LA SUPERVISIÓN HUMANA DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL DE ALTO RIESGO. APORTACIONES DESDE EL DERECHO INTERNACIONAL HUMANITARIO Y EL DERECHO DE LA UNIÓN EUROPEA

HUMAN OVERSIGHT OVER HIGH-RISK ARTIFICIAL INTELLIGENCE SYSTEMS. CONTRIBUTIONS FROM INTERNATIONAL HUMANITARIAN LAW AND EUROPEAN UNION LAW

ARITZ OBREGÓN FERNÁNDEZ[*]
GUILLERMO LAZCOZ MORATINOS[**] [***]

RESUMEN: La automatización de decisiones por parte de sistemas de inteligencia artificial es un fenómeno creciente que abarca todos los ámbitos de la vida. La Comisión Europea, consciente de los riesgos que conlleva el uso de estas tecnologías para los derechos y libertades fundamentales, propone en su "Ley de Inteligencia Artificial" que la supervisión humana sea un requisito obligatorio para el diseño y desarrollo

Fecha de recepción del trabajo: 15 de julio de 2021. Fecha de aceptación de la versión final: 23 de noviembre de 2021.

[*] Área de Derecho Internacional Público y Relaciones Internacionales, Departamento de Derecho Público y Ciencias Histórico-Jurídicas y del Pensamiento Político, Universidad del País Vasco (UPV/EHU); e-mail: aritz.obregon@ehu.eus, El texto ha sido elaborado gracias a la financiación de la ayuda del Programa Predoctoral de Formación de Personal Investigador No Doctor del Departamento de Educación del Gobierno Vasco.
[**] G.I. Cátedra de Derecho y Genoma Humano. Derecho Público, Universidad del País Vasco (UPV/EHU); e-mail: guillermo.lazcoz@ehu.eus; Código ORCID: orcid.org/0000-0001-6567-045X; El texto ha sido elaborado gracias a la financiación de la ayuda FPU 16/06314 del Ministerio de Universidades del Gobierno de España y del Departamento de Educación del Gobierno Vasco para apoyar las actividades de Grupos de Investigación del Sistema Universitario Vasco (IT 1066-16).
[***] Para el desarrollo del presente trabajo en coautoría, ambos firmantes hemos participado de forma equitativa en las fases de redacción y posteriores trabajos de revisión del manuscrito, salvo la primera redacción de las secciones II (2º coautor) y III (1er coautor), las cuales se realizaron individualmente. Asimismo, queremos agradecer a las revisoras/es sus atentos comentarios.

de las mismas. Sin embargo, la supervisión humana tiene un escaso desarrollo en el ámbito normativo europeo. Por ello, proponemos recurrir al concepto de Control Humano Significativo desarrollado en el marco del Derecho Internacional Humanitario, analizando en este articulo las aportaciones estatales, doctrinales y propuestas normativas realizadas por el Grupo de Expertos Gubernamentales de composición abierta sobre los Sistemas de Armas Autónomos Letales. Todas estas aportaciones nos permiten acercarnos desde una perspectiva novedosa al concepto de supervisión humana que propone la Comisión Europea para los sistemas de inteligencia artificial de alto riesgo. Concluimos el articulo con la búsqueda de elementos universalizables en la supervisión humana, que puedan ser aplicables a la automatización de las decisiones, independientemente del ámbito del que se trate.

ABSTRACT: The automation of decision-making by artificial intelligence systems is a growing phenomenon affecting all areas of society. The European Commission, aware of the risks that the use of these technologies entails for fundamental rights and freedoms, proposes in its Artificial Intelligence Act to introduce human oversight as a mandatory requirement for the design and development of these technologies. However, human oversight is underdeveloped in the European regulatory environment. For this reason, we propose to resort to the concept of Meaningful Human Control developed in the framework of International Humanitarian Law. In this article we analyse the state contributions, doctrinal and policy proposals made by the Group of Governmental Experts on Lethal Autonomous Weapons Systems. All these contributions allow us to approach from a novel perspective the concept of human oversight proposed by the European Commission for high-risk artificial intelligence systems. We conclude the article looking for universally applicable elements in human oversight for the automation of decision-making, irrespective of the field in question.

PALABRAS CLAVE: Inteligencia Artificial, Sistema de Armas Autónomos Letales, Supervisión humana, Control Humano Significativo, Derecho Internacional Humanitario, Derecho de la Unión Europea.

KEYWORDS: Artificial Intelligence, Lethal Autonomous Weapons Systems, Human Oversight, Meaningful Human Control, International Humanitarian Law, European Union Law.

I. INTRODUCCIÓN

Ante la proliferación de sistemas de inteligencia artificial (en adelante IA) que impulsan la automatización de la toma de decisiones en muy diversos ámbitos, tanto las instituciones como la doctrina jurídica se preguntan cómo regular su desarrollo, despliegue y uso de forma segura y garantista con los derechos y libertades de la ciudadanía. Una de las respuestas más habituales a esta cuestión ha sido la introducción de mecanismos de gobernanza basados en la supervisión humana, ya sea en forma de control, revisión o intervención, entre otras. Tal y como analizaremos, en distintas propuestas, destacando entre ellas la reciente propuesta de Reglamento "Ley de Inteligencia Artificial", la Comisión y el Parlamento europeos proponen que la supervisión humana sea un requisito normativo fundamental para cualquier sistema de IA calificado de alto riesgo. Sin embargo, y a pesar de que encontramos ejemplos de esta clase de mecanismos de gobernanza en la regulación vigente, su aplicación ha sido escasa y han pasado de forma desapercibida en la literatura que ha analizado dichas normas. La forma en la que se prodiga la relevancia de los mecanismos de gobernanza basados en la supervisión humana para hacer frente a los retos de la inteligencia artificial y la

 DOI: 10.17103/reei.42.08

automatización de la toma de decisiones, no se corresponde con su escaso desarrollo normativo.

En busca de respuestas jurídicas, acudimos al ámbito de los sistemas de IA con fines militares, expresamente excluido del ámbito de aplicación de la propuesta de Reglamento. Nos referimos al concepto de Control Humano Significativo con origen en el Derecho Internacional Humanitario (en adelante DIH), sugerido para limitar el desarrollo y la utilización de los Sistemas de Armas Autónomas Letales (en adelante SAAL). Como veremos, esta clase de mecanismo basado en la supervisión humana se ha configurado como un elemento central a la hora de satisfacer un consenso general de la comunidad internacional sobre la utilización de estos sistemas. Por ello nos preguntamos y analizamos si existe un acuerdo sobre qué significa este concepto en el Derecho internacional y sobre su fundamento y los elementos que lo configuran. Para ello nos detenemos en los consensos que han ido alcanzando los Estados y los grupos de la sociedad civil organizada que han participado en los debates internacionales institucionalizados realizados hasta la fecha, así como las aportaciones de la doctrina especializada.

De esta manera, a partir de este análisis tratamos de dilucidad si los elementos extraídos del DIH, un ámbito que por su naturaleza trata de limitar las vulneraciones de los derechos más fundamentales, pueden ser relevantes para la regulación de la IA y los sistemas autónomos con fines diversos en otros contextos normativos, en particular para la propuesta de Reglamento "Ley de Inteligencia Artificial" de la Comisión Europea. Teniendo en cuenta que los reguladores europeos consideran esencial que los sistemas de inteligencia artificial de alto riesgo no se implementen sin supervisión humana, exploramos las posibilidades de que los elementos extraídos en el anterior apartado puedan ser de utilidad a estos efectos. Como veremos, posiciones críticas bien razonadas cuestionan que la supervisión humana pueda tener un rol sustantivo en el despliegue de estos sistemas y, por ende, los mecanismos de gobernanza basados en dicha supervisión serían inútiles. A nuestro modo de ver, mientras no se defina qué debemos entender por supervisión humana a efectos normativos, estas críticas son difícilmente abordables en uno u otro sentido, de ahí el esfuerzo realizado en este texto.

DOI: 10.17103/reei.42.08

II. SUPERVISIÓN HUMANA EN EL CONTEXTO NORMATIVO EUROPEO. TAN NECESARIA COMO INEXPLORADA

En el seno de las instituciones europeas hemos visto un creciente interés por la regulación de la inteligencia artificial y los sistemas autónomos[1]. Sin duda, hemos de destacar la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se pretenden establecer normas armonizadas en materia de inteligencia artificial y modificar determinados actos legislativos de la Unión[2], lanzada en 2021 por la Comisión Europea; propuesta a la curiosamente que se ha denominado "Ley de Inteligencia Artificial" (en adelante Ley de Inteligencia Artificial). No obstante, también hemos de tener en cuenta dos precedentes muy relevantes sobre los que se concretó dicha propuesta normativa. Primero, la Comisión Europea publicó el Libro Blanco sobre Inteligencia Artificial en febrero de 2020[3]. Uno de los pilares fundamentales de esta estrategia era la de impulsar la creación de un marco normativo que proveyese de un ecosistema de confianza en el que la innovación por parte de empresas y organismos públicos se generase a partir del respeto de los derechos y la seguridad de la ciudadanía[4]. Segundo, el Parlamento Europeo aprobó en octubre de 2020 la Resolución con recomendaciones destinadas a la Comisión sobre inteligencia artificial, robótica y tecnologías conexas[5], en cuyo anexo primero se

[1] A efectos de este trabajo, tomamos como punto de partida las definiciones propuestas por el Parlamento Europeo. Respecto a los sistemas de IA, "todo sistema basado en programas informáticos o incorporado en dispositivos físicos que muestra un comportamiento que simula la inteligencia, entre otras cosas, mediante la recopilación y el tratamiento de datos, el análisis y la interpretación de su entorno y la adopción de medidas, con cierto grado de autonomía, para lograr objetivos específicos". Respeto al concepto autónomo, "todo sistema de IA que funciona interpretando determinados datos de entrada y utilizando un conjunto de instrucciones predeterminadas, sin limitarse a ellas, a pesar de que el comportamiento del sistema esté limitado y orientado a cumplir el objetivo que se le haya asignado y otras decisiones pertinentes de diseño tomadas por su desarrollador". Ver punto 1 de la Resolución del Parlamento Europeo, de 20 de enero de 2021, sobre inteligencia artificial: cuestiones de interpretación y de aplicación del Derecho internacional en la medida en que la UE se ve afectada en los ámbitos de los usos civil y militar, así como de la autoridad del Estado fuera del ámbito de la justicia penal (2020/2013(INI)), europarl.europa.eu/doceo/document/TA-9-2021-0009_ES.html.

[2] COMISIÓN EUROPEA, *Propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión*, COM(2021) 206 final, Bruselas, 21 de abril de 2021, digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[3] COMISIÓN EUROPEA, *Libro Blanco sobre la inteligencia artificial - un enfoque europeo orientado a la excelencia y la confianza*, COM(2020) 65 final, Bruselas, 19 de febrero de 2020, ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

[4] De acuerdo con el enfoque antropocéntrico que ya había comenzado a desarrollar en la Comunicación para generar confianza en la inteligencia artificial centrada en el ser humano (COM (2019) 168 final) y que también había adoptado el Grupo independiente de expertos de alto nivel sobre IA en sus Directrices Éticas para una IA fiable. Ver GRUPO DE EXPERTOS DE ALTO NIVEL SOBRE IA, *Directrices éticas para una IA fiable*, Bruselas, 8 de abril de 2019, op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[5] Resolución del Parlamento Europeo, de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas (2020/2012(INL)), europarl.europa.eu/doceo/document/TA-9-2020-0275_ES.html. En adelante Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre.

DOI: 10.17103/reei.42.08

recoge una propuesta legislativa para la tramitación de un Reglamento sobre los principios éticos para el desarrollo, el despliegue y el uso de dichas tecnologías.

El rol que se otorga a la supervisión humana en estos documentos mantiene una misma raíz a pesar de las diferencias que presentan: la supervisión humana es un principio y requisito fundamental en la regulación de los sistemas calificados de alto riesgo[6]. Por un lado, el Libro Blanco establece que la supervisión humana debe ser un requisito legal de obligado cumplimiento para las aplicaciones de alto riesgo, que solo puede alcanzarse "garantizando una participación adecuada de las personas con relación a dichas aplicaciones"[7], pudiendo variar el tipo y nivel de esta participación humana de un caso a otro[8]. La propuesta del Parlamento opta, en su artículo 7, por la inclusión de la supervisión humana integral como principio ético de obligado cumplimiento para el desarrollo, uso y despliegue de estas tecnologías que, en todo caso, permita el restablecimiento del control humano cuando sea necesario[9]. Conforme al considerando décimo de esta propuesta, la supervisión humana integral debe tener carácter significativo con independencia del mecanismo concreto por el que se opte: revisión, evaluación, intervención o control humanos. Por su parte, la Ley de Inteligencia Artificial recoge en su artículo 14 la supervisión humana como requisito obligatorio de los sistemas de IA de alto riesgo[10], debiendo éstos ser diseñados y desarrollados de forma que se garantice que puedan ser

[6] El enfoque basado en el riesgo es común también a estas iniciativas, si bien, ha variado la forma en la que se determina la calificación de "alto riesgo". Aunque, siguiendo las propuestas anteriores, en su anexo III también recoge una serie de sectores y usos específicos de alto riesgo, la Ley de Inteligencia Artificial opta por una calificación de alto riesgo que responde principalmente a la normativa armonizada por la UE. Ver notas al pie 89 y 90.

