gobernanza adecuado, a través de instituciones o foros que supervisasen el comportamiento de los agentes para gobernar sus actividades o el desarrollo de nuevas herramientas de control, como el caso de los AMA en un marco *ex post*. Dicha situación habría llevado a desarrollar la noción del "Control Humano Significativo" (*Meaningful Human Control* (MHC)) en el debate sobre los SAA (Bovens, 2007: 454; Verdiesen *et al,* 2021: 147-148).

Fue la ONG "Article 36" la primera en acuñar el término MHC, en su aportación al debate sobre los SAAL en el foro del CCW de 2014, En el 2016, la propia ONG presentó en el mismo foro de debate una serie de elementos que consideraba necesarios para que existiese un MHC amplio de los SAA: que la tecnología fuese predecible y transparente; que el usuario tuviese una información precisa; que existiese la posibilidad y la capacidad de una inter-

vención humana oportuna y; que también hubiese alguna forma de rendición de cuentas. Además, el MHC debería estar integrado a través de algoritmos durante todo el ciclo del uso del SAA en un conflicto: *ante bellum, in bello* y *post bellum*, así como en todas las fases de combate: táctica, operacional y estratégica. Una postura similar, sobre la necesidad de un MHC apropiado, también fue argumentada por otras ONG's como Amnistía Internacional o Human Rights Watch (Chengeta, 2017: 855-856; Article 36, 2014; 2016; HRW, 2016).

No obstante, el problema de fondo surge por la inexistencia de una definición consensuada del significado del término MHC en los foros internacionales, como en el CCW. Para M. Ekelhof, lo que realmente parecería crucial sería salir del estancamiento de una definición abstracta y comprender lo que dicho control significaría en el contexto de las operaciones militares. En dicho contexto no solo sería necesario un control a nivel operativo, sino que sería crucial que dicho control se ejerciese en todas las fases, incluido el desarrollo, la formación y el despliegue de los SAA, por lo que el MHC solo tendría significado si se ejerciese dentro del marco general del control distribuido que enmarca a cualquier sistema armamentístico. En dicho contexto, I. Verdiesen *et al* propusieron un "Marco Integrado de Supervisión Humana" (*Comprehensive Human Oversight Framework*) que pretende analizar las conexiones entre las diversas fases e identificar las posibles lagunas en los mecanismos de control (ver fig. 19) (Ekelhof, 2019: 347; Verdiesen *et al,* 2021: 151).

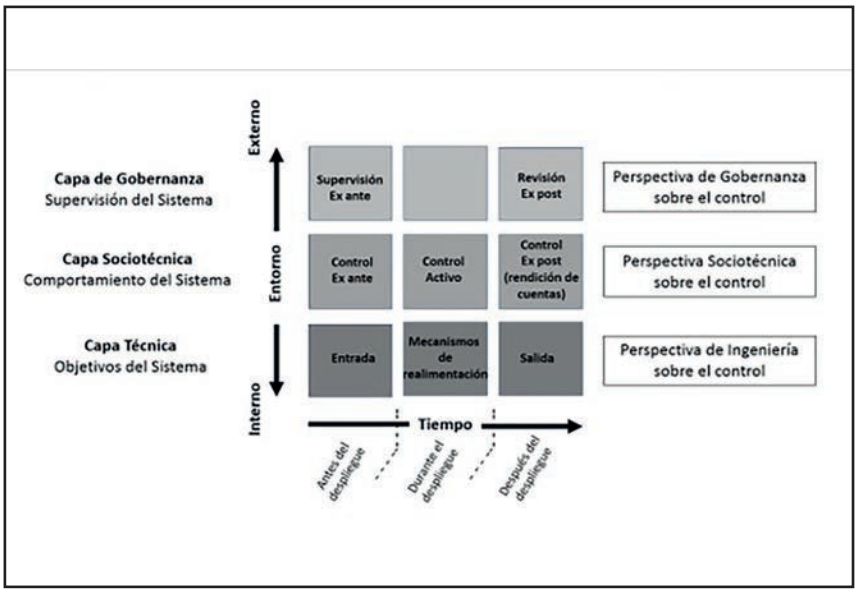


Figura 19: Esquema del Marco Integrado de Supervisión Humana, adaptado de I. Verdiesen *et al* (Verdiesen *et al*, 2021: 151)

- La capa técnica describiría las condiciones técnicas necesarias para que un SAA se mantuviese bajo control. El sistema debería recibir la información adecuada (entrada) y los mecanismos de realimentación deberían ser lo suficientemente robustos para poder verificar la diferencias entre la salida y los objetivos previstos durante el desarrollo. Después del despliegue sería posible técnicamente verificar y comprender la salida y los procesos detrás de ello.

- La capa sociotécnica describiría las condiciones psicológicas y motivadoras requeridas para que el sistema se mantuviese bajo control. En el control *ex ante*, el operador debería establecer las medidas de control adecuadas antes del despliegue y apreciar las capacidades y los límites de los

sistemas. Durante el uso debería ser capaz de llevar a cabo una interacción e intervención significativa del sistema y comprender lo que está haciendo. Después del uso debería ser capaz de inspeccionar y evaluar su comportamiento.

- La capa de gobernanza describiría las condiciones políticas e institucionales y los mecanismos de supervisión requeridos para que el sistema siguiese bajo control. En el periodo ex ante se debería establecer mecanismos políticos e institucionales a través de los foros adecuados (como por ejemplo el CCW), con definiciones claras de cómo ejercer dicha supervisión. Después del despliegue se debería establecer un proceso de revisión y asegurar que los foros establecidos tuviesen la capacidad de establecer mecanismos y sanciones de rendición de cuentas.

No obstante, según I. Verdiesen *et al,* seguiría existiendo una laguna en los mecanismos de supervisión dentro de la capa de gobernanza durante el despliegue de los sistemas SAA, lo que podría incidir en los procesos de control y de revisión ex post después del despliegue de los sistemas armamentísticos. En el caso de los SAA, debido a la idiosincrasia del estamento militar, aparte de la laguna de supervisión en dicha capa de gobernanza, también existiría una laguna dentro del "control activo" de la capa sociotécnica. Como argumentan I Verdiesen *et al*, una autonomía completa de un SAA implicaría que dicho sistema armamentístico tendría autonomía para alcanzar sus objetivos, independientemente del operador humano. Incluso, en aquellos sistemas no completamente autónomos, dependería en si los mecanismos establecidos ex ante o ex post serían los suficientes para dar al operador la capacidad y motivación

para permitir o denegar su uso (Verdiesen *et al,* 2021: 150-153, 158).

Aún con dichas limitaciones, a nuestro entender, dicho marco propuesto reforzaría la noción de un MHC y, por lo tanto, nosotros consideramos que sería una buena base en todo el proceso de desarrollo, despliegue y uso de un SAA y el establecimiento de posibles AMA adecuados, por lo que volveremos a dicho marco de referencia cuando planteemos su posible desarrollo. En todo caso, todavía quedaría pendiente como se establecerían las diversas responsabilidades de los distintos actores (individual, mandos, corporativo, Estado) que, como argumenta T. Chengeta, serían responsabilidades complementarias y no se deberían excluir unas de otras, aunque desde un punto de vista jurídico cada actor sería responsable por sus propias acciones durante las diferentes fases. Por lo tanto, para que un MHC se convirtiese en un estándar legal, se debería especificar para que actores se aplicaría dicho término. A nuestro entender, aunque el MHC fuese específico para cada capa (gobernanza, sociotécnica, técnica), su concepto debería abarcar todo el "Marco Integrado de Supervisión Humana" de una forma holística y no, únicamente, como argumenta T. Chengeta o la ONG "Article 36", a nivel individual del operador (combatiente) como usuario final del SAA. El MHC debería ejercerse en todas las fases del proceso de creación, despliegue, uso y revisión de cualquier sistema armamentístico con IA, para asegurar que cumpliese con el Derecho Internacional vigente y cualquier otra norma, estándar o buenas prácticas desarrolladas, tanto a nivel internacional, de los Estados o de la sociedad (Article 36, 2016; Chengeta, 2017: 868-869).

7.3.- EL IMPACTO DE LOS SAA SOBRE EL DERECHO INTERNACIONAL DE LOS CONFLICTOS ARMADOS (DICA)

En la actualidad, en el marco técnico de los SAA se establece el desarrollo de la fusión de sensores que permiten la imitación de las funciones del cerebro humano, combinando la información de diversos sensores individuales, al mismo tiempo que interactúan con el entorno para, por ejemplo, calcular trayectorias y movimientos. Nuevas capacidades que establecen nuevos retos para la observancia del DIH por los nuevos sistemas armamentísticos. Dos problemas principales podrían surgir: que el SAA fuese incapaz de adherirse, por sí mismo, a lo establecido en el DIH, aunque el objetivo fuese legal o; se evaluasen las funciones de dicho SAA en cuanto a sus medios (*means*), de acuerdo con lo establecido en el Artículo 36 del PAI[123], considerándose no aceptables por incumplir dicho precepto. El problema, surgiría en la forma de analizar los métodos (*methods*) (despliegue y tácticas) que fuesen implementados para un SAA determinado, con relación al cumplimiento del DIH, teniendo en cuenta que dichos métodos tendrían un impacto real en la capacidad militar del armamento. Tanto para el CICR como para los investigadores V. Boulanin, V. Sehrawat o M. Verbrugge, entre otros, lo crucial sería el incorporar el estudio del "método de combate" (*method of warfare*) establecido, como una parte intrínseca en el análisis de un eventual cumplimiento del DIH por un

123 **Artículo 36 – Protocolo Adicional I:** Armas nuevas: Cuando una Alta Parte contratante estudie, desarrolle, adquiera o adopte una nueva arma, o nuevos medios o métodos de guerra, tendrá la obligación de determinar si su empleo, en ciertas condiciones o en todas las circunstancias, estaría prohibido por el presente Protocolo o por cualquier otra norma de derecho internacional aplicable a esa Alta Parte contratante (CICR, 1977a).

SAA (Boulanin y Verbrugge, 2017: 3; CICR, 2006: 4; Sehrawat, 2017: 41-43).

En la actualidad, un importante condicionante, de la aplicación del Artículo 36 del PAI, surge por no existir ninguna estandarización a nivel internacional para llevar a cabo el análisis de dichos métodos, pues cada Estado puede desarrollar su propios procedimientos internos[124]. Tampoco existe un consenso internacional sobre que armamento sería susceptible de ser revisado de acuerdo con dicho artículo, aunque en general se establezca que estaría reservado para aquellos medios de combate destinados a causar daño a personas u objetos. Por ejemplo, el "Manual de Tallinn 2.0" define un armamento cibernético como: los medios de combate utilizados, diseñados o destinados a ser usados para causar daño o muerte a personas y/o daños a objetos. Para G. D. Brown y A. O. Metcalf sería: un artefacto desarrollado u obtenido para un uso primario de matar, mutilar, herir, dañar o destruir. Si se aceptase dicha premisa, se excluirían aquellas herramientas destinadas únicamente a obtener información (Boulanin y Verbrugge, 2017: 9-10; Brown y Metcalf, 2014: 135; OTAN, 2017).

En cuanto a la revisión de un SAA para que fuese compatible con el Artículo 36 del PAI, dejando a un lado la dimensión del nivel de control por parte de un humano sobre un SAA analizado anteriormente, habría que tener en cuenta el nivel de sofisticación del

124 Ejemplos serían: 1) Noruega: *Direktiv om folkerettslig vurdering av vapen, rigforingsmetoder og krigforingsvirkemidler* (Directiva sobre la revisión jurídica sobre Armamento, Medios y Métodos de Combate); 2) Suecia*: Förordning om folkrättslig granskning av vapenproject* (Instrucción sobre la revisión del Derecho Internacional de proyectos armamentísticos) 3) USA: *Policy for Non-Lethal Weapons* (Directiva 3000.3 de 1996) o *Weapons Review* (Departamento de la Fuerza Aérea, instrucción 51-402), entre otros (CICR, 2006: 5).

algoritmo, que determinaría el control sobre su comportamiento y como afrontaría las incertidumbres que surgiesen debido al entorno de uso. En dicho contexto, los SAA podrían ser catalogados en tres categorías (Boulanin y Verbrugge, 2017: 18):

- *Sistemas Reactivos*: El sistema seguiría una serie de reglas usando la metodología condición-acción, que explícitamente prescribiría como un sistema reaccionaría a una información recibida de un sensor. Su comportamiento sería predecible si se conociesen la reglas que utiliza;
- *Sistemas Deliberativos*: El sistema utiliza un modelo del entorno, una función de valor que provee información sobre el objetivo deseado y un conjunto de reglas potenciales que le ayuda a buscar y planificar como llevar a cabo dicho objetivo. Para decidir una acción el sistema compararía las posibles consecuencias de las acciones factibles para encontrar la más apropiada. El comportamiento no sería completamente predecible, dado que, aunque el marco de actividad lo fuese las acciones individuales puede que no;
- *Sistemas de Aprendizaje*: Dichos sistemas podrían mejorar su rendimiento a lo largo del tiempo a través de la experiencia. Aprendería a través de la abstracción de las relaciones estadísticas de los datos. El conocimiento aprendido serviría para volver a parametrizarse automáticamente y reprogramar parcialmente el sistema. En dicho contexto el comportamiento podría ser impredecible si los parámetros de aprendizaje (datos de entrada) no fuesen lo suficientemente conocidos y comprendidos por un operador.

Una tercera dimensión sería la tipología de las decisiones y las funciones que se automatizarían dentro de un SAA (movilidad, selección de objetivos, inteligencia, interoperabilidad, detección de fallos, etc.). Los parámetros de autonomía variarían en cada uno de ellos y las implicaciones jurídicas serían diferentes. Por lo tanto, según establecen M. Boulanin y V. Verbrugge, se debería establecer una lista de comprobación, con dos preguntas fundamentales, para la revisión de los medios y los métodos de combate:

- Con relación a las características técnicas, capacidades y efectos intencionados en condiciones normales de uso: ¿se podría establecer que el SAA sería capaz de cumplir con los preceptos del DIH?;
- Si el SAA pudiese seleccionar y disparar de forma autónoma: ¿en qué circunstancias podría el uso del sistema violar el DIH?

Dependiendo de la respuesta a dichas preguntas, la revisión, de acuerdo con el Artículo 36 del PAI, podría establecer restricciones de uso o proponer recomendaciones de control (hombre-máquina y mando-y-control). Al mismo tiempo, el(los) algoritmo(s) de dicho sistema se debería(n) probar y evaluar, incluyendo la revisión de los mecanismos existentes para minimizar un uso no intencionado o un ciberataque al sistema (Boulanin y Verbrugge, 2017: 20-23; CICR, 2006: 24; Sehrawat, 2017. 48-49).

7.4.- IDENTIFICACIÓN DE LA POSIBLES SOLUCIONES

Existe en la actualidad un intenso debate, entre Estados, investigadores y activistas, sobre si los SAA pueden y/o deben ser regulados y si la respuesta fuese afirmativa como se debería llevar a cabo dicha regulación. Existirían tres posibles alternativas: una prohibición completa o restricción de uso de los SAA, incluidos los SAAL; una integración de dichos sistemas en la sociedad con una regulación institucional, dado que, de todas formas, dichos sistemas serían desarrollados por los Estados o; el establecimiento de medidas de autocontrol y la incorporación de mecanismos de regulación interna, como el desarrollo, despliegue y uso de los AMA.

7.4.1.- ERRADICACIÓN O RESTRICCIÓN DE USO

En octubre de 2012 se formó una alianza de ONG's denominada "Campaña de Prohibición de Robots Asesinos" (*Campaign to Stop Killer Robots*)[125], cuyo objetivo sería trabajar para prohibir los sistemas armamentísticos completamente autónomos y mantener un control humano sobre el uso de la fuerza. La premisa sería que una máquina nunca debería tomar decisiones de vida o muerte y cuestionaría la capacidad de los SAAL completamente autónomos para cumplir el DIH, tanto en relación con los principios de distinción, proporcionalidad, responsabilidad, humanidad o precaución. Más recientemente, en 2020, la ONG Human

125 El comité directivo estaría formado por las principales ONG's a nivel mundial como: Article 36, Human Rights Watch o Amnistia Internacional, entre otras. Para más información ver: STOPKILLERROBOTS (2021): *Campaign to Stop Killer Robots,* acceso febrero 2021, en https://www.stopkillerrobots.org/members/

Rights Watch estableció la necesidad de establecer un Tratado sobre los "Robots Asesinos", en el que se prohibiese el desarrollo, producción y uso de sistemas armamentísticos sin un MHC, así como aquellos que utilizasen datos como: peso, calor o sonido para seleccionar objetivos humanos. En contraste el CICR, en su declaración ante el GGE sobre los SAAL del CCW, de 9 de abril de 2018, abogaría por una obligación positiva de control humano sobre las funciones críticas: la selección y ataque de objetivos (CICR, 2018; HRW, 2020b: 2; PAX, 2018: 6).

A nivel de los Estados, una treintena de ellos abogarían por la prohibición de los SAAL completamente autónomos que no posean un MHC[126]. En el caso de España, no apoya la prohibición total de dichos sistemas armamentísticos, pero reitera que: "todos los sistemas de armas letales dotados de algún grado de autonomía deben contar con un control humano suficiente, con una clara atribución de responsabilidad jurídica al operador de toda arma, así como a la persona que pueda ordenar su uso". Por su parte, el Secretario General de las Naciones Unidas, en su mensaje a la reunión del GGE sobre Tecnologías Emergentes en el área de SAAL, de 2019, reiteró que: "las máquinas con el poder y la discreción de tomar vidas sin una implicación humana son inaceptables políticamente, repugnantes moralmente y deberían ser prohibidas por el Derecho Internacional"

126 Argelia, Argentina, Austria, Bolivia, Brasil, Chile, China, Colombia, Costa Rica, Cuba, Djibouti, Ecuador, Egipto, El Salvador, Ghana, Guatemala, Ciudad del Vaticano, Iraq, Jordania, México, Marruecos, Namibia, Nicaragua, Pakistán, Panamá, Perú, Estado de Palestina, Uganda, Venezuela, y Zimbabue. En el caso de China, aboga por una prohibición de su uso, pero no su desarrollo o producción (HRW, 2020a: 4).

(HRW, 2020a: 4; MAEUEC, 2018; NU, 2019i).

No obstante, existen Estados que no prohíben el uso de SAAL y que, además, establecen una serie de beneficios humanitarios del desarrollo de tecnologías emergentes en dichos sistemas, como sería el caso de los USA o la Federación Rusa. Los USA argumentan que: los armamentos "inteligentes" (*smart*) que usan ordenadores y funciones autónomas para desplegar la fuerza con más precisión y efectividad, habrían demostrado que reducen el riesgo de daño de personas y objetos civiles. Además, dichos sistemas podrían asegurar mejor la rendición de cuentas o llevar a cabo un mejor seguimiento de aquella munición no explotada permitiendo una limpieza post conflicto más efectiva. Por su parte, la Federación Rusa argumenta que: el uso de tecnología altamente automatizada podría asegurar un incremento en la precisión del guiado del armamento sobre objetivos militares, contribuyendo a una menor tasa de ataques no intencionados contra objetivos civiles. También reafirmaría su compromiso de la necesidad de mantener un control humano sobre los SAAL, independientemente de cual fuese el grado de avance de dichos sistemas[127] (NU, 2018g: 2; 2019g: 4; 2019h: 2; 2019j: 6).

127 En dicho contexto, sería importante volver a destacar, en este apartado, el principio rector aprobado por el GGE del CCW sobre los SAAL, de 2019, que establece que: "El ser humano debe mantener la responsabilidad por las decisiones que se adopten sobre el uso de los sistemas de armas, ya que la obligación de rendir cuentas no puede transferirse a las máquinas. Esa consideración debería tenerse en cuenta durante todo el ciclo de vida del sistema de armas" (NU, 2019e: 11).

7.4.2.- INCORPORACIÓN A LA SOCIEDAD

El gran problema que emerge en la actualidad es si el Derecho Internacional existente sería suficiente para el control de los SAA (incluidos los SAAL). En el año 2018, los Estados de Austria, Brasil y Chile recomendaron el inicio de negociaciones para establecer un instrumento jurídicamente vinculante para asegurar un MHC sobre las funciones críticas de los SAAL. Dicha idea habría chocado con la oposición de algunos de los Estados más involucrados en el desarrollo de dicho sistemas, como sería el caso de USA y de la Federación Rusa, así como por Australia, Israel y Reino Unido, por considerarlo prematuro. Dicha oposición habría provocado un punto muerto en las discusiones del GGE del CCW sobre los SAAL, especialmente relevante en 2020 por la ausencia de la Federación Rusa de dichas discusiones, lo que habría llevado al analista D. Lewis a argumentar, que se estaría ante un punto muerto y existirían bastantes dudas sobre el futuro de dichas discusiones, dado que, por un lado, no existiría un consenso sobre definiciones, especialmente sobre que se consideraría un sistema autónomo y sus características técnicas y por otro tampoco quedaría claro si se consideraría deseable o no el establecer un uso y definición estandarizado del término MHC por parte de los diversos actores[128] (HRW, 2020a: 3-4, 9, 29, 51; Lewis, 2020; NU, 2019k; 2019l).

128 Los USA, en la reunión del GGE del CCW sobre los SAAL de 22 de septiembre de 2020, argumenta que el término "control humano" no debería ser el marco para comprender la interacción humano-máquina, dado que no creen como dicho principio ayudaría a mejorar la comprensión de los riesgos y beneficios de los SAAL o para como la tecnología podría ser utilizada para disminuir el sufrimiento durante una guerra (NU, 2020b).

Aparte del foro internacional, existirían también diversas posturas entre los investigadores sobre la posibilidad de desarrollar una normativa específica con respecto a los SAA. Para D. García, sería necesario establecer una "gobernanza de seguridad preventiva" para disipar la incertidumbre y mantener la estabilidad y el orden internacional a través de una nueva codificación de normas globales desarrolladas a partir del Derecho Internacional vigente. En dicho contexto, su opción sería una prohibición preventiva bajo el Derecho Internacional sobre el desarrollo y uso de tecnologías autónomas letales ofensivas, a través del establecimiento de un nuevo tratado internacional específico que prohibiese los SAAL preventivamente, al mismo tiempo que las NU estableciesen una serie de reglas para prevenir la guerra cibernética y los ciberataques. Contrastaría e iría más allá de la propuesta de la ONG Human Rights Watch sobre el desarrollo de un tratado para la prohibición de aquellos SAAL completamente autónomos que no tuviesen un MHC para sus funciones críticas, o la propuesta del "International Panel on the Regulation of Autonomous Weapons (IPRAW)" de la necesidad de establecer el control humano como elemento central en el uso de sistemas armamentísticos con ciertos grados de automatización (García, 2016: 95, 100, 109; HRW, 2020b: 2; IPRAW, 2021: 21).

Otros investigadores, sin embargo, como en el caso de E. Rosert y F. Sauer, en vez de considerar la necesidad de una "prohibición", abogarían por desarrollar una codificación del control humano a través de una "obligación positiva" utilizando el mismo foro existente del CCW, pero pasan-

do del principio tradicional de consenso en sus decisiones al establecimiento de un voto por mayoría. Otro modelo de control se basaría en el concepto de "poder autorizado" (*authorized power*), donde se imbricarían reglas, como las relativas al PAI y el Artículo 36, específicamente, a través del desarrollo de requisitos específicos de contratación pública, que J. Farrant y C. M. Ford argumentan sería más práctico y sustancial, aunque a nivel más técnico. Como tercera propuesta R. Crootof evocaría la necesidad de la implementación del denominado "derecho civil extracontractual de guerra" (*war torts*). Se utilizaría en el caso de violaciones contra el Derecho Internacional que constituyesen una falta, independientemente de que existiese culpa, eludiendo el concepto de culpabilidad moral, desarrollando una asignación de costes para cada incidente. El argumento sería que permitiría la aplicación del Derecho relativo a la responsabilidad del Estado en un conflicto armado, delineando aquellas violaciones especialmente significativas que requiriesen de una reparación. Además, sería una forma de reconocimiento público de que el Estado habría cometido una falta consolidando normas de comportamiento lícito y aseguraría que las victimas recibiesen una compensación (Crootof, 2016: 1386-1388; Farrant y Ford, 2017: 419-420; Rosert y Sauer, 2021: 6).

7.4.3.- AUTOCONTROL Y OTRAS SOLUCIONES ALTERNATIVAS: LOS AMA

El concepto de autocontrol englobaría elementos como declaraciones y recomendaciones, dictámenes, códigos de conducta internos o principios que, sin tener fuerza vinculante obligatoria, podrían ser utilizadas como referentes específicos, lo que en el mundo anglosajón se conocería como "*soft law*". En el ámbito del GGE del CCW sobre los SAAL, dicha postura habría sido claramente auspiciada por Alemania y Francia, en su declaración de 2018, abogando por desarrollar una Declaración Política como instrumento que guiaría los futuros desarrollos de los SAA, en línea con el Derecho Internacional y basado en estándares éticos compartidos. Propuesta que se debería tener en cuenta, según argumentó el representante de España ante dicho foro que incidió en: "apoyar medidas que, con carácter voluntario, procuren una mayor transparencia y confianza, incluyendo el intercambio de información, experiencias y mejores prácticas". Un planteamiento similar de autocontrol serían los principios rectores aprobados por el GGE del CCW sobre los SAAL de 2019, el desarrollo de procedimientos internos de revisión y prueba, incluida la revisión jurídica de los armamentos, para implementar los requerimientos del DIH, como propone USA o las "Directrices Éticas para una IA Fiable", planteadas por el Grupo de Expertos de Alto Nivel sobre IA, creado por la Comisión Europea en 2018 (MAEUEC, 2018; NU, 2018h; 2019a; 2019j).

Otra posible solución alternativa sería el desarrollo, despliegue y uso de AMA. Siguiendo con el análisis realizado en el capítulo 4º, un AMA serviría como mecanismo de regulación que aseguraría un comportamiento apropiado del sistema, rol que sería definido externamente, de acuerdo con las normas de un grupo social determinado. Así, podría configurarse tanto como una "entidad moral" pero también como un elemento práctico para la implementación de normas de autocontrol o de implementación de ciertos elementos del Derecho Internacional vigente. La ventaja de dichos sistemas, introducidos en los SAA, sería doble: por un lado, permitiría un mayor control humano, al proporcionar herramientas para poder aplicar principios como el distinción, proporcionalidad o precaución y, además, podría servir como elemento de control para una posible rendición de cuentas de su uso. En el siguiente capítulo analizaremos con mayor profundidad los desafíos a los que los AMA se enfrentarían si se optase por su desarrollo, despliegue y uso en los SAA.

Al repasar el denso análisis de este capítulo, deberíamos tener siempre presente que los SAA son artefactos que integran IA y que tendrán una autonomía cada vez más creciente, cuyo principal fin, en un conflicto armado, sería la destrucción del adversario. Por lo tanto, lo importante sería discernir el impacto de dicho uso de la fuerza en un entorno de "Realidad Mixta" y la decisión que se debería tomar sobre el tipo de control a ejercer sobre dichos sistemas: un control jurídico o a través de mecanismos de "*soft law*". Particularmente, debido al entorno geopolítico actual, nosotros nos

inclinaríamos en tomar como base los principios rectores del GGE del CCW para los SAAL (necesidad militar, humanidad, distinción, responsabilidad, proporcionalidad y precaución), un entorno de "*soft law*" pero basado en el Derecho Internacional vigente.

Ahora bien, existirían dificultades para discernir como se crean, moldean y definen dichas normas. A nuestro entender, los puntos más relevantes aún no solucionados serían:

- Dentro del DICA, como establecer el equilibrio entre la necesidad militar y la humanidad;
- Como aplicar el principio de distinción, dentro de la nebulosa de los objetos de uso dual;
- Si es posible poner de acuerdo a los Estados, en un futuro próximo, sobre la definición de lo que se considera sistema "autónomo" o "semiautónomo";
- Si es posible evaluar, dependiendo del grado de independencia de los SAAL, cuales serían sus efectos en todo momento: diseño, uso o dependiendo de la vulnerabilidad de los afectados;
- Con relación al principio de proporcionalidad, como se define el término "excesivo" del Art. 51(5)b del PAI (*ius in bello*), como se cuantifica el daño no desproporcionado de la "guerra justa" en el *ius ad bellum* o como se podría cuantificar un hipotético daño futuro en el *ius post bellum*;
- Como establecer un método de control adecuado de supervisión del desarrollo de nuevos SAAL, para que estén de

acuerdo con el DIH, según el Art. 36 del PAI;

- ¿Es posible establecer un sistema de "rendición de cuentas" *ex ante* y *ex post* adecuados?

Nuestra idea es que, en la actualidad, existen aún importantes desafíos para que un SAAL observe los distintos principios consensuados en la NU:

- Al igual que los investigadores A. K. Krishan y J. Foy o E. F. Gade, nosotros somos de la opinión que los SAAL actuales no serían capaces de asumir el principio de distinción y por lo tanto serían considerados como sistemas indiscriminados, ya que:
    - Existe un entorno computacional débil: los sensores en la actualidad no tendrían la capacidad de alta discriminación necesarias;
    - El marco de actuación es impredecible: El entorno de combate y la velocidad de los cambios crean un entorno computacional algoritmico inadecuado, necesitando de un entorno más apropiado como el de la computación cuántica;
    - Los algoritmos son débiles: la construcción modular los hace menos predecibles;
    - Los algoritmos son incapaces, en la actualidad, de distinguir entre combatientes y no combatientes

pues sería necesario que tuviesen la capacidad de "juicios de valor".

- Al igual que el Tribunal Penal Internacional sobre la ex Yugoslavia, nuestro punto de vista es que el principio de proporcionalidad existe, pero no se sabe lo que significa o como se aplicaría en un entorno cibernético;
- Existe aún un gran vacío, tanto *ex ante* como *ex post*, sobre cómo aplicar la responsabilidad y la "rendición de cuentas" sobre el uso de los SAAL en un conflicto armado. Es decir, no existe un entorno de gobernanza adecuado;
- No existe harmonización internacional sobre la aplicación del Art. 36 del PAI a los SAAL.

Ahora bien, desde un punto de vista pragmático, debido al panorama geopolítico actual, la prohibición de los SAAL tiene poco recorrido[129]. Por lo tanto, abogamos para que exista un "Control Humano Significativo" (MHC) y que se llegue a un acuerdo sobre su definición, para pasar de lo abstracto a lo concreto. Para ello, al igual que M. Ekelhoff, proponemos la creación de un "Marco Integrado de Supervisión Humana" a lo largo de todas las fases de un SAAL: creación, despliegue, uso y revisión. Es más, creemos

129 Se puede citar como ejemplos de la situación actual tanto el informe final de la Comisión de Seguridad Nacional sobre IA (*National Security Commission on AI (NSCAI))* de los USA, de 2021, que aboga para que los USA y sus aliados rechacen los llamamiento para una prohibición global de los SAA basados en la IA (NSCAI, 2021), así como el informe sobre los avances de China en armamento con IA: "*AI Weapon's in China's Military Innovation"* (Kania, 2020; NSCAI, 2021).

necesario que la aplicación del Art. 36 del PAI por los Estados debería plantear dos preguntas fundamentales antes de autorizar su despliegue: ¿el SAAL es capaz de cumplir con el DIH? y ¿el algoritmo de diseño viola el DIH? Por último, dado que los Estados abogan por medidas de carácter voluntario, quizás sería el momento de analizar la posibilidad de usar AMA como herramientas para paliar parte de las deficiencias encontradas.

CAPÍTULO 8

DESAFÍOS DE LOS AGENTES MORALES ARTIFICIALES (AMA)

Como argumentan D. G. Johnson y K. W. Miller, la tecnología estaría en parte socialmente construida y, por lo tanto, se le otorgaría un significado al mismo tiempo que se desarrolla. En dicho contexto, la pregunta a realizar sería: ¿cómo se podrían conceptualizar los sistemas informáticos que tienen un comportamiento independiente? En general se podrían seguir dos caminos: el desarrollo de modelos computacionales que intentasen capturar la realidad o; diseñar sistemas que integrasen valores morales preestablecidos. En ambos casos su desarrollo y despliegue sería llevado a cabo por humanos. El desafío vendría en el momento en que dichos sistemas desarrollasen capacidades cognitivas similares a las humanas, el alcance de la Singularidad, lo que para J. Sullins implicaría que se les debería otorgar a dichos sistemas la misma consideración moral que a los agentes morales humanos. En tal caso, al desarrollar sistemas cada vez más autónomos, sería cada vez más difícil el control humano, dado que al interaccionar con el entorno podrían modificar su comportamiento haciéndolo moralmente injustificado dentro de la sociedad que los desarrolló. Esto implicaría, como gran desafío, la necesidad de que los diseñadores estableciesen algoritmos que limitasen la capacidad de daño de dichos sistemas al

funcionar de manera autónoma, pues, como argumenta R. Picard, cuanto más libertad tenga la máquina más será necesario que tenga estándares morales (Johnson y Miller, 2008: 124-126; Picard, 1997: 9; Sullins, 2005: 145-146).

En el desarrollo de los AMA existirían dos grandes desafíos:

- Cómo decidir qué normas, reglas o principios morales a utilizar, así como el método computacional para implementarlos;
- Cómo establecer límites adecuados sobre los análisis llevados a cabo por dichos algoritmos antes de que tomasen una decisión moral en un determinado contexto.

La segunda premisa agruparía una serie de desafíos relacionados: ¿Cómo reconocería un AMA que está ante una situación éticamente significativa?; ¿Cómo podría discernir entre información esencial y no esencial?; ¿Cómo estimaría dicho algoritmo lo que se consideraría como información inicial suficiente?; ¿Qué capacidades requeriría un AMA para realizar un juicio válido en situaciones complejas, por ejemplo, en relación con el principio de distinción? (Wallach y Allen, 2013: 129)

8.1.- ENTRE TEORÍA, FICCIÓN Y REALIDAD

Como ya analizamos al desarrollar el concepto de los AMA, los desafíos teóricos para el desarrollo de dichos algoritmos estarían relacionados tanto con la problemática de la teoría moral a aplicar,

como con los límites computacionales existentes para aplicar dichas teorías. No volveremos a entrar en el debate de si una máquina pudiese llegar a ser considerada un agente moral, ampliamente analizado con anterioridad, manteniendo nuestra postura pragmática y funcional del AMA como mecanismo artificial de regulación social cuya pretensión sería maximizar aquellas acciones consideradas buenas, mientras que, simultáneamente, se minimizasen las consideradas malas por un grupo social, contextualizándolo en un espacio-tiempo cultural determinado. En dicho contexto, uno de los grandes desafíos a conseguir, sería la posibilidad de disminuir o eliminar la posibles inconsistencias de acción que pudiesen ser llevadas a cabo para un mismo tipo de objetivo, especialmente en el marco del Derecho, por una posible carencia de normas jurídicas o por diferencias e inconsistencias entre las normas ya existentes que llevarían a posibles soluciones distintas para una misma situación (Allen *et al,* 2000: 251; Capurro, 2019: 132-134).

Esto no quiere decir que no se haya intentado, de una forma teórica, desarrollar principios de gobernanza ética para la IA que podrían servir de base para el desarrollo de los AMA, como los desarrollados por la IEEE y la ACM o los principios rectores del GGE del CCW sobre los SAAL. Ahora bien, la mayor parte de dichos planteamientos teóricos chocan con su posible implementación, ya que la mayor parte serían instrumentos de Derecho indicativo no vinculante (*soft law*). Es más, como ya hemos analizado al exponer la situación de los foros internacionales, como en el caso del GGE sobre los SAAL o la posibilidad de un tratado internacional sobre la ciberseguridad, el desarrollo de nuevas normativas internacionales, se antojan extremadamente difíciles en la actualidad.

En el caso de que se hubiese solventado dicho obstáculo, el segundo desafío surgiría de los límites computacionales existentes para aplicar la posible teoría moral o las normas jurídicas elegidas. Los investigadores C. Allen *et al* argumentan que las técnicas asociativas, basadas en datos binarios, no producirían modelos satisfactorios de agencia moral, pues carecerían de la capacidad de razonamiento moral abstracto. Otro posible desafío, como argumenta J. Danaher, sería como neutralizar el desarrollo de una posible autocracia algorítmica (*algocracy*), que se relacionaría con posibles problemas de ocultación (de la forma en que se coleccionan y se utilizan los datos por un sistema) o de opacidad (de la base racional e intelectual de dichos sistemas). El posible desarrollo de sistemas de gobernanza algorítmicas, podrían parecer adecuados por su mayor velocidad, precisión y visión, pero se podrían llegar a establecer sistemas extremadamente opacos (Allen *et al,* 2000: 259; Danaher, 2016: 249).

Sin embargo, se siguen desarrollando propuestas computacionales que intentan soslayar los problemas de razonamiento abstracto, como la utilización de enfoques de ecología moral para la ética de las máquinas. Así, en el campo de la tecnología médica se han desarrollado ecosistemas evolutivos[130], como en el caso del sistema "*Global Cardiovascular Risk (GCVR)*" basado en una serie de parámetros y puntuación o en el mundo financiero desarrollando los denominados "mercados predictivos", basados en la IA. En el entorno de los vehículos autónomos (VA) se han propuesto modelos de mitigación de riesgos que se basan más en el desarrollo de algoritmos basados en actitudes morales prevalentes que en teorías

130 Para más información ver: HARMS, W., DANIELSON, P. y MacDONALD, C. (1999): *Evolving Artificial Moral Ecology,* The Centre of Applied Ethics, University of British Columbia, Vancouver.

morales. Así K. Evans *et al*, proponen una arquitectura flexible capaz de acomodar una amplia gama de políticas éticas, sin necesidad de establecer una de ellas como la correcta, la denominada: "Teoría Ética de Valencia" (*Ethical Valence Theory*), una estructura que intentaría capturar la contribución de la ética normativa en el proceso de decisión de un VA. Así, los requerimientos morales de dicho algoritmo, con relación a la moralidad a llevar a cabo en escenarios críticos, analizaría las posibles fluctuaciones del bienestar de todos los actores y como les afectaría lo correcto o lo incorrecto de una posible acción del VA. Dicho análisis se utilizaría como base para tomar una decisión, con el objetivo de maximizar la mitigación de riesgos. Análisis que podría también tener imbricado ciertas restricciones normativas específicas que se podrían incluir en dichos sistemas (Evans *et al,* 2020: 3289; Santos-Lang, 2014; Srivastava y Sengupta, 2017).

