

ENSAYOS DE FILOSOFÍA

Riesgos asociados al desarrollo de robots autónomos dotados de inteligencia artificial avanzada en contextos civil y militar

Miguel Moreno Muñoz
Universidad de Granada
mm3@ugr.es

Introducción

Existe un intenso debate acerca del potencial de las tecnologías actuales para el diseño de máquinas inteligentes, capaces de adquirir capacidades que igualan o superan a las humanas en tareas especializadas de cierta complejidad (Chan, 2008: S73). Ya no resulta descabellado asumir que ciertos dispositivos (*agentes*) dotados de inteligencia artificial (IA) podrán ampliar sus capacidades generales para orientarse y actuar en su entorno, alcanzando algo muy parecido a estados de consciencia (el *programa fuerte* de IA) en una o dos décadas, si la investigación relacionada progresa al ritmo que lo han hecho en los dos últimos años las tecnologías de *machine learning* y el desarrollo de algoritmos avanzados (Bilton, 2014). Expertos como Ben Goertzel y Cassio Pennachin no descartan plazos muchos más ajustados, de cuatro o cinco años (Goertzel & Pennachin, 2007: VII).

Cabe discutir si la inteligencia humana es verdaderamente general; pero la posición al respecto no justifica un menosprecio de aplicaciones concretas de la IA que pueden mejorar el rendimiento humano y la capacidad para tomar decisiones en dominios tan específicos como la conducción autónoma, p.ej.¹ De hecho, una ventaja de la IA es la posibilidad de prescindir de las limitaciones del diseño humano y no quedar restringida en su desarrollo a una mera emulación de capacidades y funciones previamente conseguidas en sistemas biológicos (Bostrom & Yudkowsky, 2013: 3). Este enfoque ha tenido como efecto colateral una pérdida de valor de habilidades humanas que, paulatinamente, han sido asumidas por artefactos o diseños de ingeniería incorporados en los sistemas de producción industrial automatizada.

A diferencia del carácter especulativo que tuvieron los ejercicios de prospectiva más optimistas sobre el potencial de la IA en los años ochenta y noventa (Vinge, 1993), algunas posibilidades sugeridas por Ray Kurzweil en su obra *The singularity is near: when humans transcend biology* (Kurzweil, 2005) han comenzado a resultar verosímiles gracias a la combinación de desarrollos significativos, aunque parciales, en capacidad de computación, hardware, robótica, sensores, software y algoritmos avanzados de IA aplicada.

Otros desarrollos han seguido el curso previsto por Vernor Vinge en su artículo titulado “The Coming Technological Singularity: How to Survive in the Post-Human Era” (Vinge, 1993), donde anticipa algunos escenarios compatibles con la aparición de entidades dotadas de una inteligencia superior a la humana.²

Entre enero de 2014 y diciembre de 2015 se tuvo noticia de diversos ejemplos de integración de tecnologías con resultados que igualan o superan las capacidades humanas en tareas complejas (Sharma, K, & Jawahar, 2015). Muestran de qué modo se han reajustado las expectativas iniciales, pasando del *volcado de cerebros en la nube* –como escenario futurista sugerido por Kurzweil–³ a múltiples aplicaciones en robótica industrial, transporte, control de sistemas, diseño y automatización de procesos, asistencia en diagnósticos, clasificación de imágenes, traducción y reconocimiento del habla o procesamiento del lenguaje natural.

La rápida incorporación de sistemas de IA avanzada a los procesos de toma de decisiones en dominios tan dispares como la economía, transportes y logística, desarrollo de software, ingenierías y medicina –entre otros muchos donde la capacidad para inferir e interpretar datos estadísticos es crucial–, justifica una aproximación detenida para entender el potencial de cambio social asociado. Este análisis de impacto quedaría incompleto sin considerar las aplicaciones de la IA de uso militar, puesto que ya no parece un recurso prescindible como herramienta de apoyo que aporta ventajas competitivas en el funcionamiento de los servicios de inteligencia y en el diseño de sistemas avanzados de alerta, defensa y respuesta militar (Marcus, 2013).

La versatilidad y potencial de la IA en desarrollo para fines civiles y militares explica que el debate haya ido incorporando a expertos de muchas disciplinas interesados en articular un marco de gobernanza internacional que permita regular el uso de armas robóticas autónomas (*Lethal autonomous robots*, LAR). Se trata de un proceso quizás inexorable –un país que prohíba su uso simplemente daría ventaja a otro dispuesto a utilizarlas– y continuamente incentivado en los escenarios actuales de conflicto. Por esta razón el desarrollo de IA general avanzada para aplicaciones de doble uso se considera una amenaza global del máximo nivel, que autores como Nick Bostrom encuadran en la categoría de *riesgos existenciales* (Bostrom, 2002).

