La importancia de la ética en las armas letales autónomas

¿Pueden ser los sistemas de armas autónomas transparentes? ¿Podemos confiar nuestra seguridad militar a la inteligencia artificial? ¿Son reversibles las decisiones tomadas por los sistemas de armas autónomas? Este artículo examina algunas cuestiones que plantea la IA con respecto a la reversibilidad en los drones autónomos.

¿Pueden ser los sistemas de armas autónomos transparentes?

Según un estudio sobre la confianza en la inteligencia artificial (IA) en contextos de rescate y militares (Tolmeijer et al., 2022), las capacidades autónomas pueden llegar a ser muy impredecibles, poco fiables, irreversibles e inexplicables en escenarios hostiles y cambiantes. Esto se traduce en opacidad y en la dificultad de interpretar los sistemas de *machine learning*. La estructura de código es una actividad necesaria para la implementación computacional de algoritmos. A consecuencia de ello, aparece la falta de coincidencia entre los procedimientos matemáticos de los algoritmos de aprendizaje automático y los estilos humanos de interpretación semántica. En mitad de este camino existe una opacidad que se relaciona con las técnicas específicas utilizadas en el aprendizaje automático. El aprendizaje automático en particular a menudo se describe como que sufre la 'maldición de la dimensionalidad' (Domingos, 2012).

Dentro de la era del *big data* la cantidad de datos disponibles que pueden llegar a ser analizados son inmensos. Estos datos son una parte esencial del aprendizaje automático porque la lógica interna de un algoritmo que se encuentra en un proceso de aprendizaje automático se modifica a medida que «aprende» sobre la base de datos que recibe. Si bien los conjuntos de datos pueden ser extremadamente grandes, pero claros de comprender, la interacción con la funcionalidad del algoritmo es lo que produce complejidad y, por lo tanto, produce opacidad. Uno de los problemas que plantea esta opacidad es su interpretabilidad. Encontrar formas de revelar algo de la lógica interna de un algoritmo puede abordar las preocupaciones sobre falta de "equidad" y los efectos discriminatorios.

A partir de esta información, se nos plantea la siguiente pregunta.

¿Podemos confiar nuestra seguridad militar a la inteligencia artificial?

Las limitaciones en la toma de decisiones humanas otorga una ventaja a la automatización parcial con IA en la dimensión temporal como en la calidad de la decisión. Un grupo de trabajo de investigación de la OTAN, por ejemplo, examinó la necesidad de automatización en cada paso del ciclo de inteligencia (Organización de ciencia y tecnología de la OTAN, 2020) y descubrió que la IA aplicada en sistemas de armas autónomas, al contrario que un piloto militar al no estar sometido a estrés durante el combate, es capaz de ayudar a automatizar tareas manuales, identificar patrones en conjuntos de datos complejos y acelerar el proceso de toma de decisiones en general. Dado que la recopilación de más información y perspectivas da lugar a productos de inteligencia menos sesgados, y la potencia informática aumenta la cantidad de datos que pueden ser procesados y analizados, se puede reducir el sesgo cognitivo. Por ejemplo, el sesgo de confirmación puede evitarse mediante el análisis automatizado de hipótesis contrapuestas (Dhami et al., 2019). Otras ventajas de las máquinas sobre los humanos son que permiten simulaciones escalables, realizan razonamiento lógico, tienen conocimiento transferible y un espacio de memoria expandible. Por ello, es tendencia que las máquinas cada vez posean más funciones autónomas y, en consecuencia, cada vez están

menos supervisadas por operadores humanos.

Es difícil comprobar si un agente autónomo está siguiendo un código moral

Un aspecto importante del debate actual sobre el uso de la IA para la toma de decisiones se centra en los peligros potenciales de proporcionar demasiada autonomía a los sistemas de IA, lo que puede tener consecuencias imprevistas. Una parte de la solución es proporcionar información suficiente al liderazgo sobre cómo se han diseñado los sistemas de IA, en que se basan sus decisiones (explicabilidad), qué tareas pueden estar soportadas por la automatización y cómo tratar los errores técnicos (Lever & Scheneider, 2021). De modo que cuanto mayor sea la autonomía de una máquina según afirma (Picard, 2003), mayor necesidad tendrá de normas morales.