En todo caso, el desafío de dichas propuestas computacionales vendría cuando cualquier posible solución propuesta infringiese algún principio ético aceptado e imperativamente se tuviese que tomar una decisión. Para intentar soslayar dicha situación se han desarrollado diversas propuestas a través de herramientas de software. Así, M. Anderson y S. L. Anderson elaboraron un analizador de dilemas éticos denominado "*GenEth*". Se basaría en una serie de esquemas de representación, que incluirían una serie de elementos: características; obligaciones; acciones; casos y; principios, a través de una interfaz de usuario gráfica, que utilizaría una programación lógica inductiva para inferir principios de acciones éticas. Por su parte, el MIT desarrolló el proyecto denominado "*Moral Machine*" (Máquina Moral), cuyo foco serían los VA. El proyecto estaría basado en la participación de humanos ante diversos dilemas éticos de los VA, en las que deberían seleccionar que resultado

desearían que ocurriese. Con una alta participación (3 millones de personas), dicho método obtuvo el resultado de que un VA debería hacer sacrificios propios (por ejemplo, matando al pasajero), si dicha acción salvase vidas. A nuestro entender, ambas propuestas serían difíciles de aplicar, por ejemplo, en entornos de combate fluidos, donde la rapidez en la decisión podría significar la victoria o la derrota en un enfrentamiento (Anderson y Anderson, 2014; MIT, 2017).

El desafío sería, por tanto, como soslayar el denominado "Problema del Marco" (*Frame Problem*). Para los investigadores de IA sería el poder representar los efectos de una acción en lógica computacional sin tener que representar explícitamente un gran número de soluciones obviamente no resolutivas. Es decir, si sería posible, en principio, limitar el ámbito de razonamiento necesario para encontrar las consecuencias de una acción. En el marco de un AMA sería la posibilidad de afinar la información relevante necesaria para tomar una decisión en un contexto determinado. Como argumenta M. Klincewicz, eso sería muy fácil de hacer para un humano, pero muy difícil para un ordenador, especialmente en relación con acciones de combate en un conflicto armado.

El problema radicaría en que sería imposible limitar el espacio de posibilidades futuras solo a las relevantes, ya que en un entorno de combate dichas posibilidades no estarían predefinidas y no se conocerían que inferencias serían relevantes. Pero también existiría el denominado "Problema de Representación" (*Representation Problem*). Para un AMA sería difícil establecer la diferencia entre combatiente y no combatiente, el problema de la distinción ya analizado, como, por ejemplo, si dichos combatientes se camuflasen como civiles. Dichas situaciones necesitarían de unos algoritmos

complejos de perfeccionar, siendo una posibilidad el desarrollo de AMA con modelos híbridos que incluyesen tanto concepciones de teoría ética (*top down*) como de la práctica de uso (*bottom up*), como también propondría dicho investigador (Klincewicz, 2015: 165-167; Wallach y Allen, 2013: 129).

Tampoco se debería olvidar que cualquier algoritmo se construye a partir de valores y que los parámetros operacionales serían especificados por los desarrolladores y configurados por los usuarios para obtener determinados resultados, por lo que algunos parámetros serían privilegiados sobre otros, estableciéndose posibles sesgos de comportamiento[131]. Como ya se destacó con anterioridad, la operación dentro de unos parámetros aceptados no tendría por qué justificar un comportamiento ético aceptable. Por lo tanto, el desafío estaría en determinar los posibles sesgos existentes en los algoritmos para poder establecer el impacto ético potencial y actual del mismo. Como argumentan B. D. Mittelstadt *et al,* aquellos algoritmos basados en el aprendizaje podrían introducir incertidumbre sobre la forma y la razón por la que se toman las decisiones y sería difícil establecer si una decisión problemática fue casual o un problema sistémico del algoritmo por razones de sesgos introducidos en el mismo. El creciente desfase entre diseño y operación de los posibles AMA desarrollados a través de modelos "*bottom up*" podrían tener implicaciones éticas futuras con consecuencias severas para los actores involucrados en su uso (Mittelstadt *et al,* 2016: 2),

131 Para un desarrollo en detalle del impacto de los "sesgos de comportamiento" (*bias*) de los algoritmos con relación a SAA ver la publicación de UNIDIR *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies* (UNIDIR, 2021).

8.2.- LA ÉTICA DE LA CONFIANZA Y LOS AMA

El concepto de la confianza es a la vez un elemento esencial de las interacciones al mismo tiempo que peligroso pues significa que se deben tomar riesgos. En el campo de la ética computacional, las necesidades de confianza se establecerían, según H. C. Lim *et al*, en dos dimensiones (Lim *et al,* 2010):

- *El humano hacia la máquina:* Habría que establecer las razones adecuadas para confiar en máquinas autónomas cuyas acciones conllevasen implicaciones éticas;
- *La máquina hacia el humano*: Habría que definir el alcance y las fronteras de confianza ética con las que se desearía que se construyesen e interactuasen las máquinas con los humanos.

En ambas dimensiones la construcción de la confianza sobre la IA y más concretamente sobre los AMA, relativos al entorno militar, requeriría una multiplicidad de enfoques, desde los sistemas individuales y los ámbitos de aplicación hasta los relativos a los niveles institucionales. En dicho contexto, una de las claves necesarias pero no suficientes, como argumentan A. Winfield y M. Jirotka, sería la necesidad de que existiese una gobernanza ética que inculcase comportamientos éticos en los diseñadores individuales y en las organizaciones para las que trabajan. Una forma de afrontar los asuntos éticos desde el principio y no cuando surgiesen los problemas de una forma *ad hoc*, que conllevaría, como ya analizamos en

el capítulo 5º, al desarrollo de un marco de IIR de confianza pública y el desarrollo de una reglamentación propia, como habrían propuesto dichos investigadores (Winfield y Jirotka, 2018: 2, 5-6, 8).

Ahora bien, nuestro punto de vista, debido a que en la actualidad la mayor parte de los instrumentos analizados serían de Derecho no vinculante (*soft law*), sería más práctico: abogar por el desarrollo de algoritmos para los AMA con autonomía ajustable. Se les dotaría, como ya fue argumentado por J. A. Cervantes *et al*, de la flexibilidad y fiabilidad necesarias para maximizar el rendimiento del algoritmo, transfiriendo la decisión a los humanos en situaciones imprevisibles de incertidumbre críticas. En dicho contexto, se podría utilizar el marco metodológico de autonomía ajustable propuesto por S. Zieba *et al*, que ya analizamos en profundidad en el capítulo 5º de nuestra investigación. Dicho modelo se construiría a través de un enfoque híbrido, tomando influencias adquiridas tanto de las teorías éticas (*top down*) como de la práctica (*bottom up*), a través de la construcción de una "gramática moral" (reglas morales) basada en la experiencia (Cervantes *et al,* 2020: 502; Zieba *et al,* 2010: 202).

En todo caso, el desarrollo y despliegue de dichos AMA debería abordar una serie de temas problemáticos para poder establecer una confianza adecuada sobre los mismos. Dichos temas estarían divididos, según establecen B. D. Mittelstadt *et al*, en aquellos de naturaleza epistémica y los de naturaleza normativa. En el primer caso nos referiríamos a evidencias inconclusas, inescrutables o infundadas, mientras que en el segundo se estaría ante los resultados injustos y los efectos transformadores, a lo que habría que añadir la trazabilidad de los algoritmos. El siguiente esquema resume dicha problemática (ver fig. 20) (Mittelstadt *e al,* 2016: 4):

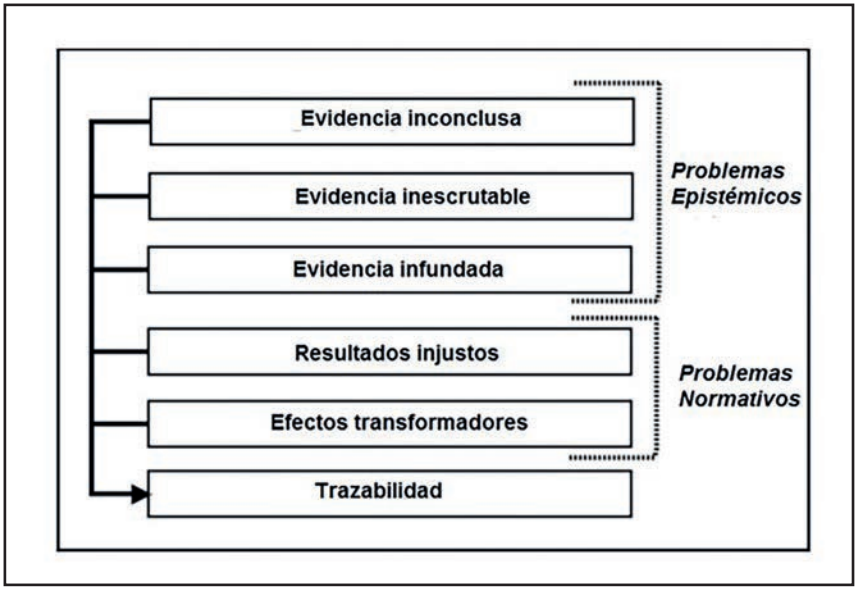


Fig. 20: Esquema de problemática ética de los algoritmos, adaptada de B. D. Mittelstadt et al (Mittelstadt et al, 2016: 4).

La evidencia inconclusa radicaría en la forma en que un algoritmo llegase a conclusiones a través de un proceso estadístico inferencial o técnicas de aprendizaje computacional establecidos a partir de unos datos determinados. Inevitablemente, al ser un conjunto de datos finito que no abarcaría todo el conjunto de datos posible, su uso llevaría a un grado de incertidumbre de los resultados obtenidos, que no serían infalibles. La evidencia inescrutable se establecería por la falta de conocimiento sobre los datos utilizados (alcance, proveniencia y calidad), pero también por la dificultad inherente para poder interpretar como influye cada elemento (sensor), utilizado en un algoritmo de aprendizaje, en la respuesta obtenida. En cuanto a la evidencia infundada, se relacionaría con el principio de "basura entrante, basura saliente", es decir, las con-

clusiones a los que llegase un algoritmo serían tan fiables, como el grado de fiabilidad de los datos de los que se obtiene. Los resultados injustos podrían provenir de los principios morales y éticos utilizados en los algoritmos y, por lo tanto, dependería del grado de imparcialidad establecido por la acción y sus efectos. Los efectos transformadores de los algoritmos estarían motivados por las acciones que transformarían el entorno en el que suceden, aunque aparentemente fuesen neutrales. Por último, la trazabilidad estaría ligada a la dificultad en la identificación de los responsables de una acción que hubiese causado daños.

Con dicha base, para Mittelstadt *et al*, la forma de añadir confianza en los algoritmos (en este caso los AMA), que nosotros subscribiríamos, estaría marcada por reducir la evidencia inconclusa para lograr hacer decrecer las acciones injustificadas, aminorar la evidencia inescrutable para que no existiese opacidad, reducir la evidencia infundada para desterrar los sesgos y los posibles prejuicios, mitigar los resultados injustos para no caer en la discriminación, reducir los efectos transformadores para mantener el algoritmo neutral con el entorno y acrecentar la trazabilidad para mejorar la responsabilidad moral (Mittelstadt *et al,* 2016: 4-10).

Dicho arduo proceso se podría aligerar concentrándose en tres brechas que se acrecientan a causa del uso de los sistemas autónomos: la semántica; la de la responsabilidad y; la de la carga. La brecha semántica representa la diferencia entre las intenciones implícitas de la funcionalidad del sistema y la especificación concreta que se utiliza para crearlo. Las causas serían: la complejidad y la impredecibilidad del entorno operativo; la complejidad y la impredecibilidad del propio sistema y; el aumento en la transferencia de la decisión al sistema. La complejidad del entorno operativo se

reduciría limitando la funcionalidad del sistema a escenarios bien definidos para los que existe una comprensión clara de los riesgos de seguridad y de las capacidades del sistema. Incluiría el uso de sensores paralelos y controles a través de la supervisión, así como una infraestructura sensorial del entorno más robusta. Además, la función de la delegación de la decisión estaría restringida para que el operador humano fuese el responsable último. Dicho proceso sería continuo y no solo en la fase de desarrollo (Burton *et al*, 2020: 1-2, 4, 7).

La brecha de la responsabilidad es una parte integral de la ética práctica y tendría que ver con la responsabilidad moral de aquel actor responsable del daño de un sistema debido a su comportamiento. Para S. Burton *et al,* debería ser una precondición para obtener la confianza pública de un sistema autónomo. Dicha brecha se produciría debido al diseño del sistema a través de dos condiciones posibles: la pérdida del control humano relevante a la acción o la carencia del conocimiento y/o comprensión de las acciones del sistema y sus consecuencias. Los investigadores S. Burton *et al* propondrían mitigar dicha brecha con la utilización del método del "equilibrio reflectivo" (*reflective equilibrium*), un método de ayuda para decidir la mejor forma de proceder en entornos de incertidumbre críticos.

Su utilización podría ser la siguiente: una serie de árbitros morales competentes, que podrían ser expertos en Derecho, con un conocimiento técnico y del entorno suficiente, trabajarían con equipos multidisciplinares de diseño realizando juicios sobre el comportamiento ético de los algoritmos en situaciones críticas. Dichos juicios se enfrentarían a las creencias morales de todos los actores. El proceso continuaría hasta alcanzar un equilibrio entre las opi-

niones de los diversos actores responsables del diseño del sistema. La principal consecuencia es que facilitaría el "consenso" sobre lo que se consideraría como "seguridad aceptable" de un sistema autónomo (en nuestro caso un AMA), aunque debería ser supervisado por algún tipo de organismo regulador[132] (Burton *et al,* 2020: 8-9).

La brecha de la carga emergería al reemplazar el control humano unido a la complejidad del AMA, que dificultaría el establecimiento de la responsabilidad de los diferentes actores sobre su uso. Significaría que, para cualquier perjuicio ocurrido, la carga y las consecuencias de la acción recaerían únicamente sobre la víctima del daño acaecido, dado que los sistemas dotados de IA, al ser artefactos, carecerían de personalidad jurídica propia y no estarían sujetos a los derechos y obligaciones de las leyes. Se podría intentar establecer la responsabilidad en el fabricante del sistema, pero para ello sería necesario probar que el sistema fuese defectuoso. En el caso de que dicho sistema hubiese seguido la reglamentación relevante para dichos productos, sería muy difícil argumentar que el sistema así lo fuese. La forma de mitigar dicha brecha, como argumentan S. Burton *et al,* sería desincentivar el desarrollo de nuevas tecnologías que no fuesen "técnicamente neutrales". Es decir,

132 Una propuesta alternativa, dentro del entorno militar, sería el desarrollo de una herramienta ética de apoyo para la toma de decisiones. Los investigadores G. S. Reed y H. Jones desarrollaron la denominada "Métrica de la Maldad" (*Metric of Evil*), que se podría considerar como un AMA y que consistiría en evaluar pares de acciones militares potenciales en el análisis de las posibles diversas formas de proceder. El modelo tendría como entrada los valores de los factores observables, con relación a la cantidad de daño (mal) asociado a una serie de acciones militares, obteniendo como salida un juicio sobre que alternativa, de las dos posibles, o ninguna sería considerada como la que ofreciese el menor de dos males (Reed y Jones, 2013: 239).

que los riesgos de posibles daños, como resultado de su uso, fuesen similares, en comparación, a los de las tecnologías precedentes. Al mismo tiempo, dichos investigadores propondrían el desarrollo de una fórmula legal para establecer la carga de la responsabilidad de los sistemas autónomos, incluidos los AMA, estableciendo la responsabilidad jurídica propia del artefacto, con el único fin de poder establecer en quién recaería la responsabilidad última de su uso (Burton, *et al,* 2020: 12).

8.3.- UN CONTROL HUMANO SIGNIFICATIVO

Como ya hemos analizado, el concepto MHC se estableció dentro del marco del foro del GGE del CCW de los SAAL, aunque dicho concepto se ha expandido hacia otros entornos. Ahora bien, cuando en nuestra investigación de los AMA nos acercamos a dicho concepto lo que pretendemos es desarrollar algoritmos de IA centrados en el ser humano, intentando conseguir un alto grado de automatización con un alto grado de control, lo que B. Shneiderman califica como aplicaciones informáticas "fiables, seguras y de confianza" (*reliable, safe and trustworthy (RST)*). Una necesidad para algoritmos críticos en los que estarían en juego vidas humanas, como en el caso de sistemas militares, médicos, vehículos autónomos, etc.

La fiabilidad se obtendría a través de buenas prácticas técnicas como la inclusión de:

- Mecanismos de auditoría y herramientas de análisis para revisar posibles deficiencias;
- Pruebas de referencia validadas y verificadas;

- Revisión continua de la calidad de los datos y pruebas anti sesgo;
- Estrategias de diseño que fomentan la confianza de los diversos actores involucrados;
- Interfaces de usuario claras y simples.

También con el desarrollo de las denominadas "culturas de seguridad" a través de estrategias de gestión que impliquen que: el liderazgo esté comprometido con la seguridad; informes abiertos sobre los posibles problemas o la existencia de organismos de supervisión. Por último, con el desarrollo de sistemas de confianza basados en la supervisión por parte de estructuras respetadas independientes, como podría ser el caso del IEEE (Shneiderman, 2020: 495-496, 498).

El desarrollo de los AMA utilizando dichas premisas podría formar la base para su desarrollo, aunque consideremos que se debería mantener una responsabilidad humana directa para todas las acciones llevadas a cabo por cualquier sistema robótico, incluso cuando la persona responsable no controlase directamente algunos aspectos del comportamiento del sistema, al igual que argumentan W. Wallach y C. Allen y que reiteramos en nuestra investigación de manera continua. Un acotamiento o delimitación del uso del algoritmo, que serviría como salvaguarda en aquellas instancias en las que (Wallach y Allen, 2013: 126-127):

- no se conociese con exactitud el entorno en el que operaría dicho AMA y, por tanto, existiesen dificultades para que el

algoritmo pudiese reconocer cuando se estuviese ante una situación éticamente significativa;

- no se comprendiese completamente las rutinas que el sistema requiriese para determinar una acción apropiada.

Esto llevaría a descartar la idea de que el desarrollo de los AMA pudiese significar un alivio sobre el control humano de los sistemas robóticos, incluidos los SAA y los SAAL o de que en algún momento un AMA pudiese ser completamente autónomo. Es más, al incrementarse la complejidad de los sistemas autónomos, los operadores humanos tendrían que poder anticipar cual sería la acción de dicho sistema en nuevas situaciones, para poder coordinar de una forma efectiva sus propias acciones, una tarea que se tornaría cada vez más compleja a la vez que aumentase la complejidad de dichos sistemas, como argumentan W. Wallach y C. Allen y que nosotros suscribimos (Wallach y Allen, 2013: 132-133).

Por lo tanto, cuando nos referimos al concepto de MHC con relación a los AMA, estaríamos ante el imperativo de un control dual humano. Por un lado, el denominado "control de diseño", un control técnico, como las especificaciones técnicas del sistema (hardware y software), que permitiese el control cuando el algoritmo estuviese en uso. Aunque el operador o sus superiores no necesariamente conociesen el sistema a nivel de software, el diseño les debería permitir comprender el por qué dicho algoritmo produciría un resultado específico. Por otro, el "control en uso", se referiría a la posibilidad de una adecuada monitorización del sistema y de su entorno operativo de una forma permanente. Esta premisa sería bastante adecuada si se desea añadir redundancia a un sistema

automatizado, dado que mitigaría los riesgos de los errores de las máquinas (Ecklund, 2019: 18; Mc Coy *et al*, 2019: 4).

Se estaría hablando de las necesidades procedimentales para poder mantener el control sobre el algoritmo en todas las fases: planificación, despliegue y operación. Dicho control sería especialmente relevante cuando se produjesen cambios en el entorno operativo que requiriese una nueva valoración moral. Es más, los controladores humanos deberían tener la posibilidad de anular la acción del algoritmo además de poderlo manipular en cualquier momento. Esto significaría, que en el "control por diseño", los modos de operación deberían permitir la intervención humana, mientras que en el "control en uso", se requeriría que los operadores humanos o sus superiores tuviesen la autoridad necesaria y la posibilidad de que pudiesen rendir cuentas de sus actos, especialmente si el AMA se construyese basado en aspectos del Derecho Penal Internacional (Eklund, 2019: 18).

8.4.- DESAFÍOS DE CONSTRUCCIÓN DE LOS AMA

Los desafíos de construcción de los AMA estarían basados sobre tres dimensiones: los diferentes tipos de teorías morales en los que se pueden basar; los aspectos no técnicos relativos a su implementación y; los posibles modelos técnicos a utilizar. En el primer caso se estaría aduciendo a la ética normativa, que juzgaría una acción a través de una teoría moral en particular. La segunda dimensión se asociaría a la metaética y los conceptos axiomáticos de las otras dos dimensiones y como se aplicarían. Por último, la tercera dimensión se establecería a través de la ética aplicada en ámbitos específicos y tendría como objetivo el lidiar con situaciones de

la vida real (Bonnemains *et al,* 2018: 42; Tolmeijer *et al,* 2020: 132.2).

En cuanto a los tipos de teorías éticas a utilizar, existirían tres posibles enfoques: la ética consecuencialista; la ética deontológica y; la ética de la virtud. Como ya analizamos en profundidad anteriormente en el capítulo 2º, la ética conscuencialista definiría que una acción sería moralmente buena si maximizase el bienestar o la utilidad. La ética deontológica la definiría así si estuviese en línea con una serie de reglas o virtudes morales aplicables. En cuanto a la ética de la virtud, una acción se consideraría moralmente buena si a través de la actuación del agente de una determinada forma, éste manifestase virtudes morales. Existiría además otra teoría ética, como exponen Tolmeijer *et al*, denominada "visión particularista", que estimaría que la vida real sería tan compleja que dependiendo de la situación se utilizaría alguna de las otras teorías o una combinación de ellas, las denominadas teorías híbridas.

La ética deontológica está basada en reglas. Un agente actuaría de acuerdo con una serie de reglas morales ya establecidas. En dicho contexto, el enfoque de la ética del AMA solo estaría basado en los deberes relativos del agente y no existiría distinción entre teorías centradas en el agente o en el sujeto pasivo. En el consecuencialismo lo único importante serían los resultados y el maximizar el bienestar general. También debería seguir aquellas reglas que maximizasen el bienestar y por lo tanto su implementación utilizaría el utilitarismo de acción. La ética de la virtud se estaría centrando en el agente, si exhiben un buen carácter moral y no en las consecuencias o en la consistencia con las reglas. En la ética particularista se utilizaría una u otra teoría según cada contexto. Habría que también indicar que en muchos casos sería de gran utilidad

utilizar una teoría ética hibrida, que podría tener dos vertientes: aquellas propuestas que enforzarían una teoría dominante sobre las demás, lo que establecería una "jerarquía específica" o; aquellas en las que no existiese dicha prevalencia, siendo determinadas "no-específicas". Pero también existirían propuestas denominadas "configurables", donde el desarrollador dejaría al implementador la teoría a utilizar y otras que, a priori, se podrían establecer como "ambiguas", que implementarían aspectos no basados en filosofías morales. En la siguiente tabla se expone una visión general de cada teoría y sus principales desafíos (ver fig. 21) (Tolmeijer *et al*, 2020: 132.8-132.9):

	Entrada	Criterio de Decisión	Mecanismo	Desafíos (ejemplos)
Ética Deontológica	Acción (estados mentales y consecuencias)	Reglas/deberes	Encaje con la regla	• Reglas en conflicto • Reglas imprecisas
Consecuencialismo	Acción (consecuencias)	Bienestar comparable	Maximización de la utilidad	• Problemas de agregación • Determinación de la utilidad
Ética de la Virtud	Propiedades de los agentes	Virtudes	Instanciación de la virtud	• Virtudes en conflicto • Concreción de las virtudes
Particularismo	Situación (contexto, características, intenciones, consecuencias)	Reglas generales, precedentes, cada situación es única	Encaje con reglas/precedentes	• Ninguna lógica universal • Cada evaluación es única

Fig. 21: Visión general de teorías éticas en el contexto de su implementación en una máquina. Adaptado de Tolmeijer et al (Tolmeijer et al, 2020: 132. 9).

Como podemos observar en la tabla, existirían desafíos inherentes a cada teoría cuando se intentan implementar en la práctica (Tolmeijer *et al*, 2020: 132.9-132.10):

- *Desafíos de la teoría deontológica*: En primer lugar, estaría el determinar que reglas aplicar, lo que implicaría una larga lista de reglas, como sería el caso si se decidiese implementar el DIH. Además, sería esencial establecer el nivel adecuado de detalle para que la implementación fuese un éxito. Si la implementación de las reglas no fuese práctica, el algoritmo sería incapaz de interpretarlas. Un segundo desafío vendría cuando existiese un conflicto entre las diversas reglas, tanto de forma general o en situaciones específicas. Aunque se podrían ponderar dando pesos específicos para cada una de ellas, podría ser difícil determinar el orden de importancia entre las mismas. También habría que tener en cuenta que dichas reglas deberían haber sido ya determinadas antes de su utilización;
- *Desafíos de la teoría consecuencialista*: En primer lugar, sería difícil determinar, a priori, las consecuencias. En situaciones reales todas las posibles consecuencias no se conocerían antes de que un AMA fuese utilizado, debido a la interdependencia causal entre ellas. El segundo desafío se relaciona con la dificultad para cuantificar las consecuencias, pues no sería fácil determinar que significa maximizar el bienestar o la utilidad dado que, dependiendo de la medida a utilizar, se podrían obtener distintos resultados. Por último, podrían existir grandes costes computacionales para llevar a cabo dicha implementación, siendo necesario

utilizar aproximaciones heurísticas que deberían ser verificadas para determinar que la solución alcanzada fuese la deseada inicialmente;

- *Desafíos de la teoría de la virtud*: El establecer que una máquina sea virtuosa sería difícil y la única forma sería que se mimetizase el comportamiento de un humano virtuoso. Ahora bien, quién decidiría que carácter se consideraría virtuoso y como se cuantificaría dicho carácter, posiblemente una posible solución teórica sería implementar una serie de reglas que estableciesen principios de virtud genéricos;

- *Desafíos de la teoría particularista*: Necesitaría tomar en cuenta todo el contexto, lo que significaría que estuviese formado para todos los casos posibles, lo que sería imposible o extrapolar sin utilizar generalidades, lo que sería un gran desafío. Para cada característica, el algoritmo debería reconocer si es relevante y como influenciaría el resultado. Los métodos basados en casos serían los más cercanos a poderse implementar;

- *Desafíos de las teorías híbridas*: En el caso de que no existiese una jerarquía entre las teorías, la interacción entre ellas sería problemática, especialmente el determinar cómo se deberían combinar para obtener un resultado moral adecuado. Además, habría que determinar cómo resolver aquellos conflictos en donde las diversas teorías chocasen. En el caso de existir una jerarquía, habría que determinar cuando el algoritmo utilizaría una u otra teoría, lo que no resulta evidente. Quizás la única solución sería el utilizar la segunda teoría cuando la primera no hubiese obtenido resultados, teniendo en cuenta que dicho resultado podría

entrar en conflicto con las bases éticas de la primera teoría.

La implementación de las diversas teorías, podrían realizarse siguiendo alguno de los enfoques establecidos por C. Allen *et al*: arriba-abajo (*top-down*); abajo-arriba (*bottom-up*) o; híbrido (*hybrid*), que ya hemos analizado en el capítulo 5º. A dichos enfoques V. Bonnemains *et al* añaden otro adicional que denominan "valores personales de sistemas éticos", que alude a la posibilidad de que dos agentes diferentes utilizasen teorías distintas para el mismo objetivo o que un agente permitiese utilizar varias teorías en la implementación del algoritmo, lo que S. Tolmeijer *et al* califican como "consideración de la diversidad". En todo caso, para todas las posibles formas de implementación se necesitaría establecer una serie de parámetros que consideraremos a continuación (Allen *et al,* 2005: 149-154; Bonnemains *et al,* 2018: 45; Tolmeijer *et al,* 2020: 132.12).

Cualquier algoritmo, incluido un AMA, se debería evaluar en virtud de su capacidad para llevar a cabo su propósito. Dichas pruebas necesitarían ser comparadas con alguna base moral que podría venir de algunos de los siguientes orígenes:

- *No-Experto*: Utilización de la moralidad tradicional, con la posibilidad de que dicha moralidad aceptada de forma generalizada podría no constituir opiniones verdaderas o aceptables;
- *Expertos*: Utilización de expertos en ética normativa, con la dificultad de saber elegir a los adecuados;

- *Leyes*: Utilización de las leyes como base moral, con la dificultad de poder implementar una larga lista de normas en un algoritmo computacional de una forma aceptable.

También se podría intentar demostrar que el algoritmo funciona a través de una verificación del propio algoritmo, para comprobar que se ajusta a las especificaciones técnicas establecidas o que su lógica, habiendo establecido una serie de parámetros, funciona como debiera. Otra fórmula podría ser a través de una evaluación informal utilizando escenarios y casos específicos como ejemplos, aunque no cubriesen todos los posibles contextos (Tolmeijer *et al,* 2020: 132.14-132.15).

Un último aspecto estaría relacionado con el tipo de modelo técnico a emplear en el desarrollo del AMA. Para ello habría que dilucidar cual sería la elección dentro de una serie de parámetros esenciales (Tolmeijer *et al,* 2020: 132.16-132.17):

- *Razonamiento lógico*: lógica deductiva; lógica no monotónica; lógica abductiva; lógica deontológica; sistemas basados en reglas; representación del conocimiento; lógica inductiva o; el cálculo de eventos;
- *Razonamiento probabilístico*: Planteamiento Bayesano; modelos de Markov o; Inferencia Estadística;
- *Aprendizaje*: lógica inductiva; árboles de decisión; aprendizaje re enforzado; redes neuronales o; computación evolutiva.

- *Optimización*: Se asignan diferentes valores a diferentes acciones basándose en fórmulas predeterminadas y se utiliza el mejor valor obtenido;
- *Razonamiento basado en casos*: Se evalúa una nueva situación basándose en una colección de casos precedentes similares y sus conclusiones se transfieren al nuevo caso.

Existirían además otros dos parámetros necesarios para la construcción del AMA. Por un lado, el determinar qué tipo de entrada a utilizar: datos de sensores; representación lógica; representación numérica o; representación por lenguaje natural, por otro el determinar a qué nivel de detalle se desea construir el algoritmo: simplemente el desarrollo de la idea; detalles de implementación o; codificación exhaustiva del algoritmo. Todos los parámetros deberían además tener una serie de elementos que facilitasen su implementación como: que hardware a utilizar; una explicación clara del algoritmo; una interfaz de usuario simple; un procesamiento automatizado que no necesite de una manipulación inicial y; un sistema de retroalimentación de la valoración del usuario del algoritmo (Tolmeijer *et al*, 2020: 132.17-132.18).

Se debería establecer especial atención a las complicaciones y dificultades de desarrollo de AMA en los "enjambres" robóticos[133]. Al incrementar el número de robots de un "enjambre" resultará más difícil el diseño de una interacción apropiada hombre-máquina. La cuestión estribaría en si se debería incluir un AMA en cada robot o

133 Una definición práctica de "enjambre" ha sido dada por UNIDIR, que los define como: "*sistemas multi robot en los que los robots coordinan sus acciones para trabajar de manera colectiva en pro de la ejecución de un objetivo*" (UNIDIR, 2020b).

utilizar un AMA único con un comportamiento preprogramado del DIH, que sería adecuado en caso de que algunos de los robots fuesen destruidos durante el despliegue. En particular, se podría establecer que la dirección de alto nivel (intervención sobre objetivos) estuviese bajo un control humano y uno a decisión del "enjambre" (UNIDIR, 2020b).

Resumiendo lo analizado en este capítulo, al igual que el investigador R. Picard, consideramos que cuanto más libertad tenga una máquina más será necesario que tenga unos principios morales a utilizar y un método computacional para llevarlo a cabo. En dicho contexto, los datos formarían una parte esencial de la capacidad de un AMA para llevar a cabo su cometido, que necesitarían de un alto grado de fiabilidad y conocimiento, pues se mantendría la máxima de que "basura entrante, basura saliente". Además, se necesitaría un proceso continuo de control sobre el AMA, para eliminar la opacidad y los sesgos, a la vez que se incrementa la trazabilidad, manteniendo el algoritmo neutral. A tal fin, la utilización de los AMA debería, en un principio, estar restringida a escenarios bien definidos, implementar infraestructuras sensoriales más robustas y, sobre todo, restringir la toma de decisiones para que el operador humano sea el responsable último.

El fin último sería establecer una gobernanza ética robusta para incentivar la confianza en la IA. A tal fin, los algoritmos de los AMA deberían mantener una autonomía ajustable e implementar una metodología híbrida (deontológica y de aprendizaje), para que sean "fiables, seguros y de confianza". Por lo tanto, se debe descartar la idea de que un AMA serviría como alivio del control humano, pues se incrementaría dicho control en dos apartados: el control del diseño, ya que todos los actores necesitarían conocerlo

y; el control de uso a través de una monitorización permanente. En definitiva, una evaluación del AMA según su capacidad para llevar a cabo su cometido.

Como corolario podemos asegurar que, debido al gran número de elementos necesarios para construir un AMA, los desafíos pueden ser extensos y complejos. En los próximos capítulos de nuestra investigación, analizaremos dicha base teórica de construcción aplicada a los SAAL, exponiendo las diversas dificultades para pasar de la teoría a la práctica en el marco actual de desarrollo de los sistemas autónomos, especialmente cuando se intenta combinar la técnica computacional con el Derecho en el desarrollo de dichos algoritmos.

PARTE IV

DESARROLLANDO LOS AMA

CAPÍTULO 9
ENTRE SOFTWARE Y HARDWARE

El diseño es una contribución esencial que define un artefacto, junto con su implementación. Para el diseño de un AMA se necesitarían dos componentes esenciales: una "ética de ingeniería" (*engineering ethics*) que determinaría de qué forma los ingenieros mantendrían un control completo del artefacto, independientemente de su complejidad y autonomía, así como una "ética de la máquina" (*machine ethics*), de cómo diseñar un artefacto inteligente para que su comportamiento autónomo se llevase a cabo de acuerdo con unos estándares éticos. Ambos componentes se podrían aunar a través de los denominados "requerimientos de ingeniería" (*engineering requirements (RE)*), que P. Zave definió como: "la rama de la ingeniería del software interesada en las funciones y restricciones de los sistemas de software, relacionados con objetivos del mundo real. También está relacionada con las especificaciones precisas del comportamiento del software y su evolución con el tiempo y a través de las diversas familias de software" (Crnkovic y Çürüklü, 2012: 62-63; Nusseibeh y Easterbrook, 2000: 4; Zave, 1997: 315).

Sin embargo, el mundo de la ingeniería debería ser cauteloso a la hora de desarrollar un AMA pues, como argumenta J. S. Gordon, existirían dos grandes tipos de errores que se suelen cometer cuando se intenta añadir una dimensión ética a una máquina. El

primero sería el desconocimiento de las cuestiones fundamentales de la ética o el asumir supuestos básicos éticos erróneos como, por ejemplo, sobrestimar la capacidad de un método ético para resolver casos morales complejos. El segundo se relacionaría con cuestiones metodológicas, que incluso suponen desafíos a los expertos éticos porque discreparían en cómo resolverlos. El progreso moral a lo largo del tiempo, como por ejemplo el desarrollo de los derechos humanos universales, no significaría que necesariamente serían aceptados por todo el mundo, por lo que sería imposible de considerar seriamente la posibilidad de que las máquinas alcanzasen una agencia ética completa. Por lo tanto, con el progreso de la IA sería necesario también un progreso en los estándares éticos introducidos en los algoritmos a través de la creación de agentes morales implícitos o explícitos adecuados (Cervantes *et* al, 2020a: 510-511; Gordon, 2020: 142, 144-145).

Para llevar a cabo dicho trabajo utilizaremos como base el borrador estándar establecido por el IEEE (P7000/D5) para abordar las inquietudes éticas en el diseño de procesos modeladores de sistemas de IA. En particular trabajaremos en este capítulo, dentro de la fase de conceptuación, estableciendo el entorno de los valores y requerimientos éticos del AMA, así como el desarrollo de gestión de riesgos a través de un proceso de gestión transparente (ver fig. 22) (IEEE, 2021):

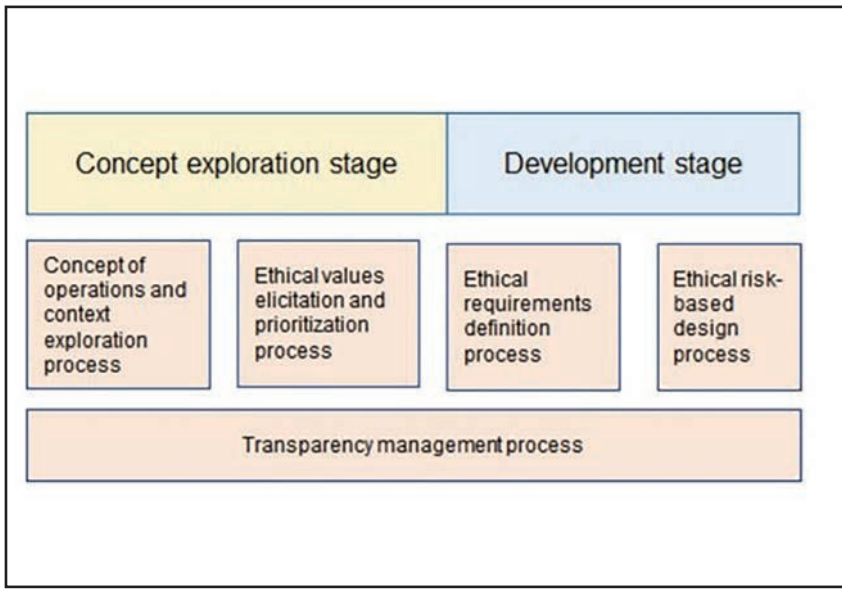


Fig. 22: Esquema descriptivo de procesos y fases del IEEE Estándar 7000 sobre la inclusión de elementos éticos en sistemas de IA (IEEE, 2021).

9.1.- INGENIERÍA, RAZONAMIENTO Y COGNICIÓN

Según el Diccionario de la lengua española, cognición vendría del latín "*cognitio*" y significaría "conocimiento" (la acción de conocer). Por lo tanto, en el dominio de la IA, las "Arquitecturas Cognitivas", serían aquellos modelos de conocimiento relativos a los AA y formarían parte de la investigación sobre la IGA, en la que se pretendería crear AA capaces de incorporar inteligencia como la de los seres humanos. La mayor parte de dichos AA actuarían como un humano o racionalmente, es decir, capaces de producir comportamientos correctos y consistentes para tareas arbitrarias. En el ámbito de los AMA estaríamos, por tanto, ante los denominados "agentes éticos explícitos" de J. H. Moor, aquellos AA que

podrían identificar y procesar información ética sobre una gran variedad de situaciones y llevar a cabo juicios éticos explícitos, pero no necesariamente completos. En todo caso, se diferenciarían de aquellos AA incapaces de diferenciar entre comportamientos éticos y no éticos, pero que en su diseño se establecerían mecanismos de seguridad para impedir comportamientos no éticos, los denominados "agentes éticos implícitos" (Moor, 2006: 19-20; Cervantes *et al,* 2020b: 118-119; RAE, 2020).