[7] COMISIÓN EUROPEA, *Libro Blanco sobre la inteligencia artificial - un enfoque europeo orientado a la excelencia y la confianza*, *op cit.*, p. 25.

[8] Cita en este sentido, de forma no exhaustiva, varios mecanismos de gobernanza a través de los cuales puede materializarse la supervisión humana: el resultado del sistema de IA no es efectivo hasta que un humano no lo haya revisado y validado; el resultado del sistema de IA es inmediatamente efectivo, pero se garantiza la intervención humana posterior; se realiza un seguimiento del sistema de IA mientras funciona y es posible intervenir en tiempo real y desactivarlo; en la fase de diseño, se imponen restricciones operativas al sistema de IA. Ver COMISIÓN EUROPEA, *Libro Blanco sobre la inteligencia artificial - un enfoque europeo orientado a la excelencia y la confianza*, *op cit.*, p. 26.

[9] La formulación propuesta por el Parlamento es la siguiente: "1. Las tecnologías de inteligencia artificial de alto riesgo, incluidos los programas informáticos, los algoritmos y los datos utilizados o producidos por dichas tecnologías, se desarrollarán, desplegarán y utilizarán de forma que se garantice en todo momento una supervisión humana integral".

2. Las tecnologías a que se refiere el apartado 1 se desarrollarán, desplegarán y utilizarán de forma que se pueda restablecer en todo momento el control humano cuando sea necesario, incluso mediante la alteración o la desactivación de dichas tecnologías". Ver art. 7 de la Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas, contenida como Anexo de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre.

[10] La versión en lengua castellana de esta propuesta ha traducido "human oversight" como "vigilancia humana". A nuestro parecer hay motivos de peso para tomar como referencia supervisión humana y no vigilancia. En primer lugar, esta traducción se aparta de los precedentes mencionados, el Libro Blanco de la propia Comisión y la propuesta del Parlamento. En segundo lugar, en la misma propuesta también se traduce "market surveillance" como "vigilancia del mercado" –a nuestro juicio esta traducción sí es ajustada–, provocando una reiteración del término vigilancia en contextos distintos. Por último, la propia traducción literal de "oversight" como "supervisión" resulta más ajustada.

DOI: 10.17103/reei.42.08

supervisados eficazmente por agentes humanos durante el uso y despliegue de los mismos. El modelo regulatorio por el que ha optado la Comisión responsabiliza a los proveedores de los sistemas del cumplimiento de los requisitos obligatorios como la supervisión humana, poniendo definitivamente el enfoque normativo sobre las fases de diseño y desarrollo de estas tecnologías que habían permanecido fuera del análisis doctrinal jurídico y de las iniciativas políticas[11].

Con carácter general, podemos afirmar que la supervisión humana ocupa un lugar central en la visión regulatoria de las instituciones europeas. Teniendo en cuenta el contenido concreto de la Ley de Inteligencia Artificial, todo apunta a que esta visión se traducirá en la prohibición de desarrollar y utilizar sistemas de IA considerados de alto riesgo que no permitan asegurar la supervisión humana. Ahora bien, los mecanismos de gobernanza basados en la supervisión humana no son necesariamente novedosos en el contexto normativo europeo. En su análisis sobre la regulación en materia de protección de datos en Europa, Leta JONES expone que, desde hace más de 50 años con las primeras regulaciones frente al desarrollo informático, se han impuesto restricciones al tratamiento totalmente automatizado para las tecnologías computacionales[12], a través de la figura del *human in the loop* y de otros mecanismos de gobernanza análogos[13]. Con carácter general, este mecanismo implica la introducción de un operador humano en la toma de decisiones con autoridad final sobre el sistema automatizado, de modo que el tratamiento automatizado no sea el fundamento exclusivo de la decisión[14]. En palabras de JONES, la cultura política europea, frente a la norteamericana, considera que "tratar a un individuo de forma totalmente automatizada […] es deshumanizar al individuo, porque una máquina sólo puede tratar a un humano de forma computacional. Por lo tanto, tratar a un ser humano de forma totalmente automatizada merma la dignidad del individuo y el restablecimiento de la misma solo puede ser proporcionado por la intervención de un operador humano en la toma de decisiones"[15].

En la normativa vigente encontramos varios ejemplos en los que se refleja esta cultura normativa. La Directiva (UE) 2016/680 relativa al tratamiento de datos personales por parte de las autoridades para fines de prevención, investigación, detección o enjuiciamiento penales, contiene en el artículo 11 el derecho a no ser objeto de una

[11] Ver OHM, P., LEHR, D., "Playing with the Data: What Legal Scholars Should Learn about Machine Learning", *UC Davis Law Review*, vol. 51, nº 2, 2017, p. 655 y ss, lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf.

[12] JONES, M. L., "The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood." *Social Studies of Science* vol. 47, nº 2, 2017, p. 220 y ss, doi.org/10.1177%2F0306312717699716.

[13] La figura del *human in the loop* tiene su origen en el enfoque centrado en el ser humano para los sistemas autónomos desarrollado por Sheridan. Ver SHERIDAN, T. B., "Human Centered Automation: Oxymoron or Common Sense?", en IEEE, *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, IEEE, Vancouver, 1995, pp. 823-828, doi.org/ 10.1109/ICSMC.1995.537867. Para una definición de los mecanismos human in/on/out the loop, ver METHNANI, L., *et al*, "Let Me Take Over: Variable Autonomy for Meaningful Human Control", *Frontiers in Artificial Intelligence*, nº 4, 2021, pp. 2-3.

[14] WAGNER, B., "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems", *Policy & Internet*, vol. 11, nº 1, 2019, p. 108, doi.org/10.1002/poi3.198.

[15] JONES, M. L., "The Right to a Human…", *op. cit.*, pp. 231-232. La traducción es nuestra.

DOI: 10.17103/reei.42.08

decisión que evalúe aspectos personales que le conciernen que se base únicamente en un tratamiento automatizado de los datos. Asimismo, la Directiva (UE) 2016/681 relativa a la utilización de datos del registro de nombres de los pasajeros, incluye en el artículo 7.6 la prohibición de tomar ninguna decisión que pudiera tener efectos jurídicos adversos para una persona o afectarle gravemente en razón únicamente del tratamiento automatizado de datos del registro. Por último, sin ánimo de ser exhaustivos, el artículo 22.1 del Reglamento (UE) 2016/679 General de Protección de Datos (en adelante RGPD) prohíbe la toma de decisiones basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en la persona interesada o le afecte significativamente de modo similar[16]. Según las directrices refrendadas por el Comité Europeo de Protección de Datos, para que una decisión no se base únicamente en el tratamiento automatizado, debe garantizarse que la supervisión de la decisión sea significativa y no un mero gesto simbólico, y se lleve a cabo por una persona autorizada y competente para modificar la decisión, analizando todos los datos pertinentes[17].

Todos estos ejemplos recogen la prohibición de llevar a cabo un determinado tratamiento automatizado sin la supervisión humana adecuada dado el riesgo que representa el propio tratamiento. Sin embargo, a pesar de que estos instrumentos forman parte del ordenamiento jurídico, hasta el momento, su aplicación e interpretación por los tribunales y desarrollo por la doctrina han sido muy escasos. Así puede observarse en el historial del artículo 22 del RGPD y su precedente, de contenido muy similar, el artículo 15 de la Directiva 95/46/EC, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos. Este precedente no pasó de ser un derecho de segunda categoría, raramente aplicado y esquivado por los destinatarios de la norma con facilidad[18]. Ni en el Tribunal de Justicia de la Unión Europea, ni en las jurisdicciones nacionales encontramos casos relevantes de aplicación del artículo 22 y su precedente[19], menos aun que centren su interés específicamente en la supervisión humana del tratamiento automatizado. Del mismo modo, ésta tampoco ha sido abordada con interés por la doctrina[20], a pesar de que el artículo 22 sí ha despertado múltiples discusiones, especialmente con relación a la existencia o no de un derecho a una explicación ante la toma de decisiones automatizadas[21].

---

[16] No podemos pasar por alto que a los sistemas de IA que utilicen datos personales será aplicable tanto el RGPD como el Reglamento al que dé lugar la propuesta de Ley de Inteligencia Artificial. El interés de esta intersección normativa será de relevancia extraordinaria, no obstante, sobrepasa los objetivos del presente trabajo.

[17] GRUPO DE EXPERTOS DE ALTO NIVEL SOBRE IA, *Directrices éticas para…*, *op. cit.*, p. 23.

[18] MENDOZA, I., BYGRAVE, L.A., "The Right Not to be Subject to Automated Decisions Based on Profiling", en SYNODINOU, T.E., JOUGLEUX, P., MARKOU, C., PRASTITOU, T. (eds), *EU Internet Law*, Springer, 2017, p. 78.

[19] Roig recoge los escasos casos resueltos en Alemania y Francia. Ver ROIG, A., *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, José María Bosch Editor, Barcelona, 2020, p. 30.

[20] HUQ, A. Z., "A Right to a Human Decision", *Virginia Law Review*, vol. 106, nº 3, 2020, p. 624, virginialawreview.org/wp-content/uploads/2020/05/106VaLRev611.pdf.

[21] A este respecto ver EDWARDS, L., VEALE, M., "Slave to the Algorithm? Why a Right to Explanationn Is Probably Not the Remedy You Are Looking For", Duke Law & Technology Review, vol. 16, 2017, pp. 18-84, scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1315&context=dltr o WACHTER, S., *et al*, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data

DOI: 10.17103/reei.42.08

En definitiva, a pesar de la indudable relevancia que se otorga a la supervisión humana a la hora de establecer las bases para la futura regulación de la IA, los mecanismos de gobernanza con fundamento en la supervisión humana que encontramos en la normativa europea vigente para regular el uso de sistemas automatizados apenas han sido legislados y aplicados. Es por ello que resulta pertinente prestar nuestra atención a otros ámbitos del Derecho que han avanzado en este punto. En este sentido, el Control Humano Significativo, encuadrado en el contexto del Derecho Internacional Humanitario, ha sido considerado por el Grupo Europeo de Ética en la Ciencia y las Nuevas Tecnologías (en adelante EGE) como un precedente a la hora de garantizar la supervisión humana para los sistemas autónomos[22]. Por su parte, el Parlamento Europeo lo ha identificado como una condición para la legalidad de los sistemas de armas autónomos[23].

III. APORTACIONES DESDE EL DERECHO INTERNACIONAL HUMANITARIO: CONTROL HUMANO SIGNIFICATIVO

El concepto Control Humano Significativo, en inglés *Meaningful Human Control*, tiene su origen en el desarrollo y despliegue de los llamados Sistemas de Armas Autónomas Letales o por su acrónimo SAAL[24]. Estas armas que, a efectos del presente trabajo,

Protection Regulation." *International Data Privacy Law*, vol. 7, nº 2, 2017, pp. 76–99, doi.org/10.1093/idpl/ipx005.

[22] Esta consideración también ha sido destacada por la doctrina más solvente, como Methnani, Wagner o Romeo Casabona. Ver EGE, *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*, Comisión Europea, Bruselas, marzo de 2018, op.europa.eu/es/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en; METHNANI, L., *et al*, "Let Me Take Over…", *op. cit.*, p. 1; ROMEO CASABONA, C. M., "Criminal Responsibility of Robots and Autonomous Artificial Intelligent Systems?", *Comunicaciones En Propiedad Industrial y Derecho de La Competencia*, nº 91, 2020, pp. 167-187 y WAGNER, B., "Liable, but Not in…", *op. cit.*, p. 119.

[23] De manera persistente el Parlamento Europeo ha solicitado la prohibición del desarrollo, la producción y la utilización de SAAL capaces de realizar ataques sin un control humano significativo. Ver considerando L y punto 3 de la Resolución del Parlamento Europeo, de 12 de septiembre de 2018, sobre los sistemas armamentísticos autónomos (2018/2752(RSP)); considerando 89 de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre, y considerandos 27, 29 y 34 de la Resolución del Parlamento Europeo, de 20 de enero de 2021, sobre inteligencia artificial: cuestiones de interpretación y de aplicación del Derecho internacional en la medida en que la UE se ve afectada en los ámbitos de los usos civil y militar, así como de la autoridad del Estado fuera del ámbito de la justicia penal (2020/2013(INI)).

