ACADEMIA DE LAS CIENCIAS
Y LAS ARTES MILITARES

Comunicaciones académicas

Ética y control humano significativo en sistemas de armas autónomos letales regidos por la Inteligencia Artificial

Juan A. Moliner González
Academia de las Ciencias y las Artes Militares
Sección de Prospectiva de la Tecnología Militar

26 de marzo de 2024

Concepto de Control Humano Significativo-CHS (*Meaningful Human Control*)

La utilización de la Inteligencia Artificial (IA) en Sistemas de Armas Letales Autónomos (SALAS) ha producido una abundante literatura en la que el concepto de «control humano significativo (CHS)» ha sido empleado de forma habitual como una de las más importantes exigencias legales y éticas de esos sistemas autónomos. Intentar profundizar y precisar el significado de este concepto es el objetivo del presente trabajo.

En cuanto al control sobre la autonomía del sistema, el CHS se refiere, de forma genérica, a la posibilidad de que el ser humano, es decir su operador, lo gobierne en todo momento y mantenga en su empleo la capacidad humana de proporcionarle *inputs*, derivados de su propio juicio «y que le permitan retener un dominio legal y éticamente aceptable de las funciones críticas que se le encomienda ejecutar» (Jiménez-Segovia, R. (2019). 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*, Núm.37, junio. DOI: 10.17103/reei.37.07).

Uno de los aspectos más relevantes que se cuestionan en relación con el CHS, y que dificulta los eventuales acuerdos jurídicos y normativos sobre el mismo, se refiere a su propia terminología. Así, el CHS es empleado y discutido en las reuniones que se llevan a cabo en la sede de Ginebra de las Naciones Unidas (Convención sobre Ciertas Armas Convencionales, Grupo de Expertos Gubernamentales, CCW/GGE) y en otros foros, mientras que el Departamento de Defensa de EE. UU. utiliza la expresión «niveles apropiados de juicio humano» (*appropriate levels of human judgement*) en su Directiva de Defensa 3000.09 (Department of Defense (2023). Directive 3000.09. Autonomy in Weapon Systems. January, 25.).

Imagen generada por Lucia Pascual y herramientas de Inteligencia Artificial

La primera vez que apareció el concepto de CHS fue en abril de 2013 en un informe de la Organización No Gubernamental británica *Article 36* (Article 36 (2013). Killer Robots: UK Government Policy on Fully Autonomous Weapons, a commentary on the UK Ministry of Defense's 2011 Joint Doctrine Note on The UK Approach to Unmanned Systems), haciendo referencia al artículo 36 del I Protocolo Adicional de 1977 a las Convenciones de Ginebra de 1949.

Dicho artículo establece que:

> En el estudio, desarrollo, adquisición o adopción de un arma nueva, medio o método de guerra, una Alta Parte Contratante tiene la obligación de determinar si su empleo estaría, en todas o alguna circunstancia, prohibido por este Protocolo o por alguna otra regla de Derecho Internacional aplicable a la Alta Parte Contratante.

En el Informe se exigía que el CHS cumpliera los tres siguientes requisitos:

1) Información: un operador humano, y otros responsables por el planeamiento del ataque, necesitan tener adecuada información contextual sobre el área objetivo de un ataque, información sobre los objetivos de la misión, e información sobre los efectos inmediatos y a largo plazo del arma que se derivarán de un ataque en ese contexto.
2) Acción: la iniciación del ataque debería requerir una acción positiva por un operador humano.
3) Rendición de cuentas [*accountability*]: aquellos responsables de evaluar la información y ejecutar el ataque necesitan responsabilizarse de los resultados del ataque.