Este enfoque se ha visto reforzado tras hacerse pública una carta abierta, firmada por destacados expertos en robótica, IA y disciplinas afines, en la que reconocen que el empleo de IA en sistemas robóticos autónomos de uso militar es una posibilidad factible en cuestión de muy pocos años (AA.VV., 2015). Sin acuerdos vinculantes para la comunidad internacional que lo impidan, el empleo de armas autónomas será la tercera gran revolución en el negocio de la guerra, tras el uso de la pólvora y el desarrollo de armas nucleares.

Un futuro inquietante para el mercado laboral

La interacción humano-máquina en entornos cooperativos de producción industrial es ya una realidad que está transformando el potencial y alcance de sectores estratégicos de la actividad económica a escala mundial (Pedrocchi, Vicentini, Matteo, & Molinari, 2013). Su impacto en el mercado internacional de trabajo está sujeto a muchas variables, cuyo efecto combinado puede traducirse tanto en una mayor demanda de trabajadores cualificados para nuevas tareas como en la extinción de sectores completos de actividad profesional, en servicios que hoy desempeñan trabajadores con una cualificación de nivel básico o intermedio.

El mercado de robots que prestan servicios específicos en la producción industrial es un sector en auge. Junto a una bajada sostenida de precios cercana al 10% anual –a medida que la demanda se

generaliza en los países industrializados—, el factor decisivo parece ser la disminución gradual de la ventaja comparativa que aporta la mano de obra humana en tareas que implican movilidad y destreza. Si bien la tendencia actual a la polarización del mercado de trabajo conlleva a corto plazo un aumento de la informatización restringida a ocupaciones claramente asociadas con baja cualificación y salarios modestos, es previsible una reasignación de trabajadores menos cualificados a tareas que requieren inteligencia creativa y habilidades sociales (Frey & Osborne, 2013: 44-45).

A pesar de los notables desafíos en materia de seguridad que plantean los entornos híbridos de cooperación humano-máquina (Pedrocchi et al., 2013: 2-8), los desarrollos en IA avanzada aplicada a la robótica industrial, combinados con una mayor destreza, precisión y capacidades sensoriales mejoradas, extenderán la dinámica de automatización a muchos tipos de tareas manuales no rutinarias. Diversos estudios e informes recientes aportan conclusiones inequívocamente pesimistas sobre la evolución previsible de los nichos ocupacionales asequibles a la mayor parte de los trabajadores en las próximas décadas (Graetz & Michaels, 2015). Este fenómeno constituye una tendencia transversal, común a diversos sectores industriales y a yacimientos tradicionales de empleo masivo (Frey & Osborne, 2013).

Quizás resulte equívoca la referencia genérica a *servicios* cuando se pretende contribuir a una comprensión más ajustada de la amenaza potencial que suponen los actuales desarrollos en robótica industrial para la empleabilidad de muchos trabajadores. Los sistemas que experimentan una expansión más rápida incluyen robots que asisten en procedimientos médicos de diversa complejidad, sistemas logísticos utilizados en fábricas y vehículos aéreos no tripulados (popularmente conocidos como *drones*). Pero la IA en desarrollo tiene entre sus objetivos emular, automatizar y reemplazar algunas funciones humanas en máquinas controladas por ordenador, con capacidad para ver, escuchar y responder a preguntas (telemarketing y teleasistencia, p.ej.). En algunos casos, su diseño incluye algoritmos que consiguen extraer conclusiones no programadas y resolver problemas con estrategias novedosas (Markoff, 2009).

El desarrollo de vehículos autónomos

El desarrollo del software y de la electrónica necesaria para posibilitar la conducción autónoma de vehículos —sin asistencia humana y en condiciones reales de circulación— constituye un ejemplo interesante que ayuda a comprender el potencial alcanzado por la IA en desarrollo. En primer lugar porque son la última fase de un largo proceso en el que múltiples funciones de los conductores han sido asumidas por dispositivos automatizados, capaces de reaccionar en menor tiempo, con mayor precisión y garantías de seguridad que la mayoría de los conductores humanos. La automatización conseguida incluye, en muchos modelos comerciales, tareas como la frenada de emergencia, maniobras de aparcamiento y, en vías rápidas, control de la conducción manteniendo la velocidad de cruce, el carril adecuado y la distancia con otros vehículos.

Esas tareas requieren sistemas avanzados para integrar la información que proporcionan radares y sensores del propio vehículo, así como señales de orientación en tiempo real mediante GPS y mapas muy precisos.⁴ Pero la tendencia en el desarrollo de vehículos autónomos se orienta hacia el diseño de vehículos que no requieren asistencia alguna de conductor humano, sin volante ni pedales, puesto

que estarían dotados de sistemas completamente automatizados para responder adecuadamente incluso en condiciones de circulación muy adversas (un entorno urbano congestionado, p.ej.).