[24] Los SAAL –*lethal autonomous weapons system*– también han sido denominados como robots asesinos –*killer robots*–, sistemas de armas autónomas –*autonomous weapon systems*–, sistemas de armas totalmente autónomos –*fully autonomous weapon systems*–, sistemas de armas robóticas, armas autónomas –autonomus weapon–, armas robotizadas, *autonomy in weapon system*, etc. En el presente trabajo atenderemos a la denominación que viene utilizando el Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales. Ver NACIONES UNIDAS, *Informe provisional del Relator sobre las ejecuciones extrajudiciales, sumarias o arbitrarias*, (A 65/321, 23 agosto 2010), párr. 17, p. 12; NACIONES UNIDAS, *Informe del Relator Especial para las ejecuciones extrajudiciales, sumarias o arbitrarias, Christof Heys*, (A HRC/23/47, 9 abril 2013), párr. 38, p. 8; NACIONES UNIDAS, *Informe del Secretario General sobre la labor de la Junta Consultiva en Asuntos de desarme de 2013*, (A 68/206, 26 julio 2013), párr. 42, p. 11; JIMÉNEZ-SEGOVIA, R., "Los sistemas de armas autónomos en la Convención sobre ciertas armas convencionales: sombras legales y éticas de una autonomía, ¿bajo el control humano?, *Revista Electrónica de Estudios Internacionales*, vol. 37, 2019, nota al píe 2, p. 3, doi.org/10.17103/reei.37.07; GUTIÉRREZ ESPADA, C., CERVELL

DOI: 10.17103/reei.42.08

podemos definir cómo sistemas de armas que en base a una programación previa son capaces de seleccionar y atacar objetivos sin la necesidad de intervención adicional de un operador humano[25], han generado grandes debates interdisciplinares que inciden de forma continua en el devenir de la comunidad internacional.

HORTAL, M. J., "Sistemas de armas autónomas, drones y derecho internacional", *Revista del IEEE*, Nº 2, 2013, pp. 29-33, revista.ieee.es/article/view/338/567 y BOULANGER, V., VERBRUGGEN, M., "Mapping the development of autonomy in weapon systems", *Stockholm International Peace Research Institute*, 2017, pp. 5-7, sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf

[25] Aunque la propuesta presentada haya sido ampliamente utilizada, cabe señalar que en estos momentos no hay un acuerdo sobre cómo definir esta clase de armas. Así, por ejemplo, otras definiciones posibles pueden ser "any system that is capable of targeting and initiating the use of potentially lethal force without direct human supervision and direct human involvement in lethal decision-making" o "any weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention"; relacionada con esta última propuesta, también se ha considerado que puede caracterizarse como autónoma un arma que realice funciones críticas por medio de sensores, ordenadores y algoritmos. Más allá de las definiciones propuestas, respectivamente, por Peter Asaro, el Comité Internacional de la Cruz Roja (en adelante CICR) y el Presidente del período de sesiones de 2021 del Grupo de Expertos Gubernamentales sobre los SAAL debe tenerse presente que la inexistencia de un acuerdo básico sobre la definición provoca contradicciones. Por ejemplo, Alemania, Francia, Estados Unidos o China sólo consideran SAAL a los sistemas completamente autónomos *–fully autonomous systems–*, mientras que el Movimiento de Países No Alineados (en adelante MPNA) y el CICR incluyen de forma expresa o tácita los sistemas semiautomáticos, es decir, los SAAL que requieren la intervención de un operador en parte del proceso. También pueden darse contradicciones en la identificación de sistemas de armas como SAAL. Así, por ejemplo, Estados Unidos excluye los softwares autónomos diseñados para ciberoperaciones como parte de los SAAL, Francia realiza la misma salvedad respecto los sistemas automatizados de defensa de misiles o Rusia respecto las municiones inteligentes. Ver NACIONES UNIDAS, *Informe del Relator Especial para las ejecuciones extrajudiciales…, op. cit.*, párr. 38, p 8; CALVO PÉREZ, J. L., "Debate internacional en torno a los sistemas de armas autónomos letales. Consideraciones tecnológicas, jurídicas y éticas*", Revista general de marina*, vol. 278, nº 3, 2020, p. 459, armada.defensa.gob.es/archivo/rgm/2020/04/rgmabril20cap5.pdf; PETER, A., "On banning autonomous weapons systems: human rights, automation, and the dehumanization of lethal decision-making", *International Review of The Red Cross*, vol. 94, nº 886, 2012, p. 690, international-review.icrc.org/sites/default/files/irrc-886-asaro.pdf; ICRC, "Autonomous weapon systems implications of increasing autonomy in the critical functions of weapons", *Expert meeting report*, 2016, p. 71, icrcndresourcecentre.org/wp-content/uploads/2017/11/4283_002_Autonomus-Weapon-Systems_WEB.pdf; GEG, *Borrador de elementos sobre posibles recomendaciones consensuadas en relación con la aclaración, consideración y desarrollo de aspectos del marco normativo y operativo sobre tecnologías emergentes en el ámbito de los sistemas de armas autónomas letales*, 20 septiembre 2021, párr. 1.a, reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2021/gge/documents/chair-paper-september.pdf; Parte II del Glosario de la Directiva 3000.09, de 21 de noviembre de 2012, del Departamento de defensa de Estados Unidos, pp. 13-14, hsdl.org/?abstract&did=726163; FRANCIA y ALEMANIA, *Documento de trabajo de Francia y Alemania de 2017*, (CCW/GGE.1/2017/WP.4, 7 noviembre 2017), p. 2; RUSIA, *Documento de trabajo de Rusia de 2018*, (CCW/GGE.1/2018/WP.6, 4 abril 2018), pp. 1-3; CHINA, *Documento de trabajo de China de 2018*, (CCW/GGE.1/2018/WP.7, 11 abril 2018), párr. 3, p. 1; MPNA, *Documento de Trabajo del Movimiento de Países No Alineados de 2020*, (CCW/GGE.1/2020/WP.5, 14 septiembre 2020), párr. 34, p. 5 y REINO UNIDO, "Join Doctrine Publication 0-30.2. Unmanned Aircraft Systems, Development", *Concepts and Doctrine Centre*, 2017, p. 14, assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/673940/doctrine_uk_uas_jdp_0_30_2.pdf.

DOI: 10.17103/reei.42.08

Aunque la automatización de las armas viene de *larga data* –la mina terrestre, por ejemplo, sería un arma automatizada–, desde apenas 10 o 15 años, los Estados están desarrollando sistemas de armas cada vez más autónomos. En el imaginario colectivo, los SAAL suelen estar representados por el dron que realiza asesinatos selectivos[26], sin embargo, hay evidencias de que sus tareas pueden ser mucho más diversas. En este sentido, se han identificado sistemas de armas que llevan a cabo labores de vigilancia a modo de centinelas fronterizos o en instalaciones críticas y bases militares, tareas de reconocimiento, seguridad en puestos de control, patrullaje de toda clase, tareas de defensa aérea o antibuques, detección de armas biológicas o químicas, destrucción de explosivos improvisados, eliminación de minas y residuos, sistemas de control de dirección de proyectiles una vez lanzados, munición "inteligente" o softwares destinados a ciberataques[27]. La información disponible hace indicar que los Estados tienen la pretensión de hacer un uso intensivo de esta clase de sistemas, sin embargo, cabe enmarcar su despliegue y uso en un marco muy delimitado, en la medida en el que existe el acuerdo general de no renunciar al control humano de las tareas que implican "hacer la guerra" en los ámbitos estratégicos y operacionales, limitando su ámbito de actuación a la dimensión táctica del conflicto[28].

En cualquier caso, desde el inicio del desarrollo de los SAAL se planteó la necesidad de que existiera un control sobre los sistemas de armas que garantizara su correcto funcionamiento, el cumplimiento de las órdenes establecidas y, en general, el respeto de las normas establecidas por el Derecho Internacional, especialmente el Derecho

[26] La ausencia de un piloto ha inducido a la confusión, en la medida en la que estos aparatos están controlados a distancia por un operador, pero es una realidad que cada vez tienen más funciones automáticas. Entre las funciones autónomas, son comunes los despegues y aterrizajes, el mantenimiento de la órbita sobre una localización GPS, el vuelo a unas coordenadas GPS, los misiles guiados por láser o las bombas guiadas por GPS. En 2021, un Grupo de Expertos de Naciones Unidas hizo público lo que algunos han considerado el primer ataque conocido de un SAAL sin supervisión humana. El ataque fue realizado en 2020 por el dron-bomba turco, Kargu-2, en el contexto de un enfrentamiento entre fuerzas del gobierno libio y del militar Jalifa Hafter. Ver FROELICH, P., "Killer drone 'hunted down a human target' without being told to", *The New York Post*, 29 de mayo de 2021, nypost.com/2021/05/29/killer-drone-hunted-down-a-human-target-without-being-told-to/; NACIONES UNIDAS, *Informe final del Grupo de Expertos sobre Libia establecido en virtud de la resolución 1973 (2011) del Consejo de Seguridad*, (CS 2021/229, 8 marzo 2021), párr. 69, p. 20; GUTIÉRREZ ESPADA, C., CERVELL HORTAL, M. J., "Sistemas de armas autónomas…", *op. cit.*, pp. 40-41; PETER, A., "On banning…", *op. cit.*, p. 690 y REINO UNIDO, "Join Doctrine Publication 0-30.2...", *op. cit.*, p. 14.

[27] Se ha llegado a plantear su uso para la defensa de civiles o combatir grupos terroristas. Ver PETER, A., "On banning…", *op. cit.*, p. 690 y RUSIA, *Documento de trabajo de Rusia de 2019*, (CCW/GGE.1/2019/WP.1, 8 marzo 2019), párr. 3, p. 1.

[28] La mayor parte de fuerzas armadas coinciden en identificar tres niveles de mando y control. En primer lugar, el estratégico, que responde a la fijación de objetivos globales, dirección y coordinación general de las fuerzas, en función del objetivo político establecido por el decisor político. En segundo lugar, el operacional, en el que los objetivos globales son concretados a los teatros de operaciones previamente definidos. Finalmente, el táctico, dónde se planifican las operaciones militares concretas en las que se lleva a cabo el uso de la fuerza al que nos referimos. Ver EKELHOF, M., PERSI PAOLI, G., "El elemento humano en las decisiones sobre el uso de la fuerza", *UNIDIR – Instituto de las Naciones Unidas de investigación sobre el desarme*, 2020, pp. 2-3, unidir.org/publication/el-elemento-humano-en-las-decisiones-sobre-el-uso-de-la-fuerza.

DOI: 10.17103/reei.42.08

Internacional Humanitario[29]. Desde la sociedad civil organizada, Organizaciones internacionales y Estados no poseedores de esta tecnología se impulsó el debate y análisis de la cuestión, logrando la celebración de reuniones informales de expertos durante el periodo 2014-2016. En 2016 estas presiones culminaron en la constitución de un Grupo de Expertos Gubernamentales de composición abierta sobre los SAAL, en el marco de la Convención sobre ciertas armas convencionales, que lleva realizando sesiones anuales desde entonces[30].

Desde el principio de los debates del grupo de expertos, comenzó a fraguarse un consenso emergente sobre la prohibición de los sistemas de armas completamente autónomos –*fully autonomous weapon systems*– en los que no medie supervisión humana alguna[31], ya que se parte de la consideración de que sin intervención humana en la toma de decisiones de los SAAL no es posible respetar el Derecho internacional, concretamente las normas del DIH[32]. Esta posición de principio, con la consiguiente voluntad de prohibir los sistemas completamente autónomos, dio lugar a la cuestión que nos ocupa: determinar el tipo de control humano necesario para que los operadores mantengan el control sobre las

[29] Los SAAL no operan al margen del Derecho internacional; están regulados y sometidos, al menos, a tres ámbitos del Derecho internacional general. En primer lugar, el *ius ad bellum*. Esto es, la norma de *ius cogens* que prohíbe de la amenaza o uso de la fuerza, el derecho inmanente a la legitima defensa como excepción a esta prohibición y el sistema de seguridad colectiva que asegura el mantenimiento de la paz y la seguridad internacionales, así como su restablecimiento cuando ésta se rompe. En segundo lugar, el *ius in bello*, es decir, los principios y normas del DIH en contextos de conflictos armados y, en otros contextos, las obligaciones derivadas del Derecho Internacional de los Derechos Humanos. Finalmente, las normas internacionales que regulan la responsabilidad internacional de los Estados y las Organizaciones internacionales para los casos en los que se produzca un uso ilícito de estas armas de los que se derive esta clase de responsabilidad. Para un análisis más profundo del marco jurídico, ver GUTIÉRREZ ESPADA, C., CERVELL HORTAL, M. J., "Sistemas de armas autónomas…", *op. cit.*, pp. 33-51. También, ver GEG, *Informe del período de sesiones de 2019 del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales*, (CCW/GGE.1/2019/3, 8 noviembre 2019), anexo IV, p. 15, principio rector a y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párrs. 2.1, 2.4 y 2.20.