Volvemos a incidir, no obstante, en la necesidad de pragmatismo a la hora de desarrollar los AMA, pues en circunstancias prácticas de ingeniería la intencionalidad y la propia voluntad serían difíciles de abordar, como ya observaron G. Dodig-Crnkovic y D. Persson, que nosotros compartimos y analizamos en profundidad en el capítulo 5º de esta investigación. Así, aunque sería importante seguir investigando, como plantean S. Cervantes *et al,* en el diseño de modelos computacionales de los AMA y que desarrollasen arquitecturas cognitivas que incluyesen funciones "morales emocionales", "agencia moral" y "autonomía completa" a medio y largo plazo, a corto plazo sería importante establecer nuevos marcos de confianza a través del desarrollo de estándares y certificaciones internacionales, pues representarían principios y valores éticos de mejores prácticas para el desarrollo de software de calidad. En este punto estaríamos de acuerdo con S. Cervantes *et al,* en la necesidad de que los desarrolladores, en el campo de la ética de las máquinas, necesitarían de estándares éticos explícitos que abordasen claramente las inquietudes éticas para el desarrollo y la pruebas a las que se deberían someter a aquellas arquitecturas cognitivas que se propusieran para los AMA. Estaríamos, por tanto, ante una definición del concepto de razonamiento alejado de su acepción filosófica, para utilizar la definición de la RAE relativa a: "la ac-

ción de ordenar y relacionar ideas para llegar a una conclusión", en este caso, estandarizar los conceptos éticos y relacionarlos para desarrollar algoritmos para los AMA adecuados para cada contexto (Cervantes *et al,* 2020b: 120-121; Dodig-Crnkovic y Persson, 2008: 165-168; RAE, 2020).

De manera generalizada, las formas primarias de implementar valores morales y sociales en una comunidad, se suele llevar a cabo a través de la aplicación de las leyes o a través de controles sociales informales, de la misma forma, la implementación de dichos valores en los AMA se podría llevar a cabo a través del desarrollo de algoritmos que replicasen el Derecho o a través de guías de buenas prácticas éticas no vinculantes, pero socialmente aceptadas. Para A. Etzioni y O. Etzioni dicha distinción sería crítica, pues aquellas normas basadas en la simple implementación del Derecho, tanto Internacional como el de los Estados, implicarían la necesidad de desarrollar algoritmos para los AMA conceptuados de "arriba abajo" (*top down*), no estando sujetos a decisiones o deliberaciones individuales. Las leyes se aplicarían independientemente del algoritmo utilizado, aunque se podrían adaptar a las particularidades técnicas del artefacto (Etzioni y Etzioni, 2017: 412-413).

Ciñéndonos al Derecho Penal Internacional, dentro del marco de los SAA, el desarrollo de AMA, siguiendo dicho criterio, estaría circunscrito, en principio, a la hipotética aplicación del DIH a través de dichos algoritmos. Como argumenta M. Sassóli, nuestra postura sería de escepticismo de que, en la actualidad y en un futuro cercano, se puedan crear algoritmos con la inteligencia contextual necesaria para adaptar de una forma completa el DIH a través de sensores, dada la gran variedad de situaciones que pueden surgir en un escenario bélico a lo largo del tiempo, lo que implicaría que

el algoritmo debería poseer un número ilimitado de escenarios, lo cual sería imposible de determinar. Otra problemática surgiría de la posible interacción de reglas del DIH que pudiesen entrar en conflicto en un contexto determinado, así como las posibles limitaciones de capacidad computacional del hardware necesario para procesar toda la información relevante en un escenario bélico, sin conocer, a priori, que información podría ser relevante o no en un contexto y periodo de tiempo determinado (Sassòli, 2014: 312).

Creemos, por tanto, que dicho enfoque debería ser descartado a corto plazo para explorar la utilización de guías éticas para la creación de algoritmos, que podrían ser tanto públicas como privadas, y que servirían como base de conocimiento para los ingenieros de software. Dichas guías normalmente tendrían un enfoque "híbrido" (*hybrid*), utilizando tanto elementos de desarrollo de "arriba abajo" (*top down*) y de "abajo a arriba", (*bottom up*), pues estarían basados en principios sociales generalmente aceptados y basados, en parte, en el Derecho existente. Recientemente, el investigador T. Hagendorff realizó un estudio de guías éticas, desarrolladas en los últimos años por diversos estamentos públicos y privados, que contendrían principios normativos y recomendaciones para encauzar las posibles repercusiones morales negativas de la IA. Entre ellas destacaríamos: la "Recomendación sobre la Ética de la IA" de las UNESCO; las "Directrices Éticas para una IA Fiable" de la Comisión Europea; el "Informe sobre el futuro de la IA" de los USA y; los "Principios de IA de Beijing", de la Academia de Beijíng de IA, apoyado por el Ministerio de Ciencia y Tecnología de China (BAAI, 2019; Bundy, 2017: 285-286; CE, 2018; Hagendorff, 2020: 101; UNESCO: 2021; USA, 2016).

Para el ámbito militar, con especial impacto en el desarrollo de

AMA par los SAA, creemos que sería también importante añadir, como base, las guías específicamente elaboradas con relación al Derecho Penal Internacional existente y otros principios morales y éticos de posible aplicación. A tal fin, consideramos que los ingenieros de software deberían tener en cuenta los siguientes: los "Manuales de Tallinn" de la OTAN; los principios rectores establecidos por el GGE del CCW sobre los SAAL de las Naciones Unidas y; el "Manual de normas Internacionales que rigen las operaciones militares" del CICR. Tampoco se deberían olvidar aquellas guías específicas de los diversos Estados, que podrían ser importantes para su desarrollo, según el Estado que lo diseñase. Destacaríamos por su especial relevancia los siguientes: el "Law of War Manual" de los USA y en el caso de España, las "Orientaciones. El Derecho de los conflictos armados (Tomo I)" del Ministerio de Defensa (CICR, 2016f; DoD, 2016; Mº Defensa, 2007; NU, 2019a; OTAN, 2013; 2017).

En el ámbito de la ingeniería, habría que subrayar la "Iniciativa Global del IEEE sobre la Ética de los Sistemas Autónomos e Inteligentes". La misión de dicha iniciativa sería la de: "asegurar que cada actor envuelto en el diseño y desarrollo de sistemas autónomos e inteligentes sea educado, formado y habilitado para priorizar consideraciones éticas, para que dichas tecnologías avancen para el beneficio de la humanidad". Una de sus misiones específicas sería el establecer un replanteamiento de los SAA, destacando que las organizaciones técnicas deberían aceptar que un MHC sobre los sistemas armamentísticos sería beneficioso para la sociedad y que el desarrollo de registros de auditoría aseguraría dicho control. Además, argumentarían que sería necesario que aquellos que creasen dichas tecnologías comprendiesen las implicaciones de su trabajo y que se estableciesen códigos profesionales éticos que

abordasen apropiadamente aquellos trabajos destinados a causar daños. Entre los trabajos de dicha Iniciativa Global, destacaríamos el informe denominado "Diseño Alineado Éticamente" (*Ethically Aligned Design*), un marco de diseño que combinaría valores humanos universales, la agencia de datos y la dependencia tecnológica, a través de un conjunto de principios y recomendaciones, cuyo objetivo sería servir de guía para los desarrolladores y usuarios de los sistemas autónomos e inteligentes (ver fig. 23) (IEEE, 2017a: 10; 2019; 2021):

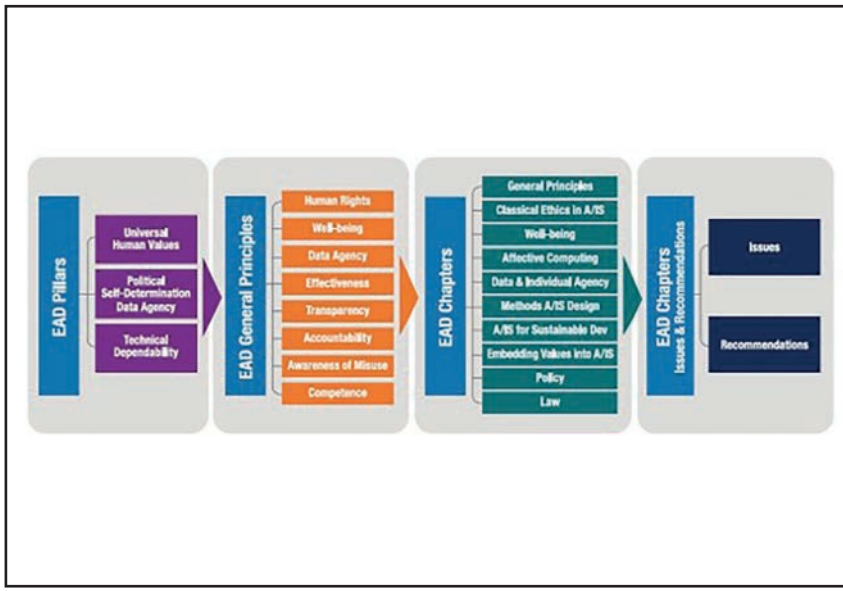


Fig. 23: Esquema del marco conceptual del informe "Diseño Alineado Éticamente" (Ethically Aligned Design) (IEEE, 2019: 15).

9.2.- MODULARIDAD E INTEROPERABILIDAD

El permitir que un AMA sea reutilizado y reconfigurado de una forma modular, tanto con relación al hardware como al software,

podría significar un abaratamiento de los costes de Investigación y Desarrollo con relación a nuevas plataformas de SAA. Implicaría también dotar dicho armamento de una flexibilidad a la hora de configurar sistemas armamentísticos que siguiesen el actual DIH. Por ende permitiría el desarrollo de AMA confiables Consideramos que dicha modularidad e interoperabilidad sería crítica para maximizar su efectivad y una implementación generalizada. A tal fin, como propone la investigadora V. Dignum, sería importante que los AMA siguiesen los principios de: Rendición de Cuentas, Responsabilidad y Transparencia (RTR) (*Accountability, Responsibility and Transparency (ART)*), establecidos para sistemas sociotécnicos de IA. El siguiente esquema lo describe (ver fig. 24) (Dignum, 2020: 7):

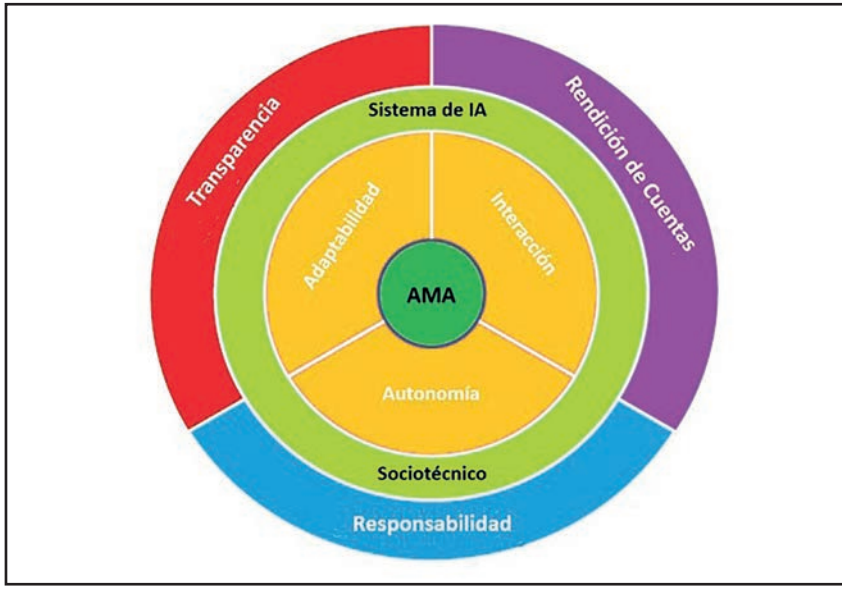


Fig. 24: Esquema de un AMA siguiendo los principios (RTR) adaptado de V. Dignum (Dignum, 2020:7)

- *Transparencia*: Con relación a un AMA sería la capacidad de describir y reproducir los mecanismos a través de los cuales toma las diversas decisiones y se adapta al entorno, incluido la proveniencia y la dinámica de los datos utilizados, los posibles sensores que utiliza y los resultados que obtiene;
- *Responsabilidad*: Incluye al conjunto de actores responsables del diseño, la construcción, el despliegue y el uso del AMA y su rol en cada una de las fases, para desarrollar un sistema que cumpla con los requisitos morales y éticos establecidos;
- *Rendición de cuentas*: La capacidad del AMA para justificar y explicar sus decisiones, a través de una serie de mecanismos como la auditoría, el desarrollo de "cajas negras" que registran sus acciones, pero también la capacidad de mantener un sistema abierto de información sobre sus capacidades y limitaciones.

Una vez establecidos los principios generales del AMA, el siguiente paso sería el desarrollar una visión conceptual de dicho sistema de IA. Dicha visión se podría concebir a través de una estructura de alto nivel. El AMA se podría construir por medio de tres elementos principales: sensores, una lógica operacional y actuadores. Los sensores coleccionan los datos "en bruto" del entorno, los actuadores establecerían las acciones necesarias para actuar en dicho entorno, mientras que la lógica operacional formaría el núcleo operativo que, de acuerdo con los objetivos planteados y basándose en

los datos obtenidos de los sensores, proporcionaría la información para los actuadores (recomendaciones, predicciones o decisiones), que influenciarían el entorno. Dicha lógica operacional estaría formada por un modelo con dos componentes: un proceso de construcción y otro de interpretación. El siguiente esquema, basado en la estructura general, de alto nivel, de un sistema de IA desarrollado por la OCDE, podría servir como base para dicha estructura (ver fig. 25) (OCDE, 2019: 7):

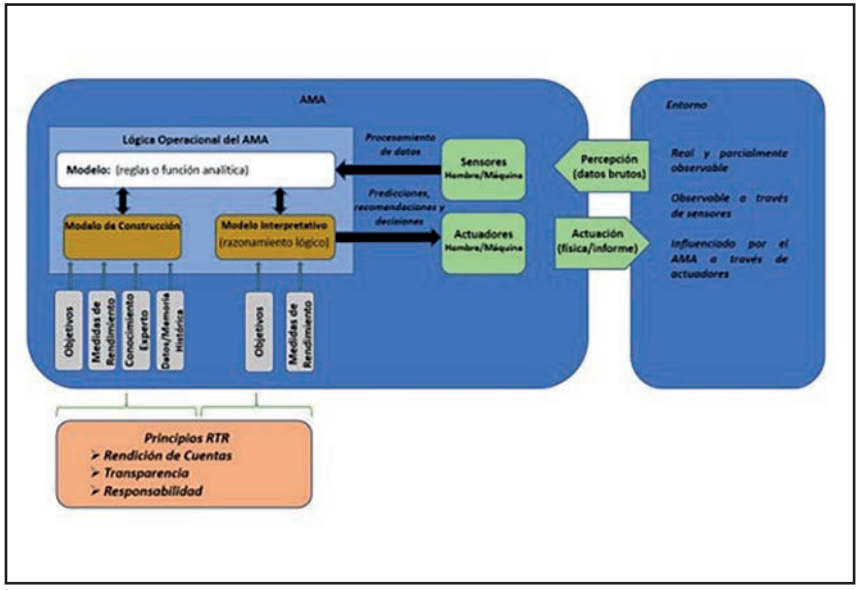


Figura 25: Estructura de alto nivel de un AMA. Adaptado del esquema de un sistema de IA de la OCDE y de los principios RTR (Dignum, 2020: 7; OCDE, 2019: 7).

La construcción del modelo se basaría en datos procesados a través de algoritmos de aprendizaje automático (*machine learning*) o de forma manual. Dichos datos serían agregados utilizando diversas formas: a través de la información recibida de expertos con conocimiento de la información a tratar, de los objetivos a alcanzar, así como otras medidas de rendimiento, como la obtenida de la formación de los operadores, el nivel de eficacia de las pruebas realizadas o del nivel de representatividad del banco de datos utilizado. Para establecer el modelo interpretativo se deberá tener en cuenta los objetivos a alcanzar, así como los resultados de las medidas de rendimiento obtenidas a través de casos de prueba, uso real previo o rendimientos teóricos establecidos. En el caso de que el AMA utilice reglas determinísticas entonces la interpretación conllevará una única recomendación, mientras que si se tratase de modelos probabilísticos el modelo podría ofrecer un abanico de recomendaciones asociadas, de acuerdo con las medidas de rendimiento establecidas como, por ejemplo, el nivel de confianza, el riesgo asumido o la robustez teórica exigida al modelo (OCDE, 2019: 8).

Una vez establecida la estructura básica del AMA, se considera importante establecer las diferentes fases del "ciclo de vida" de dicho sistema de IA. De forma generalizada, al tratarse de un desarrollo de software informático, una gran parte del ciclo de vida de un desarrollo tradicional de programación sería de aplicación, aunque con ciertas características específicas. El siguiente esquema resumiría el "ciclo de vida" teórico de un AMA, siguiendo el esquema establecido por la OCDE para un sistema de IA (ver fig. 26) (OCDE, 2019: 13):

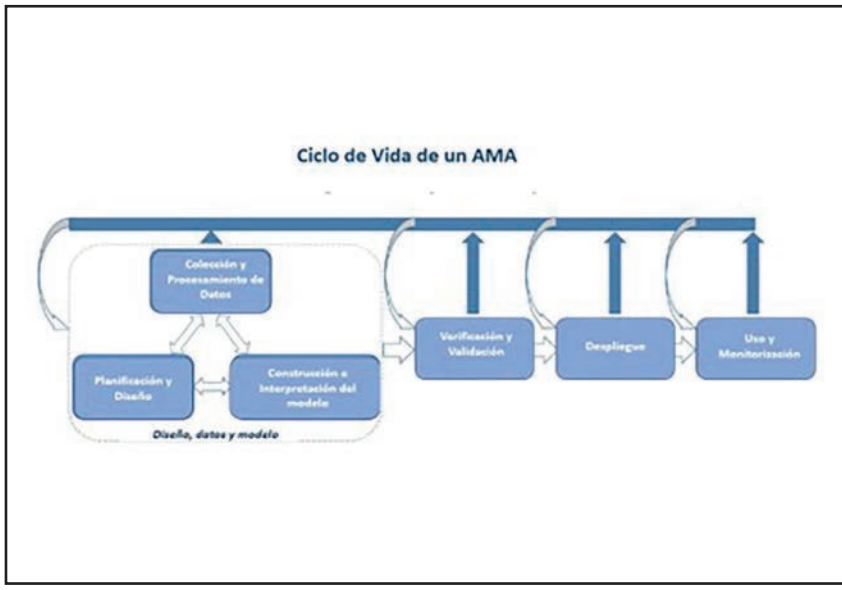


Fig. 26: Esquema del "Ciclo de Vida" de un AMA, adaptado del "Ciclo de Vida" de un Sistema de IA de la OCDE (OCDE, 2019: 13).

Las diversas fases incluirían los siguientes elementos (OCDE, 2019: 13):

- *Diseño, datos y modelización*: incluiría los siguientes elementos:
    - *Planificación y Diseño*: del AMA, incluiría la articulación del concepto y los objetivos, las hipótesis subyacentes, el contexto y sus requerimientos y la posibilidad de construir un prototipo;
    - *Colección y Procesamiento de datos*: recogida y limpieza de los datos, verificación de la calidad e integridad, documentación de las características del

bloque de datos (como se creó, composición, uso previsto, mantenimiento);

- *Construcción e Interpretación del modelo*: creación o selección de los posibles modelos/algoritmos, su calibración y pruebas a las que se le somete y su interpretación.

- *Verificación y validación*: Consiste en ejecutar y afinar el AMA, con pruebas para evaluar su rendimiento a través de diversas dimensiones y condiciones.

- *Despliegue*: producción del AMA, pilotaje, compatibilidad con sistemas precedentes si existiesen, asegurar que cumplen el ordenamiento establecido (ej.: el DIH), gestión de los posibles cambios organizativos, evaluar la experiencia del usuario.

- *Uso y monitorización*: Operatividad del AMA con una monitorización constante del impacto real (deseado y no deseado) de acuerdo con los objetivos y las consideraciones éticas. Dicha monitorización podría llevar a la revisión del sistema en cualquiera de sus fases o incluso llegar a retirar el AMA en caso necesario.

Para cada fase establecida se debería establecer un análisis y gestión de riesgos específico. En todo caso el concepto de riesgo va íntimamente ligado al de seguridad, dado que son dos conceptos opuestos. El riesgo sería la posibilidad de daño y la seguridad relativa la libertad sobre el posible riesgo. En todo caso, la seguridad nunca sería absoluta, dado que ninguna actividad, especialmente la militar, estaría exenta de riesgos. Además, como observan P. Lin *et al,* durante un conflicto la tendencia sería a maximizar el daño

del enemigo al mismo tiempo que se minimiza el propio, lo que complica la gestión de riesgos en el ámbito militar, si al mismo tiempo se debe cumplir con las normas del Derecho Internacional como el DIH. Por lo tanto, sería importante establecer, durante la planificación del diseño de un AMA, el posible nivel de riesgo de un SAA o SAAL, que se consideraría aceptable, ya que influiría en el desarrollo técnico del AMA incorporado a dichos sistemas armamentísticos. Esto implicaría, además, la necesidad de diseño de AMA que tuviesen la necesaria flexibilidad de adaptación a diversos niveles de riesgos, dependiendo de su entorno de actuación (Lin *et al,* 2008: 63-72). El siguiente esquema resume la gestión de riesgos para un AMA, adaptado del esquema de la gestión de riesgos de un sistema de IA propuesto por la OCDE (ver fig. 27) (OCDE, 2019: 16):

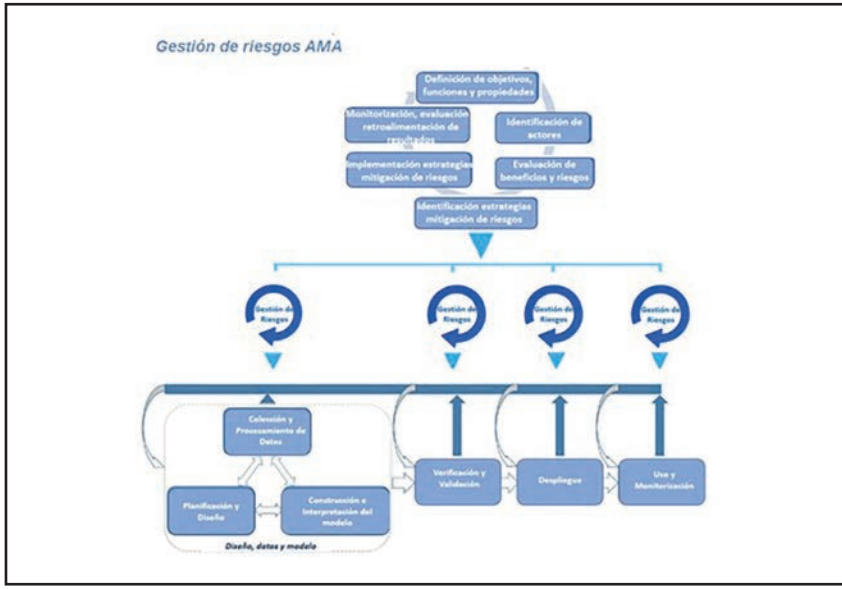


Fig. 27: Esquema de gestión de riesgos de un AMA, adaptado del esquema de gestión de riesgos de un sistema de IA propuesto por la OCDE (OCDE, 2019: 16).

Una de las principales formas de mitigación riesgos, sería incrementar la seguridad del SAA o SAAL en cuestión, pero también del AMA incorporado a un sistema armamentístico. Por lo tanto, dicho agente debería mantener un nivel de rendimiento similar al de un agente moral humano en las mismas circunstancias. Dicho objetivo se podría lograr de diversas formas: mejoras tecnológicas[133]; manteniendo una supervisión humana de su utilización (monitorización)[134] o; ser capaz de imbricar en su diseño de la capacidad de auditoria "a posteriori", esto es, incluir una especie de "caja negra" que registraría todas las acciones de un AMA durante su uso. Además, sería importante estandarizar al máximo el diseño de dichos agentes, pues permitiría una minimización de riesgos, un aumento de su seguridad y un marco conceptual de planificación y diseño estable (Sparrow, 2009: 172).

Otro elemento que se debe explorar para maximizar la seguridad y minimizar el riesgo sería evitar al máximo posible los posibles efectos secundarios del AMA. Un agente que opera en un entorno complejo debería tener en cuenta todos los condicionantes de dicho entorno antes de proceder a minimizar los riesgos de un solo elemento. Es decir, establecer el AMA como un "agente de impacto mínimo" a través de, como argumentan D. Amodei *et al,* el establecimiento de un "regulador de impacto", que controlase y

133 Implicaría la necesidad de que tanto el SAA como el AMA fuesen lo suficientemente fiables para minimizar posibles fallos (ej. fallo de la batería) e instalar componentes (hardware/software) que impidiesen el posible "hackeo" del sistema, incluso llegando a la posibilidad de mecanismos de autodestrucción (Sparrow, 2009: 173).

134 Una posibilidad sería el desarrollo de un sistema de procedimiento para supervisar la seguridad dentro del "ciclo de vida" de un AMA, similar al *Joint Capabilities Integration and Development System (JCIDS)* del DoD de los USA, que garantiza una supervisión inicial en el desarrollo de SAA en el ámbito de la seguridad (Hall, 2017: 90).

penalizase posibles cambios en el entorno. Es decir, comparando el estado del entorno una vez el AMA hubiese actuado, con el estado si no hubiese realizado la acción. Dicho estudio se realizaría para un gran número de casos distintos que servirían como casos de estudio para incrementar la seguridad minimizando los posibles riesgos (Amodei *et al,* 2016: 4-6).

9.3.- PLANIFICACIÓN Y DISEÑO

Una vez establecido un entorno de maximización de la seguridad y minimización de los riesgos, por lo menos a escala del conocimiento de los desarrolladores del AMA, sería el momento de comenzar a establecer los diversos componentes que configurarían el "ciclo de vida" de un AMA. Dentro de la cadena de valor de un algoritmo la fase de planificación y diseño se considera crítica para una correcta implementación posterior. En dicho contexto y ciñéndonos al ámbito militar, dado que nuestra investigación se centra en el posible desarrollo de AMA que tengan en cuenta el Derecho Penal Internacional en el ámbito de los conflictos armados, un primer paso en su desarrollo sería el especificar como paso previo, el modelo de diseño a utilizar. Este diseño de alto nivel tendría como objetivo establecer el entorno de trabajo del AMA, en sus diversos elementos y actores.

Como ya analizamos en el capítulo 5º, consideramos que aquellos algoritmos que tuviesen una autonomía ajustable serían los más capaces de lidiar con situaciones complejas, al poder implementar mecanismos que permitiesen a los humanos, compartir, supervisar e intervenir en su control. Dicha circunstancia sería crítica para los SAA, donde los rápidos cambios en las operaciones de combate

necesitarían que dichos sistemas estuviesen dotados de unos algoritmos éticos flexibles y fiables para maximizar su rendimiento en espacios cortos de tiempo. A tal fin, nuestra propuesta, como ya avanzamos en dicho capítulo, sería adaptar el marco metodológico de autonomía ajustable propuesto por S. Zieba a los AMA de ámbito militar. El siguiente esquema proporciona una propuesta de visión de dicha adaptación para un AMA de un SAA (ver fig. 28 (Zieba *et al*, 2010: 202):

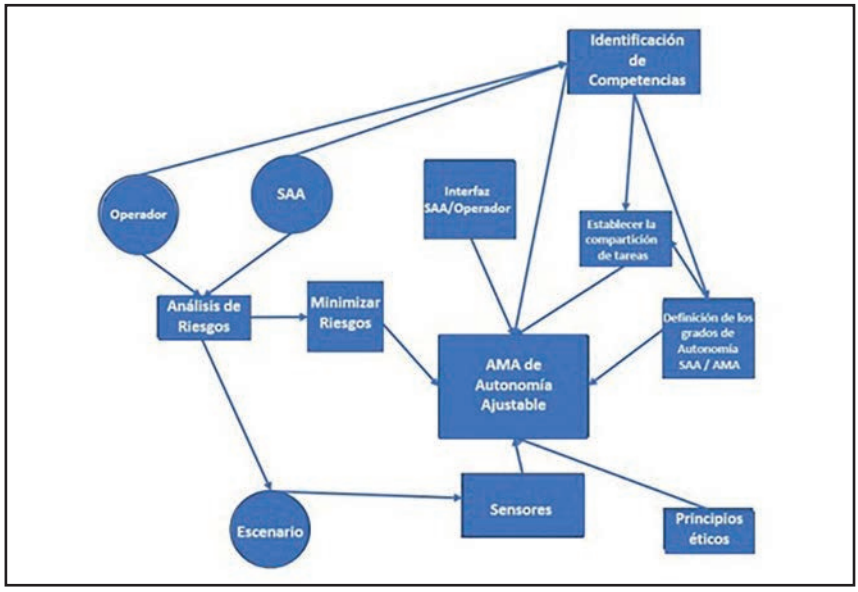


Fig. 28: Modelo de diseño de un AMA para un SAA, adaptado del modelo de autonomía ajustable de S. Zieba (Zieba et al, 2010: 202).

Los diversos elementos del modelo de diseño se descompondrían de la siguiente forma:

- *Identificación de competencias*: Elemento que identificaría a los diversos actores que formarían parte del diseño,

desarrollo, despliegue y uso del AMA. Se podría definir utilizando el esquema de valor propuesto por A. Fritz *et al* ya analizado en el capítulo 5º (ver fig. 29) (Fritz *et al*, 2020: 8):

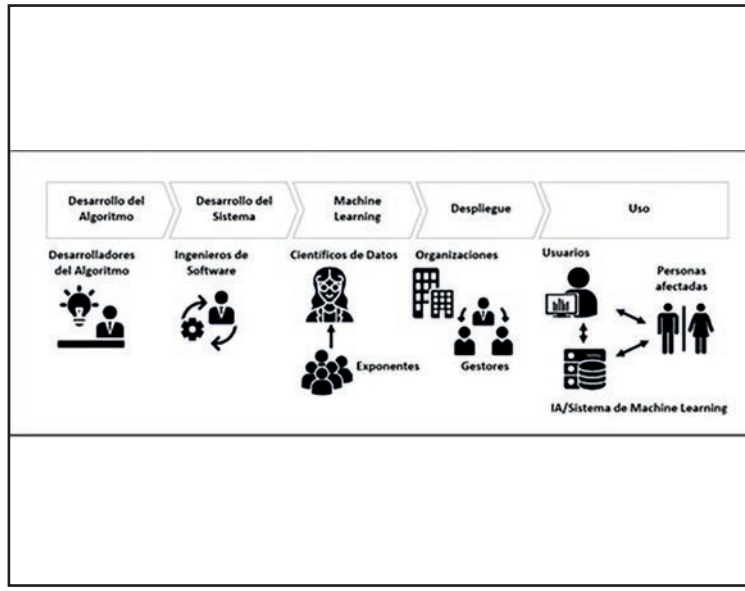


Fig. 29: Identificación de las diversas competencias de los distintos actores que conforman el esquema de valor de un AMA, de acuerdo con la propuesta de A. Fritz para un sistema de IA (Fritz et al, 2020: 8)

Los desarrolladores del algoritmo serían aquellos diseñadores de algoritmos basados en "*machine learning*" de carácter general, como los desarrolladores de redes neuronales. Los ingenieros de software serían los encargados de adaptar dichos algoritmos al sistema de software específico del AMA para el SAA en particular. Los científicos de datos serían los expertos en "*big data*" que "entrenarían"

el algoritmo, con un conjunto de datos obtenidos de los diversos sensores integrados en el SAA. En este caso, la organización sería el estamento militar/civil que autorizaría el despliegue del SAA, en el que estaría integrado el AMA. Por último, el usuario final sería el operador humano del AMA que no necesariamente tendría que ser el mismo que el del SAA.

- *Principios éticos*: Establecimiento de los objetivos éticos que se deben incluir en el AMA. Como se ha analizado anteriormente, sería imposible que todas las normas del Derecho Penal Internacional estuviesen incluidas, por lo que habría que establecer la prioridad en aquellos principios éticos que se considerasen imprescindibles en un AMA para un entorno militar. El investigador R. C. Arkin, en su propuesta de modelo de arquitectura híbrida de un AMA militar, propuso los siguientes principios: necesidad militar; humanidad o sufrimiento innecesario; proporcionalidad y; discriminación o distinción. Principios básicos dentro de las leyes de "focalización de objetivos" (*targeting*), que se corresponderían con diversos artículos del PAI y los objetivos no prohibidos del DICA, como ya analizamos extensamente en el capítulo 4º. Destacaríamos entre ellos, por su importancia, los principios de proporcionalidad y distinción, unidos al de precaución, ya que en los debates del GGE del CCW sobre los SAAL, las principales potencias mundiales (China, Rusia, USA), mostraron la necesidad de que dichos principios fuesen la base del control de los SAAL en la selección y puesta en combate (uso) de dichos sistemas (Arkin, 2007: 23; NU, 2018b; 2019k; 2019j: 2; Titiriga, 2016: 59, 76-78).

- *Escenario*: Caracterización de los posibles campos de batalla y de sus componentes, donde se desplegaría y utilizaría un SAA y consecuentemente el AMA asociado.

- *Sensores*: Existen gran variedad de factores que afectarían a qué tipo de sensores debería incorporar un AMA y que dependería del SAA utilizado. Podríamos incluir los siguientes: el tipo de función del armamento en cuestión; el tipo de objetivo; el tipo de fuerza aplicable; el contexto en el que se utilizaría (escenarios simples o atestados); la facilidad de distinción de objetivos en un contexto particular; la forma de interacción entre el humano y el SAA; la "libertad" del SAA para moverse en el espacio (fijo o móvil, área geográfica amplia o pequeña); el periodo de tiempo de acción (un solo punto y momento o sobre un largo espacio de tiempo) y; la fiabilidad y seguridad incorporadas al SAA. Por lo tanto, cada AMA utilizaría unos sensores específicos de cada SAA o SAAL, aunque algunos sensores podrían servir de forma generalizada, como por ejemplo los relativos al principio de distinción y la utilización de algoritmos de visión artificial y reconocimiento de imágenes.

- *Análisis y minimización de riesgos*: Se alinearía con el principio rector aprobado por el GGE del CCW de los SAAL de 2019: "Las evaluaciones de riesgos y las medidas de mitigación deberían formar parte del ciclo de diseño, desarrollo, ensayo y despliegue de tecnologías emergentes en cualquier sistema de armas", así como los diversos elementos analizados en el punto anterior sobre seguridad y minimización de riesgos. En el AMA se trataría de un análisis y minimización de riesgos de posibles errores tanto del ope-

rador del SAA y, por tanto, del AMA, como de los propios algoritmos del SAA y del AMA, así como de riesgos por factores externos, incluidos la posibilidad de la piratería informática (*hacking*). En todo caso, el diseño de un AMA en un SAA no sería simplemente un asunto ético sino también un asunto de seguridad[135] y estaría vinculado a la identificación de los grados de autonomía del AMA, la compartición de tareas entre los algoritmos del SAA y el AMA, así como la de los diversos actores que formarían parte del diseño, desarrollo, despliegue y uso tanto del SAA como del AMA (NU, 2019a: 15; Zieba *et al,* 2010: 202).

- *Grados de autonomía SAA / AMA*: Dependiendo del nivel de autonomía del SAA, así también se definiría el grado de autonomía del AMA y su complejidad. Dependería de los resultados obtenidos del análisis de riesgos y su minimización arriba descritos. Por lo tanto, el grado de autonomía de un AMA dependería en si se estuviese hablando de agentes implícitos, explícitos o completamente autónomos, lo que también repercutiría en el diseño de su algoritmo.

- *Compartición de tareas*: Una vez establecido el análisis y mitigación de los riesgos, la definición de los diversos grados de autonomía y teniendo en cuenta los diversos actores que influyen en las diversas etapas de desarrollo, despliegue y uso del AMA, se establecería la compartición de tareas entre los diversos actores y los distintos algoritmos.

135 Como argumenta R. C. Arkin, las consideraciones sobre seguridad, no solo incluiría el disparo erróneo o accidental de un armamento, sino también el riesgo ético potencial de una identificación errónea de un objetivo, lo que podría resultar en una "focalización de objetivo" o el disparo hacia objetivos no previstos (Arkin, 2007: 56).

- *Interfaz operador / AMA*: Definición de la interfaz entre el operador (del SAA y del AMA, que podrían ser distintos) y el propio algoritmo del AMA, que se suele denominar como "Interfaz Hombre Máquina" (IHM) (*Human Machine Interface (HMI)*). En la actualidad existen guías IHM para la seguridad de industrias críticas, como las militares, utilizando estándares específicos, como la "Sección 5.14" del MIL-STD-1472F de los USA, de 1999[136]. También existen estándares internacionales generales de interfaz de usuarios, como el ISO 9241-110[137]. En todo caso, dicha IHM debería establecer los siguientes principios: prevenir al operador del estado actual del sistema, con relación a su seguridad; ayudar al operador a establecer un modelo mental del sistema y los controles necesarios para una supervisión exacta de su seguridad y comportamiento y; remover o minimizar aquellas características del sistema que podrían inducir a errores del propio operador (Rae, 2007).

Una vez configurado el entorno del algoritmo de autonomía ajustable, a través del diseño propuesto por S. Zieba *et al*, adaptado en nuestra investigación a un SAA, el siguiente paso sería la construc-

136 Para más información ver: DoD (2007): *Department of Defense Design Criteria Standard. Human Engineering,* Department of Defense USA, Washington DC, acceso marzo 2021, en http://chassis-plans.com/PDF/MIL-STD-1472F.pdf

137 Para más información ver: INTERNATIONAL ORGANIZATION FOR STANDARIZATION (ISO) (2020): *Ergonomics of Human-System Interaction-Part 110: Interaction Principles,* ISO, Ginebra, acceso marzo 2021, en https://www.iso.org/standard/75258.html

ción del modelo del AMA, así como el establecimiento del proceso de colección y procesamiento de los datos relativos a dicho AA. Como ya hemos analizado dentro del "ciclo de vida" de un AMA, tanto con relación a la construcción e interpretación del modelo como en la obtención y proceso de los datos asociados, el diseño del AMA no podría realizarse sin un estudio de los otros elementos: los datos y el modelo a construir e interpretar.

Recapitulando, sería imprescindible establecer los requerimientos de ingeniería de los AMA, tanto a la ética de la propia ingeniería como a la de la máquina. Particularmente, no se deberían cometer errores de sobrestimación de la capacidad de un método ético para resolver casos morales complejos, ni minimizar la importancia de la metodología a utilizar. Es decir, si se crean agentes morales implícitos o explícitos.

A nivel particular de los AMA para los SAAL, al igual que M. Sassoli, somos escépticos de que en un futuro cercano se pueda adoptar el DIH completo a un algoritmo a través de sensores, pues existirían tres condicionantes: una capacidad computacional limitada; un posible conflicto interno entre reglas y; la gran variedad de situaciones operacionales existentes. En todo caso, un AMA debería observar tres principios básicos: transparencia, responsabilidad y rendición de cuentas, para lo cual necesitaría de una gran adaptabilidad, una interacción continua con el medio y una autonomía controlable por el ser humano. Dicho proceso se necesitará a lo largo de todo el ciclo de vida del AMA, así como un control de riesgos aceptable, para lograr en lo posible que se convierta en un agente de impacto mínimo. En los próximos capítulos analizaremos en profundidad dichos aspectos.