Aparte de instrumentos de medición precisos capaces de proporcionar toda la información requerida de posición y distancia a los obstáculos, los vehículos autónomos tienen que ir dotados de sistemas de IA avanzada capaces de combinar la información de sensores y cámaras o sistemas de reconocimiento de objetos en la calzada con mapas en 3D del entorno. Estos sistemas de IA deben simular los procesos perceptivos y de toma de decisiones de un conductor humano, en coordinación con el control del volante y de los sistemas de frenada o aceleración, incorporando la capa de información que aportan las señales, indicadores y luces que regulan el tráfico.

Pero el aspecto más novedoso atañe al tipo de algoritmos que requieren, puesto que no se trata de meros conjuntos de reglas tipo *si-entonces* (aunque sean miles o millones de ellas) extraídos de un manual o teoría de la conducción. Los vehículos autónomos son posibles gracias a la combinación de aprendizaje-máquina (*machine learning*) y algoritmos sofisticados aplicados al reconocimiento de patrones, cuyos resultados refuerzan o descartan ciertos elementos antes de incorporar a su estructura aquellos que proporcionan resultados más precisos.¹

En lugar de dotar a estos vehículos con sistemas cerrados de análisis capaces de proporcionar la acción correcta para cada situación (algo difícil de conseguir incluso en entornos relativamente estables y homogéneos), los ingenieros prefieren entrenar el software en situaciones de tráfico muy diversas y especificar de modo inductivo la acción correcta para cada situación, de modo que el sistema de IA encuentre por sí mismo la configuración óptima de parámetros internos y de reglas lógicas para producir correctamente las señales de control que esas situaciones u otras novedosas requieren (Anderson et al., 2014: 58-66). Un sistema de IA aplicado a la conducción autónoma es capaz de incorporar la experiencia de todos los vehículos, en sus múltiples situaciones y entornos de conducción, en un proceso continuo de mejora y sofisticación. No se trata, por tanto, de emular el funcionamiento de los mejores conductores humanos en sus entornos habituales, sino de algo totalmente diferente, puesto que los sistemas de IA pueden alimentarse con datos de una complejidad que supera la capacidad de integración de los seres humanos y extraídos de entornos de riesgo a los que la mayoría de los conductores nunca se expondría.

A pesar de los desarrollos tecnológicos pendientes, fabricantes como Toyota, Nissan y Ford esperan poder comercializar sus primeros vehículos autónomos en 2020; Audi y Tesla adelantan planes similares con algunos modelos a 2017-2018, aunque los aspectos reguladores podrían retrasar 2-3 años la llegada al mercado; Jaguar y Land-Rover sugieren el año 2024; y Daimler considera más realista 2025. Las asociaciones de entidades aseguradoras consideran muy probable que sea entre 2028 y 2032 cuando la tecnología esté madura como para ser incorporada en vehículos de producción masiva; y el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) estima que el 75% de los vehículos serán autónomos en 2040, convirtiéndose en el modo de transporte más eficiente.⁵

El desarrollo de vehículos dotados de sistemas de conducción autónoma proporciona un terreno interesante para anticipar escenarios muy complejos en lo que atañe a la evolución inevitable del

¹ Véase “Top misconceptions of autonomous cars and self-driving vehicles”, disponible en http://www.driverless-future.com/?page_id=774#programming-model; también “Google’s Self-Driving Cars Aren’t as Good as Humans—Yet”, disponible en <http://www.wired.com/2016/01/google-autonomous-vehicles-human-intervention/>.

marco regulador. La atribución de responsabilidad –en caso de fallo, emergencia o actuación no prevista en el código del dispositivo– tendrá siempre a seres humanos o a instituciones (fabricante) al final de la cadena. Pero es interesante explorar la posible transformación de prácticas, culturas de trabajo y procesos de atención que sustentan todas las actividades actuales en las que el control de los sistemas de transporte depende directamente de seres humanos, cuando se aproxima un escenario donde medios masivos de transporte totalmente automatizados pueden ofertarse como un servicio más que como un producto (Anderson et al., 2014: 23).

Si en los medios de transporte convencionales se producen miles de accidentes al día, con un alto coste en vidas humanas que en muchas ocasiones originan procesos onerosos de identificación de responsables y evitación o reducción de indemnizaciones,⁶ cabe imaginar lo que ocurriría en un contexto socio-técnico donde más del 50% de los desplazamientos se produjeran en vehículos autónomos.