[30] Las presiones llevadas a cabo por diferentes actores culminaron en 2013 con la celebración de sesiones informales de expertos durante el periodo 2014-2016 y con la constitución del grupo de expertos en 2016, que ha venido realizando hasta la actualidad diferentes periodos de reuniones en los que se ha tratado aspectos vinculados con la SAAL, siendo su éxito más concreto el enunciamiento de unos principios rectores en 2019. Para profundizar en el debate internacional en torno a esta problemática, ver el artículo de la profesora Reyes Jiménez Segovia, participante en las reuniones del Grupo de Expertos Gubernamentales por la Universidad Pablo Olavide de Sevilla. Ver JIMÉNEZ-SEGOVIA, R., "Los sistemas de armas autónomos…", *op. cit.*, pp. 1-33. También, ver GUTIÉRREZ ESPADA, C., CERVELL HORTAL, M. J., "Sistemas de armas autónomas…", *op. cit.*, pp. 27-57. Para analizar el desarrollo del proceso y debates, consultar *Reaching Critical Will*, programa de desarme de la Liga Internacional de Mujeres por la Paz y la Libertad: reachingcriticalwill.org/disarmament-fora/ccw.

[31] GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 1.b.

[32] Si bien esta posición es mayoritaria, hay algunos Estados que defienden que los SAAL son, por su propia naturaleza, incompatibles con el DIH. Por el contrario, hay otros Estados que sostienen que es posible utilizar las tecnologías emergentes en el ámbito de los SAAL para garantizar la aplicación del DIH en la medida en la que se reducirían los errores y riesgos humanos, se mejoraría la precisión, así como el posible uso de estos sistemas para desactivar o destruir minas, explosivos, etc. Ver GEG, *Coincidencias entre las observaciones nacionales sobre los principios rectores*, (CCW/GGE.1/2020/WP.1, 20 octubre 2020), párr. 8, p. 2 y párr. 17, p. 4. Ver también y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 2.18.

DOI: 10.17103/reei.42.08

decisiones de la máquina, permitiendo con ello que el DIH sea observado[33]. Entre las propuestas realizadas con este fin, destacan el control humano determinante y eficaz, la participación humana sustantiva, la participación humana, el discernimiento humano apropiado o la supervisión humana[34]. De todas estas propuestas, el concepto Control Humano Significativo, acuñado por la ONG *Article 36*[35], es el que ha ido asumiéndose, con y sin variaciones terminológicas y de fondo, por todos los actores que participan en el estudio de estos sistemas[36]. Así, en 2018, el grupo de expertos recogió los diferentes términos utilizados hasta ese momento que pretenden garantizar el control humano sobre los SAAL[37]:

(Mantener) (Garantizar) (Ejercer) (Conservar)	un nivel	(sustantivo) (significativo) (apropiado) (suficiente) (mínimo) (mínimo indispensable)	De	(participación) (implicación) (responsabilidad) (supervisión) (validación) (control) (discernimiento) (decisión)	por parte del ser humano

Tabla 1. CCW/GGE.1/2018/3.

[33] GEG, *Resumen del Presidente del debate de 2019 del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales*, (CCW/GGE.1/2019/3/Add.1, 8 noviembre 2019), párr. 19, p. 4; GEG, *Informe del período de sesiones de 2019…, op. cit.*, principios rectores a, b y c; GEG, *Coincidencias entre las observaciones nacionales…*, *op. cit.*, párr. 8-9, pp. 2-3 y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 2.5.

[34] GEG, *Informe del período de sesiones de 2017 del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales*, (CCW/GGE.1/2017/3, 22 diciembre 2017), párr. 4, anexo II, pp. 7-12.

[35] En este sentido, cabe precisar que, si bien la ONG planteó este concepto respecto los ataques individuales, es también aplicable a los sistemas de armas, funciones críticas de los SAAL, la selección de objetivos y la decisión final sobre el uso de la fuerza. Ver ARTICLE 36, "Killer Robots: UK Government Policy on Fully Autonomous Weapons", 2013, pp. 1-5, article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf; UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward", *UNIDIR Resources*, nº2, 2014, p 2, unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf y EKELHOF, M., PERSI PAOLI, G., "El elemento humano…", *op. cit.*, p. 2.

[36] ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons systems. The need for meaningful human control", nº 97 AIV, nº 26 CAVV, 2015, p. 33, advisorycouncilinternationalaffairs.nl/documents/publications/2015/10/02/autonomous-weapon-systems; ARTICLE 36, "Killer Robots: UK…", *op. cit.*, pp. 1-5; UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control…", *op. cit.*, p. 2 y JIMÉNEZ-SEGOVIA, R., "Los sistemas de armas autónomos…", *op. cit.*, p. 24.

[37] GEG, *Informe del período de sesiones de 2018 del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales*, (CCW/GGE.1/2018/3, 23 octubre 2018), párr. 22, pp. 17-18.

DOI: 10.17103/reei.42.08

Esta profusión y creatividad terminológica denota que hasta ahora no se ha logrado alcanzar un consenso global sobre la definición del Control Humano Significativo[38], pero no hay duda de que se trata de una conceptualización "intuitivamente atractiva"[39]; tanto es así que ha sido calificado como "esencial"[40] por parte de los Estados que participan en el grupo que estudia los SAAL[41].

1. Fundamento y transformación del Control Humano Significativo

Como hemos apuntado, los miembros del Grupo de Expertos Gubernamentales han llegado al convencimiento, y así lo han plasmado en sus Principios Rectores, que los SAAL deben respetar el Derecho internacional, concretamente las normas del DIH[42]. Esta posición de principio es el fundamento de la centralidad del Control Humano Significativo en los debates sobre regulación de los SAAL, ya que se entiende que es el mecanismo que garantiza que los seres humanos sigamos manteniendo la responsabilidad sobre las decisiones que se adoptan en el uso y despliegue de estos sistemas de armas, permitiendo con ello que el DIH sea observado[43]. Así, se ha llegado a sostener que los SAAL que no sean capaces de actuar conforme a este ámbito del Derecho internacional son ilegales y deben prohibirse expresamente[44]; este planteamiento va en la línea de los

[38] ICRC, "Autonomous weapon systems technical, military, legal and humanitarian aspects", *Expert meeting report*, 2014, p. 11, icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014; ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, p. 32; EKELHOF, M., PERSI PAOLI, G., "El elemento humano…", *op. cit.*, p. 2 y REINO UNIDO, *Documento de trabajo de Reino Unido de 2018*, (CCW/GGE.2/2018/WP.1, 8 agosto 2018), párr. 7, p. 2.

[39] UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control…", *op. cit.*, p. 2.

[40] GEG, *Informe del período de sesiones de 2018…, op. cit.*, párr. 5, p. 12 y REINO UNDIO, *Documento de trabajo de Reino Unido de 2020*, (CCW/GGE.1/2020/WP.6, 11 noviembre 2020), párr. 1, p 2.

[41] Los actores que apuestan por el establecimiento de un tratado internacional sobre la materia han propuesto que el "principio" del control humano significativo sea adoptado como una obligación internacional. Ver FORO DE BERLÍN, *Documento de trabajo del Foro de Berlín de apoyo al Grupo de expertos gubernamentales sobre sistemas de armas autónomas letales de 2020*, (CCW/GGE.1/2020/WP.2, 20 junio 2020), párr. 33, p. 4. Ver también MPNA, *Documento de Trabajo del Movimiento de Países No Alineados de 2018*, (CCW/GGE.1/2018/WP.1, 28 marzo 2018), párr. 4, p. 1.

[42] La centralidad de este principio se ejemplifica con el consenso emergente sobre la obligación de los Estados de asegurarse que sus documentos, procedimientos, doctrinas y entrenamiento del personal involucrado en los SAAL es acorde al DIH. Ver punto j sobre los posibles elementos para las recomendaciones de consenso contenido en el GEG, *Resumen del Presidente del período de sesiones de 2020 del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales*, (CCW/GGE.1/2020/WP.7, 21 abril 2021), párr. 15, p. 6.

[43] GEG, *Resumen del Presidente del debate de 2019…, op. cit.*, párr. 19, p. 4; GEG, *Informe del período de sesiones de 2019…, op. cit.*, principios rectores a, b y c; GEG, *Coincidencias entre las observaciones nacionales…, op. cit.*, párr. 8-9, pp. 2-3; punto d y k sobre los posibles elementos para las recomendaciones de consenso contenido en el GEG, *Resumen del Presidente del período de sesiones de 2020…, op. cit.*, párr. 15, pp. 5-6 y párr. 27, p. 8 y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 2.12.

[44] Este sería el caso, entre otros, de los sistemas de armas completamente autónomos –*fully autonomous weapon systems*–. Ver GEG, *Coincidencias entre las observaciones nacionales…, op. cit.*, párr. 8, p. 2 y punto g sobre los posibles elementos para las recomendaciones de consenso contenido en el GEG, *Resumen del Presidente del período de sesiones de 2020…, op. cit.*, párr. 15, p. 6 y párr. 7, p. 4

DOI: 10.17103/reei.42.08

ámbitos de exclusión que ha ido desarrollando la Unión Europea para los sistemas de IA de toda clase de ámbitos[45].

De forma preliminar cabe señalar que, en un principio, se planteó de forma intuitiva que el control por parte del operador humano del SAAL debía darse en la fase más cercana a la selección y ataque del objetivo[46]. En este punto cabe precisar que debe diferenciarse entre las fases que componen el ciclo de vida del arma y la dimensión en la que se produce el proceso de adopción de decisiones para el uso de la fuerza en las operaciones militares, esto es, los niveles de mando y control[47]. Tal y como puede observarse en la tabla que se expone a continuación, las fases que componen el ciclo de vida de los SAAL son: dirección política en la fase de predesarrollo; fase de investigación y desarrollo; fase de ensayo, evaluación y certificación; fase de despliegue, capacitación, mando y control; fase de uso y aborto o interrupción del uso; y fase de evaluación posterior a la utilización. De acuerdo con la concepción originaria del Control Humano Significativo, éste debía tener lugar en el solapamiento entre el mando táctico –dentro del proceso de adopción de decisiones para el uso de la fuerza en las operaciones militares – y la fase de uso y aborto –dentro del ciclo de vida del SAAL–.

[45] La propuesta de la Comisión Europea, como hemos señalado, se fundamenta en la determinación de diferentes ámbitos de riesgo a los que cabe aplicar una supervisión humana más o menos gravada y la exclusión de la toma de decisiones automatizadas de la IA en las áreas que se tiene el convencimiento que la salud, seguridad o derechos fundamentales de las personas quedarían dañados o vulnerados. Nos referimos a sistemas de IA que utilicen técnicas subliminales trascendiendo la conciencia de las personas, sistemas que se aprovechen de vulnerabilidades de un grupo de personas debido a su edad o discapacidad, sistemas al servicio del Estado que se utilicen para el control social a través de la clasificación de las personas en función de diferentes parámetros –conducta social, características personales, personalidad conocida o predicha– y algunos sistemas de identificación biométrica remota en tiempo real en lugares de acceso público para la aplicación de la ley. Ver art. 5.1 de la Ley de Inteligencia Artificial.

[46] En cualquier caso, no cabe dudar de la relevancia de este momento del ciclo de vida de los SAAL. Es la fase en la que se deciden las reglas de enfrentamiento, se programan los objetivos o los parámetros de los mismos y se decide el despliegue de los SAAL. Tanto es así, que se incide en la importancia de que los Estados aseguren que el operador humano ejerce su juicio en ámbitos concretos como selección de perfiles, el marco temporal de la operación o marco de movimiento del SAAL. Ver ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, p. 37 y punto i sobre los posibles elementos para las recomendaciones de consenso contenido en el GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 15, p. 6.

[47] Estratégico, operacional y táctico, tal y como ya hemos explicado en la nota al pie 28.