Respecto al primero de estos requisitos, la información, resulta apropiado hacer notar que ya se contempla, para todos los sistemas de armas, en el Derecho Internacional Humanitario (DIH) y es muestra de las obligaciones legales que supone. El segundo, la acción, es importante en el concepto de CHS pues pone en cuestión la propia concepción del sistema autónomo y la mayor capacidad operativa que aporta, en rapidez y precisión, esa autonomía. Finalmente, el tercero, la rendición de cuantas, no es un requisito específico o exclusivo de los sistemas autónomos pues cuando una máquina de guerra falla o produce consecuencias indeseables, pero la atribución de responsabilidad si es un problema, como se analizará más adelante.

Generalmente se acepta que el DIH es aplicable a los SALAS, por lo que una importante cuestión a resolver es si el CHS y sus requisitos satisfacen las reglas de ese derecho en lo relativo a su control y exigencia de responsabilidad. A esto se objeta que el DIH no requiere explícitamente el control humano, sino que cada

medio o método de guerra empleado cumpla con las obligaciones que ese derecho establece.

Si se evaluara un sistema autónomo y se determinara que cumple con el DIH, el CHS quedaría sin base para ser considerado como una obligación legal (Trabuco, L. (2023). What is Meaningful Human Control, anyway? Cracking the code on autonomous weapons and human judgment. *Modern War Institute*, 09.21.2023).

Pero incluso sin compromisos legales permanecerían las exigencias éticas del CHS, pues en todo caso la actuación de la máquina persigue incorporar los importantes beneficios que proporciona la autonomía regida por la IA sin desdeñar, sino al contrario, las ventajas del juicio humano.

Además, puede ocurrir que los SALAS ejecuten una conducta (acción) inesperada o indeseada y el problema que se presenta es la consiguiente asignación de responsabilidades, dificultada por la ambigüedad en su atribución.

Resulta conveniente, en consecuencia, ampliar y precisar los argumentos y razones que soportan la necesidad del CHS en los sistemas autónomos si se quiere mantener la influencia de ese juicio humano en dichos sistemas, de forma que su operación responda a las exigencias legales y razones éticas que se demandan en su funcionamiento.

Autonomía, control y condiciones del CHS en la operación de los sistemas de armas autónomos

La autonomía de los sistemas de armas autónomos está relacionada con la forma y el grado de participación, así como el nivel de control del ser humano, del operador, en el funcionamiento y resultados de las acciones del sistema.

En la literatura académica se suelen citar tres escenarios en el proceso de actuación del sistema:

1) *Human in the loop.* El ser humano es necesario para la operación del sistema y su intervención imprescindible en todo momento.
2) *Human on/over the loop.* El ser humano actúa como vigilante continuo del proceso, siendo su actuación necesaria y no imprescindible y teniendo la importante capacidad de interrumpirlo en cualquier momento.
3) *Human out of the loop.* El ser humano es incapaz de intervenir en el proceso, una vez este se ha iniciado. En este escenario la intervención humana no es posible y el sistema opera con total autonomía una vez activado.

En relación con los dos primeros escenarios, importante dilema ético que se presentan se refiere a la sensación de seguridad del controlador al no encontrarse bajo peligro por la distancia física al objetivo (en ocasiones a miles de kilómetros), que, a menudo, se convierte en distancia psicológica y emocional.

Hay autores que defienden que bajo esa percepción de que el operador no está en riesgo extremo, en cuyo caso la tendencia es usar la fuerza letal, el no estar sujeto a factores como el estrés o el miedo, hace a los operadores ser más precisos y respetar principios éticos, o incluso no emplear la fuerza letal (Erbland, 2018). Por el contrario, también se postula que en base a esa seguridad se pierde la conciencia humanitaria y la empatía, pudiéndose llevar a cabo acciones que vayan contra el principio de restricción en el combate.

Respecto al escenario *out of the loop*:

> […] es el eje central de la polémica que suscitan las armas autónomas letales, desde una perspectiva ética y jurídica. La disolución de la responsabilidad por los fallos de una máquina que carecía de control alguno (González, L. (2022). La responsabilidad del Estado por el uso de armas autónomas letales. *Revista Española de Derecho Militar*. Núm. 118, julio-diciembre).