Los estudios realizados sobre el colectivo de pilotos de aviación quizás sirvan de orientación, puesto que las aeronaves llevan años funcionando con sistemas totalmente automatizados de navegación –aunque en entornos más simples y homogéneos, asistidos por sistemas de control de tránsito aéreo para la orientación y coordinación con otras aeronaves– que permanecen activos durante la mayor parte de los trayectos en los que operan (Martinussen & Hunter, 2010: chap. 8). Esta ventaja incuestionable parece haber reducido notablemente la habilidad de los pilotos para actuar con rapidez y acierto en situaciones de emergencia.⁷ Lo mismo podría ocurrir –con una incidencia a escala mayor– en vehículos capaces de conducción autónoma pero con un diseño que no permita prescindir del sujeto humano ante un brusco cambio en las condiciones de tráfico o situaciones de emergencia. Autores como Nicholas Carr extienden a otros muchos dominios de actividad los efectos negativos sobre las habilidades humanas producidos por desarrollos tecnológicos destinados a simplificar o automatizar operaciones de cierta complejidad (Carr, 2010).

Pese a todo, el incentivo mayor para el empleo generalizado de vehículos autónomos tiene mucho que ver con la reducción de daños producidos por error humano. La incidencia de los factores humanos en los peores accidentes es un fenómeno bien documentado en la aviación civil y militar (Gilbey & Hill, 2012), donde al menos el 53% de los peores accidentes se producen por decisiones erróneas de pilotos (Martinussen & Hunter, 2010: 184). Con cifras aproximadas entre países como Australia (84%) y Estados Unidos (70%), los peores accidentes se atribuyen a factores humanos que dependen sobre todo de las habilidades y, en menor medida, de otros factores psicológicos involucrados en decisiones erróneas (33% y 29,2%, respectivamente).⁸

Riesgos específicos de la IA avanzada en robots autónomos de uso militar

Para reducir al mínimo los posibles sesgos que incrementan la posibilidad de error en los servicios de inteligencia militar, los expertos valoran cualquier herramienta que pueda servir de apoyo en los procesos de análisis y toma de decisiones. La IA avanzada puede resultar de gran ayuda en la generación de múltiples escenarios según evolucionan los parámetros iniciales, en la validación de indicadores y en la detección de sesgos frecuentes en el razonamiento humano (Heuer & Pherson, 2015). Uno de los desafíos consiste en administrar la información disponible sobre acciones militares

y maniobras del enemigo para detectar patrones de actuación que permitan adelantarse a posibles ataques (Stanton et al., 2015).

El empleo de tecnología de doble uso susceptible de ampliar el potencial de ciertas capacidades funcionales y cognitivas en humanos o emular su desarrollo en androides ha sido bien explorado en la literatura y el cine distópico de ciencia ficción.⁹ En el último lustro, además, ha despertado el interés de la academia y de los expertos en ética aplicada. Los escenarios actuales de conflicto incrementan los incentivos en el ámbito de la I+D militar para desarrollar robots autónomos, capaces de seleccionar objetivos y disparar o adoptar decisiones con consecuencias letales sin supervisión humana, fundamentalmente porque reducen el desgaste político asociado con la pérdida de vidas humanas.¹⁰

Sobre este trasfondo se delimitan con mayor claridad los riesgos que plantea el desarrollo de IA avanzada para robótica militar. Los algoritmos de reconocimiento de patrones pueden resultar igualmente efectivos tanto sobre caracteres borrosos de un texto antiguo como sobre imágenes de baja calidad en una pantalla conectada a un sensor o cámara de infrarrojos en un puesto de vigilancia fronteriza.¹¹ No es previsible que los países con una infraestructura científico-tecnológica se descuelguen voluntariamente de la carrera para diseñar robots letales autónomos (LAR) que puedan desplegarse en múltiples escenarios de conflicto.

Agencias de investigación militar como la estadounidense DARPA (*Agencia de Proyectos de Investigación Avanzados de Defensa*) contemplan, entre otras aplicaciones, el uso de exoesqueletos para aumentar la fuerza humana, previsiblemente en combinación con dispositivos que mejoren las representaciones visuales para la toma de decisiones. Pero sus proyectos destinados al diseño de robots autónomos utilizables en tareas de rescate o defensa captan gran parte de la financiación disponible.¹²

El doble uso potencial de la IA resulta obvio cuando el objetivo es dotar al personal militar de dispositivos que amplíen, precisen o mejoren sus capacidades de orientación, de detección de amenazas y de reacción en los escenarios de mayor riesgo. El margen de seguridad que podrían aportar estos dispositivos sería previsiblemente mayor si, además, pudieran configurarse para activar por sí solos los sistemas de neutralización de objetivos en condiciones desfavorables, donde el sujeto humano se desorienta con facilidad y pierde precisión o capacidad de reacción. Algunas tecnologías recientes permiten corregir el rumbo del proyectil tras el disparo y pueden mejorar notablemente el alcance y precisión de francotiradores militares, al tiempo que reducen el riesgo de ser localizados.¹³