DOI: 10.17103/reei.42.08

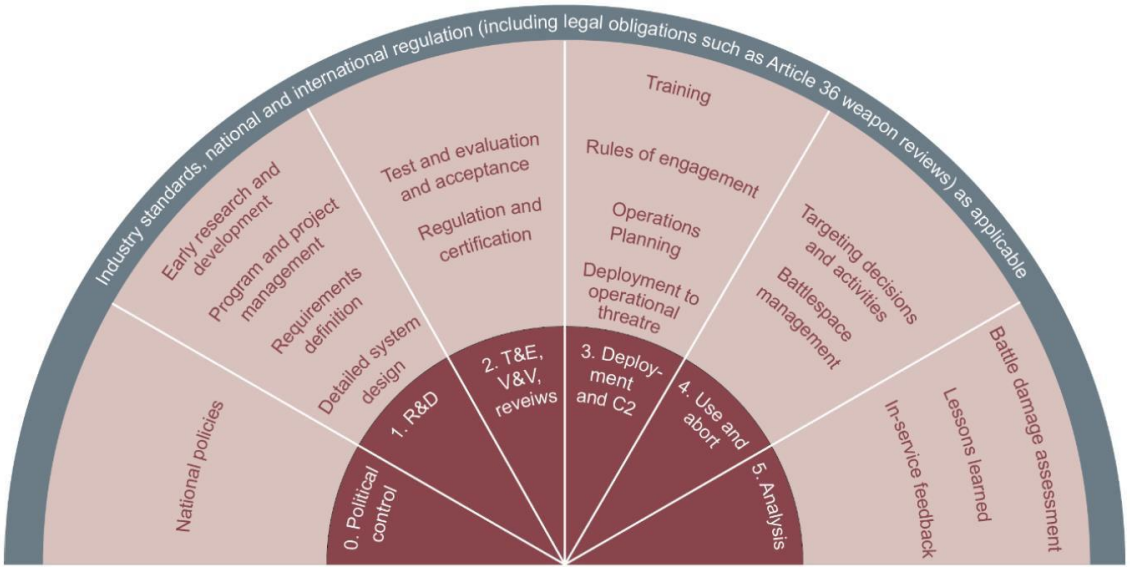

Tabla 2. CCW/GGE.1/2018/WP.8

Pronto se constató que la interacción humano-máquina se da durante todas las fases que componen el ciclo de vida del arma[48]. De esta manera, se superó una concepción originaria del Control Humano Significativo reducida a la fase de utilización e interrupción del uso por una concepción más integral y multidimensional. Desde nuestro punto de vista, esta evolución permite una doble lectura. Por un lado, la calidad y alcance del control humano que los operadores pueden llevar a cabo durante la fase más cercana a la selección y ataque del objetivo, dependen de factores que se producen tanto en distintas fases del ciclo del arma, como de distintas etapas dentro del proceso de adopción de decisiones para el uso de la fuerza en las operaciones militares. Tal y como desarrollamos a continuación, ello tiene un reflejo directo en los elementos que configuran el Control Humano Significativo. Por otro lado, en ese conjunto de factores que afectan a la calidad y alcance del control humano en la fase más cercana a la selección y ataque del objetivo, hay también una serie de interacciones humano-máquina que presentan diferentes manifestaciones, lo cual nos lleva a considerar el control humano como un proceso que va más allá del mero acto de supervisión al incluir capas de diseño y gobernanza en lo que supone tener un control efectivo[49].

[48] GEG, *Informe del período de sesiones de 2018…*, *op. cit.*, párr. 13, p. 15; GEG, *Informe del período de sesiones de 2019…, op. cit.*, principios rectores b y c y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párrs. 3.1 y 3.2.Ver también BODE, I., y WATTS, T., "Meaning-less human control: Lessons from air defence systems for lethal autonomous weapons", *Drone Wars UK – Center for War Studies*, febrero 2021, p. 17, dronewars.net/wp-content/uploads/2021/02/DW-Control-WEB.pdf.

[49] Entre otros, ver FORO DE BERLÍN, *Documento de trabajo del Foro de Berlín…*, *op. cit.*, párr. 9, p. 2; punto a sobre los posibles elementos para las recomendaciones de consenso contenido en el GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 31, p. 9 y párr. 28, p. 8 y GEG, *oincidencias entre las observaciones nacionales…*, *op. cit.*, conclusión C, párr. 21, p. 5

DOI: 10.17103/reei.42.08

2. Elementos fundamentales del Control Humano Significativo

El desarrollo del concepto del Control Humano Significativo se ha concretado en la identificación de una serie de elementos fundamentales cuya combinación permita mantener la responsabilidad humana sobre las decisiones que se adoptan en el uso y despliegue de estos sistemas de armas y, con ello, que el DIH sea debidamente observado.

En primer lugar, en lo que se refiere a la máquina o *software* y a sus características, se ha planteado que el Control Humano Significativo implica, en un primer punto, que sea técnicamente posible que el operador humano modifique parámetros en el despliegue[50]. Además, se considera que es necesario que durante el diseño del arma se tenga presente cual será el tipo de misión que lleve a cabo –vigilancia, reconocimiento, defensa, desactivación de explosivos o minas, ataque, etc.–, el medio en el que se va a desplegar –agua, tierra, aire, espacio, etc.–, la capacidad de movilidad del sistema –móvil, tipo de movilidad, fija en el espacio, etc.–, qué funciones serán automáticas y cuáles de uso manual, así como la clase de interacciones posibles entre la máquina y el operador[51]. Entender desde el inicio al sistema y al operador como un único sistema se presenta como una premisa fundamental para que pueda darse el Control Humano Significativo[52]. Posiblemente por ello hay quien ha apuntado que la consideración del humano y la maquina como parte del mismo sistema remite inevitablemente al paradigma *human in the loop*[53]. Desde nuestro punto de vista, dicha premisa sí permitiría afirmar que las armas capaces de seleccionar los objetivos y de utilizar la fuerza sin ninguna aportación o intervención humana, *human out of the loop*, deberían considerarse incompatibles con el requisito de control humano significativo[54].

El aborto o interrupción del uso es la última salvaguarda que, sin lugar a duda, se configura como un elemento nuclear del Control Humano Significativo[55]. Existe un

[50] FORO DE BERLÍN, *Documento de trabajo del Foro de Berlín…*, *op. cit.*, párr. 17, p. 3.

[51] En este sentido, por ejemplo, se ha sostenido que un posible consenso puede ser que los SAAL que no puedan actuar de forma fiable o predecible acorde a las intenciones del operador humano de cumplir las normas de DIH es inherentemente ilegal. Ver punto f sobre los posibles elementos para las recomendaciones de consenso contenido en el GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 15, p. 6. Ver también GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 29, pp. 8-9.

[52] GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 28, p. 8; REINO UNIDO, *Documento de trabajo de Reino Unido de 2018*, *op. cit.*, 11, p. 3 y anexo I, p. 7; FORO DE BERLÍN, *Documento de trabajo del Foro de Berlín…*, *op. cit.*, párr. 9, p. 2; JIMÉNEZ-SEGOVIA, R., "Los sistemas de armas autónomos…", *op. cit.*, p. 24; ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, p. 35 y EKELHOF, M., PERSI PAOLI, G., "El elemento humano…", *op. cit.*, p. 3.

[53] PETER, A., "On banning…*", op. cit.*, p. 964.

[54] Esta es la posición expresada por España. Ver comentarios de España a los 11 principios rectores en Anexo del GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 16 p. 83. Sobre cómo estas armas son, además, una vulneración del derecho positivo, ver BONTRIDDER, N., "Les armes létales autonomes et le droit international humanitaire: le nécessaire contrôle humain sur l'usage de la force", en Hervé Jacquenim (coord.), *Time to reshape the digital society. 40th anniversary of the CRIDS*, 1ª Edición, Bruselas: Larcier (Lefebvre Sarrut Group), 2021. pp. 407-440.

[55] GEG, *Informe del período de sesiones de 2018…*, *op. cit.*, párr. 19, p. 16 y FORO DE BERLÍN, *Documento de trabajo del Foro de Berlín…*, *op. cit.*, párr. 17, p. 3.

DOI: 10.17103/reei.42.08

acuerdo general en que el operador debe poder paralizar o detener por completo el proceso en todo momento a criterio exclusivo del mismo, sin que el sistema pueda impedírselo de ningún modo[56].

La justificación de esta prerrogativa la encontramos ante una posible falta de información necesaria para llevar a cabo la tarea, la activación del sistema sin tiempo para una reflexión sosegada –por ejemplo, en caso de los SAAL de carácter antiaéreo–, o la constatación de que la calidad y alcance de la intervención humana se va reduciendo su significancia a medida que se acerca el momento de realizar la acción.

Además de la posibilidad técnica de que permita el aborto de la tarea, el propio sistema debe contener salvaguardas preestablecidas para el caso en el que detecte errores o un mal funcionamiento[57]. Un ejemplo práctico de esto lo encontramos en el dron autónomo israelí Mini-Harpy: ante un ataque contra un vehículo en el que alrededor hay civiles, el sistema suspende el ataque. Aunque en este caso parece pertinente que el SAAL suspenda el ataque, plantea la cuestión sobre si esto debe ser posible y en qué casos pueden estos sistemas invalidar las órdenes de los operadores.

En cualquier caso, cabe recordar que las características y capacidades del sistema en su conjunto ha sido identificado por el Grupo de Expertos Gubernamentales como uno de los factores que determina la calidad y alcance del Control Humano Significativo [58].

En segundo lugar, en lo referente al operador humano, se requiere que éste tenga una serie de conocimientos de carácter jurídico[59], ético[60], técnico[61] y específico en función del ámbito en el que opere el sistema[62]. El conocimiento del marco jurídico se vuelve fundamental para posibilitar que el operador o usuario pueda analizar si el funcionamiento del arma es acorde al DIH, especialmente los principios de distinción, proporcionalidad

---

[56] GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 29, pp. 8-9 y SHARKEY, N., "Towards a…", *op. cit.*, p. 2

[57] REINO UNIDO, *Documento de trabajo de Reino Unido de 2018*, *op. cit.*, anexo I, pp. 11-12.

[58] GEG, *Informe del período de sesiones de 2019…, op. cit.*, principio rector c.

[59] GEG, *Coincidencias entre las observaciones nacionales…*, *op. cit.*, conclusión D, párr. 21, p. 5; punto j sobre los posibles elementos para las recomendaciones de consenso contenido en el GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 15, p. 6; ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, p. 37 y REINO UNIDO, *Documento de trabajo de Reino Unido de 2018*, *op. cit.*, anexo I, pp. 9-10.

[60] En el ámbito militar, es una cuestión esencial a todos los niveles. Ver GEG, *Coincidencias entre las observaciones nacionales…*, *op. cit.*, conclusión D, párr. 21, p. 5 y CALVO PÉREZ, J. L., "Debate internacional en torno a los sistemas…", *op. cit.*, p. 458.

[61] GEG, *Coincidencias entre las observaciones nacionales…*, *op. cit.*, conclusión D, párr. 21, p. 5; REINO UNIDO, *Documento de trabajo de Reino Unido de 2018*, *op. cit.*, anexo I, p. 8 y CALVO PÉREZ, J. L., "Debate internacional en torno a los sistemas…", *op. cit.*, p. 458.

[62] Por ejemplo, el operador de un SAAL que vaya a realizar un ataque debe conocer las reglas de enfrentamiento, esto es, las ordenes de mando que imponen restricciones al empleo de la fuerza. Ver CALVO PÉREZ, J. L., "Debate internacional en torno a los sistemas…", *op. cit.*, p. 458; PETER, A., "On banning…*", op. cit.*, p. 696; REINO UNIDO, *Documento de trabajo de Reino Unido de 2018*, *op. cit.*, anexo I, p. 9 y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 2.17.b.

DOI: 10.17103/reei.42.08

y precaución[63]; valoración que siempre será previa a la ejecución de la tarea. Respecto al conocimiento técnico, se exige que el operador conozca las especificaciones técnicas, las funciones generales y críticas[64], los efectos, las limitaciones, así como la predictibilidad y fiabilidad[65] del sistema. Estos conocimientos, junto con otros elementos, es lo que le permitirá alcanzar la conciencia contextual y situacional, a la que nos referiremos después.

En tercer lugar, nos encontramos con la información; otro de los elementos necesarios para que el operador posea conciencia contextual y situacional. El operador humano debe contar con información "suficiente y fiable"[66] para poder realizar una valoración jurídica y ética correcta, conocer el adecuado o erróneo funcionamiento del aparato en ese momento, así como el desarrollo de la misión. Con esta información, podrá deliberar y decidir no realizar ninguna intervención o modificar los parámetros del proceso, pudiendo llegar a anularlo. En este sentido, la interfaz humano-maquina es importante a efectos de que la información presentada sea relevante y necesaria, así como se haga de forma organizada y rápida[67].