CAPÍTULO 10

LA ESTRUCTURA COMPUTACIONAL

Una vez establecido el modelo de diseño del AMA para un hipotético SAA, habría que definir, a nivel teórico, su modelo de arquitectura computacional que integraría todos los elementos de diseño identificados visualizando sus posibles componentes estructurales. En todo caso, como establece R. Benjamins, las opciones técnicas escogidas impactarán e influenciarán el impacto ético y social que tenga el AMA construido, tanto con relación a las opciones técnicas digitales genéricas, como las específicas de la IA. Como hemos establecido en nuestra investigación que nuestro enfoque sería híbrido[138], nuestro objetivo será establecer una estructura computacional teórica del diseño del modelo de AMA que desarrollamos en nuestra investigación en el capítulo anterior, siguiendo el modelo de autonomía ajustable de S. Zieba. Habrá que tener en cuenta, no obstante, como argumentan A. Martinho *et al*, que la forma en que dicho AMA desarrolle técnicamente el proceso de tomar una decisión moral y la forma en que el diseñador establezca los algoritmos para dicho proceso, deberá implicar transparencia, agencia y responsabilidad moral, elementos que en la actualidad son objeto de gran controversia entre los investigadores y que hemos abordado

138 Tanto LIDA, MedEthEx y Ethical Multiple-Agent System, pueden ser considerados como sistemas híbridos, donde las excepciones son consideradas como comportamientos aceptables, ya que una acción puede ser buena o mala dependiendo de cada situación específica (Cervantes *et al,* 2020: 512).

con anterioridad en nuestra investigación (ver capítulos 2, 4, 5 y 8) (Benjamins, 2021: 50; Martinho *et al*, 2021).

Tampoco se debería obviar, dada su gran importancia, la comprensión de que los aspectos técnicos son solo una parte del dominio del ciberespacio, el espacio central pero no exclusivo que forma parte de una unidad con los espacios sociotécnico y de gobernanza, como pusieron de manifiesto con relación a los SAA los investigadores I. Verdiesen, F. Santoni de Sio y V. Dignum y que se resume en el siguiente diagrama (Ver Fig. 30) (Verdiesen *et al*, 2020:139):

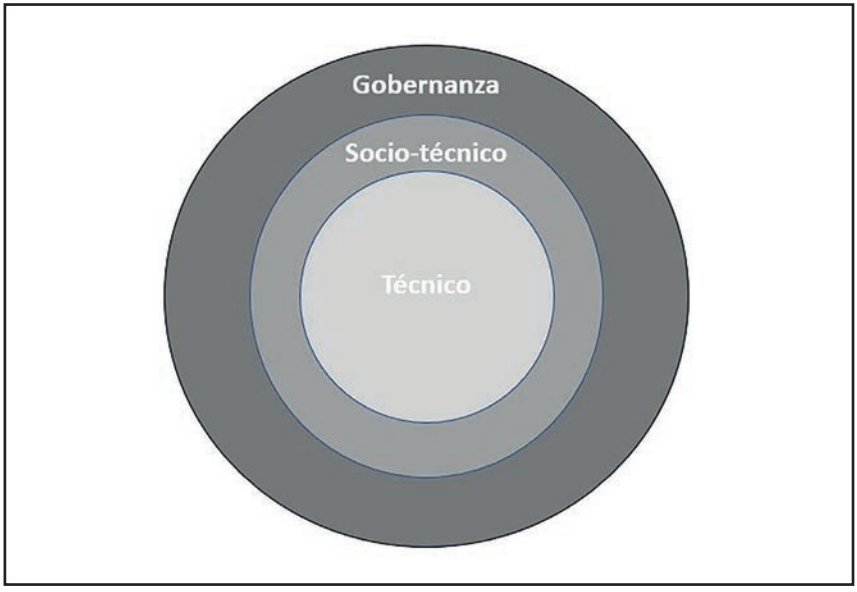


Fig. 30: Conceptualización del ciberespacio en capas, de acuerdo con I. Verdiesen et al (Verdiesen et al, 2020: 139).

10.1.– CONSIDERACIONES TÉCNICAS INICIALES

Tomando como base la visión pragmática planteada en el capítulo 5º de nuestra investigación, la estructura computacional del AMA que estamos considerando tendría como objetivo servir como mecanismo de regulación de un SAA, que aseguraría que su comportamiento se ajustase en lo posible, de acuerdo con la definición establecida, a las normas establecidas internacionalmente, que en nuestro entorno estaría descrito por el Derecho Internacional aplicable en cada momento. Por lo tanto, estaría más cercano a un ejemplo de "herramienta sensible" de F. Fossa, lo que implicaría una aculturación explícita de dicho algoritmo, dado que su objetivo sería que un SAA, en sus acciones, se ajustase al Derecho Internacional vigente (el para qué), a través de la comunidad internacional (el por quién), para aquellos Estados que utilizasen dichos SAA (el para quién). Esto significaría, como estableciera R. Capurro y que ya analizamos en dicho capítulo, que dicho algoritmo no tendría por qué ser neutral a nivel global, pues dependería del contexto cultural en el que fuese creado, como por ejemplo el entorno de las NU, aunque en dicho entorno si lo tendría que ser (Capurro, 2019: 132-134, Fossa, 2018: 123).

Uno de los principales investigadores que inicialmente plantearon la necesidad de formalizar el control del flujo de los algoritmos relativos al control ético ha sido R. C. Arkin que en compañía de los investigadores D. C. Mackenzie y J. M. Cameron desarrollaron, en 1997, una configuración del comportamiento multi agente, que sirvió para establecer una implementación de arquitectura directa para un gran número de sistemas autónomos robóticos, incluidos los militares, desarrollando un mapeo (*mapping*) de comportamiento formalizándolo a través de una metodología matemática. Así, para

cada comportamiento activo individual, se podría establecer una función de mapeo entre el ámbito de estímulo y el posible rango de respuestas, definiendo una función de comportamiento **ß** donde:

$$\beta(s) \longrightarrow r$$

Dicha función **ß** podría ser definida arbitrariamente pero siempre para todas las clases posibles de percepción ***p*** dentro de **s** y dependería del grado de fuerza del estímulo ***δ***. Así, para cada **s** (***p***, ***δ***), suponiendo que se estableciese un umbral ***π*** que debería ser excedido antes de que se produjese una respuesta **r** específica, entonces se tendría la siguiente función (Arkin, 2007: 16; Mackenzie *et al*, 1997):

$\beta(p, \delta) \longrightarrow$	para todo $\delta < \pi$	entonces si $r = \Theta$	no habría respuesta
		o si $r =$ *función arbitraria*	habría respuesta

Si nos ceñimos a los SAAL, para el desarrollo del algoritmo habría que distinguir que es lo que sería aceptable o no con relación al DIH. El conjunto de restricciones éticas, por tanto, sería el espacio donde la letalidad sería válida y permitida. Así, la aplicación de la letalidad como respuesta a uno o varios estímulos estaría limitada por el Derecho Internacional vigente antes de que se pudiese usar por un SAAL. En dicho contexto, para R. C. Arkin, una restricción cualquiera podría ser considerada de dos formas (Arkin, 2007: 18):

- Una restricción negativa de comportamiento (una prohibición), que bloquearía un comportamiento letal de la fun-

ción de comportamiento para un determinado estímulo o;

- Una restricción positiva de comportamiento (una obligación), que requeriría un comportamiento letal de dicha función para un determinado estímulo.

Si se considera a ***P*** como el conjunto de todas las posibles acciones, al conjunto de acciones letales y todas las posibles acciones éticas letales, el siguiente diagrama ilustraría aquellas respuestas permitidas que podrían ser tanto letales como no letales, cuyas restricciones vendrían por el DIH (ver Fig. 31) (Arkin, 2007: 19):

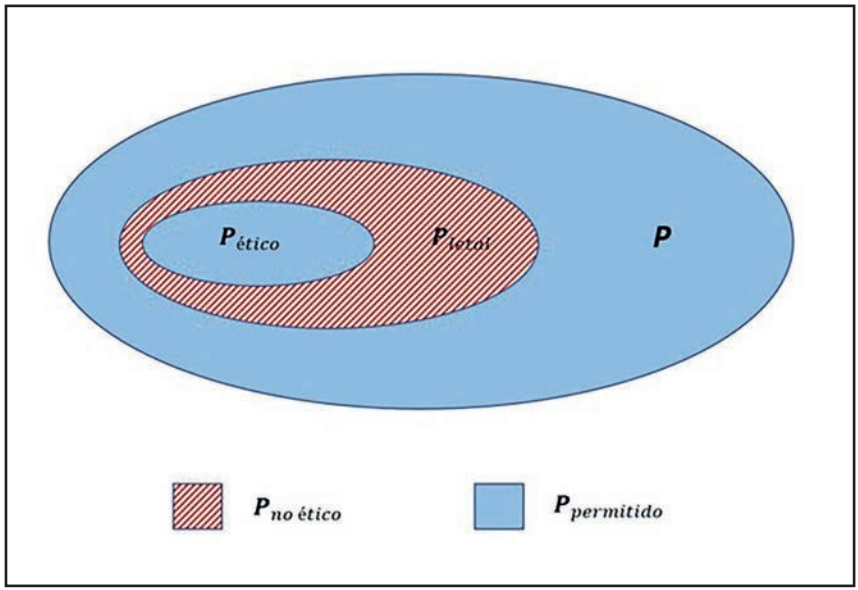


Fig. 31: Diagrama de acciones permitidas y no éticas, según R. C. Arkin (Arkin, 2007: 19)

Es decir, solo una parte de las posibles acciones letales serían consideradas éticas, que vendría establecido por el Derecho Interna-

cional vigente y, por tanto, podrían formar la base de los principios institucionalizados por los Estados al igual que otros genéricos como la privacidad, la seguridad y la fiabilidad.

Ahora bien, aparte de la selección de los principios para asegurar un impacto positivo social y ético de la IA utilizada, también existen, como alega el investigador R. Benjamins, una selección técnica específica de IA, así como una elección técnica digital genérica. Dicho investigador identifica ocho elementos técnicos relevantes, tanto específicos para la IA como genéricos, que consideramos importantes a considerar. El siguiente diagrama ilustra dichos elementos (ver Fig. 32) (Benjamins, 2021:50):

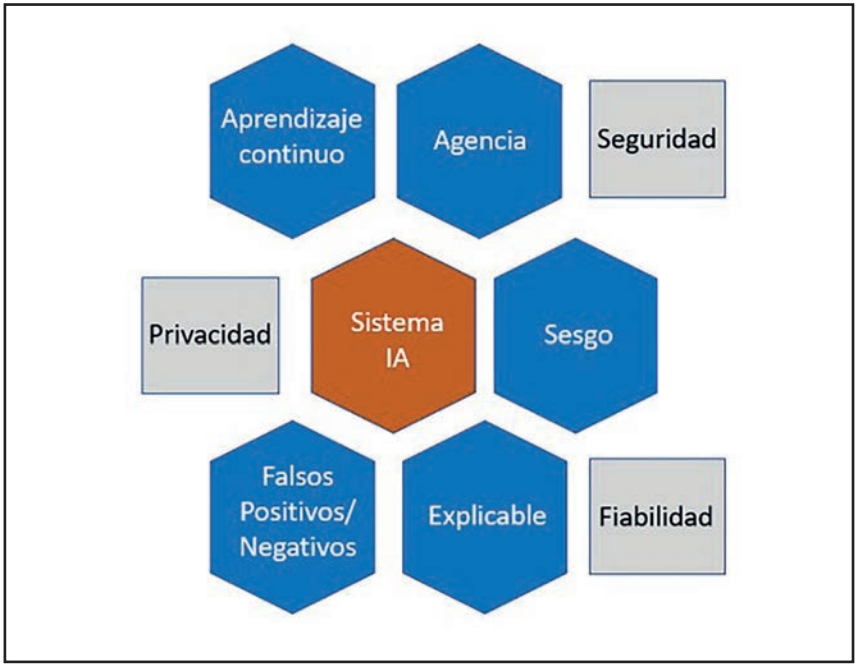


Fig. 32.- Elementos del impacto ético y social de un sistema de IA (específicos en azul y genéricos en gris). Adaptado de R. Benjamins (Benjamins, 2021: 50)

Elementos técnicos específicos de IA (Benjamins, 2021: 51):

- *Aprendizaje continuo*: El algoritmo de IA continua su aprendizaje autónomamente desde el momento de su puesta en marcha, por lo que su rendimiento evolucionará con el tiempo sin intervención humana dependiendo de los datos recibidos;
- *Agencia*: El grado de intervención humana del sistema de IA;
- *Sesgo*: Discriminación del sistema de IA por diversas causas: datos no representativos; datos de aprendizaje sensibles (raza, religión, etc.); correlación significativa entre variables;
- *Explicable*: La necesidad de comprender como el sistema de IA llega a un resultado. Existen sistemas de aprendizaje profundo que funcionan como "cajas negras" que son difíciles de comprender;
- *Falsos positivos/negativos*: Siempre existe una tasa de error. Dependerá de lo que se considere aceptable para cada caso. En el caso de una acción letal, la tasa de error debería ser mínima.

Con relación a las acciones técnicas genéricas (privacidad, seguridad, fiabilidad), son comunes a todos los sistemas digitales y no solo los de IA, por lo que el sistema tendrá que seguir los estándares y parámetros aprobados en cada caso por el estamento diseñador y fabricante.

10.2.- MARCO GENÉRICO COMPUTACIONAL

El marco genérico computacional vendrá limitado por el modelo de diseño, que en nuestro caso sería una adaptación del modelo de autonomía ajustable de S. Zieba. También hemos apostado por un modelo híbrido basado en las investigaciones de M. Klincewitz, tanto con la aplicación "arriba-abajo" de una teoría ética (en nuestro caso la teoría deontológica a través de la aplicación de las normas del DIH), con la práctica de uso "abajo-arriba", basada en el aprendizaje de casos, ya propuesta por R. C. Arkin, lo que implicaría también seguir en cierto modo la "visión particularista" de S. Tolmeijer *et al.* No obstante, mitigando los posibles sesgos del aprendizaje, a los que aludían B. D. Mittelstadt *et al,* a través de un control de riesgos y una supervisión adecuada (Arkin, 2007: 43-44, 47-50; Klincewitz, 2015: 165-167; Mittelstadt *et al,* 2016: 2; Tolmeijer *et al,* 2020: 132.8-132.9).

Como ya estableciera R. C. Arkin, la particularidad de inferir la capacidad de letalidad en un robot no difiere de cualquier otra respuesta genérica a una situación específica. En este caso la respuesta robótica estaría gobernada por elementos externos derivados del DIH y las ROE. La respuesta, en todo caso, tendría que ser considerada de tipo binario discreto: o el armamento es disparado o no. Cualquier respuesta (disparo a matar, a herir, de prueba, etc.), sería considerada como una respuesta binaria discreta específica y diferenciada. El procedimiento, siguiendo a R. C. Arkin, tendría una arquitectura computacional basada en acciones, donde la teoría ética basada en el DIH informaría al robot que acciones llevar a cabo. La ventaja es que sería una metodología consistente, completa, práctica y de acuerdo con la teoría ética propuesta. Dichas acciones seguirían tres formas primarias (Arkin, 2007: 22, 39-40):

- Las permitidas: potencialmente obligatoria pero no prohibida;
- Las obligatorias: permitidas y no prohibidas;
- Las prohibidas: ni permitidas ni obligatorias.

Según dicho investigador, sería el DIH y las ROE las que definirían aquellas acciones completamente prohibidas, mientras que las restantes deberían ser consideradas obligatorias de acuerdo con las ROE establecidas.

Ahora bien, como indican J. Galliott y J. Scholz, el desarrollo de un AMA completamente ético requeriría una ingeniería ética muy extensiva. El desarrollo de un sistema restrictivo utilizando la teoría deontológica basada en los postulados del DIH, que a su vez es complejo, y en un gran número de casos posibles relativos a las diversas ROE dependiendo del escenario en cuestión, debería analizar un gran número de acciones que necesitaría de una estructura de datos muy compleja. Esto implicaría, según dichos investigadores, que el razonamiento sobre el marco de lo que sería permisible, incluida la noción de proporcionalidad y distinción, estaría abocada a grandes dificultades. Por lo tanto, abogarían por un AMA minimalista, al que denominan como "MinAI" (Inteligencia Artificial Mínima) que, aunque siguiese siendo un sistema restrictivo, solo poseería acciones de supresión elementales, que serían activadas de acuerdo con un conjunto de restricciones muy reducido, que necesitarían menos necesidad de interpretación por el sistema. Estaría basado en prohibiciones expresas e implicaría la codificación de valores normativos en una serie de reglas y la interpretación de un número limitado de entradas a través de sensores específi-

cos obteniendo una serie de actuadores también limitados. En todo caso, cualquier AMA desarrollado necesitaría de una revisión estatal de acuerdo con lo establecido en el Art. 36 del PAI (Galliott y Scholz, 2018: 58-60, 62).

La limitación de restricciones, sensores y actuadores nos parece adecuado dada la complejidad del DIH y el sinfín de ROE posibles. Dado que nuestra propuesta considera un sistema híbrido, creemos que el desarrollo de una Red Neuronal Artificial (RNA) (*Artificial Neural Network (ANN)*), sería un marco genérico computacional adecuado, ya que al mismo tiempo que combinaría una serie de reglas limitadas, pertenecientes a los postulados del DIH, también incluiría la posibilidad de entrenamiento de escenarios específicos con diversas ROE utilizando algoritmos de "*machine learning*" (aprendizaje automático) y estadísticos. Permitiría al AMA "aprender" autónomamente sobre diferentes escenarios específicos utilizando un perceptrón multi capa, donde cada neurona estaría conectada con una neurona de la siguiente capa, teniendo un peso específico según su relevancia e importancia dentro de la red neuronal[139] (Van Erp, 2016: 8-9).

En general, la estructura de una red neuronal tendría un vector de entrada **X** con la información de entrada que sería modificada por un vector de pesos **W** a lo que se añadiría un coeficiente de sesgo **Θ,** produciendo un resultado **Y** (ver fig. 33) (Gámez Albán *et al*, 2016: 160):

139 Los investigadores G. S. Reed *et al* desarrollaron un modelo ético militar denominado REV (*Relative Ethical Violation*), en el que a una serie de principios del DIH y de las "Reglas de la Guerra" se le asignaron una serie de pesos para establecer si una acción era éticamente aceptable, utilizando un calibrado a partir de una serie de entrenamientos continuos (Reed *et al,* 2016).

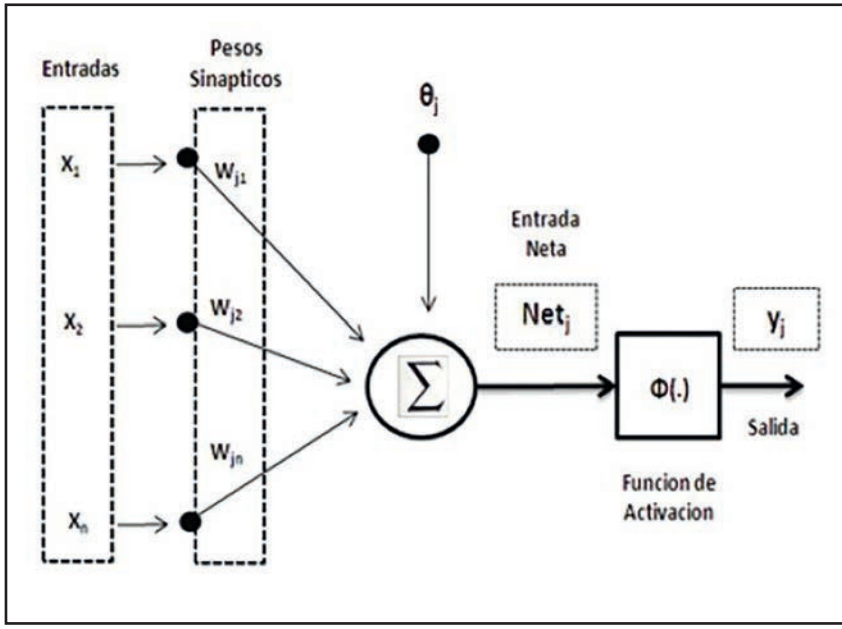


Fig. 33: Estructura general de una red neuronal artificial (Gámez Albán et al, 2016: 160).

Las entradas son: $x_1, x_2, \ldots, x_n$

Los pesos son: $w_1, w_2, \ldots, w_n$

El sumatorio ponderado sería: $x_1 w_1 + x_2 w_2 + \cdots + x_n w_n$

Añadiendo el sesgo obtendríamos: $x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + \Theta$

Se proporcionaría dicho resultado a la "función de activación"

Función de activación $(x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + \Theta)$

Dicha función sería una operación matemática que normalizaría las entradas y producirían un resultado que, dependiendo del entorno, podría ser de diversos tipos (lineal, sigmoide, etc.). Dicho resultado alimentaría las neuronas de la siguiente capa (Malik, 2019).

Normalmente la RNA propuesta sería multicapa, con una serie de capas ocultas, que en el caso de proceso de diferentes medios (textos, imágenes, números, etc.) se denominaría una "red neuronal profunda". El siguiente diagrama presenta un perceptrón multicapa con una capa oculta con sus pesos (w) y sesgo (Θ) (ver fig. 34) (Romero, 2020)

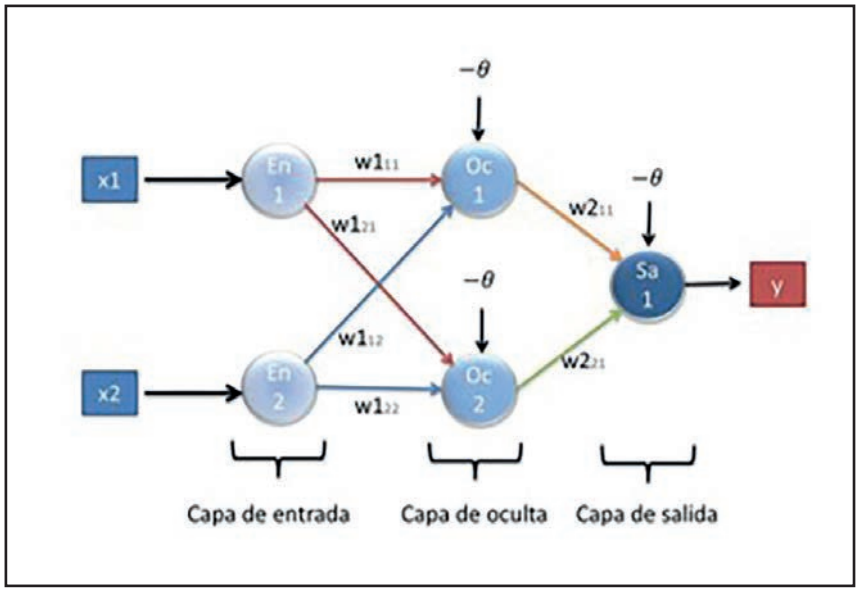


Fig. 34: Perceptrón multicapa (Romero, 2020).

Hay que tener en cuenta que tanto los pesos (w) y el sesgo (Θ) son los principales componentes de una RNA. Al inicio de la RNA los

pesos y el sesgo tendrían un valor predeterminado que se irán modificando durante el aprendizaje de los diversos casos. Los pesos establecen la importancia de un sensor en determinadas casuísticas, mientras que el sesgo es una constante o un vector de constantes que compensa un resultado hacia un estado específico. Tanto los pesos como el sesgo son definidos en el algoritmo del RNA, de acuerdo con las especificaciones establecidas previamente para dicha red.

10.3 – MARCO ESPECÍFICO COMPUTACIONAL

Como ya indicamos en el capítulo 9º, los debates dentro del GGE del CCW sobre los SAAL, identificaron los principios de proporcionalidad, distinción y precaución como la base de control de dichos armamentos en la selección y puesta en uso de dichos sistemas, con el beneplácito de las grandes potencias como China, Rusia y USA. También, dentro de los principios rectores establecidos, la responsabilidad humana sobre el armamento, el respeto al derecho internacional vigente y la rendición de cuentas no difieren mucho de los principios establecidos por R. C. Arkin o G. S. Reed *et al,* que incluyeron también la necesidad militar y la humanidad como principios básicos (Arkin, 2007:23; NU, 2018b; 2019a; 2019k; 2019j, Reed *et al,* 2016).

Siguiendo la idea de un AMA minimalista, como el propuesto por J. Galliott y J. Scholz, pero atendiendo a los principios rectores establecidos en el marco de las NU, nuestra propuesta incluiría los principios de distinción y proporcionalidad como entradas básicas ($\boldsymbol{y}$) de la RNA. El principio de distinción aludiría al Art. 85(3)(b) del PAI, mientras que el de la proporcionalidad se basaría en el Ar.

51(5)(b) del PAI. También sería necesario incluir como entradas básicas el principio de responsabilidad (Art. 51(5)(b) y 57(2) del PAI) pero también, como alude R. C. Arkin, la necesidad militar, pues en un escenario bélico la necesidad de alcanzar los objetivos, como premisa para la victoria militar es una constante que no se debe obviar. Estos principios serían las entradas (***w***) añadiendo el principio de humanidad como nuestra constante de sesgo (**Θ**), pues consideramos que se podría establecer como elemento del AMA que sirviese como compensación al principio de necesidad militar y proporcionase un grado relativo de equilibrio entre los dos principios. El siguiente esquema resumiría las entradas básicas de nuestro RNA y su conexión con el DIH (ver fig. 35) (Arkin, 2007: 23; Galliott y Scholz, 2018: 58-60; NU, 2019a: 15; Reed *et al,* 2016: 200):

Principio	RNA	DIH
Distinción	X_1	Art. 85(3)(b) del PAI
Proporcionalidad	X_2	Ar. 51(5)(b) del PAI
Responsabilidad	X_3	Art. 51(5)(b) y 57(2) del PAI
Necesidad militar	X_4	Art. 35, 36 y 52 del PAI
Humanidad	Θ	Art. 1(2) del PAI

Fig. 35: Esquema de las entradas de una RNA para un SAAL y su relación con el DIH (elaboración propia).

Como se puede observar no se han incluido ni los valores de los diversos pesos () ni el valor de la constante de sesgo (**Θ**). Creemos que dependería tanto de la ROE específica, así como la acumulación de aprendizaje y la consecuente base de datos disponible. Además, dicha constante de sesgo eliminaría la delegación del principio de humanidad a una máquina, dado que el impacto de uso de un SAAL sobre la dignidad humana sería independiente del nivel de sofisticación de la tecnología, adhiriéndose además a la necesidad de mantener los conceptos establecidos por la "Clausula Martens" dentro de dichos sistemas armamentísticos. En todo caso, podemos construir un procedimiento básico para los principios éticos, que se basaría en la propuesta de R. C. Arkin, con ciertas modificaciones, según el siguiente esquema (ver fig. 36) (Arkin, 2007: 59; Taddeo y Blanchard, 2021: 3):

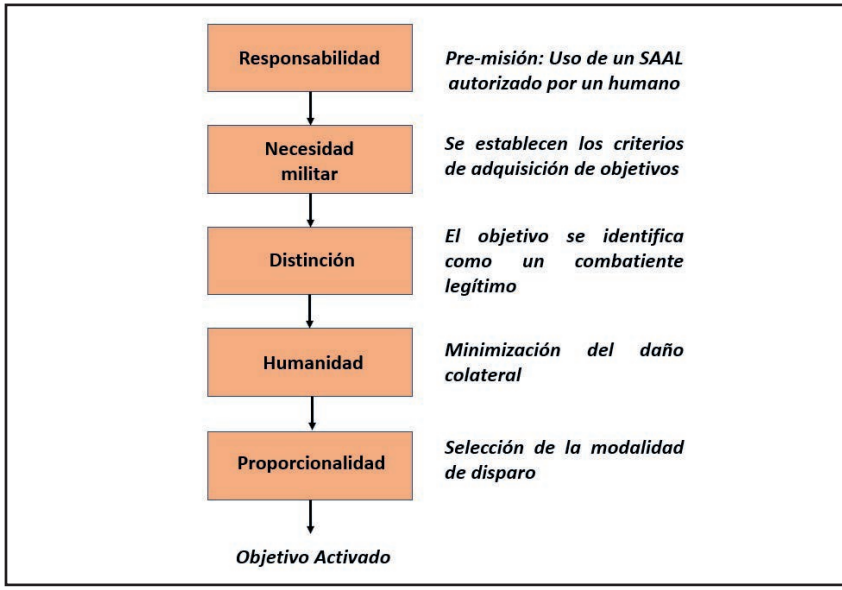


Fig. 36: Procedimiento ético básico adaptado de R. C. Arkin (Arkin, 2007: 59 y elaboración propia).

Algunos de los principios podrían ser refinados con los siguientes requisitos (Arkin, 2007: 57):

- *Distinción:*
    - Distinción entre civil y combatiente
    - Distinción entre combatiente y combatiente rendido
    - Fuerza directa solo hacia objetivos militares
- *Proporcionalidad:*
    - Utilización únicamente de armamento legal
    - Uso de un nivel apropiado de fuerza
- *Humanidad:*
    - Minimizar el daño colateral
    - La defensa propia no justifica el tomar vidas civiles

Para que el objetivo sea activado se necesitaría que el resultado del RNA fuera positivo, además de necesitar que un humano supervisara dicho resultado y también lo permitiese. Esto implicaría que el grado de autonomía del SAAL siempre estaría supeditado a un control efectivo humano en todas las fases (antes, durante y después). Así, se debería establecer, como ya desarrollamos en profundidad en el capítulo 9º, un "Marco Integrado de Responsabilidad Humana", según lo establecido por I. Verdiesen *et al*., para dicho procedimiento ético básico anteriormente descrito. Entraríamos el principio de responsabilidad en la "supervisión *ex ante*" dentro de la capa de gobernanza y los principios de humanidad y proporcionalidad a través de las capas Sociotécnica (control *ex ante* y control activo) y Técnica (entrada y mecanismos de realimentación), mientras que el principio de distinción estaría también en las capas

Sociotécnica y Técnica, con la particularidad de que en esta última y con relación a los mecanismos de realimentación estaría presente a través de los pesos (**w**) de cada entrada del RNA. Así, nuestro "Marco Integrado de Responsabilidad Humana" se iría construyendo y seguiría el siguiente esquema, que sería completado según avancemos en la construcción del AMA para un SAAL, para lograr un MHC aceptable. En nuestro desarrollo sustituiría a las funciones que R. C. Arkin otorgaba al "Gobernador Ético" y al "Asesor de la Responsabilidad" (ver fig. 37) y a la infografía desarrollada por UNIDIR sobre el elemento humano en las decisiones sobre el uso de la fuerza (Arkin, 2007: 61; UNIDIR, 2020a; Verdiesen *et al,* 2021: 151):

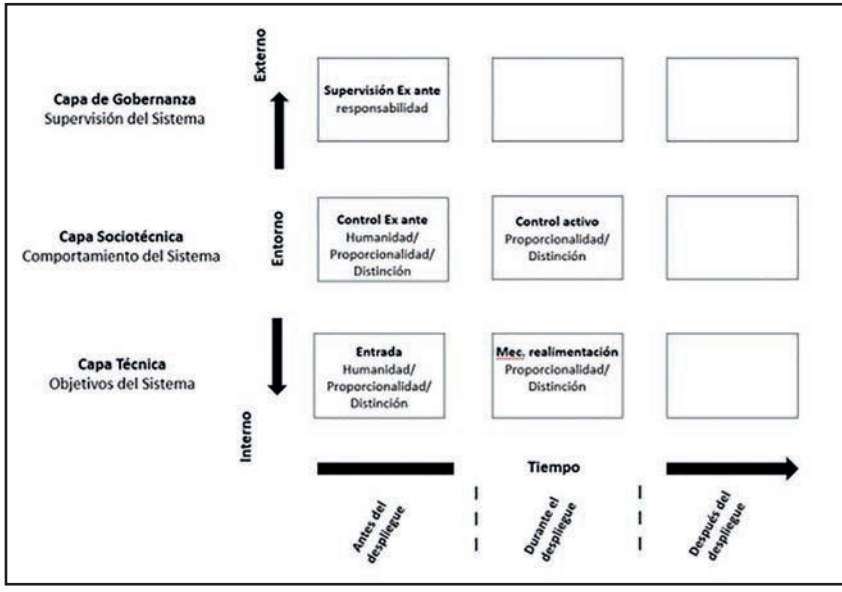


Fig. 37: Marco Integrado de Responsabilidad Humana (parcial) para un AMA de un SAAL, basado en el Procedimiento ético básico de la Fig. 7. Adaptado de I. Verdiesen et al (Verdiesen et al, 2021: 151 y elaboración propia).

Habrá también que tener en cuenta, de acuerdo con el modelo de S. Zieba, que no necesariamente los mismos actores serían responsables de cada fase de supervisión, dado que la "Identificación de Competencias", seguiría el esquema de valor propuesto por A. Fritz *et al* ya analizado en los capítulos 5º y 9º (Fritz *et al*, 2020: 8).

A la vista de estos nuevos requisitos, no exhaustivos, podemos intuir la necesidad de crear un perceptrón multicapa. Analicemos teóricamente lo que esto supondría de incremento de dificultad para la concepción y desarrollo del algoritmo de nuestro AMA, tomando como ejemplo el principio de distinción. El siguiente esquema resumiría, en una primera aproximación simplificada, el posible desarrollo de dicho principio dentro de nuestro RNA (ver fig. 38):

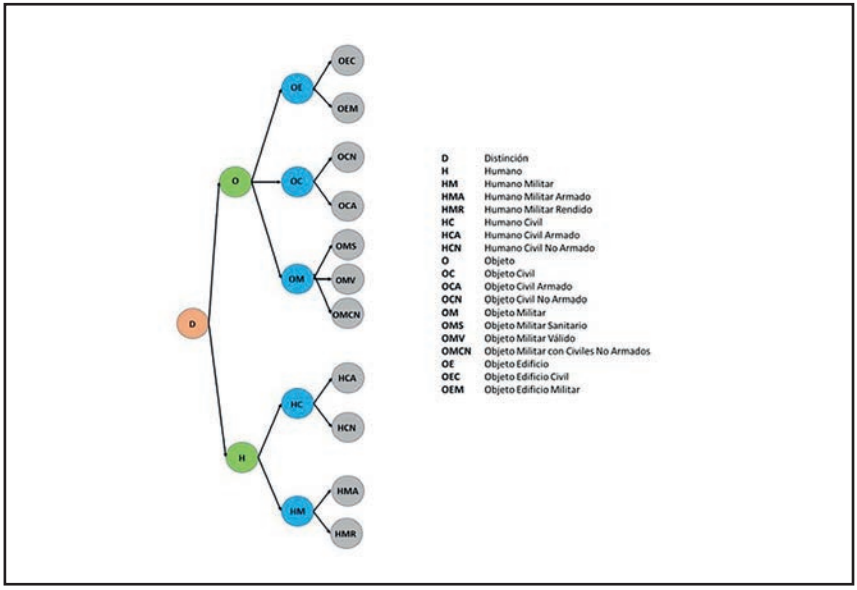


Fig. 38: Desarrollo del sensor del principio de distinción de un RNA multicapa para un AMA de un SAAL, primera aproximación no exhaustiva (elaboración propia).

Como podemos observar en el esquema la complejidad es importante, ya que se podría afinar aún más para cada neurona. Esto requeriría una sensor de reconocimiento de imágenes potente y a la vez rápido. También es importante destacar que ni los pesos ni el sesgo se han calculado. No se debe olvidar que dicho RNA multicapa debería suscribir un nivel legal de discriminación aceptable, de acuerdo con los argumentos de K. Anderson y M. Waxman ya analizados en el capítulo 7º y que nosotros suscribimos. Por ejemplo, todavía existiría inconcreción para distinguir aquellos edificios civiles (OEC) que a su vez tienen un uso dual, como sería una fábrica de microchips. Esto implicaría la necesidad de añadir otro nivel oculto incrementando, a su vez, la complejidad del RNA. En caso de que no añadiésemos otro nivel, consideraríamos la necesidad de una supervisión humana a través de un control activo. En aras de la simplificación tampoco se habría incluido, entre otros, una neurona relativa a los seres vivos no humanos, que podrían también ser considerados, en ciertas circunstancias (ej. perro con explosivos adheridos) como objetivo militar (Anderson y Waxman, 2013: 41, 46; Geib y Lahmann, 2012: 382-383).

En resumen, el AMA sería un algoritmo híbrido, compuesto por una serie mínima de reglas y restringiendo el número de sensores y actuadores, utilizando una RNA (perceptrón multicapa), con sus correspondientes pesos y sesgo, que tendría que observar un nivel legal de discriminación aceptable. En particular para un SAAL, la entrada estaría formada por los principios rectores esenciales establecidos por la NU, así como por un Marco Integrado de Responsabilidad Humana, que aportaría el control efectivo humano durante todo el ciclo de vida del AMA.

Configurado el entorno del algoritmo computacional y el Marco

Integrado de Responsabilidad Humana, siguiendo nuestro diseño basado en el propuesto por S. Zieba *et al,* el siguiente paso sería el desarrollo de los módulos de aprendizaje y control, que completaría la supervisión humana *ex post* y el control de riesgos. En el próximo capítulo analizaremos en profundidad dichos aspectos. Así, el AMA para un SAAL estaría compuesto, en este punto de su desarrollo, de dos módulos: una red neuronal artificial y un marco integrado de responsabilidad humana. El siguiente esquema plasma dicho desarrollo (ver fig. 39):

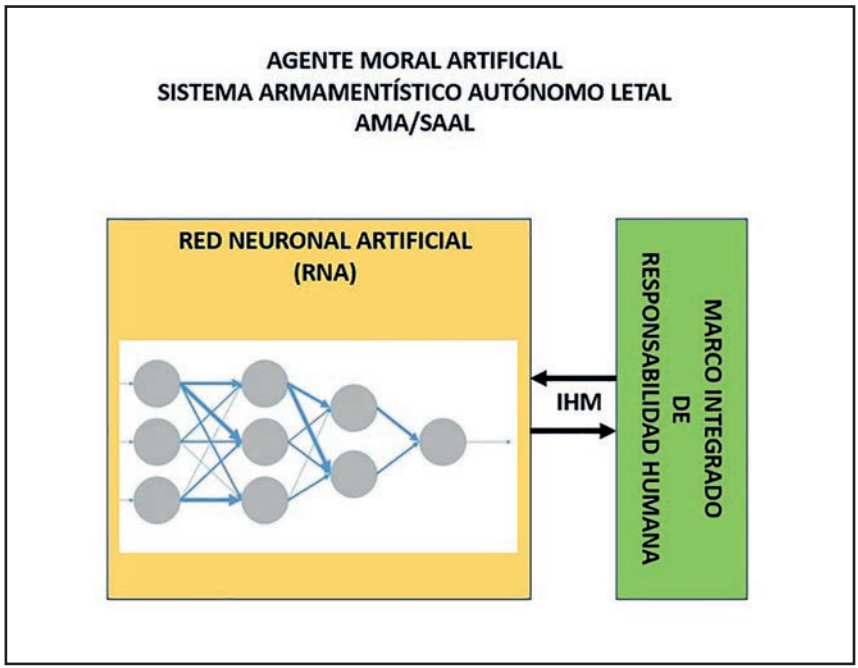


Fig. 39: Esquema de estructura de un AMA de un SAAL. Módulos computacionales del RNA y del Marco Integrado de Responsabilidad Humana (Xeridia y elaboración propia) (Xeridia, 2021).