La tecnología que hace posible el diseño de armas autónomas, capaces de seleccionar y atacar objetivos sin intervención humana, está ya disponible, aunque requiera integrar conocimientos e ingeniería de campos muy diversos. Pero esa capacidad está sobradamente demostrada por parte de las instituciones que sustentan la I+D civil y militar en los países industrializados. Tras más de una década en la que ejércitos de varios países han estado empleando vehículos aéreos no tripulados (UAV, utilizado aquí como sinónimo de RPAS: *Remotely Piloted Aircraft Systems*), probablemente a nadie sorprendería ya la presentación –en una feria de equipamiento militar, p.ej.– de cuadricópteros armados dotados de sensores y sistemas de reconocimiento con IA avanzada, capaces de buscar y eliminar a objetivos humanos que cumplan con ciertos criterios pre-definidos. Los obstáculos imaginables parecen más bien de tipo legal que ingenieril o tecnológico.¹⁴

Un aspecto crucial es el elevado número de accidentes que se producen en los UAV (*drones*) operados por el ejército y la fuerza aérea estadounidenses en lugares como Irak, Siria, Afganistán, Somalia, Yemen, Libia, Malí y Camerún, entre otros. La demanda de este tipo de cobertura es tal que no se puede satisfacer con el sistema actual de reclutamiento y adiestramiento de pilotos para ocuparse del control remoto, siendo habitual el recurso a compañías privadas.¹⁵ Descartados problemas mecánicos y fallos de comunicación, la escasez de personal adecuado y de recursos complementarios parece ser un factor decisivo en el elevado número de accidentes producidos en 2015 (20 de ellos involucraron a los modelos más costosos y de mayor tamaño, el ya antiguo *MQ-1 Predator*, del cual se han accidentado más de la mitad; y el *MQ-9 Reaper*, de mayor alcance y con mayor capacidad para equipar armas).¹⁶

Según diversos registros, desde 2001 se han producido 237 accidentes graves (*clase A*) con aviones militares no tripulados que destruyeron la aeronave o causaron al menos 2 millones de dólares en daños.¹⁷ En el mismo período, al menos 47 de ellos (puesto que se desconocen los detalles de otros muchos incidentes *clasificados*) se han producido en territorio estadounidense. Los riesgos son aún mayores en ciertas zonas donde se usan las mismas instalaciones para operaciones militares y para control aéreo de aviación civil (la mayor parte de los UAV de uso militar no disponen de sistema de radar anticolidión).¹⁸

Conclusión

Sobre la IA avanzada se proyectan ahora expectativas y temores similares a los suscitados durante las dos últimas décadas en el debate social acerca de algunas biotecnologías y sus posibles aplicaciones. Otros escenarios y dominios de problemas analizados recientemente a propósito de la biología sintética (Menuz, Hurlimann, & Godard, 2013; Schmidt, Meyer, & Cserer, 2013) podrían resultar útiles para contextualizar los estudios de prospectiva ética, social y legal referidos a la IA avanzada, puesto que pueden orientar sobre obstáculos similares en el proceso de regulación y control social.

El debate en curso sobre la convergencia de tecnologías con fines de mejora humana proporciona también abundante literatura (Kurzweil, 2005; Bostrom & Yudkowsky, 2013; Holland, 2010; Khushf, 2007; Stock, 2002; Hughes, 2012; Zylinska, 2010, entre otros) que puede resultar útil para analizar la problemática de la IA avanzada de doble uso y diversos escenarios de riesgo.

Dada la orientación especulativa y la falta de distancia crítica que es frecuente encontrar en algunos trabajos de *visionarios* o *gurús* de la tecnología, merecen un reconocimiento especial aportaciones mucho más rigurosas como las de Nicholas Agar en *Humanity's end* (Agar, 2010: 77-81) y la de Nick Bostrom en *Superintelligence* (Bostrom, 2014). Evitando cualquier enfoque alarmista, aportan elementos para contribuir a un debate informado y riguroso sobre escenarios de riesgos verosímiles, sin minimizar las dificultades para articular un marco regulador que facilite el apoyo social a las aplicaciones con fines inequívocamente beneficiosos.