En cuanto al origen de la información, no se aprecian inconvenientes en que pueda provenir del mismo aparato o se apoye en otras fuentes como pueden ser los propios operadores, otras personas u otros sistemas de vigilancia o de recolección de datos[68]. De hecho, en el caso del uso de sistemas de armas tradicionales, ya sean manuales o automatizados, no se exige tener toda la información, por lo que exigir ese grado de conocimiento a los SAAL resultaría desmesurado[69]. En otras palabras, en la medida en la que la confianza en la información aportada por el arma es un elemento muy arraigado en

[63] ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, pp. 35-36 y PETER, A., "On banning…*", op. cit.*, p. 695.

[64] No existe un acuerdo general sobre el contenido de las funciones críticas de los SAAL. El CICR entienden que entrañan tareas de obtención, rastreo, selección y ataque de objetivos. Ver ICRC, "Autonomous weapon systems technical…", *op. cit.*, p. 11. Ver también BODE, I., y WATTS, T., "Meaning-less human control…", *op. cit.*, pp. 11-12; GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 29-30, pp. 8-9 y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 3.4.b.

[65] UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control…", *op. cit.*, p. 5; GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 27, p. 8 y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 2.17.a.

[66] GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 30, p. 9; GEG, *Informe del período de sesiones de 2018…*, *op. cit.*, párr. 18, p. 16; GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 3.4.b y REINO UNIDO, *Documento de trabajo de Reino Unido de 2018*, *op. cit.*, anexo I, p. 10.

[67] ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, pp. 34-36; BODE, I., y WATTS, T., "Meaning-less human control…", *op. cit.*, p. 17 y PETER, A., "On banning…*", op. cit.*, p. 695.

[68] ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, p. 34.

[69] ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, p. 32.

DOI: 10.17103/reei.42.08

las formas de hacer la guerra, no resulta razonable exigirle a los SAAL que aporten todos los datos necesarios[70].

Dentro de la información que pueda considerarse suficiente y fiable a la hora de tomar decisiones en el uso y despliegue, existen ciertos factores exógenos que pueden afectar al grado de Control Humano Significativo. El más importante de ellos se refiere al contexto operacional[71], esto es, si se trata de un espacio estático y ordenado, como el espacio aéreo, o un espacio dinámico y complejo, como un entorno urbano, así como la amplitud y dimensión del mismo[72]. Junto a este elemento nos encontramos con las situaciones circunstanciales, como pueden ser las condiciones meteorológicas[73].

Además de la información disponible, el tiempo se presenta como otro de los factores a tener en cuenta. El operador debe tener el tiempo suficiente para poder reflexionar y tomar la decisión de intervenir o no hacerlo[74]. Es por ello por lo que el sistema debe permitir técnicamente que el operador pueda llevar a cabo un juicio sobre el proceso a fin de determinar si es necesaria una intervención. Hay quien defiende que en ese tiempo el operador debe poder deliberar sobre la naturaleza del objetivo, la necesidad y la adecuación del ataque, así como realizar una valoración de la probabilidad y posibles incidentes y accidentes durante la ejecución de la labor encomendada al SAAL[75].

Si el SAAL dispone de las características técnicas exigidas, se presenta la información de la forma adecuada, el operador dispone de la capacidad cognitiva y conocimientos necesarios, así como el tiempo suficiente para deliberar, es posible alcanzar lo que se ha venido a llamar conciencia contextual y situacional o capacidad operativa o contextual. Se trata del estado en el que el operador percibe y puede reaccionar ante una disfunción del plan prestablecido o imprevisto sobrevenido[76], mediante una decisión "informada y consciente"[77]. En esa situación el operador podrá modificar los parámetros que considere para el cumplimiento de la tarea encomendada y si esto no es posible, su suspensión.

---

[70] En este sentido, ver GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 2.6.
[71] GEG, *Informe del período de sesiones de 2019…, op. cit.*, principio rector c; GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 29, pp. 8-9 y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párr. 3.4.b.
[72] JIMÉNEZ-SEGOVIA, R., "Los sistemas de armas autónomos…", *op. cit.*, pp. 24-25.
[73] REINO UNIDO, *Documento de trabajo de Reino Unido de 2018*, *op. cit.*, anexo I, p. 9.
[74] El Presidente del período de sesiones de 2021 del Grupo de Expertos Gubernamentales sobre los SAAL sostuvo que debía tratarse de una deliberación realizada de buena fe y una decisión informada. Ver y GEG, *Borrador de elementos sobre posibles recomendaciones…, op. cit.*, párrs. 2.6 y 3.4.a, y PETER, A., "On banning…*", op. cit.*, p. 695.
[75] SHARKEY, N., "Towards a…", *op. cit.*, p. 2.
[76] GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, párr. 30, p. 9; GEG, *Informe del período de sesiones de 2018…*, *op. cit.*, párr. 18, p. 16; CALVO PÉREZ, J. L., "Debate internacional en torno a los sistemas…", *op. cit.*, p. 461; ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, p. 35; SHARKEY, N., "Towards a…", *op. cit.*, p. 2 y UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control…", *op. cit.*, p. 5.
[77] ADVISORY COUNCIL ON INTERNATIONAL AFFAIRS, ADVISORY COMMITTEE ON ISSUES OF PUBLIC INTERNATIONAL LAW, "Autonomous weapons…", *op. cit.*, p. 34.

DOI: 10.17103/reei.42.08

En definitiva, la concurrencia de todos los elementos expuestos posibilitaría que el funcionamiento del SAAL sea conforme a las normas del DIH y del Derecho internacional en general, fundamento de esta conceptualización.

IV. LA SUPERVISIÓN HUMANA DE LOS SISTEMAS DE IA A LA LUZ DEL CONTROL HUMANO SIGNIFICATIVO DE LOS SAAL, ¿UN CONCEPTO UNIVERSALIZABLE?

Las aportaciones extraídas de los debates en torno al Control Humano Significativo sobre los SAAL nos han permitido esbozar el fundamento y elementos que configuran este concepto en el ámbito del Derecho Internacional Humanitario. Tal y como señalábamos al final de la segunda sección, se trata de un aportación de carácter cualificado a la hora de garantizar la supervisión humana en los sistemas autónomos[78]. Por su parte, la Comisión Europea en su propuesta recoge que la supervisión humana sea un requisito de obligado cumplimiento para los sistemas de IA de alto riesgo, de manera que dichos sistemas deban ser diseñados y desarrollados con el fin de que pueda darse una supervisión efectiva[79]. En esta sección exponemos la propuesta de la Comisión a la luz de los elementos extraídos en la sección tercera y tratamos de determinar si el concepto de Control Humano Significativo puede ser de utilidad para garantizar la supervisión humana de los sistemas autónomos en otros ámbitos del Derecho.

Para realizar esta comparativa es preciso, eso sí, aclarar los diferentes estadios en los que encontramos el desarrollo jurídico de estos dos conceptos. Por un lado, en lo que concierne al Control Humano Significativo, hemos visto que se está trabajando en alcanzar consensos en lo que algunos Estados han denominado el grado "mínimo indispensable de control humano"[80] sobre los que elaborar medidas normativas concretas, sin que esto último haya llegado a materializarse. Por otro lado, como ya hemos señalado, la supervisión humana como requisito obligatorio para los sistemas de IA de alto riesgo sí se ha materializado en una propuesta normativa europea concreta que establece obligaciones para los proveedores en la fase de diseño y desarrollo[81], debiendo los usuarios utilizar estos sistemas conforme a las instrucciones aportadas por los proveedores en la fase de uso y despliegue[82]. En el momento de redacción de este artículo la propuesta de Reglamento de la Comisión Europea está en fase de primera lectura.

[78] Ver nota al pie 22.
[79] Art. 14.1 de la Ley de Inteligencia Artificial.
[80] GEG, *Informe del período de sesiones de 2018…*, *op. cit.*, párr. 19, p. 16.
[81] Art. 16 de la Ley de Inteligencia Artificial.
[82] Art. 29 de la Ley de Inteligencia Artificial.

DOI: 10.17103/reei.42.08

1. Fundamento de la supervisión humana

En primer lugar, hemos de reparar en el fundamento aducido para la introducción del control o supervisión humanos en el ordenamiento jurídico. Tal y como hemos expuesto anteriormente, el Grupo de Expertos sobre los SAAL entiende que los sistemas autónomos deben cumplir con las disposiciones de Derecho internacional, concretamente las normas de DIH, y para ello, es necesario introducir el Control Humano Significativo, garantizando que los seres humanos sigamos manteniendo la responsabilidad sobre las decisiones que se adoptan en el uso y despliegue de estos sistemas[83]. En su propuesta, la Comisión considera que las tecnologías sirvan a las personas, previniendo o minimizando el riesgo para la salud, la seguridad o los derechos fundamentales[84]. De manera que, en función de la finalidad del sistema de IA y, por ende, en función de la actividad que realice, el ámbito de la vida de las personas que quede afectado o en la medida de que incida en bienes jurídicos protegidos, existirá un marco jurídico concreto y determinado que el sistema de la IA debe respetar. Como ya hemos señalado, garantizar el cumplimiento de este marco que define los bienes jurídicos que han de protegerse en cada ámbito será el fundamento bajo el que se introduce la supervisión humana[85].

No obstante, este fundamento ha sido objeto de críticas en la doctrina. A medida que los sistemas automatizados evolucionan y se hacen más eficaces, se produce una creciente delegación de tareas acompañada de una mayor dependencia en la propia delegación más que en la supervisión del sistema[86]. Ante este paradigma, parte de la doctrina ha puesto de manifiesto que introducir seres humanos en la toma de decisiones podría empeorar los

[83] Cabe precisar que, mantener la responsabilidad sobre las decisiones autónomas a través de mecanismos de supervisión humana como el Control Humano Significativo, no implica que la responsabilidad jurídica del uso y despliegue de dichos sistemas recaiga, necesariamente, sobre el agente humano al que se encomienda dicha supervisión. Es decir, mantener la responsabilidad humana sobre el cumplimiento del ordenamiento jurídico a través de la supervisión humana como fundamento es independiente del régimen de responsabilidad jurídica.

[84] En este sentido resulta pertinente recordar que el DIH también contienen normas que pretenden proteger la salud, la vida y, en general, la seguridad de las personas. En esencia, las dos regulaciones pretenden salvaguardar la dignidad humana. Ver art. 7 de la Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas, contenida como Anexo de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre; Considerandos 2 y 10 de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre; punto A.III.I de las Recomendaciones detalladas respeto a la propuesta del Parlamento Europeo, contenida como como Anexo de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre; art. 14.2 de la Ley de Inteligencia Artificial; punto 4 de los Comentarios de China a los 11 Principios Rectores en Anexo del GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, p. 32; punto 8 de los Comentarios de Colombia a los 11 Principios Rectores en Anexo del GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, p. 33; Punto 1.a.v de los Comentarios de Mauricio a los 11 Principios Rectores en Anexo del GEG, *Resumen del Presidente del período de sesiones de 2020…*, *op. cit.*, p. 58, y GEG, *Informe del período de sesiones de 2019…, op. cit.*, principio rector c.

[85] GEG, *Informe del período de sesiones de 2019…, op. cit.*, principio rector c; considerandos 68 y 69 de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre, y considerando 10 de la Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas, contenida como Anexo de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre.

[86] TADDEO, M.R., "Trusting Digital Technologies Correctly.", *Minds and Machines*, vol. 27, nº 4, p. 566.

DOI: 10.17103/reei.42.08

resultados inicialmente arrojados por esta clase de sistemas autónomos[87], o que la cuestión de la justicia en la toma de decisiones debería ser separada de cualquier clase de derecho a la intervención o supervisión humana[88]. A nuestro juicio, estas críticas ponen de relieve que la supervisión humana se limita y transforma con la evolución de los sistemas autónomos, y lo seguirá haciendo. De ahí la evolución que hemos constatado en el concepto de Control Humano Significativo, como un proceso que va más allá del mero acto de supervisión al incluir capas de diseño y gobernanza en lo que supone tener un control efectivo. Por ello, entendemos que la evolución de los sistemas autónomos continuará transformando la forma en la que hacer efectiva la supervisión humana, lo cual no implica que esta no pueda responder al fundamento que motiva su inclusión en el ordenamiento jurídico.