Se considera que a menos que la IA que dirige los SALAS esté programada con una «moral artificial» (que de momento no existe y parece muy poco probable se pueda desarrollar), determinar qué representa una amenaza y tomar la decisión de eliminarla debe requerir necesariamente supervisión humana.

En cuanto a las condiciones esenciales para asegurar un CHS en los sistemas de armas autónomos, encontramos (Horowitz, M.C. y Scharre, P. (2015). *Meaningful Human Control in Weapon Systems: A Primer*. Washington: Center for a New American Security, March):

1) Los operadores toman decisiones conscientes y correctamente informadas a la hora de decidir la utilización de los sistemas autónomos.
2) Para ese empleo su información debe ser suficiente para asegurar la legalidad de las acciones que toman en relación con el objetivo, el arma y el contexto de la acción.
3) El sistema de armas se ha diseñado, evaluado y comprobado, y los operadores está entrenados adecuadamente para asegurar un control efectivo sobre el empleo del sistema.

Lo expuesto hasta ahora nos lleva a la necesidad de considerar como el juicio humano debe incorporarse en todo el ciclo de vida del sistema y no solo en la fase de operación de los SALAS, haciendo del CHS un concepto más integrador y comprensivo.

Necesidad de integrar el CHS en todo el ciclo de vida de los sistemas autónomos

Actualmente es bastante general la necesidad de ampliar el CHS a todas las fases de un sistema de armas autónomo. A partir de la consideración de que la propuesta del *Article 36* no parece suficientemente adecuada para captar todos los aspectos que influyen en el control y juicio humano sobre un SALAS, se debería incluir no solo la participación de los usuarios y operadores finales del sistema, sino de todos aquellos que intervienen en las fases anteriores a la activación del mismo y que también deben tenerse en cuenta para calificar y validar el CHS (Trabuco, 2023).

Se presenta, así, el problema de la responsabilidad legal y moral de quién toma la decisión: ¿el diseñador, los desarrolladores, el *hardware*, los algoritmos de IA, el comandante de la operación, los operadores, la propia máquina dotada de entidad jurídica?

Así, en la fase de diseño y desarrollo son humanos los que, a través de sistemas de IA (*machine learning, deep learning, neural networks*), deben tener en cuenta en la arquitectura del sistema los estímulos y las circunstancias ambientales inesperadas que pueden surgir. La superior capacidad de las máquinas, gracias a la IA, en velocidad y precisión sobre la capacidad cognitiva humana no puede obviar la cuestión de la responsabilidad en caso de mal funcionamiento, que pudiera llegar a la comisión de un crimen de guerra.

En la fase de planeamiento operacional la decisión de emplear sistemas autónomos debe tener muy presente el entorno operativo en que van a funcionar, pues no es lo mismo hacerlo en entornos urbanos, donde la distinción entre combatientes y no combatientes es un reto y de una complejidad enorme, que su funcionamiento en entornos abiertos como amplias superficies marinas o extensos terrenos llanos, donde esa discriminación parece ser más factible. Otro tipo de decisiones humanas que deben ser tenidas en cuenta en el planeamiento, para el empleo operacional de los sistemas autónomos, son las relativas al estilo de mando y liderazgo, que puede influir en la mayor o menor asunción de riesgos, llegando a un posible fuego amigo para las tropas propias.

Finalmente, en la fase de planeamiento táctico y combate se debe considerar que el sistema autónomo sea capaz de tener en cuenta la capacidad de supervisión a la hora de atacar un objetivo, algo que puede variar entre diferentes operadores humanos. Por ejemplo, con la desigual percepción individual ante la presentación de información (datos o imágenes), o las distintas respuestas cognitivas que podrían dar varios operadores si el sistema estableciera que hay un 95% de que el objetivo sea enemigo, a diferencia de si indicara que hay un 5% de que no sea enemigo.

