La gran pantalla como laboratorio y espejo para la roboética

José Miguel Biscaia Fernández*

Universidad Europea de Madrid

Recibido 13 de diciembre de 2021; aprobado 10 de junio de 2022

Resumen

La roboética es una rama de la ética aplicada interesada por los desafíos éticos y sociales de la robótica y la inteligencia artificial. En el presente ensayo utilizo el cine de ciencia-ficción para reflexionar sobre esta disciplina, habida cuenta de la doble función de este género como "espejo" de la vanguardia tecno-científica y como "laboratorio" para la especulación futurista y la influencia social. De entre los muchos temas ético-sociales relacionados con estas tecnologías, centro los esfuerzos en analizar (1) los riesgos y medidas de seguridad asociadas, (2) los derechos y deberes inherentes y (3) las consecuencias relacionales de la díada humano-máquina. Como conclusión, cabe destacar que el cine de ciencia-ficción se ha aproximado sin disimulo (con diferente grado de plausibilidad) a todas estas cuestiones, las cuales se antojan de máxima urgencia si consideramos el impacto bio-psico-social que la robótica y la cibernética ya tienen (y tendrán) en nuestras vidas.

Palabras Clave: roboética | ciencia-ficción | robótica | inteligencia artificial | singularidad tecnológica | transhumanismo

The Big Screen as a Laboratory and Mirror for Roboethics

Abstract

Roboethics is a branch of applied ethics interested in the ethical and social challenges of robotics and artificial intelligence. In this essay I use the science-fiction cinema to reflect in this discipline, taking into account the dual function of this genre as "mirror" of the techno-scientific avant-garde and as a "laboratory" for futuristic speculation and social influence. Among the ethical-social issues related to these technologies, I focus my efforts on analysing (1) the risk and associated security measures, (2) the inherent rights and duties, and (3) the relational consequences of the human-machine dyad. In conclusion, it should be noted that the science-fiction cinema has approached without dissimulation (with a different degree of plausibility) all these questions, which seem of the utmost urgency if we consider the bio-psycho-social impact that robotics and cybernetics already have (and will have) in our lives.

Key words: roboethics | science-fiction | robotics | artificial intelligence | technological singularity | transhumanism

Introducción

El cine de ciencia-ficción puede ser un excelente escenario de exploración al respecto de la vanguardia tecnológica. Además de dedicarse a la especulación científica, las obras fílmicas suelen plantear controvertidos escenarios sociales que invitan a la reflexión ética y filosófica. En el caso de la robótica y la inteligencia artificial (IA), dicha reflexión se enmarcaría dentro de la roboética, novedosa subdisciplina de la ética aplicada que nace como consecuencia de la preocupación frente a los riesgos y desafíos bio-psico-sociales que una tecnología tan sobrecogedora plantea (Langman et al., 2021). Por otro lado, como medio de masas global, el séptimo arte tiene la capacidad de ejercer una poderosa influencia social y cultural (Mangot y Vasantmeghna, 2017), por lo que sus propuestas tienen un espíritu pragmático susceptible de ser sometido a análisis y crítica (Biscaia y Mohedano, 2018).

El presente estudio se subdivide en tres grandes apartados: en primer lugar, (1) se describirán sucintamente las bases epistemológicas sobre las que descansa la roboética. A continuación, (2) se discutirá al respecto del sentido y la conveniencia de utilizar el cine de ciencia-ficción como medio y como fin en este ensayo. Para concluir, se analizarán cinematográficamente tres aspectos de la agenda roboética: (3) los riesgos (y mecanismos de seguridad vinculados) de la tecnología cibernética y de la robótica; (4) los derechos y deberes asociados a la implantación de la IA y la robótica; y (5) las consecuencias ético-sociales del advenimiento de una IA interactiva.

Así pues, bajo la hipótesis de que el cine de ciencia-ficción ha especulado abiertamente sobre estas cues-

* josemiguel.biscaia@universidadeuropea.es

tiones centrales para la roboética, el presente estudio tiene por objetivo:

1. Reflexionar sobre el modo en el que lo ha hecho (en cuanto a plausibilidad y significado).
2. Analizar los temas más candentes de la ética aplicada a la robótica y la IA apuntados más arriba.

Por supuesto, teniendo en cuenta la amplitud de aspectos de la agenda roboética, el presente ensayo no tiene la aspiración de ser un estudio cerrado y total. Lo pretendido aquí es, simplemente, señalar los elementos más relevantes de algunos de los escenarios intelectuales que mayor interés han suscitado por sus implicaciones pragmáticas presentes y futuras.