2. Supervisión humana basada en el riesgo

Hemos podido observar que la supervisión humana no se exige para el uso y despliegue de cualquier tipo de sistema autónomo, sino para aquellos que representan un alto riesgo para la protección de los bienes jurídicos definidos por el ordenamiento jurídico en cuestión[89]. En la sección segunda mencionábamos que este enfoque basado en el riesgo ha sido un mínimo común denominador en las distintas propuestas de la Comisión y el Parlamento, de forma que la supervisión humana como requisito obligatorio queda relegada a aquellos ámbitos y usos que entran dentro de la consideración de alto riesgo[90]. En lo que concierne al Control Humano Significativo en el DIH, no puede obviarse que el debate en torno a este concepto está vinculado con el uso y despliegue de las tecnologías

[87] ALMADA, M. "Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems.", en *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL, 2019, p. 5.

[88] HUQ, A. Z., "A Right to a…", *op cit.*, p. 656.

[89] El Parlamento definió el alto riego como "riesgo significativo, derivado del desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas, de causar lesiones o daños a las personas o a la sociedad, vulnerando los derechos fundamentales y las normas de seguridad establecidas en el Derecho de la Unión, teniendo en cuenta su uso o finalidad específicos, el sector en el que se desarrollan, despliegan o usan y la gravedad de las lesiones o daños que cabe esperar que se produzcan". Ver art. 4 de la Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas, contenida como Anexo de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre.

[90] A pesar de que la propuesta del Parlamento pretendía ser aplicable también al uso de los SAAL, este ámbito y uso de los sistemas de IA no fue finalmente incluido en el texto de la Comisión. En dicha propuesta del Parlamento se consideró que los ámbitos que requieren una especial protección son el "empleo, educación, asistencia sanitaria, transporte, energía, sector público (asilo, migración, controles fronterizos, sistema judicial y servicios de seguridad social), seguridad y defensa, finanzas, bancos y seguros". Por otro lado, que los usos que implican un alto riesgo abarcan actividades como la "contratación, clasificación y evaluación de estudiantes, asignación de fondos públicos, concesión de préstamos, comercio, corretaje, fiscalidad, tratamientos y procedimientos médicos, procesos electorales y campañas políticas, conducción automatizada, gestión del tráfico, producción y distribución de energía, gestión de residuos, control de emisiones y sistemas militares autónomos", así como las "decisiones del sector público que tienen un impacto significativo y directo en los derechos y las obligaciones de las personas físicas o jurídicas". Ver Anexo de la Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas, contenida como Anexo de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre.

DOI: 10.17103/reei.42.08

emergentes en el ámbito de los sistemas de armas autónomos letales y no de cualquier sistema autónomo que pueda utilizarse en el ámbito militar, lo cual pone de manifiesto la aplicación de un enfoque basado en el riesgo también para este concepto.

Incluso, esta graduación del riesgo puede implicar que, dentro de los sistemas considerados de alto riesgo, determinados ámbitos y usos puedan requerir de una supervisión humana agravada[91], en la medida en la que es la única vía mediante la cual se pueden salvaguardar los derechos fundamentales. Asimismo, en los casos en los que el funcionamiento del sistema, ya sea un SAAL u otro tipo de IA, impida el cumplimiento del marco regulador protector de derechos es razonable prohibir la implementación de la toma de decisiones automatizadas[92]. Sin embargo, en los casos en los que es posible la observancia del marco jurídico determinado creemos factible la identificación de los elementos mínimos requeridos para la supervisión humana, independientemente del ámbito o uso y de si este es de alto riesgo o presenta otro nivel de riesgo.

3. La supervisión humana a lo largo del ciclo de vida de la IA

De forma preliminar, cabe reiterar que la supervisión humana, independientemente del sistema ante el que nos encontremos, debe garantizarse durante todo el ciclo de vida del mismo. En este sentido, es destacable la superación de la concepción originaria mencionada en la sección tercera, donde los debates sobre el Control Humano Significativo sobre los SAAL comenzaron centrándose en la intervención humana en la fase de utilización e interrupción del uso, hasta reconocer que el control humano debía tenerse en cuenta durante todo el ciclo de vida del sistema de armas[93]. Ello implica que, que en función de la fase en la que se encuentre el sistema, los elementos a los que nos referiremos a continuación se manifestarán de forma diferente[94]. En cuanto a la supervisión humana y los sistemas de IA de alto riesgo, tal y como recogemos en la sección segunda, el Parlamento optó por la inclusión de la supervisión humana integral como principio ético de obligado cumplimiento para las fases de desarrollo, uso y despliegue de estas tecnologías. Sin embargo, el enfoque de la Ley de Inteligencia Artificial es algo más limitada. Conforme a lo mencionado arriba, la Comisión establece obligaciones sobre los proveedores en las fases de diseño y desarrollo, de manera que los

[91] La propuesta legislativa de la Comisión Europea exige la intervención de dos personas en la decisión, recomendación o información de sistemas de IA dedicados a la identificación biométrica remota en tiempo real o posterior de personas físicas. Ver art. 14.5 de la Ley de Inteligencia Artificial y punto 1.a del Anexo III de la Ley de Inteligencia Artificial.

[92] En este sentido, en el marco de los debates en torno a los SAAL, los Estados afirmaron que el DIH se aplica de forma plena a esta clase de armas, teniendo que determinar si su uso, en algún o todo momento, contraviene el Derecho internacional. Así, la utilización de SAAL que no cumplan con las normas del DIH se pueden considerar ilegales. En el contexto europeo, el Parlamento Europeo ha afirmado que debe excluirse las decisiones automatizadas de determinados sectores como en el sector público cuando tengan un impacto directo y significativo en los derechos y obligaciones de los ciudadanos o en la aplicación de la ley como es el caso de la emisión de sentencias, ámbito que jamás debería sustraerse del control humano Ver GEG, *Informe del período de sesiones de 2019…, op. cit.*, principios rectores a y e, y considerandos 70 y 71 de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre.

[93] GEG, *Informe del período de sesiones de 2019…, op. cit.*, principios rectores b y c.

[94] GEG, *Coincidencias entre las observaciones nacionales…*, *op. cit.*, conclusión C, párr. 21, p. 5 y *Documento de trabajo de Reino Unido de 2020, op. cit.*, párr. 13, p 4.

DOI: 10.17103/reei.42.08

sistemas de IA deban poder ser supervisados de forma efectiva por los usuarios del sistema durante la fase de uso y despliegue.

Esto implica que la propuesta de la Comisión Europea no establece obligaciones jurídicas concretas para la supervisión humana en la fase de uso y despliegue –más allá de seguir las instrucciones del proveedor–. Ahora bien, es posible que para determinados usos y despliegues de sistemas de IA el "usuario" del sistema[95] esté obligado por otra norma a introducir un supervisor humano en esta fase bajo determinadas particularidades. Este sería el caso de los mecanismos de gobernanza mencionados en la sección segunda[96], así, por ejemplo, el RGPD obliga en virtud del artículo 22.1 a la introducción de intervención humana por parte del responsable del tratamiento para las decisiones que produzcan un efecto jurídico o afecten de forma significativa a la persona interesada. Lo cual implica que en la fase de uso y despliegue de los sistemas de IA con tratamiento de datos personales, el Reglamento obliga a introducir un operador humano para tomar determinadas decisiones[97].

Por lo tanto, consideramos que el modelo de gobernanza que adopta la Ley de Inteligencia Artificial, estableciendo obligaciones para los proveedores en las fases de diseño y desarrollo para garantizar una supervisión efectiva en la fase de uso y despliegue de los sistemas, reafirma esa visión de la supervisión humana durante todo el ciclo de vida de un sistema que hemos constatado tanto en los documentos del Grupo de Expertos sobre los SAAL como en los precedentes de esta última propuesta de la Comisión para regular los sistemas de IA.

4. Vertiente cualitativa de la supervisión humana: elementos mínimos a considerar

Finalmente, hemos podido constatar que la supervisión humana no es, simplemente, la introducción de un operador humano en la toma de decisiones con autoridad final sobre el sistema autónomo[98]. El mismo concepto de Control Humano Significativo hace referencia al aspecto cualitativo del control, esto es, su carácter significativo. Más arriba hemos podido constatar cómo en el seno del Grupo de Expertos sobre los SAAL se han tratado de identificar aquellos elementos que posibilitan llevar a cabo esta clase de control

[95] En la línea de lo establecido por la Ley de Inteligencia Artificial, el usuario del sistema será quien lo utiliza bajo su autoridad en la fase de uso y despliegue del mismo. Esto es, "toda persona física o jurídica, autoridad pública, agencia u organismo de otra índole que utilice un sistema de IA bajo su propia autoridad, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional". Ver art. 3.4 de la Ley de Inteligencia Artificial.

[96] A saber, art. 11 de la Directiva (UE) 2016/680 relativa al tratamiento de datos personales por parte de las autoridades para fines de prevención, investigación, detección o enjuiciamiento penales, de 27 de abril de 2016; art. 7.6 de la Directiva (UE) 2016/681 relativa a la utilización de datos del registro de nombres de los pasajeros, de 27 abril de 2016, y art. 22 del Reglamento (UE) 2016/679 General de Protección de Datos, de 27 de abril de 2016.

[97] En otros casos, es posible que el uso y despliegue de sistemas de IA considerados de alto riesgo no estén obligados por ninguna norma a introducir mecanismos de supervisión humana, eso sí, en cualquier caso, los sistemas serán diseñados y desarrollados para ser supervisados de forma efectiva, estando los usuarios obligados a utilizar los sistemas conforme a las instrucciones de los proveedores.

[98] Lo cual podríamos definir como una visión formalista del *human in the loop*.

DOI: 10.17103/reei.42.08

humano. Esta vertiente cualitativa está también presente en esa visión regulatoria de las instituciones europeas, entendiendo igualmente que la supervisión humana no es un concepto meramente formal. Así, el Parlamento Europeo propuso una supervisión humana significativa con independencia de la forma concreta que adoptase dicha supervisión: revisión, evaluación, intervención o control humanos[99]. Aunque la Comisión no adopta el término "significativo" en la propuesta de Reglamento "Ley de Inteligencia Artificial", indica que el diseño y desarrollo de los sistemas debe permitir una supervisión "efectiva".

En el marco de la propuesta de la Comisión, dicha supervisión efectiva se define a partir de una serie de capacidades que el operador humano debe ser capaz de desplegar, en función de las circunstancias, durante la fase de uso y despliegue del sistema de IA. Es decir, los proveedores deben garantizar[100] que, en la fase de uso y despliegue, los operadores humanos pueden entender por completo las capacidades y limitaciones del sistema, ser conscientes del sesgo de automatización, interpretar correctamente la información de salida del sistema, desestimar, invalidar o revertir dicha información o interrumpir el sistema accionando un botón específicamente destinado a tal fin, entre otros[101].

Desde nuestro punto de vista, en base al análisis realizado en la sección tercera, dicha supervisión efectiva definida a partir de las capacidades del operador humano para entender, interpretar e intervenir en el funcionamiento del sistema conlleva, inevitablemente, la integración de unos elementos mínimos en las distintas medidas que el proveedor deberá adoptar. En este punto, entendemos que los elementos extraídos de los debates del grupo de expertos en relación con el Control Humano Significativo de los SAAL, pueden ser de utilidad a la hora de considerar qué elementos deben integrar los proveedores. Así, será necesario integrar elementos que hemos sistematizado en tres grandes bloques.

---

[99] Considerando 10 De la Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas, contenida como Anexo de la Resolución 2020/2012(INL) del Parlamento Europeo, de 20 de octubre.

[100] Definiendo e integrando, cuando sea técnicamente posible, las medidas que sean adecuadas. Ver art. 14.3 de la Ley de Inteligencia Artificial.

[101] "Las medidas mencionadas en el apartado 3 permitirán que las personas a quienes se encomiende la vigilancia humana puedan, en función de las circunstancias: a) entender por completo las capacidades y limitaciones del sistema de IA de alto riesgo y controlar debidamente su funcionamiento, de modo que puedan detectar indicios de anomalías, problemas de funcionamiento y comportamientos inesperados y ponerles solución lo antes posible; b) ser conscientes de la posible tendencia a confiar automáticamente o en exceso en la información de salida generada por un sistema de IA de alto riesgo («sesgo de automatización»), en particular con aquellos sistemas que se utilizan para aportar información o recomendaciones con el fin de que personas físicas adopten una decisión; c) interpretar correctamente la información de salida del sistema de IA de alto riesgo, teniendo en cuenta en particular las características del sistema y las herramientas y los métodos de interpretación disponibles; d) decidir, en cualquier situación concreta, no utilizar el sistema de IA de alto riesgo o desestimar, invalidar o revertir la información de salida que este genere; e) intervenir en el funcionamiento del sistema de IA de alto riesgo o interrumpir el sistema accionando un botón específicamente destinado a tal fin o mediante un procedimiento similar." Art. 14.4 de la Ley de Inteligencia Artificial.