Por otro lado, son las tareas de alta complejidad las que no son aptas para la automatización por el momento, estas son aquellas en las que los humanos superan a las máquinas (Blair, D et al., 2021). Por lo tanto, el debate sobre la IA responsable debe tener también en cuenta las fortalezas humanas. En la práctica, los sistemas de IA no pueden trabajar de forma aislada, sino que deben trabajar en equipo con los responsables humanos de la toma de decisiones.

Respecto a la cadena de toma de decisiones existen ciertos tipos de robots militares de IA sometidos a un control y a un juicio humano en el que su uso puede estar legitimado con fines de autodefensa. Además, un punto que conviene destacar, es que la colaboración entre humanos e IA podría conducir a una toma de decisiones más rápida y adecuada bajo presión e incertidumbre, y los sistemas de IA podrían utilizarse ampliamente para el entrenamiento adaptativo del personal militar, permitiendo reducir los sesgos en la toma de decisiones, por ejemplo, mediante la toma detección de la somnolencia o la fatiga a partir de señales neurometricas en el cerebro (Weelden et al., 2022).

La reversibilidad de las armas letales autónomas

Por otro lado, puede darse un problema basado en la irreversibilidad del disparo. El problema de la irreversibilidad proviene de la dificultad de dar marcha atrás si surgen objeciones morales dentro del funcionamiento de un agente artificial autónomo. Y es que, es difícil comprobar si un agente autónomo está siguiendo un código moral. Esto es debido al rápido desarrollo de la IA, que ha sido utilizada por los ejércitos modernos para que se actualicen y mejoren continuamente la precisión y la velocidad de las capacidades de sus armas para que su ejército continúe siendo competitivo (Zenko, 2013; Beard, 2014). A su vez, los incentivos y la competitividad con otras potencias aumentan la necesidad de automatizar procesos por motivos de eficiencia, seguridad y versatilidad, así como por razones políticas y económicas. Los procesos para los que más se requiere el uso de algoritmos basados en datos para funcionar bien son aquellos que son complejos, dinámicos y no estructurados. Como ya hemos mencionado, este tipo de procesos crean opacidad inherente que a su vez genera nuevos desafíos, tanto operativos como legales. De modo que, cuando el operador no comprende cómo estos factores interactúan con el arma, solo se pueden tomar ciertas precauciones efectivas. Estas obligaciones son difíciles de cumplir si se desconoce la causa del problema, lo que impide la reparación y el seguimiento del resultado hasta la persona considerada más responsable. Junto a este problema, existe también la falta de previsibilidad que puede frustrar las pruebas de intención y causalidad requerida para la responsabilidad personal. Obstaculizando el seguimiento hasta la persona

considerada más responsable. Por ello, los estados deben hacerse responsables de garantizar que los principios del Derecho internacional humanitario (DIH) se respeten en el campo de batalla.

En circunstancias de alto riesgo como la guerra, la interpretabilidad es eminentemente necesaria (Doshi-Velez & Kim, 2017). Como tal no existe una prohibición explícita en el DIH sobre el uso de las capacidades de una IA opaca, y esperar a nuevas leyes que estén acorde a las necesidades actuales no es realista dada la dificultad de obtener consenso multilateral, y mucho menos global. Con todo, es necesario implementar la interpretabilidad desde el inicio de la fase de desarrollo y prueba. Sin embargo, esto perjudica la versatilidad causando un costo de oportunidad por parte de incentivar la interpretabilidad y la transparencia.

Las tecnologías que poseen autonomía deben poder ser reversibles

Este es uno de los problemas centrales de la ética de las máquinas. La palabra autónomo, describe dentro de este contexto, una máquina que funciona sin control humano efectivo. Del mismo modo, la máquina puede funcionar bajo control humano efectivo, perdiendo su autonomía. Sin embargo, una máquina puede también funcionar sin control humano efectivo, pero requiriendo en determinadas condiciones solicitar de forma activa la intervención de un agente humano.