Se puede apreciar que lo relevante no solo es que los SALAS puedan seguir los principios de proporcionalidad, necesidad y discriminación del DIH, sino también conocer quién es responsable de las decisiones a lo largo de todo el ciclo del *targeting*, teniendo en cuenta que la intervención humana en el mismo puede reducir las ventajas funcionales que proporciona su autonomía tecnológica.

Siguiendo la doctrina de las Fuerzas Armadas españolas, se entiende por *target* (objetivo):

> Cualquier entidad (instalación, persona, estructura virtual, equipo u organización) que cumple una función para el adversario y sobre la que se pueden realizar acciones letales y no letales para crear efectos físicos, psicológicos o virtuales (PDC-02. Marco legal para el empleo de las FAS, junio 2021).

Elegir, seleccionar y ejecutar la operación contra determinados objetivos exige evaluar los que se pueden atacar, emplear los medios de localización adecuados y establecer las formas y medios más idóneos de combate.

Solo los seres humanos pueden ser sujetos de responsabilidad legal y moral ante las actuaciones indeseadas de los SALAS (bajas de civiles no combatientes, ataques no deseados, pérdida de control sobre los sistemas, *hackeos* o adquisición por grupos terroristas, entre otros), aunque no se intervenga en su ciclo del *targeting*, y, además, esa asunción de responsabilidad es mayor cuanto mayor sea la autonomía de los sistemas.

Por estas razones se indica que es importante conceptualizar e insertar los requisitos del CHS en el sistema autónomo, y ello debe hacerse en todas las fases del ciclo de vida del sistema y no solo pensar que el CHS afecta exclusivamente al empleo táctico de los sistemas de armas letales autónomos (Trabuco, 2023).

En esta línea de tener en cuenta las exigencias éticas en todas las fases de la vida de los SALAS hay quien advierte que «el concepto de ´control humano significativo` es un concepto ilusorio y que se debería avanzar hacia el de ´certificación humana significativa` de los sistemas de armas autónomos» (Cummings, M.L. (2029). Lethal Autonomous Weapons: Meaningful Human Control or Meaningful Human Certification? *IEEE Technology and Society Magazine* 38(4): 20-26).

Se razona esta afirmación por la incapacidad de los humanos, en comparación con las máquinas, en tomar decisiones de vida y muerte dadas las dificultades de evaluar una información siempre imperfecta, un reducido tiempo de reacción y la enorme complejidad presente y futura del campo de batalla.

> La más significativa forma de control humano en el empleo de sistemas de armas autónomos ofensivos es decidir *a priori* que objetivos serán establecidos y bajo qué condiciones (*Ibídem*, 7).

Conclusiones

Las cuestiones planteadas sobre el control humano significativo, ante las posibles disfunciones en los SALAS, hacen notoria y urgente la necesidad de que se establezca y precise el ser humano moral y legalmente responsable en el uso de un sistema de armas letal autónomo no solo en su empleo en el campo de batalla, sino desde la primera fase de su desarrollo e investigación.

Por tanto, antes de ese uso operativo en guerras y conflictos, los SALAS deben ser integrados, tanto si se interviene como si no se hace en el ciclo del *targeting*, como una parte extendida del mecanismo humano de toma de decisiones, las cuales, teniendo en cuenta las condiciones y situaciones ambientales en que actúan los sistemas de armas, deben ser relevantes a la legalidad nacional e internacional y a las restricciones éticas que se plantean.

Para evitar los dilemas morales y legales que se han presentado en las reflexiones anteriores, el futuro, que ya es presente, exige que tanto los investigadores como los militares usuarios de estos sistemas de armas incorporen y tengan presentes los requerimientos del CHS y del DIH desde las fases iniciales en el diseño y desarrollo de los sistemas y hasta su finalidad última, que es la acción militar que representa la utilización de estas armas. Armas que, como todas las que se han empleado y se seguirá haciendo en la guerra, son letales y, no se puede olvidar, producen destrucción y muerte. ■