El desarrollo de tecnología militar con IA avanzada plantea riesgos específicos (éticos, políticos y estratégicos) que no pueden quedar al margen del escrutinio y del debate públicos. Aunque el uso de robots autónomos puede reducir las bajas en los nuevos escenarios de conflicto, también puede

contribuir a que estos se originen con mayor facilidad. Es necesario tomar en serio el riesgo de una nueva carrera armamentística, donde las armas autónomas sean el medio preferido para neutralizar o destruir objetivos. Una vez disponibles, será difícil evitar que caigan en las manos equivocadas (Leveringhaus & Giacca, 2014).

Referencias

AA.VV.

2015 *Autonomous Weapons: An Open Letter From AI & Robotics Researchers*. Disponible en <http://futureoflife.org/open-letter-autonomous-weapons/>

Agar, N.

2010 *Humanity 's End. Why We Should Reject Radical Enhancement*. Cambridge, Massachusetts / London, England: A Bradford Book - The MIT Press.

Anderson, J. M., Kalra, N., Stanley, K. D., Sorensen, P., Samaras, C., & Oluwatola, O. A.

2014 *Autonomous Vehicle Technology: a Guide for Policymakers*. Santa Monica, California: RAND Corporation.

Baum, S. D., Goertzel, B., & Goertzel, T. G.

2011 "How long until human-level AI? Results from an expert assessment". *Technological Forecasting and Social Change*, 78(1), 185–195.

doi:<http://dx.doi.org/10.1016/j.techfore.2010.09.006>

Bell, L.

2015 "Humans vs robots: Driverless cars are safer than human driven vehicles", *The Inquirer*, 23/09/2015. Disponible en <http://www.theinquirer.net/inquirer/feature/2426988/humans-vs-robots-driverless-cars-are-safer-than-human-driven-vehicles>

Bilton, N.

2014 "Artificial Intelligence as a Threat". *Nytimes.com*, Nov., 5. Disponible en http://www.nytimes.com/2014/11/06/fashion/artificial-intelligence-as-a-threat.html?_r=0

Bonnefon, J.-F., Shariff, A., & Rahwan, I.

2015 "Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?" Disponible en <http://arxiv.org/abs/1510.03346>

Bostrom, N.

2002 "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, 9(March 2002), 1–30. Disponible en <http://www.jetpress.org/volume9/risks.html>

Bostrom, N.

2014 *Superintelligence: Paths, Dangers, Strategies*. Oxford, United Kingdom: Oxford University Press.

Bostrom, N., & Yudkowsky, E.

2013 "The ethics of artificial intelligence. In *Cambridge Handbook of Artificial Intelligence*" (chap. 15).

Carr, N.

2010 *The shallows : what the Internet is doing to our brains*. New York: W.W. Norton & Company.

Chan, S.

- 2008 "Humanity 2.0? Enhancement, evolution and the possible futures of humanity". *EMBO Reports*, 9(SPECIAL ISSUE), S70–S74.
- Frey, C. B., & Osborne, M. A.
2013 *The future of employment: how susceptible are jobs to computerisation?* Disponible en http://arche.depotoi.re/autoblogs/wwwinternetactunet_8a3fe3331e0ad7327e18d9fe6ec3f0ad04dcea58/media/3722fa7d.The_Future_of_Employment.pdf
- Gilbey, A., & Hill, S.
2012 "Confirmation Bias in General Aviation Lost Procedures". *Applied Cognitive Psychology*, 26(5), 785–795. doi:[10.1002/acp.2860](https://doi.org/10.1002/acp.2860)
- Goertzel, B., & Pennachin, C.
2007 *Artificial general intelligence*. Berlin: Springer.
- Graetz, G., & Michaels, G.
2015 *Robots at Work*. London. Disponible en <http://cep.lse.ac.uk/pubs/download/dp1335.pdf>
- Heuer, R., & Pherson, R.
2015 *Structured analytic techniques for intelligence analysis*. Washington, DC: Washington, DC: CQ Press.
- Holland, S.
2010 "Human Enhancement" (Edited by Julian Savulescu and Nick Bostrom). *Analysis*, 70(2), 398–401. doi:[10.1093/analys/ang011](https://doi.org/10.1093/analys/ang011)
- Hughes, J. J.
2012 "The politics of transhumanism and the techno-millennial imagination, 1626-2030". *Zygon*, 47, 757–776. doi:[10.1111/j.1467-9744.2012.01289.x](https://doi.org/10.1111/j.1467-9744.2012.01289.x)
- Khushf, G.
2007 "Open questions in the ethics of convergence". *The Journal of Medicine and Philosophy*, 32(3), 299–310. doi:[10.1080/03605310701397057](https://doi.org/10.1080/03605310701397057)
- Kurzweil, R.
2005 *The Singularity is Near. When Humans Transcend Biology*. New York: Viking.
- Leveringhaus, A., & Giacca, G.
2014 *Robo-Wars: The Regulation of Robotic Weapons*. Oxford. Disponible en <http://www.oxfordmartin.ox.ac.uk/downloads/briefings/Robo-Wars.pdf>
- Marcus, G.
2013 "Why We Should Think About the Threat of Artificial Intelligence". Disponible en <http://www.newyorker.com/tech/elements/why-we-should-think-about-the-threat-of-artificial-intelligence>
- Markoff, J.
2009 "The coming superbrain". Disponible en <http://www.nytimes.com/2009/05/24/weekinreview/24markoff.html>
- Martinussen, M., & Hunter, D.
2010 *Aviation psychology and human factors*. CRC Press/Taylor & Francis.
- Menuz, V., Hurlimann, T., & Godard, B.
2013 "Is Human Enhancement also a Personal Matter?" *Science and Engineering Ethics*, 19(1), 161–177. doi:<http://dx.doi.org/10.1007/s11948-011-9294-y>
- Pedrocchi, N., Vicentini, F., Matteo, M., & Molinari, L.
2013 "Safe Human-Robot Cooperation in an Industrial Environment". *International Journal of Advanced Robotic Systems*, 10(27), 1–13. doi:[10.5772/53939](https://doi.org/10.5772/53939)
- Schmidt, M., Meyer, A., & Cserer, A.