De este modo, la reversibilidad se trata de un problema estructuralmente similar al denominado dilema de Collingridge (Lierbert & Schmidt, 2010). Afirma que el desarrollo y la aplicación de la tecnología es difícil de controlar debido a la falta de información y la baja velocidad de reacción humana, comparados al funcionamiento de una máquina autónoma. Las tecnologías que poseen autonomía deben poder ser reversibles, ya que cualquier elección puede ser falible, y comprometerse irreversiblemente con una versión concreta de la tecnología equivaldría a afirmar sobre la certeza de su idoneidad. Sin embargo, este problema choca con la necesidad de un despliegue práctico para su afianzamiento tecnológico. Sin embargo, el afianzamiento tecnológico puede limitar e incluso impedir su reversibilidad.

Conclusión

La conclusión que se desprende de las AWS es que el cumplimiento de un código moral debe demostrarse mediante pruebas empíricas para garantizar que estas armas son fiables y pueden utilizarse de forma permisible. La permisibilidad moral de las AWS depende de su capacidad para superar otros umbrales éticos. Sin embargo, los dilemas que plantea este debate deben ir resolviéndose de uno en uno.

Blair, D., Chapa, J., Cuomo, S. y Hurst, J. (2021): *Humans and hardware: an exploration of blended tactical workflows using John Boyd's OODA loop*. In R. Johnson, M. Kitzen, & T. Sweijs (Eds.), *The conduct of war in the 21st century : Kinetic, connected and synthetic* (pp. 93–115). Taylor & Francis Group.

Dhami, M. K., Belton, I. K. y Mandel, D. R. (2019): «The "analysis of competing hypotheses" in intelligence analysis». *Applied Cognitive Psychology*, 33(6), 1080–1090.

Doshi-Velez, F. y Kim, B.): «Towards A Rigorous Science of Interpretable Machine Learning» en *arXiv,* 2017. Disponible en: https://doi.org/10.48550/arXiv.1702.08608

Eidelman, S. y Crandall, C. S. (2012): «Bias in favor of the status quo». *Social and Personality Psychology Compass*, 6(3), 270–281.

Hanska, J. (2020): *War of time: Managing time and temporality in operational art*. Londres, Palgrave Macmillan.

Kwik, J. y Van Engers, T.: «Algorithmic fog of war: When lack of transparency violates the law of armed conflict» en *IOS Press*, 2(1-2), 43-66. 2021. Disponible en: https://doi.org/10.3233/frl-200019

Lever, M. y Schneider, S.: «Decision augmentation and automation with artificial intelligence: Threat or opportunity for managers?» en *Business Horizons*, 64(5), 711–724. 2021. Disponible en: https://doi.org/10.1016/j.bushor.2021.02.026

Liebert, W. y Schmidt, J. C.: «Collingridge's dilemma and technoscience» en *Poiesis & Praxis*, 7(1), 55-71. 2010. Disponible en: https://doi.org/10.1007/s10202-010-0078-2

NATO Science and Technology Organization: «Automation in the intelligence cycle» en *NATO,* 2020. Disponible en: https://www.sto.nato.int/Lists/STONewsArchive/displaynewsitem.aspx?ID=552

Domingos, P.: «A few useful things to know about machine learning» en *Commun,* 2012. Disponible en: https://doi.org/10.1145/2347736.2347755

Picard, R.W.: «Affective computing: challenges» en *Int J Hum Comput Stud,* 2003. Disponible en: https://doi.org/10.1016/S1071-5819(03)00052-1

Weelden, E. V., Alimardani, M., Wiltshire, T. J., & Louwerse, M. M.: «Aviation and neurophysiology; A systematic review» en *Applied Ergonomics*, 2022. Disponible en: https://doi.org/10.1016/j.apergo.2022.103838