CAPÍTULO 11

APRENDIZAJE Y CONTROL

Siguiendo el borrador estándar IEEE (P7000/D5) para abordar las inquietudes éticas, que analizamos en el capítulo 9º, con la idea de mantener un proceso de gestión transparente del AMA nos quedaría por analizar el concepto operacional dentro de la fase conceptual operativa y el proceso de diseño de riesgos éticos dentro de la fase de desarrollo. Una vez analizados dichos puntos habríamos completado el P7000/D5 del IEEE específico para el AMA de un SAAL. En dicho momento se podría, por tanto, resumir dicho estándar para el AMA.

Los valores éticos vendrán representados a través de la teoría ética deontológica representada en el AMA por el DIH. Los requerimientos éticos en la fase de desarrollo se habrían establecido a través del RNA y los principios minimalistas del DIH propuestos, que sería una guía ética basada en los principios básicos adoptados por el GGE del CCW de las NU para los SAAL (analizados en el capítulo 10º). El Marco Integrado de Responsabilidad Humana, como proceso de gestión transparente, se habría desarrollado parcialmente también en dicho capítulo 10º, aunque quedaría por definir el control *ex post* y el análisis de riesgos que abordaremos en este capítulo. Por último, la concepción operativa del AMA, al ser un sistema militar SAAL, vendría dado por el ROE específico para cada operación. El siguiene esquema resumiría dicho contenido (ver fig. 40) (IEEE, 2021):

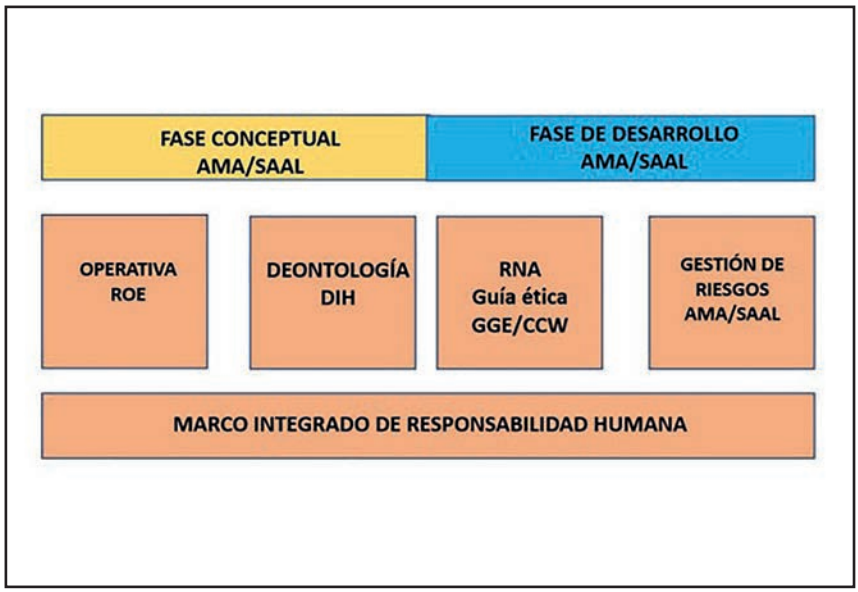


Fig.40: Esquema específico de uso del IEEE P7000/D5 sobre la inclusión de elemento éticos en un SAAL a través de un AMA (IEEE, 2021 y elaboración propia)

11.1- APRENDIZAJE

Dicha propuesta mantendría nuestra apuesta de un sistema híbrido compuesto de una parte de desarrollo de "arriba-abajo", a través del marco deontológico del DIH y la guía ética del GGE del CCW para los SAAL de las NU, potenciado por otra parte ("abajo-arriba") a través de una serie de casos basados en las diversas ROE que se establecerían para cada acción militar en la que un SAAL estaría envuelto. Es importante destacar que es a través de las ROE como se institucionaliza la fuerza por parte de los Estados y son un elemento clave del principio rector de la responsabilidad, establecido por el GGE del CCW para los SAAL, para deslindar las

posibles conductas contrarias a Derecho, pero también jugarían un papel fundamental para garantizar, como establece R. Lorenzo Ponce de León, la: "legitimidad política, jurídica, moral y militar de los Estados …" (Lorenzo Ponce de León, 2012: 169-170).

Dichas ROE servirían como base para el aprendizaje de la RNA del AMA, constituyendo una base de datos que, para cada una de las acciones, establecerían los pesos y sesgo específico de las diferentes neuronas que compondrían dicha RNA. El siguiente esquema resume dicha propuesta (ver fig. 41) (Arkin, 2007: 41):

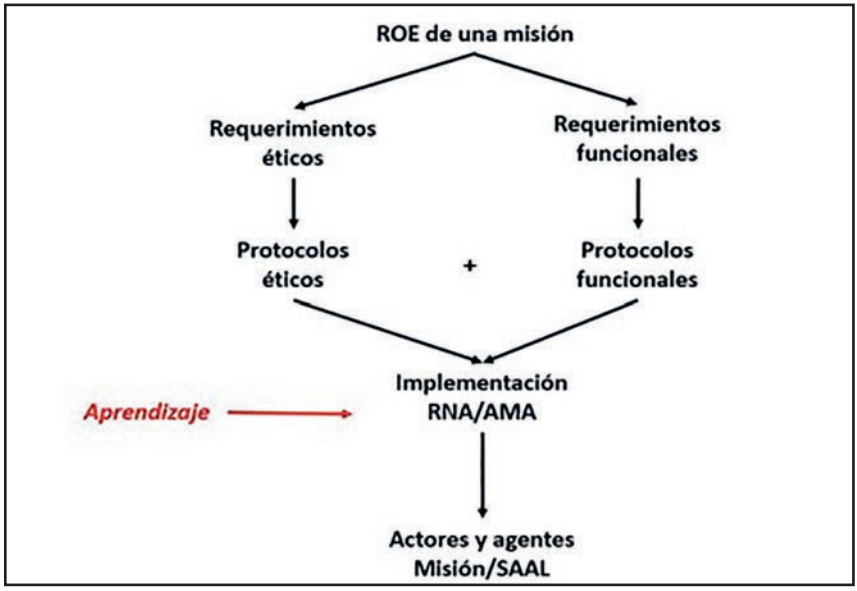


Fig.41: Esquema de uso de los ROE como base de aprendizaje de un AMA para un SAAL (Arkin, 2007: 41 y elaboración propia)

Se estaría hablando de un aprendizaje adaptativo a través de casos ilustrativos basados en las ROE para una determinada acción de un SAAL. Destacaríamos también, como argumentan I. Bode y H. Huelss, que la continuidad de dicho proceso alimentaría la idea de que serviría también para crear, dar forma y definir nuevos normas de Derecho consuetudinario. Los enlaces ponderados a través de pesos (**w**) se ajustarían creando una distribución específica para cada ROE. Dicho aprendizaje no solo se realizaría previo al despliegue del AMA sino que seguiría aprendiendo a lo largo de la actividad del SAAL después de completado el periodo de entrenamiento inicial. En cuanto a la constante del sesgo (**Θ**) del principio de "humanidad", tendría un valor dado al inicio del entrenamiento del AMA, que solo podría ser modificado por sus desarrolladores aplicando los requisitos de gobernanza del AMA específico, en caso de necesidad después de un análisis *ex post* del funcionamiento del AMA de forma continuada. Sería un aspecto del "control de riesgos" aplicado, que analizaremos en detalle en otro apartado de este capítulo (Bode y Huelss, 2018: 396; Matich, 2001: 9).

Habría que tener en cuenta que, en el AMA propuesto, el RNA llevaría a cabo un aprendizaje profundo (*deep learning*), subdivisión del aprendizaje automático, realizando una aproximación a través de un procesamiento de datos propios. Esto implicaría usar gran cantidad de datos de entrenamiento *a priori* para realizar dichas predicciones. Nuestro pensamiento sería que, en la actualidad, la capacidad de procesamiento y rapidez necesaria para un AMA de un SAAL en un entorno bélico solo sería posible a través de un procesamiento de datos cuántico, dado el gran número de capas posibles en un RNA del AMA (Lazzeri, 2021).

Dada la propia naturaleza de las ROE, para cada acción del AMA/SAAL existiría un registro dentro de la base de datos de aprendizaje, con sus pesos y sesgo específicos. En el caso que una acción bélica del SAAL estuviese ya codificada en la base de datos, el RNA sería nutrido de los pesos y sesgo específicos para dicha acción extraídos de dicha base de datos. En el caso de que no estuviese en la base de datos se añadiría un registro nuevo que sería codificado en la base de datos de aprendizaje a través del IHM del AMA, utilizando el Marco Integrado de Responsabilidad Humana establecido. Todo el proceso se especificaría en un módulo específico del AMA para el SAAL que vendría encriptado, tanto en el contenido como en el acceso a sus datos para mantener la seguridad e integridad ante posibles "hackeos"[140].

Ahora bien, los posibles datos que se alimentarían estarían expuestos a una serie de limitaciones. El investigador A. Holland Michel ha descrito los principales problemas relativos a los datos que pueden influir en un SAAL y por tanto en los datos suministrados a un hipotético AMA adscrito. Estos serían (Holland Michel, 2021: 4):

- *Datos incompletos*: cuando la información necesaria para que el sistema tome una acción apropiada no está presente en los datos, con posible clasificación errónea de objetos;
- *Datos de baja calidad*: los datos pueden contener errores o ambigüedades, lo que puede resultar en respuestas inapropiadas (por ejemplo, una baja resolución de imagen de un objeto);

140 De especial relevancia dentro del ámbito de la computación cuántica y la necesidad de garantizar una autentificación segura dentro del ámbito militar a través de contraseñas fiables (Steinwandt, R. *et al*, 2021).

- *Datos incorrectos o falsos*: Sensores defectuosos o mal calibrados, datos proporcionados por humanos con errores o datos deliberadamente erróneos (por ejemplo, poner una cruz roja sobre un vehículo armado), para confundir al sistema;
- *Datos discrepantes*: Inconsistencia entre los datos establecidos durante el proceso de aprendizaje con el obtenido en tiempo real. Nuevos registros de entrada fuera del rango establecido durante el diseño, anomalías que también pueden surgir cuando los datos no encajan completamente en las categorías estructuradas (por ejemplo, un niño portando un arma, si no estuviese previsto inicialmente o desplegar un sistema en escenarios para los que no estaba diseñado.

Las causas pueden ser variadas: condiciones adversas en los escenarios (polvo, humo, camuflaje, degradación por uso, etc.), acciones del adversario (contra los sensores, bloqueo de señales, "hackeos" del sistema, alimentación de datos falsos, etc.); escenarios bélicos complejos y variables (a mayor diversidad más probabilidad de errores en los datos), que algunos elementos solo se manifestarían en entornos reales y no durante el entrenamiento; existencia de escenarios singulares que pueden provocar un "desfase de datos" (*data drift*) (comportamiento humano impredecible, cambios físicos del escenario bélico durante el combate, nuevas tácticas de combate del enemigo no previstas, etc.)[141] (Holland Michel, 2021: 6-9). El siguiente esquema resume la estructura computacional del

141 Para más información sobre el importante rol de los datos en el desarrollo de los procesos de decisión algorítmicos en el ámbito militar, ver el informe de UNIDIR *The Role of Data in Algorithmic Decision-Making* (UNIDIR, 2019).

AMA con dicho módulo (ver fig.42):

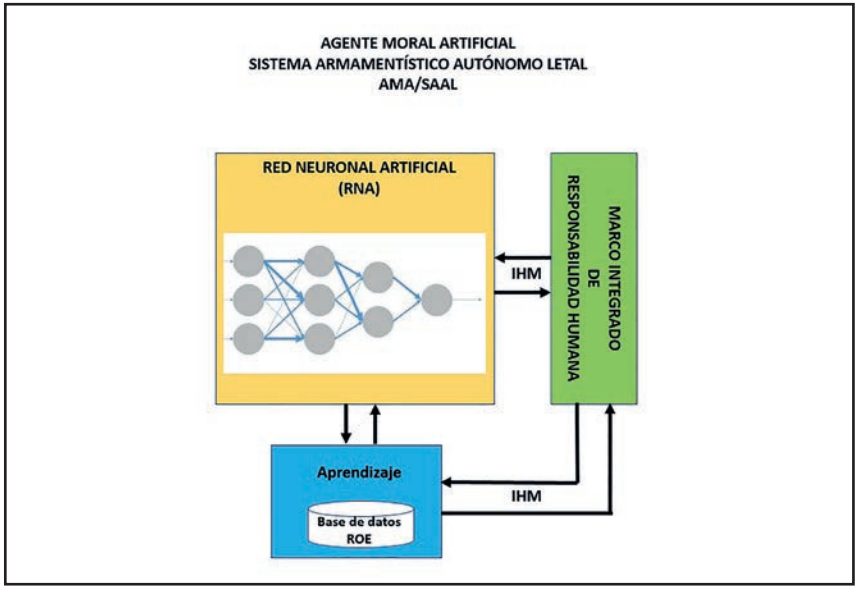


Fig. 42: Esquema de estructura de un AMA de un SAAL. Módulos computacionales del RNA, Marco Integrado de Responsabilidad Humana y Aprendizaje ((Xeridia y elaboración propia) (Xeridia, 2021).

Sería importante destacar, como hemos observado, que el módulo de aprendizaje y la utilización de los ROE como base, también alteran la responsabilidad de control y, por tanto, el Marco Integrado de Responsabilidad Humana. El siguiente esquema resume la modificación de dicho "Marco" (ver fig. 43) (Verdiesen *et al,* 2021: 151):

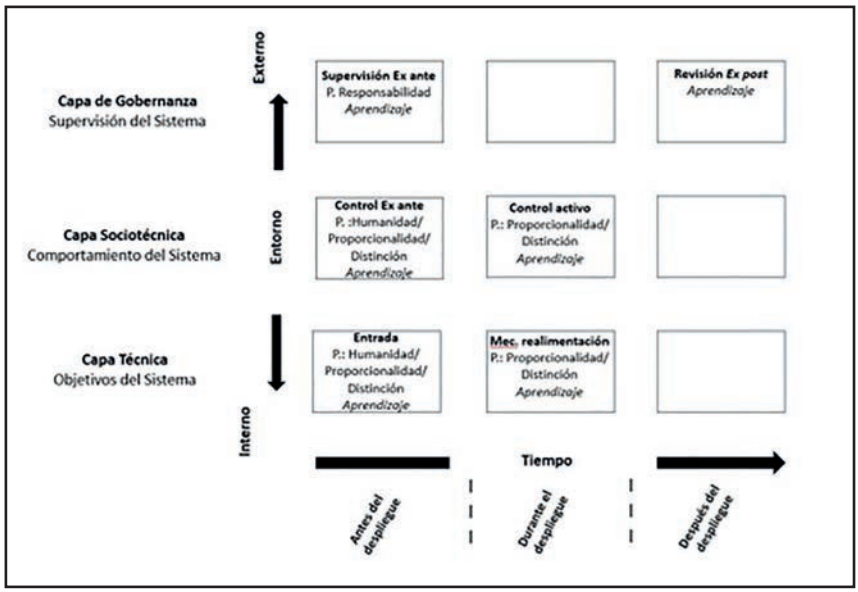


Fig. 43: Marco Integrado de Responsabilidad Humana (parcial) para un AMA de un SAAL, añadiendo el aprendizaje. Adaptado de I. Verdiesen et al (Verdiesen et al, 2021: 151 y elaboración propia).

Así, el control del aprendizaje afectaría tanto a la gobernanza *ex ante* y *ex post* del AMA, pero también tanto al control *ex ante* y activo de su comportamiento en la capa sociotécnica. Es más, el control técnico del entrenamiento (aprendizaje) previo al despliegue del AMA se considera imprescindible, así como el desarrollo de los mecanismos de realimentación de la base de datos de aprendizaje dentro de los objetivos del sistema para adaptar dicha base de datos, que alimentarán a la RNA del AMA, de acuerdo con las nuevas ROE que apareciesen para el SAAL en cuestión.

11.2- RENDICIÓN DE CUENTAS

La inexistencia, en la actualidad, de algoritmos que permitiesen un discernimiento completo del principio de distinción y por tanto con la incapacidad del AMA para seguir las directrices del DIH, daría pie a los argumentos de a E. Rosert y F. Sauer, así como de *Human Rights Watch,* de que los SAAL serían incapaces de discernir entre combatientes y civiles, violando el principio de distinción y siendo por tanto considerado como un sistema indiscriminado. Nosotros, sin embargo, somos de la opinión de la necesidad de ser pragmáticos y no maximalistas, aunque el AMA desarrollado debería tener la capacidad de seleccionar objetivos con un nivel de discriminación aceptable, al igual que habrían propuesto K. Anderson y M. Waxman. Una forma de calibrar el efecto de la utilización de un SAAL y del impacto del AMA sobre él, sería el establecer un mecanismo de "rendición de cuentas", íntimamente ligado al concepto de control del SAAL, que permitiese el mantenimiento de la responsabilidad por parte humana, siguiendo los principios rectores del GGE del CCW para los SAAL de las NU[142], dentro del denominado "Control Humano Significativo (MHC) (Anderson y Waxman, 2013: 41, 46; Bovens, 2007: 454; HRW, 2012: 30-31; NU, 2018a: 4; Rosert y Sauer, 2019: 370, 372-373).

A nuestro entender, una forma adecuada sería el desarrollo de un "registrador" de actividades" del AMA, que incluiría los resultados de la RNA, así como cualquier actividad de control humano sobre el AMA para una acción específica. Una especie de "caja negra" (*black box*), como la utilizada en las aeronaves o los tacógrafos digitalizados del transporte motorizado terrestre. Se crearía

142 Un análisis pormenorizado de dicho argumento se establece en el capítulo 7º de nuestra investigación.

una base de datos en un soporte extraíble (ej. chip de silicio) con condiciones indestructibles. Es importante tener en cuenta, que no solo se registraría la actividad del AMA y del SAAL, sino también cualquier interacción humana *ex ante*, durante y *ex post*, Incluiría las manipulaciones sobre la base de datos de aprendizaje, las órdenes de los comandantes y operadores del SAAL, los controladores del AMA, etc. Por lo tanto, dicho elemento del proceso de "rendición de cuentas", formaría parte del Marco Integrado de Responsabilidad Humana del AMA, por también del SAAL. El siguiente esquema resume dicho módulo registrador y las conexiones a los elementos principales de la estructura computacional del AMA/SAAL (ver fig. 44):

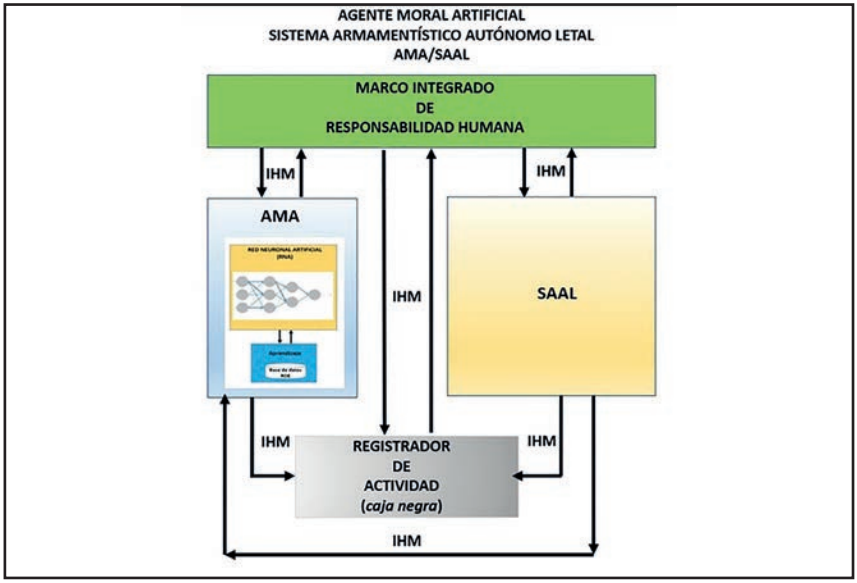


Fig. 44: Esquema de conexión del módulo "Registrador de Actividad" y las conexiones con el AMA, el SAAL y el Marco Integrado de Responsabilidad Humana (Xeridia y elaboración propia) (Xeridia, 2021).

Como se puede observar, las conexiones entre módulos se realizarían a través de una "Interfaz Hombre/Máquina" (IHM) apropiada y estandarizada común. Se ha omitido la conexión directa sentido AMA a SAAL permitiendo el contrario. La razón es que el nivel de desarrollo actual de los algoritmos aconseja que las decisiones del AMA sean analizadas por un control humano y siempre existiría una supervisión adecuada. Una gobernanza de seguridad preventiva que seguiría las propuestas de los Estados sobre un control holístico en todas las fases de desarrollo de los SAAL, reflejadas en sus intervenciones en el GGE del CCW, como en el caso de España (García, 2016: 95, 100, 109; NU, 2021e). Una vez creada la estructura modular del "registro de actividad" se modificaría el Marco Integrado de Responsabilidad Humana para incluir dicho control (ver fig. 45):

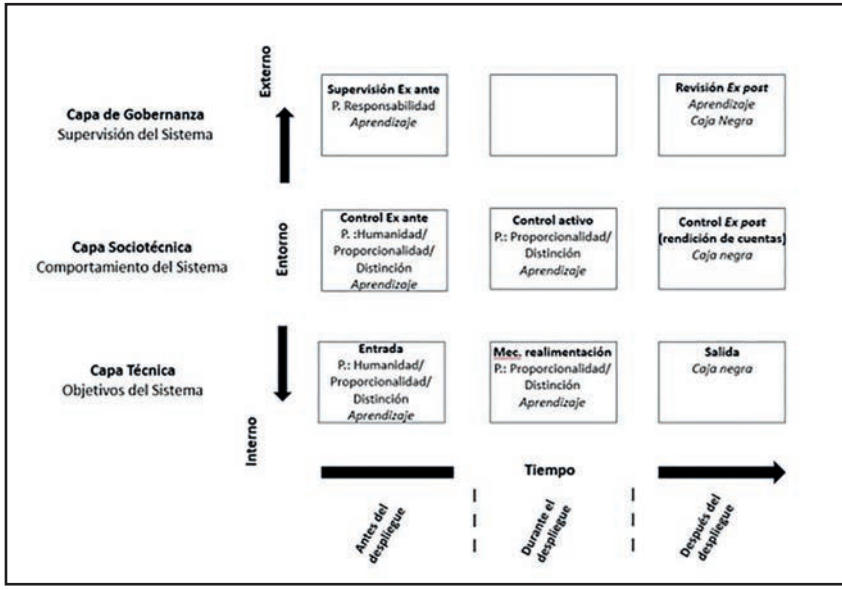


Fig. 45: Esquema del Marco Integrado de Responsabilidad Humana (parcial) con la inclusión del módulo de "Registro de actividad" (caja negra) (elaboración propia).

Observando el nuevo Marco de Responsabilidad Humana, en todas las capas (gobernanza, sociotécnica, técnica), la inclusión de una "caja negra" de "Registro de actividad" se considera imprescindible. En la capa de gobernanza ya que es una buena forma de supervisar el sistema y su funcionalidad, en la capa sociotécnica, pues muestra el comportamiento del AMA ante una ROE específica y los efectos del aprendizaje en la capa técnica, ya que permite observar si los objetivos del AMA se han cumplido.

11.3- CONTROL DE RIESGOS

Como ya analizamos en el capítulo 9º, una de las principales formas de mitigación riesgos, sería incrementar la seguridad del SAAL y del AMA incorporado. Dicho objetivo se podría lograr de diversas formas: mejoras tecnológicas para aumentar la fiabilidad y minimizar el posible "hackeo" del sistema (haciendo uso de elementos criptográficos); manteniendo una supervisión humana de su utilización (monitorización), a través de un procedimiento estandarizado de supervisión durante todo el ciclo de vida del AMA e; incorporando una auditoria "*ex post*", para lo cual sería extremadamente importante la existencia de un "Registro de actividad". Se añadiría la posibilidad de modulación y estandarización computacional, pues permitiría una minimización de riesgos, un aumento de su seguridad y un marco conceptual de planificación y diseño estable al mismo tiempo que reutilizable. El AMA, como argumentan D. Amodei *et al,* debería actuar, por lo tanto, como un "agente de impacto mínimo", estableciendo en cada fase "reguladores de impacto" sobre la base de datos de los ROE, los resultados de la "caja negra" y el estudio de un gran número de casos, tanto de entrenamiento como los reales de las acciones realizados. (Amodei

et al, 2016: 4-6; Sparrow, 2009: 172).

¿Cuál sería por tanto el nivel de riesgos aceptable de un AMA para un SAAL? Como observaron P. Lin *et al,* dependería tanto del nivel de riesgo aceptable del propio SAAL, las directrices establecidas en las ROE para cada acción, pero también por el impacto del DIH sobre sobre cada acción. Esto es, se tendrían en cuenta los objetivos a alcanzar, los resultados de las medidas de rendimiento obtenidas a través de casos de prueba para ROE's similares, usos reales previos o rendimientos teóricos establecidos, pero también el análisis del DIH para dicha ROE y el posible impacto del Derecho consuetudinario sobre dicha acción. Así, para completar el Marco Integrado de Responsabilidad Humana, se debería incluir el control de riesgos (minimización) en cada una de las fases (ver fig. 46) (CICR, 2005; Lin *et al,* 2008: 63-82):

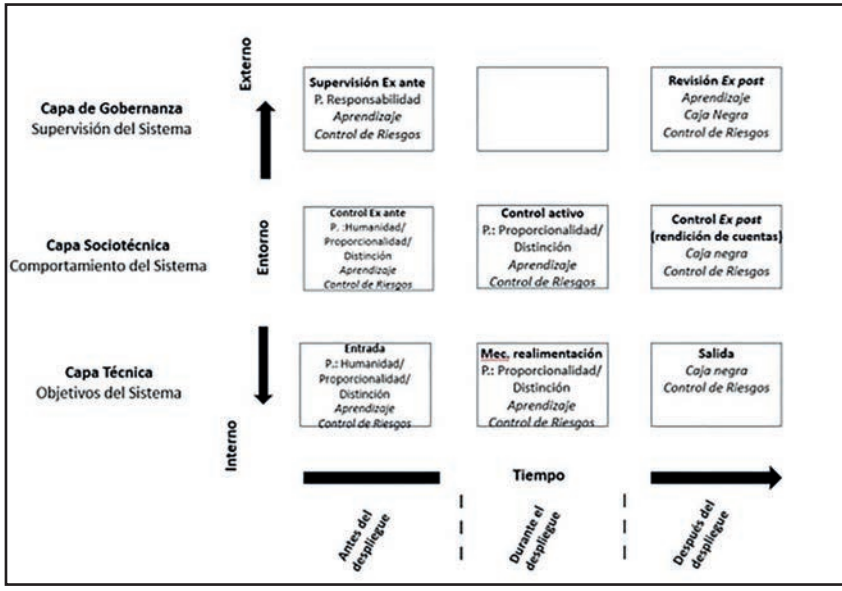


Fig. 46: Esquema del Marco Integrado de Responsabilidad Humana (completo) con la inclusión del Control de Riesgos en cada fase (elaboración propia).

Este desarrollo, cumpliría con la premisa de un control humano completo a través de todo el ciclo de vida del AMA, con excepción de una supervisión del sistema, dentro de la capa de gobernanza, durante el despliegue del AMA dentro del SAAL. Su desarrollo dependería del grado de autonomía que se desea tenga el SAAL. Cuanto más autonomía, menor será la capacidad de supervisión durante el despliegue. En el otro extremo, si el grado de autonomía desapareciese entonces se tendría una supervisión del sistema completo. En definitiva, lo que se pretende con el control de riesgos es tanto: conocer profundamente los elementos de control de cada fase, anticiparse a posibles problemas y tener una respuesta en caso de que surjan. Un control de riesgos que implicaría un control holístico tanto durante la fase de desarrollo, el entrenamiento o la experiencia real de los escenarios (Holland Michel, 2021: 12-16).

11.4- SÍNTESIS DE LA ESTRUCTURA COMPUTACIONAL

La estructura computacional propuesta de un AMA para un SAAL se ha desarrollo utilizando la estructura de diseño de S. Zieba, que se ha ido construyendo de forma teórica durante la Parte V de esta investigación. Como base se han adaptado estándares propuestos por entidades de reconocida experiencia, como la IEEE o la OCDE, siguiendo principios de modularidad y adaptabilidad (RTR) y los trabajos continuos en las NU para adaptar el Derecho internacional, en la práctica el DIH, a través de los GGE del CCW para los SAAL. El análisis ha llevado a la conclusión que el uso de la IA sería la base de la estructura computacional propuesta, a través de un marco genérico computacional basado en redes neuronales artificiales (RNA). Nuestra propuesta, además, ha consistido en un planteamiento híbrido, con una parte de estructura de “arriba-abajo”, con

el desarrollo de principios de carácter deontológico basados en los principios rectores del GGE del CCW sobre los SAAL, para la aplicación del DIH, como base de nuestra RNA, unido a otra estructura de "abajo-arriba", basada en las reglas de enfrentamiento ROE, que alimentarían la base de datos de aprendizaje de dicha red.

En paralelo, se ha desarrollado, teóricamente, un Marco Integrado de Responsabilidad Humana, que ha servido para establecer el control holístico humano, en todas las fases (*ex ante,* durante y *ex post*) del desarrollo del AMA, analizando las necesidades de control y el control de riesgos. Como resultado se ha obtenido una propuesta de AMA para un SAAL, que se resume en los dos siguientes esquemas complementarios (ver figs. 47 y 48):

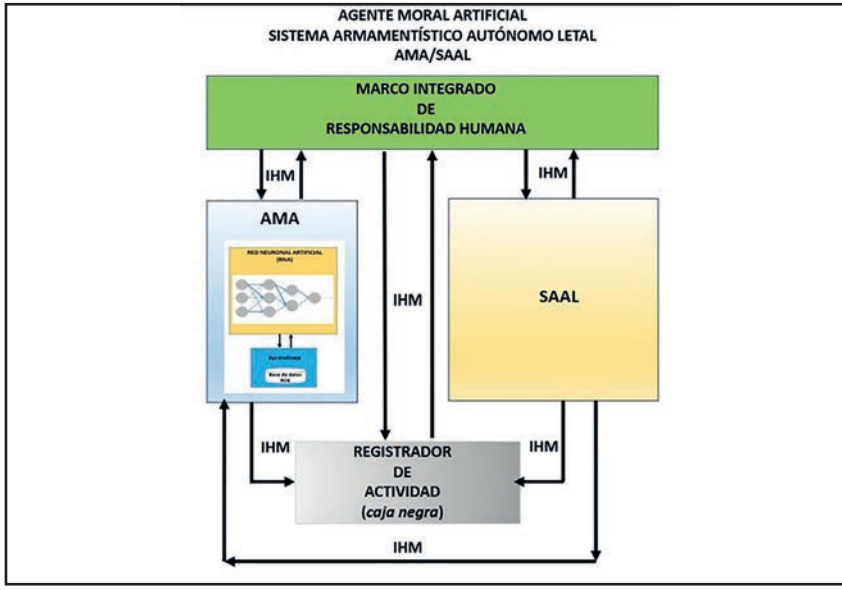


Fig. 47: Esquema teórico integrado completo de un AMA par un SAAL (elaboración propia).

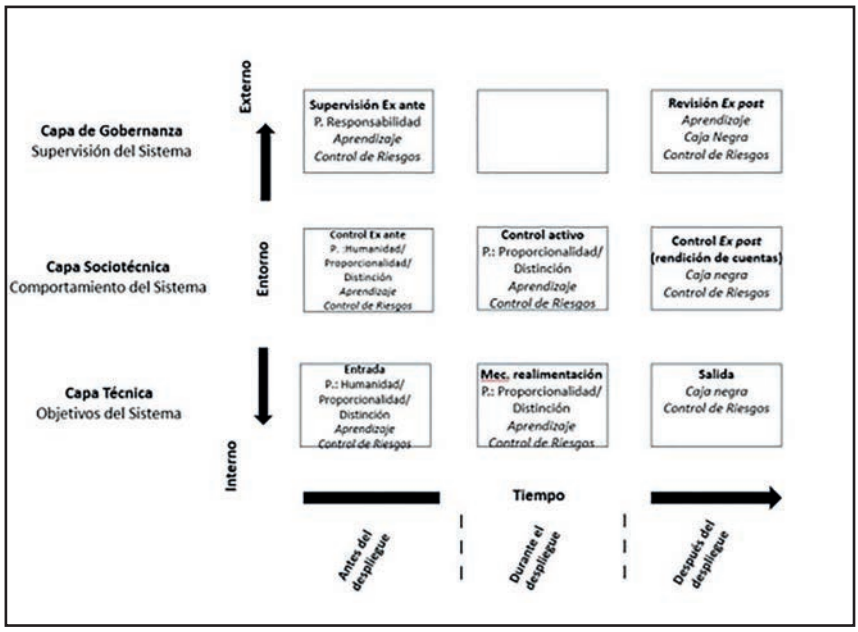


Fig. 48: Esquema completo del Marco Integrado de Responsabilidad Humana para un AMA de un SAAL (elaboración propia).

El primer esquema resume el AMA con sus principales módulos y su conexión con el SAAL. El siguiente, complementario y de carácter fundamental, desarrolla el Marco Integrado de Responsabilidad Humana para dicho AMA, en todas sus fases. En dicho Marco resaltaríamos los siguientes elementos:

- Las ROE institucionalizan la fuerza de los Estados y son un elemento clave en el principio rector de la responsabilidad y por lo tanto lo consideramos como una pieza esencial en el proceso de aprendizaje adaptativo, a través de casos, de la RNA propuesta;

- Aún restringiendo el número de sensores y actuadores, el desarrollo de la RNA propuesta sería complejo y no exento de ambigüedades, lo que implicaría una adecuación del control de riesgo aceptable para el AMA;
- El riesgo aceptable de un AMA dependerá de los objetivos a alcanzar, el resultado del aprendizaje, tanto en la fase de entrenamiento como en la real y el impacto del DIH sobre cada ROE específico;
- Consideramos imprescindible el establecimiento de un registrador de actividades (caja negra) tanto del AMA como del SAAL, al ser una pieza fundamental del proceso de rendición de cuentas y de minimización del riesgo.

Como corolario final recalcamos que dicha estructura computacional sería eminentemente teórica y de alto nivel. Tendría en cuenta las dificultades de plasmar el DIH en un algoritmo de IA para un AMA de un SAAL, y las restricciones computacionales existentes hasta la utilización generalizada de la computación cuántica, por lo que llevaría a una solución pragmática de una construcción computacional basada en principios rectores básicos de "*soft law*" consensuados internacionalmente, que tomarían como referencia elementos principales del DIH y las ROE, lo que aportaría una primera base de responsabilidad jurídica a los SAAL utilizando como herramienta un AMA.

PARTE V

CONCLUSIONES

CAPÍTULO 12
CONCLUSIONES

La investigación llevada a cabo ha tenido como premisa de fondo explorar las dificultades y desafíos encontrados, tanto a nivel jurídico como computacional, en la aplicación del Derecho Internacional en el ciberespacio y más concretamente en los SAA y lo SAAL. A tal fin, hemos configurado un triple abordaje complementario: un estudio del marco conceptual: una panorámica actualizada de la aplicación de la responsabilidad jurídica internacional ha dicho marco armamentístico y; una propuesta de utilización de los AMA como herramienta de uso. Dicho proceso ha identificado una serie de complejidades, dificultades y necesidades que lastran una adecuada implementación del Derecho Internacional en los sistemas armamentísticos que utilizan la IA como base y que a continuación resumimos:

I

El ciberespacio y la IA han propiciado la creación de un nuevo paradigma: la ciberética, más acorde con la ética aplicada computacional qué con la ética normativa tradicional. Dicho cambio de paradigma ha estresado la capacidad de aplicación de los principios morales (reglas) existentes surgidos de la ética normativa a dicho entorno.

II

El cambio de paradigma ético no significa que se deba sustituir la ética normativa por la ética aplicada, sino que debemos ser capaces de establecer una relación simbiótica entre la ética tradicional y la ciberética, manteniendo por un lado el enfoque reglamentario, es decir, los principios morales consensuados internacionalmente ya existentes, pero adaptándolos al nuevo paradigma. Se añadirían a dicho marco, aquellos nuevos principios éticos surgidos del nuevo dominio y consecuentemente las nuevas normas derivadas de la ciberética, surgidas especialmente por la introducción de la IA, para paliar los posibles vacios conceptuales, normativos y de diseño..

III

Particularmente, en el entorno de la responsabilidad jurídica internacional de los SAA y los SAAL, plasmada en el DIH y los DD. HH., nuestra postura sería mantener las normas existentes, pero identificando y corrigiendo aquellos vacíos conceptuales surgidos de la ciberética aplicada a la ciberguerra. Debido a la situación geopolítica internacional actual, que descarta el desarrollo de nuevos tratados de Derecho Internacional, se debería identificar y solucionar los posibles vacíos de ética normativa surgidos de la ciberética a través de mecanismos no estrictamente vinculantes (*soft law*), que en un futuro puedan incorporarse al Derecho consuetudinario.

IV

En las guerras posmodernas, los medios cibernéticos tendrán cada vez un rol más importante en los conflictos armados y el impacto de las computadoras y la IA introducirán importantes cambios del rol humano en las guerras y en su complejidad. Como consecuencia, dicho marco de referencia tendrá una incidencia cada vez mayor sobre el Derecho aplicable a los conflictos armados (DICA) y el DIH, especialmente al aplicar las nuevas teorías de la disuasión o la ampliación de la guerra híbrida.

V

En dicho contexto, dentro de la ambigüedad, consideramos de especial relevancia el uso por parte de los Estados de las "guerras por delegación", que extienden el "anonimato" de las operaciones cibernéticas y el uso frecuento de los denominados "*hackers*". Eventos, dentro del marco del *ius ad bellum*, que, aunque incompatibles con el Derecho Internacional, son considerados aceptables desde el punto de vista operativo por parte de los Estados. Particularmente, cuando se habla de acciones preventivas de ciberdefensa que serían contrarias al "derecho de legítima defensa", de acuerdo con el Art. 51 de la Carta de las NU, y que tienen especial impacto sobre el principio de "proporcionalidad".