DOI: 10.17103/reei.42.08

En primer lugar, habrán de tenerse en cuenta elementos referidos a la máquina o *software*. Anteriormente nos hemos referido a la necesidad de que sea técnicamente posible que el operador humano modifique parámetros en el despliegue. El conjunto de características técnicas que un sistema debe reunir para permitir la supervisión humana teniendo en cuenta la clase de interacciones posibles entre la máquina y el operador, así como el medio y fines con los que se despliegan, es relevante, entre otros, a la hora de poner solución a anomalías, problemas de funcionamiento y comportamientos inesperados[102], de decidir no utilizar el sistema o revertir su información de salida[103], o de intervenir en su funcionamiento[104]. Asimismo, la existencia de un mecanismo que permita al operador el aborto o interrupción de uso del sistema, elemento concebido como fundamental para el Control Humano Significativo, es también un elemento necesario para interrumpir el funcionamiento de un sistema de IA[105] . Los elementos referidos a la máquina o *software* deberán ser integrados técnicamente por el proveedor, lo cual no obsta para que deban ser también definidos por el mismo como parte de las medidas adoptadas para garantizar la supervisión humana[106], permitiendo al usuario del sistema conocer apropiadamente el funcionamiento de estos elementos[107].

Asimismo, debemos referirnos a los elementos que giran en torno a la capacidad del propio operador humano para llevar a cabo dicha supervisión. En el debate sobre los SAAL, hemos identificado como elemento mínimo del Control Humano Significativo los conocimientos de carácter jurídico, ético, técnico y específico en función del ámbito en el que opere el sistema. De forma análoga, estos conocimientos son esenciales para que pueda concurrir una supervisión humana efectiva de los sistemas de IA conforme a los fines establecidos, previniendo o minimizando el riesgo para la salud, la seguridad o los derechos fundamentales[108]. Los conocimientos técnicos del operador humano son esenciales para entender las capacidades y limitaciones del sistema y controlar debidamente su funcionamiento[109], para interpretar correctamente la información de salida del mismo[110], o para intervenir en su funcionamiento[111]. En cuanto a los conocimientos éticos y jurídicos, cobran una especial relevancia a la hora de decidir no utilizar el sistema o desestimar su información de salida[112]. Desde nuestro punto de vista,

[102] Art. 14.4.a de la Ley de Inteligencia Artificial.
[103] Art. 14.4.d de la Ley de Inteligencia Artificial.
[104] Art. 14.4.e de la Ley de Inteligencia Artificial.
[105] Art. 14.4e de la Ley de Inteligencia Artificial.
[106] Art. 14.3.a de la Ley de Inteligencia Artificial.
[107] La necesidad de integrar unas apropiadas características técnicas de un sistema con un conocimiento apropiado de las mismas por el operador humano nos remite, inevitablemente, a la investigación de Martin-Castro y Úcar acerca del metrónomo de Beethoven. Ante las distintas hipótesis acerca de por qué las partituras de Beethoven están marcadas con tempos inusualmente rápidos, en su investigación proponen que, siendo uno de los primeros usuarios de este instrumento, no sabía cómo leer la marca del metrónomo y sobreestimó todos los tempos en 12 pulsos por minuto. Ver MARTIN-CASTRO, A., UCAR, I., "Conductors' tempo choices shed light over Beethoven's metronome", *Plos one*, vol, 15, nº 12, pp. 1-16, doi.org/10.1371/journal.pone.0243616.
[108] Art. 14.2 de la Ley de Inteligencia Artificial.
[109] Art. 14.4.a de la Ley de Inteligencia Artificial.
[110] Art. 14.4.c de la Ley de Inteligencia Artificial.
[111] Art. 14.3.e de la Ley de Inteligencia Artificial.
[112] Art. 14.4.d de la Ley de Inteligencia Artificial.

DOI: 10.17103/reei.42.08

estos elementos que se refieren al operador humano deben recogerse entre las medidas definidas por el proveedor para que las lleve a cabo el usuario, antes de la introducción del sistema de IA de alto riesgo en el mercado[113].

Por último, hemos de tener en cuenta los elementos referidos a la información y tiempo del que dispone el operador humano en el uso y despliegue del sistema. Dentro del concepto de Control Humano Significativo, hemos observado la importancia de contar con una información suficiente y fiable para la toma de decisiones, información tanto de carácter endógeno –refiriéndonos a la información que se percibe a partir del propio sistema automatizado– como exógeno –información que se percibe al margen del sistema–. Asimismo, se ha destacado el tiempo del que dispone el operador para ejecutar sus funciones para poder adquirir plena capacidad operativa o contextual. Una información suficiente y fiable es esencial a la hora de detectar indicios de anomalías, problemas de funcionamiento y comportamientos inesperados en el sistema[114], de interpretar correctamente la información de salida del sistema[115], o de intervenir en su funcionamiento[116]. Incluyendo también la información de carácter exógeno que puede ser particularmente relevante para que el operador humano sea consciente de la posible tendencia a confiar automáticamente o en exceso en el sistema[117] . También el tiempo del que dispone el operador para evitar el sesgo de automatización, así como para la capacidad de entender, interpretar e intervenir en el sistema de IA. La disponibilidad de información suficiente y fiable de carácter endógeno, así como del tiempo suficiente para alcanzar capacidad operativa o contextual deberá integrarse técnicamente por el proveedor[118], mientras que deberán definirse las medidas necesarias para que el usuario recabe la información de carácter exógeno necesaria para un uso y despliegue efectivamente supervisado[119].

En definitiva, a nuestro juicio, queda constatado que, tanto de los consensos surgidos en los debates sobre los SAAL y de las propuestas realizadas en el marco de la Unión Europea, emergen una serie de elementos mínimos comunes que, de cumplirse, permiten alcanzar el nivel suficiente de supervisión humana sobre las decisiones automatizadas de los sistemas de IA de alto riesgo, posibilitando con ello, el respeto del bien jurídico protegido determinado.

---

[113] Art. 14.3.b de la Ley de Inteligencia Artificial. En desarrollo del Considerando 48 de la misma, estas medidas deben incluir la garantía de que: "(…) las personas físicas a quienes se haya encomendado la vigilancia humana posean las competencias, la formación y la autoridad necesarias para desempeñar esa función". Aunque en este aspecto sea necesario que el usuario del sistema asegure que las personas encomendadas para dicha función cumplan con estos requisitos, entendemos necesario que desde la fase de diseño y desarrollo del sistema se definan cuáles son esos requisitos y se comuniquen a los usuarios del sistema.

[114] Art. 14.4.a de la Ley de Inteligencia Artificial.

[115] Art.14.4.c de la Ley de Inteligencia Artificial.

[116] Art. 14.4.e de la Ley de Inteligencia Artificial.

[117] Art. 14.4.b de la Ley de Inteligencia Artificial.

[118] Art. 14.3.a de la Ley de Inteligencia Artificial.

[119] Art. 14.3.b de la Ley de Inteligencia Artificial.

DOI: 10.17103/reei.42.08

V. CONCLUSIONES

Si bien es cierto que hasta hace relativamente poco tiempo la Unión Europea no se ha centrado en regular los sistemas de IA de alto riesgo, se constata que desde hace décadas existe la conciencia que la única manera de mantener la dignidad humana, como expresión de los derechos y libertades fundamentales de la ciudadanía, es mediante la supervisión humana sobre las decisiones automatizadas. Reflejo de esto es la existencia de cláusulas, dispersas en la normativa, que protegen a los ciudadanos de la Unión de esta automatización, las normas de *soft law* sobre los sistemas de IA y las propuestas legislativas del Parlamento y la Comisión, siendo la Ley de Inteligencia Artificial su expresión más desarrollada. La Comisión, a través de esta propuesta de Reglamento, plantea una supervisión humana efectiva para el uso de sistemas de IA de alto riesgo mediante el establecimiento de obligaciones jurídicas a los proveedores de las mismas en las fases de diseño y desarrollo.

Asimismo, como hemos constatado, el contexto europeo no es el único que ha prestado atención a esta clase de sistemas. La comunidad internacional lleva cerca de una década dedicada al control de los sistemas de armas que de forma autónoma pueden llevar a cabo ataques. Lejos de ser una realidad futura, los SAAL constituyen unas herramientas de hacer la guerra de importancia e implementación creciente en los diferentes conflictos armados. Los debates en el contexto del Grupo de Expertos Gubernamentales de composición abierta sobre los SAAL, orientados a que estos sistemas respeten el Derecho Internacional Humanitario, han evidenciado que el Control Humano Significativo es esencial para garantizar que el ser humano mantiene bajo su control estos sistemas de armas y, por tanto, hace viable respetar las normas de este ámbito del Derecho internacional. También se ha concluido que se trata de un proceso que se ejerce sobre todo el ciclo de vida del SAAL, es decir, desde la dirección política en la fase de predesarrollo hasta la fase de evaluación posterior a la utilización.

En este sentido, aunque, por el momento, no hay un acuerdo general sobre el concepto, hemos identificado una serie de elementos que lo conforman. A saber, la posibilidad técnica de modificar el sistema para tomar cualquier tipo de decisión, incluyendo el aborto o la interrupción del uso, capacidad cognitiva y conocimiento jurídico, ético, técnico –general y especifico– del operador para tomar la decisión adecuada, la existencia de información suficiente y fiable y del tiempo necesario para tomar una decisión. La concurrencia de estos elementos posibilitaría lo que se ha venido a llamar Control Humano Significativo.

Asimismo, el contraste de los consensos del Grupo sobre los SAAL y las diferentes propuestas de la UE han constatado la voluntad de prohibir aquellos sistemas de IA que no puedan ser objeto de supervisión humana. Esta consideración se fundamenta en el axioma de que el marco jurídico que regula el bien jurídico protegido sobre el cual incide el sistema debe ser respetado. Así, la supervisión humana, que se dará en todo caso sobre todas las fases de vida del sistema, será más o menos acusada en función del riesgo en cuanto a la probabilidad de lesionar el bien jurídico protegido.

DOI: 10.17103/reei.42.08

No se trata, por tanto, de la simple presencia de un operador humano durante todo el ciclo de vida de la máquina, sino de una participación cualificada, siendo definida como significativa y/o efectiva. Este grado de supervisión se logra mediante la concurrencia de una serie de elementos que hemos identificado tanto en los consensos sobre los SAAL como en las propuestas de legislación de la UE. En primer lugar, existencia de unos parámetros técnicos, integrados por el proveedor, que permita al usuario ejercer su supervisión humana tomando una serie de decisiones: no modificar los parámetros, modificarlos, suspender o e interrumpir la decisión automatizada. En segundo lugar, la tenencia por parte del operador de conocimientos jurídicos, éticos, técnicos y específicos en función del ámbito en el que opere el sistema para conocer la previsibilidad del sistema, interpretar la información de salida del mismo, detectar los posibles errores e intervenir de forma adecuada en su funcionamiento. Finalmente, la información endógena y exógena necesaria, así como el tiempo mínimo para poder procesar los datos, deliberar y tomar una decisión. A nuestro juicio, la concurrencia de estos elementos posibilitará la existencia de una supervisión humana de cualquier sistema de IA de alto riesgo.

En conclusión, las evidencias de uso de SAAL sin intervención humana[120] o la toma de decisiones automatizadas por parte de sistemas de IA en ámbitos tan diversos como la medicina, el acceso a protección social o la aplicación de la ley[121], ponen de relieve la urgencia de clarificar el contenido de la supervisión humana y su implementación. Si bien se ha realizado una labor encomiable, es necesario seguir trabajando en la delimitación de este concepto para posibilitar la existencia de decisiones automatizadas antropocéntricas y antropogénicas.

Esta labor se presenta primordial para garantizar que los derechos y libertades fundamentales de la ciudadanía son observados. En este sentido, en la medida en la que el fundamento último de esto es la protección de determinados bienes jurídicos fundamentales, consideramos que la vía adecuada para su protección son normas de *hard law* en la línea en la que persigue la propuesta de Ley Artificial, aunque somos conscientes que en algunos casos solo será posible aspirar a normas de *soft law*, como códigos de conducta.

[120] FROELICH, P., "Killer drone…, *op. cit.* y NACIONES UNIDAS, *Informe final del Grupo de Expertos*…, *op. cit.*, párr. 69, p. 20.

[121] Ver los estudios sectoriales que se han desarrollado en amplios informes como los de la Agencia de los Derechos Fundamentales de la Unión Europea o la plataforma AlgorithmWatch. FRA, *Getting the future right – Artificial intelligence and fundamental rights*, Publications Office of the European Union, Luxemburgo, 2020 y CHIUSI, F., *et al.* (Eds.), *Automating Society Report 2020*, AlgorithmWatch gGmbH, Berlín, 2020, automatingsociety.algorithmwatch.org.

DOI: 10.17103/reei.42.08