- 2013 "The Bio:Fiction film festival: Sensing how a debate about synthetic biology might evolve". *Public Understanding of Science (Bristol, England)*. doi:[10.1177/0963662513503772](https://doi.org/10.1177/0963662513503772)
- Sharma, R. A., K, P. S., & Jawahar, C.
2015 "Fine-Grain Annotation of Cricket Videos. Multimedia; Computation and Language; Computer Vision and Pattern Recognition". Disponible en <http://arxiv.org/abs/1511.07607>
- Stanton, A., Thart, A., Jain, A., Vyas, P., Chatterjee, A., & Shakarian, P.
2015 "Mining for Causal Relationships: A Data-Driven Study of the Islamic State". doi:[10.1145/2783258.2788591](https://doi.org/10.1145/2783258.2788591)
- Stock, G.
2002 "Redesigning Humans". Disponible en <http://www.gregorystock.net/redesigninghumans.asp>
- Vinge, V.
1993 "The Coming Technological Singularity: How to Survive in the Post-Human Era". *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace. Proceedings of a symposium cosponsored by the NASA Lewis Research Center and the Ohio Aerospace Institute. Westlake, Ohio, March 30-31, 1993*. Disponible en <http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855.pdf#page=23>
- Zylinska, J.
2010 "Playing God, Playing Adam: The Politics and Ethics of Enhancement". *Journal of Bioethical Inquiry*, 7(2), 149–161. doi:[10.1007/s11673-010-9223-9](https://doi.org/10.1007/s11673-010-9223-9)

Notas

¹ Véase, p.ej. Bell, 2015. Para hacerse una idea del tipo de dilemas involucrados en el diseño de coches dotados con sistemas avanzados de conducción autónoma, véase Bonnefon, Shariff, & Rahwan, 2015.

² “[...] we are on the edge of change comparable to the rise of human life on Earth. The precise cause of this change is the imminent creation by technology of entities with greater than human intelligence. There are several means by which science may achieve this breakthrough [...]: Large computer networks (and their associated users) may "wake up" as a superhumanly intelligent entity. Computer/human interfaces may become so intimate that users may reasonably be considered superhumanly intelligent.” Cfr. Vinge, 1993: 12.

³ Y criticado de modo consistente por Nicholas Agar, entre otros (Agar, 2010).

⁴ Véase <http://www.wired.com/2014/12/nokia-here-autonomous-car-maps/>.

⁵ Véase: http://www.driverless-future.com/?page_id=384.

⁶ Un lamentable ejemplo de cómo las instituciones responsables del transporte público (con frecuencia asistidas por las compañías aseguradoras) emplean cualquier tipo de artimañas para evitar, diluir o minimizar su responsabilidad incluso en accidentes con decenas de víctimas lo tenemos en la gestión del accidente de metro ocurrido en la comunidad autónoma de Valencia (España), que costó la vida a 43 personas y heridas de diversa gravedad a 47 más, el 3 de julio de 2006. Aunque inicialmente se determinó que se produjo por un exceso de velocidad, la investigación posterior puso de manifiesto las carencias en materia de seguridad que habrían podido solucionarse con equipos no especialmente caros, disponibles en otras vías similares y capaces de interaccionar automáticamente con el sistema de frenado si la reacción humana no se produce. El caso terminó cerrándose sin concretar responsables políticos ni técnicos (*causas fortuitas*), hasta que un grupo de periodistas mostraron en diversos medios y en un programa de televisión de la cadena de cobertura nacional *La Sexta* cómo fueron aleccionados los testigos antes de declarar y sobornados los familiares de las víctimas con ofertas de trabajo o dinero a cambio de silencio y renuncia a reclamar en los tribunales una compensación más justa. Ante las evidencias de manipulación informativa, ocultación de pruebas e influencia indebida sobre testigos y técnicos, el Comité de Peticiones del Parlamento Europeo aceptó ocuparse del caso y el 13 de enero de 2015 comunicó a la Asociación de Víctimas del accidente que iba a investigar la manipulación política e informativa del accidente. Véase <http://www.elmundo.es/comunidad-valenciana/2015/07/03/55966a7e46163fde298b457b.html>;