VI

En todo caso, consideramos que la aplicación del Derecho Internacional en el ciberespacio es una lucha continua en "las fronteras de la ley", existiendo zonas obscuras y ambiguas que los Estados explotan a su conveniencia, especialmente cuando se entra en territorios inexplorados como la guerra híbrida o el desarrollo de nuevos artefactos cada vez más autónomos y con más IA.

VII

Particularmente, existe en la actualidad una debilidad del sistema de gobernanza, pues asistimos a una "gobernanza jurídica asimétrica" ya que, aunque las NU en sus resoluciones han establecido que el Derecho Internacional y la Carta de las Naciones Unidas son de aplicación al ciberespacio, la problemática surge cuando se intenta establecer la forma de su aplicación y si para ello es necesario establecer un nuevo instrumento jurídico. Teniendo en cuenta los intercambios entre Estados y los debates en la NU, nuestro punto de vista es que será difícil que a corto o medio plazo se establezca un nuevo tratado, pues los Estados prefieren medidas internas a nivel estatal y solo contemplan el intercambio de información entre Estados de manera voluntaria (China, Rusia, USA y España, entre otros).

VIII

Consideramos que es, por tanto, esencial el desarrollo de instrumentos jurídicos no vinculantes (*soft law*), que ofrezcan respuesta a la falta de nuevos tratados. Tanto los trabajos de la OTAN, a través de los "Manuales de Tallinn" o los del CICR sobre el Derecho consuetudinario son de gran interés, al proporcionar un marco interpretativo estable. Ahora bien, no se debe obviar que dichos instrumentos presentan diferencias para una misma acción, como por ejemplo las diferencias de interpretación de la definición de "ataque". Es por ello, que los trabajos institucionales a nivel internacional son de especial relevancia. Los trabajos del GGE del CCW sobre los SAAL y el desarrollo de principios rectores son fundamentales, para una aplicación consensuada del Derecho Internacional en los sistema armamentísticos cibernéticos. Hay que tener en cuenta que los principios de humanidad, necesidad, proporcionalidad y distinción forman el núcleo principal del DIH.

IX

La complejidad del ciberespacio y las dificultades de control jurídico de los nuevos artefactos SAA y SAAL, hacen imprescindible, según nuestro criterio, que se debe implementar una prohibición de los denominados "armamentos indiscriminados", que no siguen el principio de distinción, así como la vigencia actual de la "Clausula Martens" en aquellas instancias donde el Derecho Internacional vigente fuese de difícil aplicación. Es más, la reacción ante la

ambigüedad debería establecer el principio de precaución como la base para el desarrollo de los SAA y los SAAL, ya desde la fase de investigación y desarrollo.

X

Particularmente, quedan aún por solucionar aspectos importantes de la responsabilidad jurídica internacional dentro del marco cibernético. Los más relevantes serían:

- *La definición de "autonomía"*: La falta de un consenso internacional sobre lo que se entiende por "autonomía" de un sistema armamentístico o lo que significa un objeto como un SAA o un SAAL, hace que, en la actualidad, existan múltiples definiciones de dichos términos, a nivel de los Estados, que dificultan una aplicación estable de la responsabilidad jurídica internacional. Dicha referencia auna a los sistemas semiautónomos y los completamente autónomos, únicamente posibles cuando se alcance la Singularidad;
- *La definición de "ataque armado" según el Derecho Internacional*: Todos los ataques armados son usos de la fuerza, pero no todos los usos de la fuerza son considerados ataques armados. Sería necesario reconceptualizar la noción de severidad dentro del dominio del ciberespacio;
- *La definición de "objeto"*: No están resueltos los aspectos de visibilidad y tangibilidad, la consideración de los datos como "objetos virtuales" o aquellos objetos de doble uso;

- *La definición de "necesidad militar"*: Existen graves discrepancias entre los Estados y las instituciones internacionales. Nuestro punto de vista se alinea con el expresado por la Comisión de Derecho Internacional (CDI/ILC): la necesidad no es admisible para no aplicar el DIH;
- *La aplicación práctica del Derecho Internacional y del Derecho consuetudinario*: Consideramos imprescindible pasar de la teoría a la práctica y establecer mecanismos de implementación en los nuevos sistemas armamentísticos, especialmente en los SAA y SAAL. Establecer criterios internacionales sobre el principio de responsabilidad y los criterios de rendición de cuentas, así como un control de riesgos holístico de los sistemas armamentísticos, basado en los principios de Derecho Internacional vigentes, para lo que se necesitará una adecuada formación de todos los actores y el trabajo conjunto de diversas disciplinas (jurídica, científica, militar, etc.).

XI

A nuestro entender será imperativo desarrollar un control y supervisión humana holística, utilizando todos las herramientas tecnológicas posibles, para no permitir "ataques indiscriminados" incontrolados de "entidades" artificiales. Significaría la adquisición por parte de un SAA o un SAAL de una "moralidad operativa", donde la IA subyacente de un AMA teórico funcionaría únicamente según su diseño, que sería modulable y escalable a través del aprendizaje y la validación, pero no permitiendo que dicho aprendizaje resul-

tase en que el AMA alcanzase la condición de AMAA, una moralidad propia de la máquina, no necesariamente similar a la humana, cuyas consecuencias serían imprevisibles.

XII

Dichos AMA podrían servir como mecanismos de implementación de la responsabilidad jurídica internacional en los SAA y los SAAL. Serían, desde nuestro punto de vista, algoritmos representativos de una herramienta social y no un "agente" en sentido estricto, ya que la única forma de serlo sería poder pasar un "Test de Turing Moral", donde no se podría identificar la diferencia entre máquina y humano con concepto de "ser" que habría alcanzado la Singularidad . Maximizaría lo bueno frente a lo malo, pero siempre dentro de un contexto en un espacio-tiempo determinado, que dependería del exterior a través de una normas de grupo, pero también de la situación real operativa en un determinado momento. Dependería de un proceso de aculturación (valores, normas e instituciones) y necesitaría de una gobernanza ética por parte de diseñadores y organizaciones responsables de su despliegue. Dicha proceso de aculturación y de gobernanza debería surgir de un consenso internacional amplio, por lo que consideramos que los principios rectores desarrollados y consensuados por el GGE del CCW para los SAAL de las NU, serían una buena base para un hipotético algoritmo para los AMA.

XIII

A tal fin, para el desarrollo de hipotéticos AMA, para los SAA o los SAAL, nuestra valoración apuntaría al uso de una metodología de autonomía adaptable, variable en el espacio-tiempo, a través de un enfoque híbrido, en un proceso de "arriba-abajo" (*top-down*), basado en la ética deontológica (valores adquiridos) y expresado por el DIH o por los principios rectores consensuados internacionalmente basados en él. Se complementaría, por un mecanismo de aprendizaje "abajo-arriba" (*bottom-up*) basado en la experiencia, que podría surgir del campo operacional, a través de las ROE. Dicho proceso estaría sujeto a una complejidad creciente, debido a la propia complejidad del Derecho Internacional y la infinidad de situaciones operacionales posibles, pero también por el uso creciente de la IA, como las RNA.

XIV

No obstante, se debería tener siempre presente que los SAA y los SAAL son artefactos que integran IA con una autonomía cada vez más creciente, cuyo principal fin, en un conflicto armado, será la destrucción del adversario. Por lo tanto, lo crucial será establecer el impacto de dicho uso de la fuerza, en un entorno de "Realidad Mixta", y la capacidad que pueda tener una herramienta de control, como los AMA, cuando utilizase los principios rectores propuestos por las NU, como base deontológica algorítmica. Particularmente, consideramos que aún existen dificultades para discernir como se

crean, moldean y definen los principios rectores de las NU para los SAA y los SAAL. A nuestro entender, los puntos más relevantes aún no solucionados serían:

- Dentro del DICA, como establecer el equilibrio entre la necesidad militar y la humanidad;
- Cómo aplicar el principio de distinción, dentro de la nebulosa de los objetos de uso dual;
- Si es posible lograr, en un futuro próximo, un consenso de los Estados sobre la definición de lo que se considera sistema "autónomo" o "semiautónomo";
- Si es posible evaluar, dependiendo del grado de independencia de los SAAL, cuales serían sus efectos en todo momento: diseño, uso o sobre la vulnerabilidad de los afectados;
- Con relación al principio de proporcionalidad, cómo se define el término "excesivo" del Art. 51(5)b del PAI (*ius in bello*), cómo se cuantifica el daño no desproporcionado de la "guerra justa" en el *ius ad bellum* o cómo se podría cuantificar un hipotético daño futuro en el *ius post bellum*;
- Cómo establecer un método de control adecuado de supervisión del desarrollo de nuevos SAAL, para que estén de acuerdo con el DIH, según el Art. 36 del PAI;
- ¿Es posible establecer un sistema de "rendición de cuentas" *ex ante* y *ex post* adecuados?

XV

Además, nuestra idea es que, en la actualidad, existen aún importantes desafíos para que un SAA o un SAAL observen los distintos principios consensuados en la NU y menos aún el Derecho Internacional vigente:

- Los SAA y los SAAL actuales no serían capaces de asumir el principio de distinción y por lo tanto serían considerados como sistemas indiscriminados, ya que:
    - Existe un entorno computacional débil: los sensores en la actualidad no tendrían la capacidad de alta discriminación necesarias;
    - El marco de actuación es impredecible: El entorno de combate y la velocidad de los cambios crean un entorno computacional algorítmico inadecuado, necesitando de un entorno más apropiado como el de la computación cuántica;
    - Los algoritmos son débiles: la construcción modular los hace menos predecibles;
    - Los algoritmos son incapaces, en la actualidad, de distinguir entre combatientes y no combatientes pues sería necesario que tuviesen la capacidad de "juicios de valor".
- Al igual que el Tribunal Penal Internacional sobre la ex Yugoslavia, nuestro punto de vista es que el principio de proporcionalidad existe, pero no se sabe lo que significa o como se aplicaría en un entorno cibernético;

- Existe aún un gran vacío, tanto *ex ante* como *ex post*, sobre cómo aplicar la responsabilidad y la "rendición de cuentas" sobre el uso de los SAAL en un conflicto armado. Es decir, no existe un entorno de gobernanza adecuado;
- No existe harmonización internacional sobre la aplicación del Art. 36 del PAI a los SAAL.

Abogamos, por tanto, que exista un "Control Humano Significativo" (MHC) y que se llegue a un acuerdo sobre su definición, para pasar de lo abstracto a lo concreto. Para ello proponemos, la creación de un "Marco Integrado de Supervisión Humana" a lo largo de todas las fases de un SAA o un SAAL: creación, despliegue, uso y revisión. Es más, creemos necesario que la aplicación del Art. 36 del PAI por los Estados debería plantear dos preguntas fundamentales antes de autorizar su despliegue: ¿el SAA o el SAAL son capaces de cumplir con el DIH? y ¿el algoritmo de diseño viola el DIH? En dicho entorno, los AMA serían una de las posibles herramientas, pero no la única, para paliar parte de las deficiencias encontradas.

XVI

No debemos olvidar, que cuanto más libertad tenga una máquina más será necesario que tenga unos principios morales, así como un método computacional apropiado para llevarlo a cabo. El fin último será establecer una gobernanza ética robusta para incenti-

var la confianza en la IA y, por lo tanto, se debe descartar la idea de que un AMA serviría como alivio del control humano, pues se incrementaría dicho control en dos apartados: el control del diseño, ya que todos los actores necesitarían conocerlo y; el control de uso a través de una monitorización permanente. En definitiva, una evaluación continua del AMA según su capacidad para llevar a cabo su cometido. Dicho control, también serviría para eliminar signos de opacidad y de sesgos a la vez que se incrementaría la trazabilidad. En dicho contexto, los datos formarían una parte esencial de la capacidad de un AMA para llevar a cabo su cometido, datos que necesitarían de un alto grado de fiabilidad y conocimiento, pues se mantendría la máxima de que "basura entrante, basura saliente".

Particularmente, debido al gran número de elementos para construir un AMA, consideramos que los desafíos pueden ser extensos y complejos, por lo que su utilización, en un principio, debería estar restringida a escenarios bien definidos, implementar infraestructuras sensoriales robustas y, sobre todo, restringir la toma de decisiones por el artefacto, para que el ser humano sea el responsable último en la toma de decisiones.

XVII

Nuestra visión, por lo tanto, se basaría en la idea de que sería imprescindible establecer los requerimientos de ingeniería de los AMA, tanto de la ética de la propia ingeniería como la de la máquina. Particularmente, no se deberían cometer errores de sobres-

timación de la capacidad de un método ético para resolver casos morales complejos, ni minimizar la importancia de la metodología a utilizar.

XVIII

A nivel particular de los AMA para los SAAL, consideramos que somos escépticos de que en un futuro cercano se pueda adoptar el DIH completo a un algoritmo a través de sensores, pues existirían tres condicionantes: una capacidad computacional limitada; un posible conflicto interno entre reglas y; la gran variedad de situaciones operacionales existentes. En todo caso, un AMA debería observar tres principios básicos: transparencia, responsabilidad y rendición de cuentas, para lo cual necesitaría de una gran adaptabilidad, una interacción continua con el medio y una autonomía controlable por el ser humano. Dicho proceso se necesitará a lo largo de todo el ciclo de vida del AMA, así como un control de riesgos aceptable para lograr, en lo posible, que se convierta en un agente de impacto mínimo.

XIX

Dada la estructura híbrida propuesta, el algoritmo del AMA mantendría dicho marco, pero estableciendo un desarrollo que limitaría al mínimo las reglas a utilizar, restringiendo el número de sensores y actuadores, para hacer que el algoritmo fuese manejable y fiable. Se utilizaría una RNA (perceptrón multicapa) como base computacional con sus correspondientes pesos y sesgo, que tendría que

observar un nivel legal de discriminación aceptable. En particular para un SAA o un SAAL, la entrada estaría formada por los principios rectores esenciales establecidos por la NU, así como por un Marco Integrado de Responsabilidad Humana, que aportaría el control efectivo durante todo el ciclo de vida del AMA.

XX

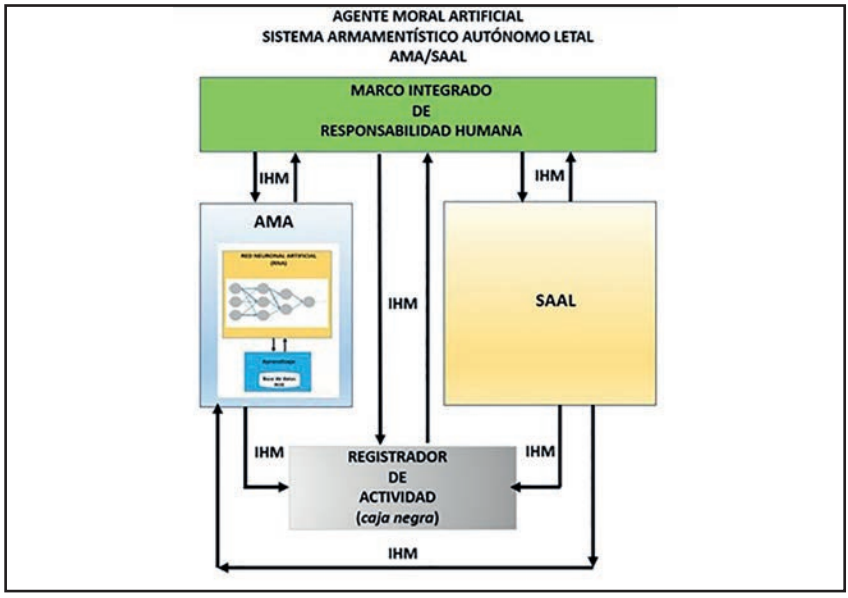
El proceso híbrido se completaría a través del aprendizaje en tiempo real del SAA o el SAAL, por medio de la creación de una base de datos que reflejaría las diversas ROE operacionales posibles. En dicho marco resaltaríamos los siguientes elementos:

- Las ROE institucionalizan la fuerza de los Estados y son un elemento clave en el principio rector de la responsabilidad y por lo tanto lo consideramos como una pieza esencial en el proceso de aprendizaje adaptativo por casos de la RNA propuesta;
- Aunque se restrinjan el número de sensores y actuadores, el desarrollo de la RNA propuesta sería complejo y no exento de ambigüedades, lo que implicaría la necesidad de una adecuación continua del control de riesgo aceptable para el AMA;
- El riesgo aceptable de un AMA dependerá de los objetivos a alcanzar, el resultado del aprendizaje, tanto en la fase de entrenamiento como en la real y el impacto del DIH sobre cada ROE específico;
- Consideramos imprescindible el establecimiento de un re-

gistrador de actividades (caja negra) tanto del AMA como del SAA o del SAAL, al ser una pieza fundamental del proceso de rendición de cuentas y de minimización del riesgo.

XXI

En síntesis, la estructura computacional de un AMA, para un SAA o SAAL, podría tener la siguiente forma:



Recalcaremos que dicha estructura computacional es una propuesta teórica y de alto nivel. Tendría en cuenta las dificultades de plasmar el DIH en un algoritmo de IA para un AMA de un SAAL, y las restricciones computacionales existentes, hasta la hipotética utilización generalizada de la computación cuántica, por lo que

llevaría a una solución de autonomía adaptable, que dependería de los avances a todos los niveles: éticos, sociales, jurídicos, militares, tecnológicos, etc. Una construcción computacional basada en principios rectores básicos de "*soft law*" consensuados internacionalmente de las NU (elementos principales del DIH) y las ROE, unido a un Marco Integrado de Responsabilidad Humana holístico, lo que aportaría una primera base de responsabilidad jurídica a los SAAL aplicando como herramienta un AMA.

PARTE VI

BIBLIOGRAFÍA

BIBLIOGRAFÍA

ABNEY, K. (2012): "Robotics, Ethical Theory, and Metaethics: A guide for the Perplexed", en P. Lin, K. Abney y G. A. Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge – Londres, 35-52.

AGENCIA EFE (EFE) (2020): "Los drones campearon en los cielos de Karabaj (10 octubre 2020)", Agencia EFE, Madrid, acceso febrero 2021, en https://www.efe.com/efe/america/destacada/los-drones-campearon-en-cielos-de-karabaj/20000065-4364853

AGUAYO, P. (2011): "La teoría de la abducción de Pierce: Lógica, metodología e instinto", *Ideas y Valores,* 145, 33-53.

ALA-PIETILÄ, P. *et al* (2019): *Directrices Éticas para una IA Fiable,* N. Smuha (ed.), Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial, Comisión Europea, Bruselas, acceso septiembre 2020, en https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

ALÍA PLANA, J. M. (2009): "Las Reglas de Enfrentamiento (ROE)", *monografías.com*, acceso octubre 2021, en https://www.monografias.com/trabajos71/reglas-enfrentamiento-roe/reglas-enfrentamiento-roe.shtml

ALLEN, C. (2008): "Machine Morality: bottom-up and top-down approaches for modelling human moral faculties", *Artificial Intelligence and Society,* 22(4), 565-582.

ALLEN, C., VARNER, G. y ZINSER, J. (2000): "Prolegomena to any future artificial moral agent", *Journal of Experimental Theory in Artificial Intelligence,* 12, 251-261.

ALLEN, C., SMIT, I. y WALLACH, W. (2005): "Artificial morality: Top-down, bottom-up and hybrid approaches", *Ethics and Information Technology,* 7, 149-155.

AMODEI, D., OLAH, C., STEINHARDT, J., CHRISTIANO, P., SCHULMAN, J. y MANÉ, D. (2016): *Concrete Problems in AI Safety,* Cornell University, , Ithaca- NY, acceso abril 2021, en https://arxiv.org/pdf/1606.06565.pdf

ANDERSON, J. R., BOTHELL, D., BYRNE, D., DOUGLASS, S., LEBIERE, C. y QIN, Y. (2004): "An Integrated Theory of the Mind", *Psychological Review,* 111, 4, 1036-1060.

ANDERSON, K. y WAXMAN, M. (2013): "Law and Ethics for Robot Soldiers", *Policy Review,* 176, 35-49.

ANDERSON, M. y LEIGH ANDERSON, S. (2007): "Machine Ethics: Creating an Ethical Intelligent Agent", *Artificial Intelligence Magazine,* 28(4), 15-26.

ANDERSON, M. y LEIGH ANDERSON, S. (2014): "GenEth: A general ethical dilemma analyzer", *Paladyn Journal of Behavioral Robotics,* 9, 337-357.

ARKIN, R. (2007): *Governing Lethal Behaviour: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture,* Technical Report GIT-GVU-07-11, Georgia Institute of Technology, Atlanta, acceso febrero 2021, en https://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf

ARKIN, R. (2013): "Lethal Autonomous Systems and the Plight of the Non-Combatant", *AISB Quaterly,* acceso febrero 2021, en https://www.cc.gatech.edu/ai/robot-lab/online-publications/aisbq-137.pdf

ARQUILLA, J. y RONFELDT, D. (1993): "Cyberwar is Coming!", *Comparative Strategy,* vol. 12, 2, 141-165.

ARTICLE 36 (2014): *Key areas for debate on autonomous weapons systems. Memorandum for delegates at the Convention on Certain Conventional Weapons (CCW). Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) Geneva, 13-16 May 2014*, Article 36, Londres, acceso febrero 2021, en https://article36.org/wp-content/uploads/2014/05/A36-CCW-May-2014.pdf

ARTICLE 36 (2016): *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons. Briefing paper for delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, 11-15 April 2016*, Article 36, Londres.

ASARO, P. M. (2012): "", *International Review of the Red Cross* 94(886), 687-709.

ASARO, P. M. (2016): "The Liability Problem for Autonomous Artificial Agents", *2016 AAAI Spring Symposium Series – Ethical and Moral considerations in Non-Human Agents,* Association for the Advancement of Artificial Intelligence, Palo Alto, 190-194.

ASSOCIATION FOR COMPUTER MACHINERY US PUBLIC COUNCIL (USACM) (2017): *Statement on Algorithmic Transparency and Accountability,* ACM US Public Policiy Council, Washington, acceso enero 2021, en https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

AYALEW, Y. E. (2015): "Cyber Warfare: A new hullaballoo under International Humanitarian Law", *Beijing Law Review,* 6, 209-223.

BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE (2019): *Beijing AI Principles,* Beijing Academy of Artificial Intelligence, Pekín, acceso marzo 2021, en https://www.baai.ac.cn/news/beijing-ai-principles-en.html

BARRETT, L.F., MESQUITA, B., OCHSNER, K. N. y GROSS, J. J. (2007): "The Experience of Emotion", *Annual Review of Psychology,* 58, 373-403.

BATHAEE, Y. (2018): "The Artificial Intelligence Black Box and the failure of intent and causation", *Harvard Journal of Law & Technology,* 31(2), 889-938.

BEAUCHAMP, T. L. Y BOWIE, N. E. (2001): *Ethical Theory and Business (6th Edition),* Prentice Hall, Upper Saddle River-NJ.

BENDIEK, A. y METZGER, G. (2015): "Deterrence Theory in the Cyber-century: Lessons from a state-of-the-art literatura review", en D. W. Cunningham *et al* (eds), *Informatik 2015*, GI-Edition Lecture Notes in Informatics Proceedings, Bonn, 553-570.

BENJAMINS, R. (2021): "A choices framework for the responsable use of AI", *Artificial Intelligence and Ethics,* 1, 49-53.

BERKEBILE, R. (2018): "New Generation Warfare and the Just War Tradition", *InterAgency Journal,* 9(3), 17-33, acceso enero 2021, en https://www.researchgate.net/publication/327253483_New_Generation_Warfare_and_the_Just_War_tradition

BERKOWITZ, B. D. (1997): "Warfare in the Information Age",en J. Arquilla y D. F. Ronfeldt (eds.), *In Athena's Camp: Preparing for Conflict in the Information Age,* RAND Corporation, 175-189, acceso diciembre 2020, en https://www.rand.org/pubs/monograph_reports/MR880.html

BIRNBACHER, D. (2016): "Are Autonomous Weapons Systems a threat to human dignity?", en N. Bhuta (ed.), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, Cambridge, 105-121.

BODE, I. y HUELSS, H. (2018): "Autonomous weapons systems and changing norms in international relations", *Review of International Studies,* 44(3), 393-413.

BOLETÍN OFICIAL DEL ESTADO (1980): *Instrumento de adhesión de 2 de mayo de 1972, del Convenio de Viena sobre el Derecho de los Tratados, adoptado en Viena el 23 de mayo de 1969*, BOE nº 142, 13 junio 1980, páginas 13099-13110.

BOLETÍN OFICIAL DEL ESTADO (1989): *Instrumentos de Ratificación de los Protocolos I y II adicionales a los Convenios de Ginebra de 12 de agosto de 1949, relativos a la protección de las víctimas de los conflictos armados internacionales y sin carácter internacional, hechos en Ginebra el 8 de junio de 1977*, BOE nº 177, 26 julio 1989, páginas 23828-23863.

BOLETÍN OFICIAL DEL ESTADO (2002): *Instrumento de Ratificación del Estatuto de Roma de la Corte Penal Internacional, hecho en Roma el 17 de julio de 1998,* BOE nº 126, 27 mayo 2002, páginas 18824-18860.

BOLETÍN OFICIAL DEL ESTADO (2019): *Orden PCI/487/2019, de 26 de abril, por la que se publica la Estrategia Nacional de Ciberseguridad 2019, aprobada por el Consejo de Seguridad Nacional*, Ministerio de la Presidencia, Relaciones con las Cortes e Igualdad, BOE nº 103 -Sección I-Página, 43437.

BONNEMAINS, V., SAUREL, C. y TESSIER, C. (2018): "Embedded ethics: some technical and ethical challenges", *Ethics and Information Technology,* 20, 41-58.

BOOTHBY, W. (2009): *Weapons and the Law of Armed Conflict,* Oxford University Press, Oxford.

BOSTROM, N. (2003): "Ethical Issues in Advanced Artificial Intelligence", *PhilArchive,* Centre for Digital Philosophy, Western University Ontario, London-Ontario, acceso enero 2021, en https://philpapers.org/archive/BOSEII.pdf

BOULININ V. y VERBRUGGEN, M. (2017): *Article 36 Reviews. Dealing with the challenges posed by emerging technologies,* Stockholm International Peace Resarch Institute, Estocolmo.

BOVENS, M. (2007): "Analysing and Assessing Accountability: A Conceptual Framework", *European Law Journal,* 13(4), 447-468.

BOYD, D. y CRAWFORD, K. (2011): "Six Provocations for Big Data", *Symposium: A Decade in Internet Time: The Dynamics of the Internet and Society (september 2011)*, acceso abril 2020, en https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

BRAGA, A. y LOGAN, R. K. (2017): "The Emperor of Strong AI Has no Clothes: Limits to Artificial Intelligence", *MDPI Information*, 8, 156-177.

BREY, P. (2000): "Disclosive Computer Ethics", *Computers and Society,* December, 10-16.

BREY, P. (2014): "From Moral Agents to Moral Factors: The Structural Ethics Approach", en P. Kroes *et al* (eds.), *The moral status of technical artifacts,* Springer Nature, Nueva York-Heidelberg, 125-142.

BRITISH BROADCASTING CORPORATION (2020): "SolarWinds: Why de Sunburst hack is so serious (16/12/2020)", British Broadcasting Corporation, Londres, acceso diciembre 2020, en https://www.bbc.com/news/technology-55321643

BROEKENS, J. (2010): "Modeling the Experience of Emotion", *International Journal of Synthetic Emotions,* 1, 1-17.

BROWN, G. D. (2017): "International Law applies to Cyber Warfare! Now What?", *Southwestern Law Review,* 46-3, 355-377.

BROWN, G. D. y METCALF, A. O. (2014): "Easier Said than Done: Legal Reviews of Cyber Weapons", *Journal of National Security Law and Policy,* 7, 115-138.

BROZEK, B. y JAKUBIEC, M. (2017): "On the legal responsability of autonomous machines", *Artificial Intelligence Law,* 25, 293-304.

BRYSON, J. J. y KIME, P.P. (2011): "Just an artifact: why machines are perceived as moral agents", en T. Walsh (ed.), *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (Barcelona 16-22 July 2011)*, AAAI Press, Menlo Park-CA, acceso enero 2021, en https://www.cs.bath.ac.uk/~jjb/ftp/BrysonKime-IJCAI11.pdf

BUMILLER, E. y SHANKER, T. (2011): "War Evolves with Drones, Some Tiny as Bugs", *The New York Times (19/6/2011),* Nueva York, acceso septiembre 2020, en https://www.nytimes.com/2011/06/20/world/20drones.html

BUCHANAN, B. (2016): *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations,* Oxford University Press, Nueva York.

BUNDY, A. (2017): "Preparing for the future of Artificial Intelligence (Book Review)", *Artificial Intelligence & Society,* 32, 285-287.

BURTON, J. (2018): "Cyber Deterrence: A comprehensive approach?", NATO Cooperative Cyber Defense Centre of Excellence (CCDCOE), Tallinn, acceso diciembre 2020, en https://ccdcoe.org/library/publications/cyber-deterrence-a-comprehensive-approach/

BURTON, S., HABLI, I., LAWTON, T., McDERMID, J. MORGAN, P. y PORTER, Z. (2020): "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical and legal perspective", *Artificial Intelligence,* 279, 1-15, acceso marzo 2021, en https://www.sciencedirect.com/science/article/pii/S0004370219301109

BYNUM, T. (2015): "Computer and Information Ethics", en E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition),* acceso febrero 2020, en https://plato.stanford.edu/entries/ethics-computer/

BYRNES, M. W. (2014): "Nightfall. Machine Autonomy in Air-to-Air Combat", *Air & Space Power Journal,* May-June issue, 48-75.

CAPURRO, R. (2019): "Enculturing Algorithms", *Nanoethics,* 13, 131-137.

CARD, D. (2017): "The 'black box' metaphor in machine learning", *Towards Data Science*, Medium, San Francisco, acceso febrero 2021, en https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0

CARSON, J.-F. (2020): "Defining semi-autonomous, automated and autonomous weapons systems in order to understand their ethical challenges", *Digital War,* 1, 173-177.

CENTRE DELÀS D'ESTUDIS PER LA PAU (2019): *Nuevas armas contra la ética y las personas. Drones armados y drones autónomos,* Centre Delàs D'Estudis per la Pau, Informe 39, Barcelona, acceso diciembre 2020, en http://arxiu.centredelas.org/images/INFORMES_i_altres_PDF/informe39_DronesArmados_CAST_web_DEF.pdf

CERVANTES, J. A., LÓPEZ, S. RODRÍGUEZ, L. F., CERVANTES, S. CERVANTES, F., RAMOS, F. (2020a): "Artificial Moral Agents: A Survey of the Current Status", *Science & Engineering Ethics,* 26, 501-532.

CERVANTES, J. A., LÓPEZ, S. y CERVANTES, J. A. (2020b): "Towards ethical cognitive architectures for the development of artificial moral agents", *Cognitive Systems Research,* 64, 117-125.

CHALMERS, D. (2010): "The Singularity: A Philosophical Analysis", *Journal of Consciousness Studies,* 17, 9-10, 7-65.

CHENGETA, T. (2017): "Defining the emerging notion of "Meaningful Human Control" in Weapons Systems", *International Law and Politics,* 49, 833-890.

CHOPRA, S. y WHITE, L. F. (2004): "Artificial Agents – Personhood in Law and Philosophy", en R. López de Mantaras *et al* (eds.), *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI-2004,* IOS Press, Amsterdam, acceso febrero 2021, en http://www.sci.brooklyn.cuny.edu/~schopra/agentlawsub.pdf

CHOPRA, S. y WHITE, L. F. (2011): *A Legal Theory for Autonomous Artificial Agents,* University of Michigan Press, Ann Arbor.

CLARK, M. (2020): *Russian Hybrid Warfare,* Military Learning and the Future of War Series, Institute for the Study of War, Washington D.C.

CLAUSEWITZ, C. V. (1968): *On War (first published 1832)*, Rapaport, A. (ed.), Pelican Classics, Penguin Books, Londres.

COINTE, N., BONNET, g. y BOISSIER, O. (2016): "Ethical judgement of agent's behaviors in Multi-Agent Systems", en N. Osman y C. Sierra (eds.), *Autonomous Agents and Multi-Agent Systems International Conference (AAMAS, 9-10 mayo 2016, Singapur),* Springer, 1106-1114.

COMISIÓN DE DERECHO INTERNACIONAL (CDI) (1980): *Yearbook of the International Law Commission 1980 Volume II Part Two Report of the Commission to the General Assembly*, Naciones Unidas, Nueva York, acceso enero 2021, en https://legal.un.org/ilc/publications/yearbooks/english/ilc_1980_v2_p2.pdf

COMISIÓN EUROPEA (2016): *Comunicación Conjunta al Parlamento Europeo y al Consejo sobre la lucha contra las amenazas híbridas. Una respuesta de la Unión Europea* (JOIN (2016) 18 final), Comisión Europea, Bruselas.

COMISIÓN EUROPEA (2017): *Comunicación Conjunta al Parlamento Europeo y al Consejo sobre Resiliencia, disuasión y defensa: fortalecer la ciberseguridad de la UE* (JOIN (2017) 450 final), Comisión Europea, Bruselas.

COMISIÓN EUROPEA (2018): *Directrices Éticas para una IA Fiable,* Comisión Europea, Bruselas, acceso febrero 2021, en https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

COMISIÓN EUROPEA (2020): *Comunicación Conjunta al Parlamento Europeo y al Consejo sobre La Estrategia de Ciberseguridad de la Unión Europea para la Década Digital* (JOIN (2020) 20 final), Comisión Europea, Bruselas.

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (1949): *III Convenio de Ginebra relativo al trato debido a los prisioneros de guerra, 1949,* Comité Internacional de la Cruz Roja, Ginebra, acceso enero 2021, en https://www.icrc.org/es/doc/resources/documents/treaty/treaty-gc-3-5tdkwx.htm#TTULOI-DISPOSICIONESGENERALES2

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (1977a): *Protocolo I adicional de los Convenios de Ginebra de 1949 relativo a la protección de las víctimas de los conflictos armados sin carácter internacional 1977,* Organización de las Naciones Unidas, Ginebra, acceso enero 2021, en https://www.icrc.org/es/document/protocolo-i-adicional-convenios-ginebra-1949-proteccion-victimas-conflictos-armados-internacionales-1977#NORMA-FUNDAMENTAL

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (1977b): *Protocolo II adicional de los Convenios de Ginebra de 1949 relativo a la protección de las víctimas de los conflictos armados sin carácter internacional 1977,* Organización de las Naciones Unidas, Ginebra, acceso enero 2021, en https://www.icrc.org/es/doc/resources/documents/misc/protocolo-ii.htm

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (1987): *Commentary on the Additional Protocols of June 1977 to the Geneva Convention of 1949,* Comité Internacional de la Cruz Roja, Martinus Nijhoff Publishers, Ginebra, acceso enero 2021, en https://www.loc.gov/rr/frd/Military_Law/pdf/Commentary_GC_Protocols.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (1997): "La Cláusula Martens y el derecho de los conflictos armados", *Revista Internacional de la Cruz Roja,* Comité Internacional de la Cruz Roja, Ginebra, acceso en enero 2021, en https://www.icrc.org/es/doc/resources/documents/misc/5tdlcy.htm

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2003): *XXVIII Conferencia Internacional de la Cruz Roja y de la Media Luna Roja (Ginebra 26-30 noviembre 2003): Declaración, Programa de Acción Humanitaria, Resoluciones*, Comité Internacional de la Cruz Roja, acceso diciembre, 2020, en https://www.icrc.org/es/doc/assets/files/other/icrc_003_1103.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2006): *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare. Measures to Implement Article 36 of Additional Protocol I of 1977*, Comité Internacional de la Cruz Roja, acceso febrero 2021, en https://shop.icrc.org/a-guide-to-the-legal-review-of-new-weapons-means-and-methods-of-warfare-pdf-en

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2007): *XXX Conferencia Internacional de la Cruz Roja y de la Media Luna Roja (Ginebra 2-6 diciembre 2003), Consejo de Delegados del Movimiento Internacional de la Cruz Roja y de la Media Luna Roja (Ginebra 23-24 noviembre 2007): Resoluciones*, Comité Internacional de la Cruz Roja, acceso diciembre, 2020, en https://www.icrc.org/es/doc/assets/files/other/icrc_003_1108.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2005); *Customary International Humanitarian Law (Volume II),* J-M Henckaerts y L. Doswald-Beck (eds.), Comité Internacional de la Cruz Roja, Cambridge University Press, Cambridge, acceso enero 2021, en https://www.icrc.org/en/doc/assets/files/other/customary-international-humanitarian-law-ii-icrc-eng.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2007); *El derecho internacional humanitario consuetudinario (Tomo I),* J-M Henckaerts y L. Doswald-Beck (eds.), Comité Internacional de la Cruz Roja, Centro de Apoyo en Comunicación para América Latina y el Caribe, Buenos Aires, acceso enero 2021, en https://www.icrc.org/es/doc/assets/files/other/icrc_003_pcustom.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2009): *Interpretative Guidance on the notion of Direct Participation in Hostilities under International Humanitarian Law,* Comité Internacional de la Cruz Roja, Ginebra, acceso enero 2021, en https://www.icrc.org/en/doc/assets/files/other/icrc-002-0990.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2011): *El Derecho Internacional Humanitario y los desafíos de los conflictos armados contemporáneos. Informe para el XXXI Congreso Internacional de la Cruz Roja y de la Media Luna Roja (28 noviembre – 1 diciembre 2011),* Comité Internacional de la Cruz Roja, acceso diciembre 2020, en https://www.icrc.org/es/doc/assets/files/red-cross-crescent-movement/31st-international-conference/31-int-conference-ihl-challenges-report-11-5-1-2-es.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2011b): *31st International Conference of the Red Cross and Red Crescent. International Humanitarian Law and the challenges of contemporary armed conflicts. Report,* Comité Internacional de la Cruz Roja, Ginebra, acceso enero 2021, en

https://www.icrc.org/en/doc/assets/files/red-cross-crescent-movement/31st-international-conference/31-int-conference-ihl-challenges-report-11-5-1-2-en.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2012): *Los Convenios de Ginebra de 12 de agosto de 1949,* Comité Internacional de la Cruz Roja, Ginebra, acceso enero 2021, en https://www.icrc.org/es/doc/assets/files/publications/convenios-gva-esp-2012.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2013): *Cyberwarfare and international humanitarian law: The ICRC's position*, Comité Internacional de la Cruz Roja, Ginebra, acceso enero 2021, en https://www.icrc.org/en/doc/assets/files/2013/130621-cyber-warfare-q-and-a-eng.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2014): *New Technologies and the modern battlefied Conference Cycle*, Comité Internacional de la Cruz Roja, acceso septiembre 2020, en https://www.icrc.org/en/cycle-new-technologies

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2016a): *Expert Meeting: Autonomous Weapons Systems. Implications of Increasing Autonomy, in the Critical Function of Weapons (Versoix, 15-16 marzo 2016)*, Comité Internacional de la Cruz Roja, acceso diciembre 2020, en https://shop.icrc.org/autonomous-weapon-systems-implications-of-increasing-autonomy-in-the-critical-functions-of-weapons-print-en

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2016b): *Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapons Systems,* Comité Internacional de la Cruz Roja, acceso diciembre 2020, en https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2016c): *Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field. Geneva, 12 August 1949. Commentary of 2016. Article 2: Application of the Convention,* Comité Internacional de la Cruz Roja, Ginebra, acceso enero 2021, en https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=BE2D518CF5DE54EAC1257F7D0036B518#96_B

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2016d): *Commentary on the Additional Protocols of 8 june 1977 to the Geneva Convention of 12 August 1949,* Comité Internacional de la Cruz Roja, acceso enero 2021, en https://www.loc.gov/rr/frd/Military_Law/pdf/Commentary_GC_Protocols.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2016e): *The principle of proportionality in the rules governing the conduct of hostilities under International Humanitarian Law,* Comité Internacional de la Cruz Roja-Universidad de Laval, Quebec, acceso enero 2021, en https://reliefweb.int/sites/reliefweb.int/files/resources/4358_002_Expert_meeting_report_WEB_1.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2016f): *Manual de Normas Internacionales que rigen las operaciones militares,* Comité Internacional de la Cruz Roja, Ginebra, acceso febrero 2016, en https://www.icrc.org/es/publication/manual-de-normas-internacionales-que-rigen-las-operaciones-militares

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2018): *Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems9–13 April 2018, GenevaStatement of the International Committeeof the Red Cross (ICRC)*, acceso febrero 2021, en https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/9April_ICRC.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2020): *Expert Meeting: Autonomous Weapons Systems. Technical, Military, Legal and Humanitarian Aspects (26-28 marzo 2014),* Comité Internacional de la Cruz Roja, Ginebra, acceso diciembre 2020, en https://shop.icrc.org/expert-meeting-autonomous-weapon-systems-technical-military-legal-and-humanitarian-aspects-pdf-en

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2020b): *ICRC Commentary on the "Guiding Principles of the CCW GGE on Lethal Autonomous Weapons Systems,* Comité Internacional de la Cruz Roja, Ginebra, acceso en enero 2021, en https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2020/gge/documents/ICRC_2020.pdf

COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2021): *Military Necessity,* Comité Internacional de la Cruz Roja, Ginebra, acceso enero 2021, en https://casebook.icrc.org/glossary/military-necessity

CONSEJO DE LA UNIÓN EUROPEA (CUE) (2014): *Rome Declaration on Responsible Research and Innovation in Europe (21 noviembre 2014),* Consejo de la Unión Europea, Bruselas, acceso febrero 2021, en https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf

CORTE INTERNACIONAL DE JUSTICIA (2003): *Separate Opinion of Judge Simma,* Corte Internacional de Justicia, La Haya, acceso en enero 2021, en https://www.icj-cij.org/public/files/case-related/90/090-20031106-JUD-01-10-EN.pdf

CRNKOVIC, G. D. y ÇÜRÜKLÜ, B. (2012): "Robots: Ethical by Design", *Ethics and Information Technology,* 14(1), 61-71.

CROOTOF, R. (2016): "War Torts: Accountability for Autonomous Weapons", *University of Pennsylvania Law Review,* 164(6), 1347-1402.

CUBEIRO CABELLO, E. (2018): "Guerra Híbrida y Ciberespacio", *III Jornadas de Ciberdefensa 2018*, Mando Conjunto de Ciber Defensa (MCCD), acceso noviembre 2020, en https://jornadasciberdefensa.es/documents/22_05_00_Conferencia_Guerra_hibrida_y_ciberespacio.pdf

CULLEN, P. J. y REICHBORN- KJENNERUD, E. (2017): *Understanding Hybrid Warfare,* MCDC Countering Hybrid Warfare Project, Multinational Capability Development Campaign (MCDC), Gov.UK, acceso noviembre 2020, en https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/647776/dar_mcdc_hybrid_warfare.pdf

D'ARCY, J. y HERATH, T. (2011): "A review and analysis of deterrence theory in the IS security literature: making sense of the disparate findings", *European Journal of Information Systems,* 20, 643-658.

DANAHER, J. (2016): "The Threat of Algocracy: Reality, Resistance and Accomodation", *Philosophy and Technology,* 29(3), 245-268.

DANCY, C. L. y RITTER, F. E. (2017): "A Standard Model of the Mind Needs a Body", *A Standard Model of the Mind: AAAI Technical Report FS-17-05,* Association for the Advancement of Artificial Intelligence, Palo Alto-California, 316-320.

DARK, M., EPSTEIN, R., MORALES, L., COUNTERMINE, T., YUANG, Q., ROSE, M. y HARTER, N. (2007): "A framework for Information Security Ethics Education", *Proceedings of the 10th Colloquium for Information Systems Security Education,* Cerias Tech Report 2007-87, Center for Education and Research Information Security and Assurance, Purdue University, West Lafayette.

DE VRIES, K. (2010): "Identity, profiling algorithms and a world of ambient intelligence", *Ethics and Information Technology*, 12(1), 71–85.

DEACON, T. (2012): *Incomplete Nature: How Mind Emerged from Matter*, WW Norton and Company, Nueva York.

DEPARTMENT OF DEFENSE (DoD) (2010): *Joint Terminology for Cyberspace Operations*, Departamento de Defensa -Estados Unidos de América, Washington DC.

DEPARTMENT OF DEFENSE (DoD) (2012): *Department of Defense Directive on Autonomy of Weapon Systems,* Departamento de Defensa-Estados Unidos de América, Washington DC, acceso en diciembre 2020, en https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

DEPARTMENT OF DEFENSE (DoD) (2016): *Law of War Manual,* General Counsel of the Department of Defense, Washington, acceso enero 2021, en https://dod.defense.gov/Portals/1/Documents/pubs/DoD%20Law%20of%20War%20Manual%20-%20June%202015%20Updated%20Dec%202016.pdf?-ver=2016-12-13-172036-190

DESJARDINS, J. R. y McCALL, J. J. (2000): *Contemporary Issues in Business Ethics,* Wadsworth, Belmont-CA.

DIGNUM, V. (2020): *Humane AI Ethical Framework,* HumanE AI Consortium, Knowledge 4 All Foundation Ltd., Redhill, Surrey, acceso abril 2021, en https://www.humane-ai.eu/wp-content/uploads/2019/11/D13-HumaneAI-framework-report.pdf

DINNISS, H. H. (2013): "Participants in conflict – Cyber Warriors, patriotic hackers and the laws of war", en D. Saxon (ed.), *International Humanitarian Law and the Changing Technology of War,* Martinus Nijhoff Publishers-Brill Publishers, Leiden, 251-278.

DODIG-CRNKOVIC, G. y PERSSON, D. (2008): "Sharing Moral Responsability with Robots: A Pragmatic Approach", en A. Holst *et al* (eds.), *Proceedings of the 2008 conference on Tenth Scandinavian Conference on Artificial Intelligence: SCAI 2008*, IOS Press, Amsterdam, 165-168, acceso enero 2021, en

https://www.researchgate.net/publication/221275166_Sharing_Moral_Responsibility_with_Robots_A_Pragmatic_Approach

ECKLUND, A. M. (2019): *Meaningful Human Control of Autonomous Weapons Systems. Institutional definitions in the light of International Humanitarian Law and International Human Rights Law,* Tesis de Máster en Derecho, Umeâ Universitet, Umeâ-Suecia.

EGELAND, K. (2014): *Machine Autonomy and the Uncanny. Recasting, Ethical, Legal and Operational Implications of the Development of Autonomous Weapons Systems,* Trabajo de Fin de Máster en Ciencias Políticas, Departamento de Ciencias Políticas, Universidad de Oslo, Oslo.

EKELHOF, M. (2019): "Moving beyond semantics on Autonomous Weapons: Meaningful Human Control in Operation", *Global Policy,* 10(3), 343-348.

ELIOT, L. B. (2020): "The Neglected Dualism Of Artificial Moral Agency And Artificial Legal Reasoning In AI For Social Good", *AI for social good Workshop,* acceso febrero 2021, en https://crcs.seas.harvard.edu/files/crcs/files/ai4sg_2020_paper_38.pdf

ENISA (2017): *ENISA overview of Cybersecurity and related terminology (version 1),* Agencia de Seguridad de las Redes y de la Información de la Unión Europea, Heraklion-Grecia.

ESS, C. (2006): "Ethical pluralism and global information ethics", *Ethics and Information Technology,* 8 (4)-November 2006, 215-226

ETZIONI, A. y ETZIONI, O. (2016): "AI assisted ethics", *Ethics and Information Technology,* 18(2), 149-156.

ETZIONI, A. y ETZIONI, O. (2017): "Incorporating Ethics into Artificial Intelligence", *Journal of Ehics,* 21, 403-418.

EVANS, K., DE MOURA, N., CHAUVIER, S. CHATILA, R. y DOGAN, E. (2020): "Ethical Decision Making in Autonomous Vehicles: The AV Ethics Project", *Science & Engineering Ethics,* 26, 3285-3312.

FARRANT, J. y FORD, C. M. (2017): "Autonomous Weapons and Weapons Reviews: The UK Second International Review Forum", *International Law Studies,* 389, vol. 93, 389-422.

FEIGENBAUM, E. A. (2003): "Some Challenges and Grand Challenges for Computational Intelligence", *Journal of the Association for Computing Machinery,* 50, 32-40.

FIESER, J. (2009): "Ethics", *Internet Encyclopedia of Philosophy (A Peer-reviewed Academic Rsources),* acceso febrero 2020, en https://www.iep.utm.edu/ethics/

FINLAY, C. J. (2018): "Just War, Cyber War and the Concept of Violence", *Philosophy & Technology,* Springer, 31, 357-377.

FISCHERKELLER, M. P. y HARKNETT, R. J. (2019): "Deterrence is Not a Credible Strategy for Cyberspace (and What Is)", Institute for Defense Analysis (IDA), acceso diciembre 2020, en https://www.ida.org/-/media/feature/publications/w/we/welch-awards-2018-research-notes-fall-2019/welch-awards-2018-research-notes-fall-2019-article-1.ashx?la=en&hash=C725B2340ABA96463DBAF3D298E7671A

FISHER, J. (2004): "Social Responsability and Ethics: Clarifying the Concepts", *Journal of Business Ethics,* 52, 4, 381-390.

FLORIDI, L. (1999): "Information Ethics: On the philosophica foundation of computer ethics", *Ethics and Information Technology,* 1, 33-52.

FLORIDI, L. y SANDERS, J. W. (2004): "On the Morality of Artificial Agents", *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science,* 14(3), 349–379.

FOLTZ, A. C. (2012): "Stuxnet, Schmitt Analysis, and the Cyber 'use of force' Debate", *Joint Force Quaterly,* 67, 4th quarter, 40-48, acceso enero 2021, en https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-67/JFQ-67_40-48_Foltz.pdf

FORMOSA, P. y RYAN, M. (2020): "Making moral machines: Why we need artificial moral agents", *Artificial Intelligence & Society,* noviembre 2020 (on line), acceso enero 2021, en https://link.springer.com/article/10.1007/s00146-020-01089-6

FOSSA, F. (2018): "Artificial Moral Agents: moral mentors or sensible tools?", *Ethics and Information Technology,* 20, 115-126.

FOX, J. y SCHULMAN, C. (2010): "Superintelligence Does Not Imply Benevolence", en Mainzer, K. (ed.), *ECAP 10: VIII European Conference on Computing and Philosophy*, 456—462.

FOY, J. (2014): "Autonomous Weapons Systems: Taking the human out of International Humanitarian Law", *Delhousie Journal of Legal Studies,* 23, 47-70.

FRITZ, A., BRANDT, W., GIMPEL, H. y BAYER, S. (2020): "Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI)", *De Ethica*, 6(1), 3-22.

FUCHS, C., BICHLER, R. M. y RAFFL, C. (2009): "Cyberethics and Co-operation in the Information Society", *Scientific and Engineering Ethics,* 15, 447-466.

GADE, E. K. (2010): "Defining the Non-Combatant: How do we determine who is worthy of protection in violent conflicto", *Journal of Military Ethics,* 3, 219-242.

GALLIOTT, J. y SCHOLZ, J. (2018): "Artificial Intelligence in Weapons. The moral imperative for Minimally-Just Autonomy", *Journal of Indo-Pacific Affairs,* Winter, 57-67.

GALVIN, J. (1992): "Conflict in the Post-Cold War Era", en E. Corr y S. Sloan (eds.), *Low Intensity Conflicts. Low Threats in a New World,* West View Press, Boulder-Colorado, Chapter 4.

GÁMEZ ALBÁN, H. M., SALAS, O. y BRAVO BASTIDAS, J. J. (2016): "Aplicación de mapas de Kohonen para la priorización de zonas de mercado: una aproximación práctica", *EIA,* XIII, 13, 25, 157-169.

GARCÍA, D. (2016): "Future arms, technologies and international law. Preventive security governance", *European Journal of International Security,* 1(1), 94-111.

GAT, A. (2006): *War in Human Civilization,* Oxford University Press, Nueva York.

GEISS, R. y LAHMANN, H. (2012): "Cyber warfare: applying the principle of distinction in an interconnected space", *Israel Law Review,* 45(3), 381-399.

GERT, B. (1999): "Common morality and computing", *Ethics and Information Technology,* 1, 57-64.

GILES, K. (2012): "Russia's public stance on cyberspace issues", en *2012 4th International Conference on Cyber Conflict (CYCON 2012) Tallinn-Estonia [5-8 June 2012],* IEEE, Nueva York, 63-75, acceso marzo 2019, https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6233114

GILADI, R. (2008): "The jus ad bellum/jus in bello Distinction and the Law of Occupation", *Israeli Law Review,* 41, 246-301.

GILES, K. y HAGESTAD, W. (2013): "Divided by a common language: Cyber definitions in Chinese, Russian and English", en *2013 5th International Conference on Cyber Conflict (CYCON 2013) Tallinn-Estonia [4-7 June 2013]*, IEEE, Nueva York.

GOEL, S. y HONG, Y (2015): "Cyber War Games: Strategic Jostling Among Traditional Adversaries", en S. Jajodia *et al, Cyber Warfare. Building the Scientific Foundation*, Springer, Cham-Heidelberg-NewYork-Dordrecht-London, 1-14.

GOOD, I. J. (1965): "Speculations Concerning the First Ultraintelligent Machine", en F. Alt y M. Rubinoff (eds.), *Advances in Computers – Volume 6,* Academic Press-Elsevier, Cambridge-Massachusetts, 31-88.

GORDON, J.-S. (2020): "Building Moral Robots: Ethical Pitfalls and Challenges", *Science and Engineering Ethics,* 26, 141-157.

GOTTERBARN, D. (1991): "Computer Ethics: Responsibility Regained", *National Forum: The Phi Beta Kappa Journal,* 71, 3, 26-31.

GRAY, C. H. (1997): *Postmodern War: The New Politics of Conflict,* Guildford Press, Nueva York.

GULICIUC, V. (2014): "Technological Singularity in the age of surprise facing complexity", *European Journal of Science and Technology,* Vol. 10, nº 4, 79-88.

HAGENDORFF, T. (2020): "The Ethics of AI Ethics: An Evaluation of Guidelines", *Minds and Machines,* 30, 99-120.

HAKIMI, M. (2018): "The *jus ad bellum's* Regulatory Form", *American Journal of International Law,* 112-2, 151-190.

HALL, B. K. (2017): "Autonomous Weapons Systems Safety", *Joint Forces Quaterly,* 3rd. Quarter July 2017, 86, 86-93.

HAQUE, A. A. (2017): "Whose Armed Conflict? Which Law of Armed Conflict?", *Georgia Journal of International & Comparative Law,* 45-3, 475-493.

HEATH, T.R., GUNNESS, K. y COOPER, C.A. (2016): *The PLA's and China's rejuvenation. National Security and Military Strategies, Deterrence concepts and Combat Capabilities,* RAND Corporation, Santa Monica-California.

HERBACH, J. D. (2012): "Into the Caves of Steel: Precaution, Cognition and Robotic Weapons Systems Under the Internationa Law of Armed Conflict", *Amsterdam Law Forum,* 4(3), 3-20.

HEYNS, C. (2016): "Autonomous Weapons Systems: living a dignified life and dying a dignified death", en N. C. Butha *et al* (eds.), *Autonomous Weapons Systems. Law, Ethics, Policy (ebook),* Cambridge University Press, Cambridge, 4-20.

HIMMA, K. E. (2009): "Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?", *Ethics & Information Technology,* 11, 19-29.

HOFFMAN, F. G. (2020): *Conflict in the 21st Century: The Rise of Hybrid Wars,* Potomac Institute for Policy Studies, Arlington-Virginia.

HOFSTADTER, D.R. y DENNETT, D.C. (1982): *The Mind's I: Fantasies and reflections of self and soul,* Bantam, Toronto-Nueva York.

HOLLAND MICHEL, A. (2021): *Known Unknowns. Data Issues and Military Autonomous Systems,* United Nations Institute for Disarmament Research (UNIDIR), Ginebra.

HUFF, F. y FINHOLT, T. (1994): *Social Issues in Computing. Putting Computing in its place*, Mc. Graw-Hill College, Nueva York.

HUGHES, R. (2010): "A Treaty for Cyberspace", *International Affairs,* 86(2), 523-541.

HUMAN RIGHTS WATCH (HRW) (2012): *Losing Humanity. The case against killer robots,* Human Rights Watch, Nueva York, acceso febrero 2021, en https://www3.nd.edu/~dhoward1/Losing%20Humanity-The%20Case%20against%20Killer%20Robots-Human%20Rights%20Watch.pdf

HUMAN RIGHTS WATCH (HRW) (2016): *Killer Robots and the Concept of Meaningful Human Control. Memorandum to Convention on Conventional Weapons (CCW) Delegates, April 2016,* Human Rights Watch, Nueva York, acceso febrero 2021, en

https://www.hrw.org/sites/default/files/supporting_resources/robots_meaningful_human_control_final.pdf

HUMAN RIGHTS WATCH (HRW) (2020a): *Stopping Killer Robots. Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control,* Human Rights Watch, Nueva York, acceso enero 2021, en https://www.hrw.org/sites/default/files/media_2020/08/arms0820_web_0.pdf

HUMAN RIGHTS WATCH (HRW) (2020b): *New Weapons, Proven Precedent. Elements of and models for a Treaty on Killer Robots,* International Human Rights Clinic – Human Rights Program at Harvard Law School, Harvard, acceso febrero 2021, en http://hrp.law.harvard.edu/wp-content/uploads/2020/10/New-Weapons-Proven-Precedent.pdf

HURKA, T. (2005): "Proportionality in the Morality of War", *Philosophy & Public Affairs,* 33(1), 34-66.

HUTTER, M. (2012): "Can Intelligence explode?", *Journal of Consciousness Studies*, 19, 1-2, 143-166.

INSTITUTE OF ELECTRICAL AND ELECTRONIC ENGINEERS INCORPORATED (IEEE) (2017): *Ethically Aligned Design – Overview (v1),* IEEE, Nueva York.

INSTITUTE OF ELECTRICAL AND ELECTRONIC ENGINEERS INCORPORATED (IEEE) (2019): *Ethically Aligned Design (1st Edition -EAD1e)*, IEEE, Nueva York, acceso enero 2021, en https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

INSTITUTE OF ELECTRICAL AND ELECTRONIC ENGINEERS INCORPORATED (IEEE) (2021): *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,* IEEE, Nueva York, acceso marzo 2021, en https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

INTERNATIONAL BUSINESS MACHINES (IBM) (2020): "Building Trust on AI", en *What's Next for AI*, IBM, Armonk-Nueva York, acceso en septiembre 2020, en https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html

INTERNATIONAL PANEL ON THE REGULATION OF AUTONOMOUS WEAPONS (IPRAW) (2021): *Building blocks for a regulation on LAWS and Human Control,* IPRAW – German Institute for International and Security Affairs, Berlin.

JANCZEWSKI, L. J. y COLARIK, A. M. (2008): *Cyber Warfare and Cyber Terrorism,* Information Science Reference, Hershey-Nueva York.

JENSEN, E. T. (2017): "The Tallinn Manual 2.0: Highlights and Insights", *Georgetown Journal of International Law*, 48, 735-778.

JIANG, B. y ORMELING, F. J. (1997): "Cybermap: the Map for Cyberspace", *The Cartographic Journal,* Vol. 34-2, 111-116.

JOHNSON, D. G., (2005): "Computer Ethics", en R. G. Frey y C. H. Wellman (eds.), *A Companion to Applied Ethics,* Blackwell Publishing Ltd., Malden-Oxford-Victoria-Berlin, 608-619.

JOHNSON, D. G. (2006): "Computer Systems: Moral entities but not moral agents", *Ethics and Information Technology,* 8, 195-204.

JOHNSON, D. G. y MILLER, K. W. (2008): "Un-making artificial moral agents", *Ethics and Information Technology,* 10, 123-133.

JONAS, H. (1953): "Cybernetics and purpose: A critique", *Social Research*, XX(2), 172–192.

KADDU, S. B. (2007): "Information Ethics: a student's perspective", *International Review of Information Ethics,* vol. 7, acceso febrero 2020, en http://www.i-r-i-e.net/inhalt/007/35-kaddu.pdf

KAGAN, S. (1998): *The Limits of Morality,* Oxford Scholarship Online, Oxford University Press, Oxford, acceso febrero 2021, en https://oxford.universitypressscholarship.com/view/10.1093/0198239165.001.0001/acprof-9780198239161

KALDOR, M. (2002): *New and Old Wars: Organized Violence in a Global Era,* Polity Press, Cambridge.

KANIA, E. B. (2020): *"AI Weapon's" in China's Military Innovation",* The Brookings Institution, acceso octubre 2021, en https://www.brookings.edu/research/ai-weapons-in-chinas-military-innovation/

KASS, K. (2015): "Autonomous Weapons and Accountability: Seeking solutions in the Law of War", *Loyola of Los Angeles Law Review,* 48, 1017-1068.

KASTAN, B. (2013): "Autonomous Weapons Systems: A coming legal 'Singularity'?", *Journal of Law, Technology & Policy,* 45, 45-82.

KELSEY, J. T. (2008): "Hacking into International Humanitarian Law: The Principles of Distinction and Neutrality in the Age of Cyber Warfare", *Michigan Law Review,* 106(7), 1427-1452.

KILOVATY, I. (2014): "Cyber Warfare and the Jus Ad Bellum Challenges: Evaluation in the Light of the Tallinn Manual on the International Law Applicable to Cyber Warfare", *American University National Security Law Brief,* 5-1, 91-124.

KIRKPATRICK, K. (2015): "The Moral Challenges of Driverless Cars", *Communications of the ACM,* 58 (8), 19-20.

KLINCEWICZ, M. (2015): "Autonomous Weapons Systems, the Frame Problem and Computer Security", *Journal of Military Ethics,* 14(2), 162-176.

KNOPF, A. A. (1962): "M. G. Singer 'Generalization in Ethics? (Book Review", *Analytic Philosophy,* 3(1), 18-21.

KNOPF, J. W. (2010): "The Fourth Wave in Deterrence Research", *Contemporary Security Policy,* 31-1, 1-33.

KOLB, J. (1997): "Origin of the twin terms jus ad bellum/jus in bello", Comité Internacional de la Cruz Roja, Ginebra, acceso en enero 2021, en https://www.icrc.org/en/doc/resources/documents/article/other/57jnuu.htm

KOSTADINOV, D. (2014): "Jus in Cyber Bello: How the Law of Armed Conflict Regulates Cyber Attacks. Part I", *Infosec Resources,* acceso enero 2021, en https://resources.infosecinstitute.com/topic/jus-cyber-bello-law-armed-conflict-regulates-cyber-attacks-part/

KRISHNAN, A. (2009): *Killer Robots. Legality and Ethicality of Autonomous Weapons,* Routledge, Londres.

KURZWEIL, R. (2000): *The age of spiritual machines: When computers exceed human intelligence,* Viking-Pinguin, Nueva York.

KURZWEIL, R. (2005): *The Singularity is Near. When humans trascend Biology,* Viking-Pinguin, Londres.

LAIRD, J., LEBIERE, C. y ROSENBLOOM, P. S. (2017): "A Standard Model of the Mind: Towards a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience and Robotics", *Artificial Intelligence Magazine,* 38-4, 13-26.

LAUKYTE, M. (2017): "Artificial agents among us: Should we recognize them as agents proper?", *Ethics and Information Technology,* 19, 1-17.

LAZZERI, F. (2021): "Aprendizaje profundo frente a aprendizaje automático en Azure Machine Learning", *Microsoft,* acceso octubre 2021, en https://docs.microsoft.com/es-es/azure/machine-learning/concept-deep-learning-vs-machine-learning

LEGG, S. y HUTTER, M. (2007): "A collection of definitions of intelligences", en *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms – Frontiers in Artificial Intelligence and Applications,* IOS Press, Amsterdam, vol. 157, 17-24.

LEGG, S. y VENESS, J. (2013): "An approximation of the Universal Intelligence Measure", en D. L. Dowe (eds.), *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence,* Lecture Notes in Computer Science, vol. 7070, Springer, Berlin-Heidelberg.

LEWIS, D. (2020): "An Enduring Impasse on Autonomous Weapons", *Just Security,* Reiss Center on Law and Security-New York University School of Law, Nueva York, acceso en enero 2021, en

https://www.justsecurity.org/72610/an-enduring-impasse-on-autonomous-weapons/

LIBICKI, M. C. (2009): *Cyberdeterrence and Cyberwar,* Rand Project Air Force, Rand Corporation, Santa Monica-CA.

LIM, H. C., STOCKER, R. y LARKIN, H. (2010): "Review of Trust and Machine Ethics Research: Towards a Bio-Inspired Computational Model of Ethical Trust (CMET)", en M. Murata *et al* (eds.), *Proceedings of the 3d International ICST Conference on Bio-Inspired Models of Network, Information, and Computing Systems (25-28 November 2008, Hyogo-Japan,* ICST, acceso marzo 2021, en https://eudl.eu/doi/10.4108/icst.bionetics2008.4728

LIN, P., BEKEY, G. y ABNEY, K. (2008): *Autonomous Military Robotics: Risk, Ethics and Design,* Ethics & Emerging Science Group, California Politechnic State University, San Obispo-CA, acceso abril 2021, en https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1001&context=phil_fac

LIU, H.-Y. (2016): "Defense and Military Policy: Autonomous Weapons Systems", en D. A. Berfield y M. J. Dubnick (eds.), *Encyclopedia of Public Administration and Public Policy (3rd Edition)*, Taylor and Francis, Nueva York, 833-838.

LOGAN, R. K. y TANDOC, M. (2018): "Thinking in Patterns and the Pattern of Human Thought as contrasted with AI Data Processing", *MDPI Information,* 9, 83-99.

LÓPEZ DE TURISO Y SÁNCHEZ, J. (2012): "La evolución del conflicto. Hacia un nuevo escenario bélico", en A. Gómez de Agreda (coord.), *El Ciberespacio. Nuevo escenario de confrontación,* Centro Superior de Estudios de la Defensa Nacional, Secretaría General Técnica - Ministerio de Defensa, Madrid, Monografías del CESEDEN 126, 117-166.

LÓPEZ DE TURISIO Y SÁNCHEZ, J. (2018): "Evolución del concepto de Ciberdefensa", en Mando Conjunto de Ciberdefensa (MCCD), *Operaciones Militares en el Ciberespacio - Jornadas de Ciberdefensa 2018 del Mando Conjunto de Ciberdefensa (22-24 de mayo 2018)*, Ministerio de Defensa-Mando Conjunto de Ciberdefensa, Madrid.

LORENZO PONCE DE LEÓN, R. (2012): "Las Reglas de enfrentamiento (ROE) como paradigma del Estado de derecho en operaciones militares", *Revista Española de Derecho Militar,* 99, 37-220.

LUKES, S. (2005): *Power. A radical view*, Palgrave Macmillan Houndmills, Basingstoke-Hampshire & New York.

LUPOVICI, A. (2011): "Cyberwarfare and Deterrence: Trends and Challenges in Research", *Military and Strategic Affairs,* vol. 3-3, 49-62.

MACÁK, K. (2015): "Military Objectives 2.0: The Case for Interpreting Computer Data as Objects under International Humanitarian Law", *Israel Law Review,* 48(1), 55-80.

MACKENZIE, D. C., ARKIN, R. C. y CAMERON, J. M. (1997): "Multiagent Mission Specification and Execution", *Autonomous Robots,* 4(1), 29-52.

MALIK, F. (2019): "Neural Network Activation Function Types", *Medium,* acceso octubre 2021, en https://medium.com/fintechexplained/neural-network-activation-function-types-a85963035196

MALLE, B. y SCHEUTZ, M. (2014): "Moral competence in social robots", *IEEE International Symposium on Ethics in Engineering, Science and Technology,* Chicago, acceso enero 2021, en https://hrilab.tufts.edu/publications/mallescheutz14ieee.pdf

MALLORY, K. (2018): *New Challenges in Cross-Domain Deterrence,* RAND Corporation,

MANCERA CASTAÑO, J. M. (2014): "La ciberguerra china desde la lógica de la guerra irrestricta", *Ciencia y poder aéreo,* Vol. 9-1, 89-96.

MANER, W. (1999): "Is computer ethics unique?", *Ética & Política /Ethics & Politics*, Universitá Degli Studi di Trieste, I/2.

MARÍN MARTÍNEZ, A. P. (2018): *Los Mercenarios en el Mediterráneo Antiguo e Iberia (siglos V-III a.C.),* Signifer Libros – Monografías y Estudios de Antigüedad Griega y Romana, 53, Salamanca.

MARÍN MARTÍNEZ, A. P. (2019): *La naturaleza cambiante del paradigma de la ciberdefensa. Hacia un nuevo marco holístico, ético y jurídico,* Trabajo de fin del Máster de Investigación en Ciberseguridad (inédito), Departamento de Matemáticas – Universidad de León, León, 2019.

MARCHANT, G. E. y ALLENBY, B. (2017): "Soft law: new tolos for governing emerging technologies", *Bulletin of the Atomic Scientists,* 73(2), 108-114.

MARTINHO, A., POULSEN, A, KROESEN, M., CHORUS, C. (2021): "Perspectives about artificial moral agents, *AI and Ethics,* acceso julio 2021, en https://link.springer.com/content/pdf/10.1007/s43681-021-00055-2.pdf

MASSACHUSETTS INSTITUTE OF TECHNOLOGY (2017): *Moral Machine*, MIT Media Lab, Cambridge, acceso marzo 2021, en https://www.moralmachine.net/

MASON, R. O. (1986): "Four Ethical Issues of the Information Age", *Management Information Systems Quaterly (MISQ),* 10, 1.

MATICH, D. J. (2001): *Redes Neuronales: Conceptos básicos y aplicaciones,* Universidad Tecnológica Nacional – Grupo de Investigación Aplicada a la Ingeniería Química, Rosario, acceso octubre 2021, en https://www.frro.utn.edu.ar

MAY, L., ROVIE, E., y VINER, S. (2006): "Carl von Clausewitz, On the Art of War", en L. May *et al* (eds.), *The Morality of War: Classical and Contemporary Readings,* Pearson Education, Upper Saddle River, 115-121.

MAZARR, M. J. (2018): *Understanding Deterrence,* Perspectives, Rand Corporation, Santa Mónica – CA, acceso noviembre 2020, en https://www.rand.org/pubs/perspectives/PE295.html

McCORMACK, T. (2018): "International Humanitarian Law and the Targeting of Data", *International Law Studies,* 94, 222-240.

McCOY, L., BURKELL, J., CARD, D., DAVIS, B., GICHOYA, J., LE PAGE, S. y MADRAS, D. (2019): *On Meaningful Human Control in High Stakes Machine-Human Partnerships,* AI Pulse, Universidad de California, Los Ángeles.

MELLO, P. (2010): "In Search of New Wars: The Debate about the Transformation of War", *European Journal of International Relations,* 16(2), 297-309.

MEYER, J.-J. C. (2006): "Reasoning about Emotional Agents", *International journal of intelligent systems*, 21, 601-619.

MIAL, H., RAMBOTHAN, O., WOODHOUS, T. (1999): *Contemporary Conflict Resolution. The prevention, management and transformation of deadly conflicts*, Malden, Polity Press, United Kingdom.

MINISTERIO DE ASUNTOS EXTERIORES FEDERACIÓN RUSA (2016): *Doctrine of Information Security of the Russian Federation (646 – 5 diciembre 2016),* Ministerio de Asuntos Exteriores de la Federación Rusa, Moscú, acceso marzo 2019, http://www.mid.ru/en/foreign_policy/official_documents/-/asset_publisher/CptICkB6BZ29/content/id/2563163

MINISTERIO DE ASUNTOS EXTERIORES, UNIÓN EUROPEA Y COOPERACIÓN (MAEUEC) (2018): *Intervención del Embajador de España. Delegado ante la Conferencia de Desarme. Convención sobre Ciertas Armas Convencionales. Grupo de Expertos Gubernamentales sobre Sistemas de Armas Autónomos Letales (Ginebra, 9 de abril de 2018),* Ministerio de Asuntos Exteriores, Unión Europea y Cooperación, Ginebra, acceso febrero 2021, en https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/9April_Spain.pdf

MINISTERIO DE DEFENSA (2007): *Orientaciones. El Derecho de los Conflictos Armados. Tomo I,* Ministerio de Defensa, Madrid, acceso enero 2021, en http://www.cruzroja.es/pls/portal30/docs/PAGE/DIH/MINISTERIO_DEFENSA/OR7_004.PDF.TOMO%20I.PDF

MINSKY, M. (1966): "Artificial Intelligence", *Scientific American,* 215, 3, 246-263.

MITTELSTADT, B. D., ALLO, P., TADDEO, M., WACHTER, S. y FLORIDI, L. (2016): "The ethics of algorithms: Mapping the debate", *Big Data & Society,* 3(2) July-December, 1-21.

MONAGHAN, S. (2019): "Countering Hybrid Warfare: So What for the Joint Force?", *Prism,* Vol. 8 (2) (October 2019), acceso noviembre 2020, en https://ndupress.ndu.edu/Journals/PRISM/PRISM-8-2/

MOOR, J. H. (1985): "What is Computer Ethics", *Metaphilosophy,* vol. 16(4), 266-275.

MOOR, J. H. (2006): "The Nature, Importance and Difficulty of Machine Ethics", *IEEE Intelligent Systems*, July-August, 21(4), 18-21.