http://ccaa.elpais.com/ccaa/2013/04/11/valencia/1365694355_535232.html;
http://www.lasexta.com/programas/salvados/noticias/accidente-metro-valencia-historia-tragedia-silenciada_2013042500132.html.

⁷ Véanse los estudios de la NASA y de la Administración Federal de Aviación de Estados Unidos (FAA) sobre el *efecto de una alta automatización en cabina*, cuyas conclusiones recogen estos enlaces:

<http://www.xatakaciencia.com/tecnologia/los-pilotos-de-avion-cada-vez-lo-hacen-peor-por-culpa-de-las-maquinas>;
<http://lat.wsj.com/articles/SB10224152129988824652004580311401269113906>;
<http://avionypiloto.es/secciones/instrumentos/efecto-de-una-alta-automatizacion-en-cabina/>.

⁸ Véase Martinussen & Hunter, 2010: 186.

⁹ Entre las más logradas, cabe citar *Metrópolis* (Fritz Lang, 1927); *2001: Una odisea del espacio* (Stanley Kubrick, 1968); *Blade Runner* (Ridley Scott, 1982); *Matrix* (A. Wachowski y L. Wachowski, 1999); *Ex machina* (Alex Garland, 2015).

¹⁰ Véase <http://www.bbc.com/future/story/20121206-moral-machines-killer-questions>.

¹¹ En la frontera de Corea del Sur ya opera un robot centinela, llamado SGR-1, equipado con sensores de calor y movimiento y capacidad de identificar objetivos potenciales a más de dos millas de distancia. Aunque configurado para actuar bajo control humano, es obvio que existe la capacidad tecnológica para que entre en funcionamiento sin esta limitación. Véase <http://www.bbc.com/news/technology-32334568>. Otro sistema similar, la torreta Super aEGIS II, con un alcance de cuatro kilómetros y una ametralladora de gran calibre capaz de detener vehículos pesados, se encuentra instalado en diversas localidades de Oriente Medio y en instalaciones industriales (aeropuertos, plantas de energía, oleoductos) de Emiratos Árabes Unidos, Abu Dhabi y Qatar, además de en muchas bases aéreas militares. Con un coste aproximado de 40 millones de dólares, este sistema fue diseñado inicialmente para actuar sin necesidad de intervención humana y modificado después para depender del control humano, a petición de algunos clientes que requerían una salvaguarda adicional. Véase <http://www.bbc.com/future/story/20150715-killer-robots-the-soldiers-that-never-sleep>.

¹² Véase un listado de proyectos recientes en <https://en.wikipedia.org/wiki/DARPA>.

¹³ Véase <http://www.economist.com/news/science-and-technology/21678766-new-technology-improving-military-sharpshooters-range-and-accuracy-enemy>.

¹⁴ Véase <http://www.theguardian.com/technology/2015/jul/27/musk-wozniak-hawking-ban-ai-autonomous-weapons>;
<http://www.bbc.com/news/technology-33686581>; <http://www.bbc.com/news/technology-33629465>.

¹⁵ Véase https://www.washingtonpost.com/world/national-security/how-crashing-drones-are-exposing-secrets-about-us-war-operations/2015/03/24/e89ed940-d197-11e4-8fce-3941fc548f1c_story.html.

¹⁶ Véase <https://www.washingtonpost.com/news/checkpoint/wp/2016/01/19/more-u-s-military-drones-are-crashing-than-ever-as-new-problems-emerge/>.

¹⁷ Véase <http://www.washingtonpost.com/sf/investigative/2014/06/20/when-drones-fall-from-the-sky/>.

¹⁸ Véase https://www.washingtonpost.com/world/national-security/miscues-at-us-counterterrorism-base-put-aircraft-in-danger-documents-show/2015/04/30/39038d5a-e9bb-11e4-9a6a-c1ab95a0600b_story.html;
<http://www.washingtonpost.com/sf/investigative/2014/06/20/when-drones-fall-from-the-sky/>.