MOSTAFA, S. A., AHMAD, M. S. y MUSTAPHA, A. (2019): "Adjustable autonomy: A systematic literature review", *Artificial Intelligence Review,* 51(2), 149–186

MOYNIHAN, H. (2019): *The Application of International Law to State Cyberattacks. Sovereignty and Non-Intervention,* Research Paper-International Law Programme, Chatham House- The Royal Institute of International Affairs, Londres, acceso enero 2021, en https://www.chathamhouse.org/sites/default/files/publications/research/2019-11-29-Intl-Law-Cyberattacks.pdf

MUMFORD, D. (2019): "Can an artificial intelligence machine be conscious?", acceso noviembre 2019, en http://www.dam.brown.edu/people/mumford/blog/2019/conscious.html

NACIONES UNIDAS (1945): *Carta de las Naciones Unidas y Estatuto de la Corte Internacional de Justicia,* Naciones Unidas, Washington, acceso diciembre 2020, en https://www.un.org/es/charter-united-nations/

NACIONES UNIDAS (1981): *Declaración sobre la inadmisibilidad de la intervención y la injerencia en los asuntos internos de los Estados,* Asamblea General de las Naciones Unidas (91 sesión plenaria – 1981), [consultado el 27 de abril 2019], https://undocs.org/es/A/RES/36/103

NACIONES UNIDAS (2001): *Yearbook of the International Law Commission 2001 – Volume II Part II*, Naciones Unidas, acceso abril 2019, en http://legal.un.org/docs/?path=../ilc/publications/yearbooks/english/ilc_2001_v2_p2.pdf&lang=EFSRAC

NACIONES UNIDAS (2005): *Resolución 60/1. Documento Final de la Cumbre Mundial 2005,* Organización de las Naciones Unidas, Nueva York, acceso en enero 2021, en https://www2.ohchr.org/spanish/bodies/hrcouncil/docs/gaA.RES.60.1_Sp.pdf

NACIONES UNIDAS (2010): *Informe Provisional del Relator Especial sobre las ejecuciones extrajudiciales, sumarias o arbitrarias (A/65/321)*, Organización de las Naciones Unidas, Ginebra, acceso enero 2021, en https://undocs.org/es/A/65/321

NACIONES UNIDAS (2013): *Informe del Relator Especial sobre las ejecuciones extrajudiciales, sumarias o arbitrarias (A/HRC/23/47),* Organización de las Naciones Unidas, Ginebra, acceso enero 2021, en https://undocs.org/es/A/HRC/23/47

NACIONES UNIDAS (2015): *Grupo de Expertos Gubernamentales sobre los Avances en la Información y las Telecomunicaciones en el Contexto de la Seguridad Internacional. Organización de las Naciones Unidas (A/70/174)*, Ginebra, acceso en enero 2021, en https://undocs.org/es/A/70/174

NACIONES UNIDAS (2017a): *Informe de 2017 del Grupo de Expertos Gubernamentales sobre Sistemas Armamentísticos Autónomos Letales (SAAL)*, Organización de las Naciones Unidas, Ginebra, acceso en diciembre 2020, en https://undocs.org/pdf?symbol=es/CCW/GGE.1/2017/3

NACIONES UNIDAS (2017b): *Examination of various dimensions of emerging technologies in the area of lethal autonomous weapons systems, in the context of the objectives and purposes of the Convention*, Organización de las Naciones Unidas, Ginebra, acceso en diciembre 2020, en https://undocs.org/ccw/gge.1/2017/WP.2

NACIONES UNIDAS (2017c): *Towards a definition of lethal autonomous weapons systems*, Organización de las Naciones Unidas, Ginebra, acceso en diciembre 2020, en https://undocs.org/ccw/gge.1/2017/WP.3

NACIONES UNIDAS (2018a): *Informe de 2018 del Grupo de Expertos Gubernamentales sobre Sistemas Armamentísticos Autónomos Letales (SAAL)*, Organización de las Naciones Unidas, Ginebra, acceso en diciembre 2020, en https://undocs.org/es/CCW/GGE.1/2018/3

NACIONES UNIDAS (2018b): *Position Paper submitted by China,* Organización de las Naciones Unidas, Ginebra, acceso en diciembre 2020, en https://www.unog.ch/80256EDD006B8954/(httpAssets)/E42AE83BDB3525D0C125826C0040B262/$file/CCW_GGE.1_2018_WP.7.pdf

NACIONES UNIDAS (2018c): *Categorizing lethal autonomous weapons systems – A technical and legal perspective in understanding LAWS*, Organización de las Naciones Unidas, Ginebra, acceso en diciembre 2020, en https://www.unog.ch/80256EDD006B8954/(httpAssets)/FD148A6783DAC304C12582F30032F633/$file/2018_GGE+LAWS_August_Working+Paper_Estonia+and+Finland.pdf

NACIONES UNIDAS (2018d): *Human Machines Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles,* Organización de las Naciones Unidas, Ginebra, acceso diciembre 2020, en https://www.unog.ch/80256EDD006B8954/(httpAssets)/050CF806D90934F5C12582E5002EB800/$file/2018_GGE+LAWS_August_Working+Paper_UK.pdf

NACIONES UNIDAS (2018e): *Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems,* Organización de las Naciones Unidas, Ginebra, acceso diciembre 2020, en https://www.unog.ch/80256EDD006B8954/(httpAssets)/2E16E59C0AB73F2FC12582F30055113C/$file/2018_GGE+LAWS_August_Working+Paper_France.pdf

NACIONES UNIDAS (2018f): *Resolución aprobada por la Asamblea General el 22 de diciembre de 2018 sobre Promoción del comportamiento responsable de los Estados en el ciberespacio en el contexto de la seguridad internacional*, Organización de las Naciones Unidas, Ginebra, acceso en enero 2021, en https://undocs.org/pdf?symbol=es/A/RES/73/266

NACIONES UNIDAS (2018g): *Humanitarian benefits of emerging technologies in the area of lethal autonomous weapon systems. Submitted by the United States of America*, Organización de las Naciones Unidas, Ginebra, acceso febrero 2021, en https://ogc.osd.mil/LoW/practice/DoDDocuments/US_Working_Paper-Humanitarian_benefits_of_emerging_technologies_in_the_area_of_LAWS-CCW_GGE.1_2018_WP.4_E.pdf

NACIONES UNIDAS (2018h): *Joint statement by the delegations of France and Germany on agenda item "Possible options for addressing the humanitarian and international security challenges, GGE on LAWS, 29 August 2018,* Organización de las Naciones Unidas, Ginebra, acceso febrero 2021, en https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/29August_France_Germany.pdf

NACIONES UNIDAS (2019a): *Informe del periodo de sesiones de 2019 del Grupo de Expertos Gubernamentales sobre las Tecnologías Emergentes en el ámbito de los Sistemas Armamentísticos Autónomos Letales*, Organización de las Naciones Unidas, Ginebra, acceso en diciembre 2020, en https://undocs.org/pdf?symbol=es/CCW/GGE.1/2019/3

NACIONES UNIDAS (2019b): *Potential opportunities and limitations of military uses of lethal autonomous weapons systems*, Organización de las Naciones Unidas, Ginebra, acceso en diciembre 2020, en https://www.unog.ch/80256EDD006B8954/(httpAssets)/B7C992A51A9FC8BFC12583BB00637BB9/$file/CCW.GGE.1.2019.WP.1_R+E.pdf

NACIONES UNIDAS (2019c): *Posición de España respecto al papel de la ciencia y la tecnología en el contexto de la seguridad internacional y desarme,* Naciones Unidas, Nueva York, acceso diciembre 2020, en https://www.un.org/disarmament/wp-content/uploads/2019/09/spain-74-122.pdf

NACIONES UNIDAS (2019d): *Declaración de España en la reunión de la CCW de 26 de marzo de 2019,* Naciones Unidas, Ginebra, acceso diciembre 2020, en https://conf.unog.ch/digitalrecordings/index.html?guid=public/61.0500/158251CC-F1AB-4612-922A-FD57F70B0CBE_10h06&position=9205

NACIONES UNIDAS (2019e): *Informe final de la Reunión de las Altas Partes Contratantes en la Convención sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales que Puedan Considerarse Excesivamente Nocivas o de Efectos Indiscriminados*, Organización de las Naciones Unidas, Ginebra, acceso en enero 2021, en https://undocs.org/es/CCW/MSP/2019/9

NACIONES UNIDAS (2019f): *Chair's Summary. Informal consultative meeting of the Group of Governmental Experts (GGE) on Advancing responsable State behaviour in cyberspace in the context of international security,* Organización de las Naciones Unidas, Ginebra, acceso en enero 2021, en https://www.un.org/disarmament/wp-content/uploads/2019/12/gge-chair-summary-informal-consultative-meeting-5-6-dec-20191.pdf

NACIONES UNIDAS (2019g): *Potential Opportunities and limitations of military uses of lethal autonomous weapons systems. Submitted by the Russian Federation,* Organización de las Naciones Unidas, Ginebra, acceso en enero 2021, en https://www.unog.ch/80256EDD006B8954/(httpAssets)/489AAB0F44289865C12583BB0063B977/$file/GGE+LAWS+2019_Working+Paper+Russian+Federation_E.pdf

NACIONES UNIDAS (2019h): *Working Paper of the Russian Federation National Implementation of the Guiding Principles on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, acceso enero 2021, en https://documents.unoda.org/wp-content/uploads/2020/09/Ru-Commentaries-on-GGE-on-LAWS-guiding-principles1.pdf

NACIONES UNIDAS (2019i): *Secretary General's message to Meeting of Group of Governmental Expert son Emerging Technologies in the área of Lethal Autonomous Weapons Systems,* Organización de las Naciones Unidas, Ginebra, acceso febrero 2021, en https://www.un.org/sg/en/content/sg/statement/2019-03-25/secretary-generals-message-meeting-of-the-group-of-governmental-experts-emerging-technologies-the-area-of-lethal-autonomous-weapons-systems

NACIONES UNIDAS (2019j): *Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems . Submitted by the United States of America*, Organización de las Naciones Unidas, Ginebra, acceso febrero 2021, en https://undocs.org/en/CCW/GGE.1/2019/WP.5

NACIONES UNIDAS (2019k): *Government of the Russian Federation. Statement to the Convention on Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Organización de las Naciones Unidas, Ginebra, acceso febrero 2021, en https://conf.unog.ch/digitalrecordings/index.html?guid=public/61.0500/70E5CC90-B100-4658-95BA-8E8C0D-4D581E_15h14&position=2576

NACIONES UNIDAS (2019l): *Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS) Opening Statement as Delivered by Ian McKayGeneva, April 9, 2018*, acceso febrero 2021, en https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/9April_US.pdf

NACIONES UNIDAS (2020a): *Commentaries on National Implementation on the guiding principles of LAWS (Gobierno de España)*, Organización de las Naciones Unidas, Ginebra, acceso en enero 2021, en https://documents.unoda.org/wp-content/uploads/2020/07/20200706-Spain.pdf

NACIONES UNIDAS (2020b): *United States. Agenda Item 5(a). Possible options for addressing the humanitarian and international security challenges posed by emerging technologies in the área of LAWS. Statement (19/9/2020),* Organización de las Naciones Unidas, Ginebra, acceso enero 2021, en https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2020/gge/statements/24Sept_US2.pdf

NACIONES UNIDAS (2020c): *U.S. Commentaries of the Guiding Principles,* Organización de las Naciones Unidas, Ginebra, acceso enero 2021, en https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2020/gge/documents/US_2020.pdf

NACIONES UNIDAS (2021a): *Considerations for the report of the Group of Governmental Experts of the High Contracting Parties to the Convention on Certain Conventional Weapons on emerging technologies in the area of Lethal Autonomous Weapons Systems on the outcomes of the work undertaken in 2017-2021*, acceso octubre 2021, en https://meetings.unoda.org/section/ccw-gge-2021_documents_14090_documents_14570/

NACIONES UNIDAS (2021b): *China's Comments on the Working Recommendations of the Group of Governmental Experts on LAWS,* acceso octubre 2021, en https://meetings.unoda.org/section/ccw-gge-2021_documents_14090_documents_14570/

NACIONES UNIDAS (2021c): *U.S. Proposals,* acceso octubre 2021, en https://meetings.unoda.org/section/ccw-gge-2021_documents_14090_documents_14570/

NACIONES UNIDAS (2021d): *Documents Reflecting U.S. Practice Related to Emerging Technologies in the Area of Lethal Autonomous Weapons Systems,* acceso octubre 2021, en https://meetings.unoda.org/section/ccw-gge-2021_documents_14090_documents_14570/

NACIONES UNIDAS (2021e): *Comentarios de España para la reunión del GGE del CCW para los LAWS 2021,* acceso octubre 2021, en https://meetings.unoda.org/section/ccw-gge-2021_documents_14090_documents_14570/

NACIONES UNIDAS (2021f): *Contribution by the International Committee of the Red Cross submitted to the Chair of the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems as a proposal for consensus recommendations in relation to the clarification, consideration and development of aspects of the normative and operational framework,* acceso octubre 2021, en https://meetings.unoda.org/section/ccw-gge-2021_documents_14090_documents_14570/

NAGEL, T. (1987): "The Fragmentation of Value", en *Moral Dilemmas,* C. Gowans (ed.), Oxford University Press, Oxford, 174-187.

NATIONAL SECURITY COMMISSION ON ARTIFICIAL INTELLIGENCE (NSCAI) (2021): *Final Report,* NSCAI, Arlington-Virginia, acceso octubre 2021, en https://www.nscai.gov/2021-final-report/

NELSON, A. L. (2013): "Artificial life and machine consciousness", en *AAAI Fall Symposium Series: AAAI Technical Report FS13-02,* 52-57.

NGUYEN, R. (2013): "Navigating *Jus Ad Bellum* in the Age of Cyber Warfare", *California Law Review,* 101- 4, 1079-1130.

NUSEIBEH, B. y EASTERBROOK, S. (2000): "Requirements Engineering. A Roadmap", *Proceedings of the International Conference on Software Engineering (ICSE 2000)*, ACM Press, Limerick, 4-11.

O'CONNELL, M. E. (2014): "Banning Autonomous Killing. The lethal and ethical requirements that Human make Near-Time Lethal Decisions", en M. Evangelista y H. Shue (eds.), *The American Way of Bombing, Changing Ethical and Legal Norms, from B-17s to Drones*, Cornell University Press, Ithaca-Londres, 224-236, acceso diciembre 2020, en https://www.law.upenn.edu/live/files/3802-oconnell-mary-banning-autonomous-killing-the-legal

ORGANIZACIÓN DEL TRATADO DEL ATLÁNTICO NORTE (OTAN) (2013): *Tallinn Manual on the International Law Applicable to Cyber Warfare – Prepared by the National Group of Experts at the invitation of the NATO-CCDCOE*, Cambridge University Press, Cambridge – Reino Unido, acceso enero 2021, en https://www.peacepalacelibrary.nl/ebooks/files/356296245.pdf

ORGANIZACIÓN DEL TRATADO DEL ATLÁNTICO NORTE (OTAN) (2017): *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations – Prepared by the National Group of Experts at the invitation of the NATO-CCDCOE*, Cambridge University Press, Cambridge – Reino Unido,

ORGANIZACIÓN DEL TRATADO DEL ATLÁNTICO NORTE [OTAN] (2019): *NATOTerm – The Official NATO Terminology Database,* Organización del Tratado del Atlántico Norte, acceso mayo 2019, https://nso.nato.int/natoterm/Web.mvc

ORGANIZACIÓN PARA LA COOPERACIÓN Y DESARROLLO ECONÓMICO (OCDE) (2019): *Scoping the OECD AI Principles. Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)*, OECD Digital Economy Papers, OCDE, Paris, acceso abril 2021, en https://www.oecd-ilibrary.org/docserver/d62f618a-en.pdf?expires=1617937963&id=id&accname=guest&checksum=CA9A2DFEE2FCB8F7D6A957D609144661

OXFORD DICTIONARIES (2019): "Cyberwarfare definition", Oxford Dictionaries, acceso marzo 2019, https://en.oxforddictionaries.com/definition/cyberwarfare

PADMANABHAN, V. M. (2013): "Cyber Warriors and the *jus in bello*", *International Law Studies -U.S. Naval War College*, 89, 288-308.

PARLAMENTO EUROPEO (2015): *Understanding Hybrid Threats* (At a Glance June 2015), Parlamento Europeo, Bruselas, acceso noviembre 2020, en https://www.europarl.europa.eu/RegData/etudes/ATAG/2015/564355/EPRS_ATA(2015)564355_EN.pdf

PARTHEMORE, J. y WHITBY, B. (2013): "What makes any agent a moral agent?: Reflections on machine consciousness and moral agency", *International Journal of Machine Consciousness,* 5(2), 105-129.

PAX (2018): *Where to draw the line? Increasing autonomy in weapon systems – Technology and trends,* Pax, Utrecht, acceso febrero 2021, en

https://www.paxforpeace.nl/publications/all-publications/where-to-draw-the-line

PETERSEN, S. (2012): "Designing people to serve", en Lin *et al* (eds.), *Robot ethics: the ethical and social implications of robotics,* MIT Press, Cambridge, 283-298.

PICARD, R. (1997): *Affective Computing,* MIT Media Lab. Technical Report 231, MIT Press, Cambridge, acceso febrero 2021, en http://www.macs.hw.ac.uk/~yjc32/project/ref-social%20media%20campaign/1995-affective%20computing.pdf

PICARD, R. (2003): "Affective computing: challenges", *International Journal of Human-Computer Studies,* 59, 55-64.

PIPYRO, K., Thraskias, C., Mitrou, L., Gritzalis, D. y Apastolopoulos, T. (2018): "A new strategy for improving cyber-attacks evaluation in the context of Tallinn Manual", *Computer's & Security*, 74, 371-383, acceso enero 2021, en

https://www.infosec.aueb.gr/Publications/COSE%20SI%20Tallinn%20Website.pdf

POHL, J. (2015): "Artificial Superintelligence: Extinction or Nirvana?", *InterSymp 2015 – 27th International Conference on Systems Research, Informatics and Cybernetics (August 3-8 2015)*, Baden-Baden-Germany.

POOLE, D., MACKWORTH, A. y GOEBEL, R. (1998): *Computational Intelligence. A logical approach,* Oxford University Press, Nueva York – Oxford.

PRATT, T. C., CULLEN, F. T., BLEVIS, K. R., DAIGLE, L. E. y MADENSEN T. D. (2006): "The empirical status of deterrence theory: a meta-analysis", F. T. Cullen *et al* (eds.), *The Status of Criminological Theory,* Transaction Publishers, New Brunswick, NJ, 37-76.

QUACKENBUSH, S. L. y ZAGARE, F. C. (2016): *Modern Deterrence Theory: Research Trends, Policy Debates and Methodological Controversies*, Oxford Handbooks Online, Oxford University Press, acceso noviembre 2020, en https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199935307.001.0001/oxfordhb-9780199935307-e-39?rskey=UDetKK&result=10

RAE, A. (2014): *Helping the Operator in the Loop: Practical Human Machine Interface Principles for Safe Computer Controlled Systems,* Queensland, acceso marzo 2021, en https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.152.9880&rep=rep1&type=pdf

RAMADHAN, A., SENSUSE, D. A., ARYMURTHY, A. N. (2011): "e-Government Ethics: a Synergy of Computer Ethics, Information Ethics and Cyber Ethics", *Internationa Journal of Advanced Computer Science and Applications,* vol. 2 (8), 82-86.

RAMASWAMY, S. y JOSHI, H. (2009): "Automation and Ethics", en S. Y- Nof (ed.), *Springer Handbook of Automation,* Springer, Berlin-Heidelberg, 809-833.

REAL ACADEMIA ESPAÑOLA (RAE) (2020): *Diccionario de la lengua española,* Real Academia Española, Madrid, acceso marzo 2021, en https://dle.rae.es/

REED, G. S. y JONES, N. (2013): "Toward Modeling and Automating Ethical Decision Making: Design, Implementation, Limitations and Responsabilities", *Topoi,* 32, 237-250.

REED, G. S., PETTY, M. D., JONES, N., MORRIS, A., BALLENGER, J. y DELUGACH. H. (2016): "A principles-based model of ethical considerations in military decision making", *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology,* 13(2): 195-211.

REISENZEIN, R. HUDLICKA, E. DASTANI, M., GRATCH, J., HINDRIKS, K., LORINI, E. y MEYER, J.-J. (2013): "Computational Modeling of Emotion: Towards improving the Inter-and Intradisciplinary Exchange", *IEEE Transactions on Affective Computing*, vol. 4-3, 246-266.

REISMAN, W. M. (1985): "Criteria for the Lawful Use of Force in International Law", *Yale Journal of International Law,* 10, 279-285.

REMUS, T. (2013): "Cyber Attacks and International Law of Armed Conflicts: A '*jus ad bellum*' Perspective, *Journal of International Commercial Law and Technology,* 8-3, 179-189.

REUTERS (2020): "Yemenis Houthis say they hit Soudi oil facility in drone, missile attack", *Reuters Aerospace and Defence (July 13 2020),* Thompson Reuters, Londres, acceso en noviembre 2020, https://www.reuters.com/article/us-saudi-security-yemen-idUSKCN24D0U6

ROBERTS, A. y GUELF, R. (1989): *Documents in the Laws of War. 28th. Ed.,* Clarendon Press, Londres.

ROBINSON, M., JONES, K. y JANICKE, J. (2015): "Cyber Warfare: Issues and Challenges", *Computers & Security,* Vol. 49, 70-94.

RODEN, D. (2012): "The Disconnection Thesis", en A. H Eden, J.H. Moor, J. H. Soraker y E. Steinhart (eds.), *Singularity Hypothesis. A Scientific and Philosophical Assessment,* The Frontiers Collection- Springer, Berlin, 281-298.

RODRÍGUEZ ÁLVAREZ, J. (2019): "Social Challenges of Artificial Intelligence: The Case of Lethal Autonomous Systems", *TSU Journal of Law,* nº 1 (2018), Tiblisi State University Press, 244-268.

ROFF, H. M. (2015a): "Lethal Autonomous Weapons and Jus Ad Bellum proportionality", *Case Western Reserve Journal of International Law,* 47-1, 37-52.

ROFF, H. M. (2015b): "Autonomous or 'Semi' Autonomous Weapons? A distinction without difference", Huffington Post, Verizon Media , Nueva York, acceso febrero 2021, en https://www.huffpost.com/entry/autonomous-or-semi-autono_b_6487268?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAKy2urj259d0Nuvumyok-4se64-JYLFivAEYmHu8eV_KfqGsWH6J1y37or3Lwxo5d04zf5wjMhvl4j33ewAM3iliRZ_ckYvYixkwoqkTj3_tq_f_s2VnEBsdsfSpEPFU3pf53uOnG_vW-9MFbqYH8_2qn5izrV06JNG98tKGwR3a

ROGERS, J. L. (2014): "Legal Judgement Day for the Rise of Machines: A National approach to Regulating Fully Autonomous Weapons, *Arizona Law Review,* 56, 4, 1258-1272.

ROMERO, V. (2020): "Redes Neuronales Artificiales: El perceptrón multicapa", *VRElectronía,* Mexico, acceso octubre 2021, en https://vrelectroniq.wixsite.com/vrelectroniq/post/redes-neuronales-artificiales-el-perceptr%C3%B3n-multicapa

ROSCINI, M. (2010): "World Wide Warfare – 'Jus ad Bellum' and the Use of Cyber Force", *Max Planck Yearbook of United Nations Law,* 14, 85-130.

ROSERT, E. y SAUER, F. (2019): "Prohibiting Autonomous Weapons: Put Human Dignity First", *Global Policy,* 10(3), 370-375.

ROSERT, E. y SAUER, F. (2021): How (not) to stop killer robots: A compartive análisis of humanitarian disarmamente campaign strategies", *Contemporary Security Policy,* 42(1), 4-29.

ROUSSEAU, J.-J., (1762 / 2004): *The Social Contract,* traducción M. Cranston, Penguin, Londres.

SÁNCHEZ MEDERO, G. (2010): "Los Estados y la Ciberguerra", *Boletín de Información*, 317, 63-76.

SANTOS-LANG, C. C. (2015): "Moral Ecology Approaches to Machine Ethics", en S. van Rysewyk *et al* (eds.), *Machine Medical Ethics*, Intelligent Systems, Control and Automation: Science and Engineering, Springer, vol. 74.

SARTOR, G. y OMICINI, A. (2016): "The autonomy of technological systems and responsabilities for their use", en N. Bhuta (ed.), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, Cambridge, 39-74.

SASSÒLI, M. (2014): "Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified", *International Law Studies/War College,* 90, 308-340.

SAXTON, A. (2016): "(Un)Dignified Killer Robotos? The Problem with the Human Dignity Argument", *Lawfare,* acceso febrero, 2021, en https://www.lawfareblog.com/undignified-killer-robots-problem-human-dignity-argument

SCHMITT, M. N. (1999): "Computer Network Attack and the Use of Force in International Law: Thoughts on a Normative Framework", *Columbia Journal of Transnational Law*, 37, 885-937.

SCHMITT, M. N. (2012a): "International Law in Cyberspace: The Koh Speech and Tallinn Manual Juxtaposed", *Harvard International Law Journal,* 54, 13-37, acceso enero 2021, en https://harvardilj.org/wp-content/uploads/sites/15/2012/12/HILJ-Online_54_Schmitt.pdf

SCHMITT, M. N. (2012b): "'Attack' as a Term of Art in International Law: The Cyber Operations Context", en C. Czosseck, R. Otis y K. Ziolkowski (eds.), *2012 4th International Conference on Cyber Conflict,* Organización del Tratado del Atlántico Norte (OTAN) – CCDCOE, Tallinn, 283-293.

SCHMITT, M. N. (2012c): "Cyber Operations and the *jus in bello*: Key issues", en R. A. Pedrozo y D. P. Wollschlaeger (eds.), *International Law and the Changing Character of War,* International Law Studies, 87, 89-110.

SCHMITT, M. N. (2013): "Reaction. Cyberspace and International Law: The Penumbral Myst of Uncertainty, *Harvard Law Review Forum,* vol. 126, 176-180.

SCHMITT, M. N. (2013b): "Autonomous Weapons Systems and International Humanitarian Law: A reply to the critics", *Harvard National Security Journal,* 4, 1-37.

SCHMITT, M. N (2017): "Grey Zones in the International Law of Cyberspace", *Yale Journal of International Law Online,* 42(2), acceso en enero 2021, en https://cpb-us-w2.wpmucdn.com/campuspress.yale.edu/dist/8/1581/files/2017/08/Schmitt_Grey-Areas-in-the-International-Law-of-Cyberspace-1cab8kj.pdf

SCORNAVACCHI, M. (2015): *Superintelligence, Humans and War,* Master Thesis, National Defense University-Joint Forces Staff College-Joint Advanced Warfighting School, Norfolk-Virginia.

SEHRAWAT, V. (2017): "Autonomous weapon system: Law of armed conflicto (LOAC) an other legal challenges, *Computer Law and Security Review,* 33, 38-56.

SHACKELFORD, S. J. (2009): "From Nuclear War to Net War: Analogizing Cyber Attacks in International Law", *Berkeley Journal of International Law,* 27-1, 192-251.

SHARKEY, A. (2017): "Can robots be responsable moral agents? And why should we care?", *Connection Science,* 29(3), 210-216.

SHARKEY, A. (2019): "Autonomous Weapons Systems, killer robots and human dignity", *Ethics & Information Technology,* 21, 75-87.

SHER, J. B. (2016): "Anonymous Armies: Modern 'Cyber-Combatants' and their Prospective Rights Under International Humanitarian Law", *Pace International Law Review,* 28-1, 233-275.

SHNEIDERMAN, B. (2020): "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy", *International Journal of Human-Computer Interaction,* 36(6), 495-504.

SINGER, M. G. (1971): *Generalization in Ethics,* Atheneum Press, Nueva York.

SIPONEN, M. T. (2001): "Five Dimensions of Information Security Awareness", *Computers and Society,*June, 24-29.

SOLOMONOFF, R. J. (1985): "The time scale of Artificial Intelligence: Reflections on social effects", *Human Systems Management,* 5, 149-153.

SPARROW, R. (2004): "The Turing Triage Test", *Journal of Ethics and Information Technology,* 6, 4, 203-213.

SPARROW, R. (2009): "Building a better WarBot: Ethical issues in the Design of Unmanned Systems for Military Applications", *Science and Engineering Ethics,* 15, 169-187.

SPINELLO, R. A. (2003): *Cyberethics, morality and law in Cyberspace*. Jones and Bartlett, Sudbury-Massachusetts.

SRIVASTAVA, A. y SENGUPTA, I. (2017): "Predicting Stock Market: An approach with Artificial Intelligence", *SMS Varanasi,* XIII(2), 73-77.

STAHL, B. C. (2004): "Information, Ethics, and Computers: The Problem of Autonomous Moral Agents", *Minds and Machines,* 14, 67-83.

STAHL, B.C. (2012): "Morality, Ethics and Reflection: A categorization of Normative IS Research", *Journal of the Association for Information Systems,* 13, 8, 636-656.

STAHL, B.C., TIMMERMANS, J. y MITTELSTADT, B. D. (2016): "The Ethics of Computing: A Survey of Computer-Oriented Literature", *ACM Computing Surveys,* 48 (4), Article 55.

STAHN, C. (2006): "'Jus ad bellum', 'jus in bello' . . . 'jus post bellum'? – *Rethinking the Conception of the Law of Armed Force*", *European Journal of International Law,* 17(5), 921-943.

STEINWANDT, R., GONZÁLEZ VASCO, M., PÉREZ DEL POZO, A. y SUÁREZ CORONA, A. (2021): "Password-Authenticated Key Establishment in the Advent of Scalable Quantum Computing", *JMM AMS-MAA (January 6-9, 2021) – AMS Special Session on Mathematics in Security & Defense, I,* acceso octubre 2021, en https://meetings.ams.org/math/jmm2021/meetingapp.cgi/Paper/2911

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE (SIPRI) y COMITÉ INTERNACIONAL DE LA CRUZ ROJA (CICR) (2020): *Limits on Autonomy in Weapon Systems. Identifying Practical Elements of Human Control,* Stockholm International Peace Research Institute, Solna-Suecia.

STOPKILLERROBOTS (2021): *Campaign to Stop Killer Robots,* acceso febrero 2021, en https://www.stopkillerrobots.org/members/

STÜCKELBERGER, C. (2018): "Cyber Society: Core Values and Virtues", en C. Stückelberger y P. Duggal (eds.), *Cyber Ethics 4.0 Serving Humanity with Values,* Globalethics.net, Global Series 17, Ginebra, 23-54.

SULLINS, J. P. (2005): "Ethics and artificial life: Fromo modeling to moral agents, *Ethics and Information Technology,* 7, 139-148.

SUSSMAN, B. (2020): "Cyber War vs Traditional War: The Difference is fading", *Secureworld (December 27, 2020),* Seguro Group Inc., Portland, acceso noviembre 2020, https://www.secureworldexpo.com/industry-news/cyber-war-vs-traditional-war

SWAN, K. y VALLIER, K. (2012): "The Normative significance of Conscience", *Journal of Ethics & Social Philosophy,* 6, 3, 1-21.

SZPAK, A. (2017): "Legal classification of the armed conflicto in Ukraine in the light of international humanitarian law", *Hungarian Journal of Legal Studies,* 58(3), 261-280.

TADDEO, M. y BLANCHARD, A. (2021): *A Comparative Analysis of the Definitions of Autonomous Weapons Systems,* SSRN Papers -Elsevier, acceso octubre 2021, en https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3941214

TANENBAUM, S. (2018): "Kinetic War vs Cyber War: The potential battlefields ahead", CyberCecurity, Denver-Aspen (Colorado), acceso noviembre 2020, en https://www.msspalert.com/cybersecurity-breaches-and-attacks/cyber-war-vs-kinetic-war-explained/

TAVANI, H. T. (2013): "Cyberethics", en A. L. C. Runehov y L. Oviedo (eds.), *Encyclopedia of Sciences and Religions,* Springer, Dordrecht, acceso febrero 2020, en https://link.springer.com/referenceworkentry/10.1007%2F978-1-4020-8265-8_279

TITIRIGA, R. (2016): "Autonomy of Military Robots: Assessing the Technical and Legal ("Jus In Bello") Thresholds, 32 J. Marshall J. Info. Tech. & Privacy L. 57 (2016)", *The John Marshall Journal of Information Technology & Privacy Law,* 32-2, 57-88.

TOLMEIJER, S., KNEER, M., SARASUA, C., CHRISTEN, M. y BERNSTEIN, A. (2020): "Implementations in Machine Ethics: A Survey", *ACM Computing Surveys,* 53(6), 132, acceso marzo 2021, en https://dl.acm.org/doi/pdf/10.1145/3419633

TORRANCE, S. (2008): "Ethics and Consciousness in Artificial Agents", *Artificial Intelligence and Society,* 22, 495-521.

TRIBUNAL PENAL INTERNACIONAL PARA LA EX YUGOSLAVIA (ICTY) (2000): *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia*, ICTY, La Haya, acceso febrero 2021, en https://www.icty.org/en/press/final-report-prosecutor-committee-established-review-nato-bombing-campaign-against-federal

TURING, A. M. (1950): "Computing Machinery and Intelligence", *Mind. A Quarterly Review of Psychology and Philosophy*, LIX, 236, 433-460.

ULAM, S. (1958): "Tribute to John von Neumann", *Bulletin of the American Mathematical Society,* 64,3 part 2, 1-49.

UNI GLOBAL UNION (2017): *Top 10 Principles for Ethical Artificial Intelligence,* UNI Global Union, Nyon-Suiza, acceso enero 2021, en http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION (UNESCO) (2021): *Proyecto de Recomendación sobre la Ética de la Inteligencia Artificial*, UNESDOC Biblioteca Digital, acceso noviembre 2021, en https://unesdoc.unesco.org/ark:/48223/pf0000378931_spa

UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH (UNIDIR) (2018): *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies*, UNIDIR Resources, acceso octubre 2021, en https://unidir.org/publication/algorithmic-bias-and-weaponization-increasingly-autonomous-technologies

UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH (UNIDIR) (2019): *The Role of Data in Algorithmic Decision-Making*, UNIDIR Resources, acceso octubre 2021, en https://unidir.org/publication/role-data-algorithmic-decision-making

UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH (UNIDIR) (2020a): *The human element in decisions about the use of force,* UNIDIR -Ginebra, acceso octubre 2021, en https://unidir.org/publication/human-element-decisions-about-use-force

UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH (UNIDIR) (2020b): *Robótica de Enjambre. Informe de Investigación,* UNIDIR, Ginebra, acceso octubre 2021, en https://unidir.org/publication/robotica-de-enjambre-informe-de-investigacion

UNITED NATIONS OFFICE OF DISSARMAMENT AFFAIRS (UNODA) (2014): *Convention on Certain Conventional Weapons,* Oficina de la Naciones Unidas, Ginebra, acceso en enero 2021, en https://unoda-web.s3-accelerate.amazonaws.com/wp-content/uploads/assets/publications/more/ccw/ccw-booklet.pdf

UNITED STATES OF AMERICA (USA) (2016): *Artificial Intelligence, Automation and Economy,* Executive Office of the President, Washington DC, acceso marzo 2021, en https://laedc.org/wp-content/uploads/2016/12/Artificial-Intelligence-Automation-Economy-12-27-16.pdf

USCYBERCOM (2018): "Achieve and Maintain Cyberspace Superiority. Command Vision for US Cyber Command", US Cyber Command, Fort Meade- Maryland, acceso diciembre 2020, en https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20Vision%20April%202018.pdf?ver=2018-06-14-152556-010

VACURA, M. (2015): " The History of Computer Ethics and its Future Challenges", *Information Technology and Society Interaction and Interdependence. Proceedings of 23rd Interdisciplinary Information Management Talks (IDIMT 2015- Linz, 9-11, septiembre, 2015)*, 325-333.

VAN DEN HOVEN, J. (2010): "The use of normative theories in Computer Ethics", en L. Floridi (ed.), *The Cambridge Handbook of Information and Computer Ethics,* Cambridge University Press, Cambridge, 59-76.

VAN ERP, D. (2016): *Creating artificial moral agents for surveillence robots that can identify care and harm,* Bachelor Thesis in Artificial Intelligence, Radboud University, Nijmegen.

VERDIESEN, I., SANTONI DE SIO, F. y DIGNUM, V. (2021): "Accountability and Control over Autonomous Weapons Systems: A Framework for Comprehensive Human Oversight", *Minds & Machines,* 31: 137-163.

VINGE, V. (1993): "The Coming Technological Singularity: How to Survive in the PostHuman Era", *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, acceso noviembre 2019, http://www.rohan.sdsu.edu/faculty/vinge/misc/singularity.html

VITHOULKAS, G. y Muresano, DF. (2013): "Conscience and Consciousness: a definition", *Journal of Medicine and Life,* 7, 1, 104-108.

WADHWA, V. (2021): "Killer Flying Robots are here. What Do We Do Know?", *Foreign Policy,* acceso octubre 2021, en https://foreignpolicy.com/2021/07/05/killer-flying-robots-drones-autonomous-ai-artificial-intelligence-facial-recognition-targets-turkey-libya/

WAGNER, M. (2014): "The Dehumanization of International Humanitarian Law: Legal, Ethical and Political Implications of Autonomous Weapons Systems, *Vanderbilt Journal of Transitional Law,* 47, 1371-1424.

WALKER, P. W. (2019): *War without oversight: challenges to the deployment of Autonomous Weapons Systems,* Tesis Doctoral, School of Humanities – University of Buckingham, Buckingham-Reino Unido.

WALLACH, W. (2008): "Implementing moral decision making faculties in computers and robots", *Artificial Intelligence and Society,* 22, 463-475.

WALLACH, W. (2010): "Robot minds and human ethics: the need for a comprehensive model of moral decision making", *Ethics and Information Technology,* 12, 243-250.

WALLACH, W., ALLEN, C. y SMIT, I. (2008): "Machine morality: bottom-up and top-down approaches for modelling human moral faculties", *Artiificial Intelligence and Society,* 22, 4, 565-582.

WALLACH, W. y ALLEN, C. (2009): *Moral Machines. Teaching Robots Right from Wrong,* Oxford University Press, Nueva York.

WALLACH, W. y ALLEN, C. (2011): "Moral Machines: Contradictions in Terms, or Abdication of Human Responsability?", en P. Lin et al (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, acceso enero 2021, en https://www.researchgate.net/publication/257931212_Moral_Machines_Contradiction_in_Terms_or_Abdication_of_Human_Responsibility

WALLACH, W. y ALLEN, C. (2013): "Framing Robot Arms Control", *Ethics and Information Technology,* 15, 125-135.

WALLACH, W. y MARCHANT, G. (2019): "Towar the Agile and Comprehensive International Governance of AI and Robotics", *Proceedings of the IEEE,* 107(3), 505-508.

WANG, P. (1995): *On the Working Definition of Intelligence,* Center for Research on Concepts and Cognition – Indiana University, Bloomington – IN.

WALZER, M. (1977): *Just and unjust wars: A Moral argument with Historical Illustrations,* Basic Books, Nueva York.

WATKIN, K. (2016): *Fighting at the Legal Boundaries: Controlling the Use of Force in Contemporary Conflict,* Oxford University Press, Oxford – UK.

WATTS, S. (2012): "The Notion of Combatancy in Cyber Warfare", en C. Czosseck *et al* (eds.), *2012. 4th International Conference on Cyber Conflict,* Organización del Tratado del Atlántico Norte (OTAN), Tallinn, 235-249.

WESTHUES, A. (2020): *Sistemas de Armas Autónomas Letales: ¿Autónomas o Automatizadas?*, Trabajo de Fin de Máser, Máster Universitario en Paz, Seguridad y Defensa, *Instituto Universitario General Gutiérrez Mellado,* Madrid.

WHITBY, B. (2008): "Computing Machinery and Morality", *Artificial Intelligence & Society,* 22, 551-563.

WHITTLESTONE, J., NYRUP, R., ALEXANDROVA, A., DIHAL, K. y CAVE, S. (2019): *Ethical and societal implications of algorithms, data and artificial intelligence: a road map for research,* Nuffield Foundation, Londres.

WIENER, N. (1948): *Cybernetics; or Control and Communication in the Animal and the Machine,* Technology Press, J. Wiley & Sons, Inc., Nueva York.

WINFIELD, A. F. T. y JIROTKA, M. (2018): "Ethical governance is essential to building trust in robotics and artificial intelligence systems", *Philosophical Transactions of the Royal Society A,* The Royal Society Publishing, acceso febrero 2021, en https://royalsocietypublishing.org/doi/full/10.1098/rsta.2018.0085

WIZENBAUM, J. (1976): *Computer Power and Human Reason: From judgement to calculation,* W. H. Freeman & Co. Ltd., Nueva York.

XERIDIA (2021): "Redes Neuronales Artificiales: Qué son y cómo se entrenan", Xeridia, León-Madrid-Londres, acceso noviembre 2021, en https://www.xeridia.com/blog/redes-neuronales-artificiales-que-son-y-como-se-entrenan-parte-i

YAMPOLSKIY, R. V. (2012): "What to do with the Singularity Paradox", en V. C. Muller (ed.), *Philosophy and Theory of Arificial Intelligence,* SAPERE 5, Springer, Heidelberg-Nueva York-Dordrecht-Londres, 397-413.

YAMPOLSKIY, R. V. (2015): "On the Limits of Recursively Self-Improving AGI", en J. Bieger, B. Goertzel y A. Potapov (eds.), *Proceedings Artificial General Intelligence: 8th International Conference AGI 2015, Berlin (22-25 July 2015*), Berlin, 394-403.

YOUNG, J. E., SHARLIN, E. y IGARASGHI, T. (2011): "What is mixed reality anyway? Considering the boundaries of mixed reality in the context of robots", en X. Wang (ed.), *Mixed Reality and Human Robot Interaction,* Springer, Londres-Nueva York , 1-11, acceso febrero 2021, en https://www.researchgate.net/publication/226611963_What_Is_Mixed_Reality_Anyway_Considering_the_Boundaries_of_Mixed_Reality_in_the_Context_of_Robots

YU, H., SHEN, Z., MIAO, C., LEUNG, C. LESSER, V.R., y YANG, Q. (2018): "Building Ethics into Artificial Intelligence", *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18),* 5527-5533.

YUDKOWSKY, E. (1996): "Staring at the Singularity", acceso noviembre 2019, en http://yudkowsky.net/obsolete/singularity.html

ZAVE, P. (1997): "Classification of Research Efforts in Requirements Engineering, *ACM Computing Surveys,* 29(4), 315-321.

ZENIL, H. (2013): "Complejidad y aleatoridad", *Ciencia – Revista de la Academia Mexicana de Ciencias*, Vol. 64, 4 (octubre-diciembre), 56-63.

ZIEBA, S. POLET, P., VANDEHAEGEN, F. y DEBERNARD, S. (2010): "Principles of adjustable autonomy: a framework for resilient human-machine cooperation", *Cognition, Technology & Work,* 12, 193-203.

ZOHORA, S.E., KHAN, A. M., SRIVASTAVA, A. K., NGUYEN, N. G. y DEY, N. (2016): "A Study of the State of the Art in Synthetic Emotional Intelligence in Affective Computing", *International Journal of Synthetic Emotions,* 7, 1, 1-12.

ZWITTER, A. (2014): "Big Data Ethics", *Big Data & Society,* July-December, 1-6.