1. Roboética: Ética aplicada a los desarrollos robóticos y en IA

La roboética es una subdisciplina académica híbrida y reciente, con un corpus de conocimiento todavía en construcción, que se interesa por los límites éticos, legales y sociales en el diseño, desarrollo y utilización no sólo de la robótica sino también de la IA (Langman et al., 2021)[1]. En palabras de Carme Torras, investigadora española del Instituto de Robótica e Informática Industrial (IRI), la roboética centra sus esfuerzos en tres grandes líneas de acción: "adaptar la ética humana a la robótica, incrustar un código moral en los robots mismos, y pensar qué ética surgiría de una posible conciencia futura de los robots"[2].

Hay cierto consenso establecido de que el término "roboética" se utiliza por primera vez en un contexto académico en 2004, en la Escuela de robótica de San Remo (Italia), en el marco del primer simposio internacional sobre la materia titulado Roboethics (Veruggio, 2005). No obstante, entre sus antecedentes inmediatos (y valedores actuales) se encuentran varias áreas de conocimiento que ya mostraron su interés por las implicaciones filosóficas, éticas y sociales de la robótica y de la IA: además de la filosofía de la mente y de las ciencias cognitivas (desde una perspectiva más epistémica y ontológica), quizá quien lo ha hecho con más fuerza haya sido el transhumanismo tecnológico (con su visión claramente pragmática, entre la técnica y la filosofía práctica, fundamentalmente en forma de ética aplicada). Dicho movimiento se caracteriza por proponer una mejora en la condición humana basada en la utilización de la tecnología disponible (Biscaia, 2021b; Diéguez, 2017). Bajo esta concepción, el uso de la robótica y de la IA supondrá un progreso socio-económico gracias a novedosas máquinas industriales y vehículos autónomos, al reconocimiento de voz, la visión artificial o la planificación y el aprendizaje autónomo, entre otros ingenios; supondrá, igualmente, un considerable avance en el tratamiento de patologías o en la mejora de determinadas capacidades físicas y psicológicas del ser humano; incluso, implicaría la creación de una suerte de híbrido humano-máquina, el conocido como *Homo ciborg* (Biscaia, 2021a), o, más especulativo aun, supondrá tal vez el advenimiento de la denominada como "singularidad tecnológica", lo cual implica la creación de una IA (robótica o no) con capacidades cognitivas humanas o incluso superiores a las nuestras[3] (Vinge, 1993).

La llegada de una super-IA es propuesta por el científico norteamericano Raymond Kurzweil[4], quien cree que se producirá cuando el primer sistema super-inteligente sea capaz de perfeccionarse a sí mismo o de fabricar otros iguales[5]. Cree, además, que en 2029 una máquina logrará pasar el test de Turing[6], y que la singularidad llegará en torno a 2045, momento en el que entraríamos en una civilización postbiológica (Kurzweil, 2012). En definitiva, aunque muchos científicos ortodoxos son reacios a admitir estas ideas futuristas (Stone y Hirsh, 2006), pues parece altamente improbable dadas las actuales dificultades técnicas (Grosz et at., 2016), el experto en transhumanismo, Antonio Diéguez (2017), considera que "no hay razones irrefutables para pensar que la creación de una super-inteligencia artificial no es ni será jamás posible" (p. 77). En esta misma línea, el científico Max Tegmark (2018) propone un escenario al que denomina "Vida 3.0", en el que se dará un crecimiento sin precedentes en la tecnología robótica y cibernética. Como puede apreciarse, no existe consenso entre los expertos al respecto del advenimiento de la singularidad, debido a las dificultades tecnológicas y conceptuales de la cuestión (Biscaia, 2021a).

De las muchas áreas por las que se ha interesado la roboética, en el presente ensayo me centraré en el análisis de las tres que a mi juicio suscitan una mayor urgencia intelectual y pragmática; en relación a desarrollos tecnológicos ya presentes, aunque, también, con respecto a lo que la especulación científica y filosófica propone como plausible (aun siendo, en algunos casos, ciertamente improbable), a saber:

- Los riesgos en el desarrollo e implantación de la robótica y la IA y las medidas adoptadas para el uso seguro y éticamente comprometido de estas tecnologías.
- Los derechos y deberes derivados de la implementación de la IA y la robótica en nuestra sociedad.
- Los nuevos escenarios ético-sociales que surgirán de la relación entre humanos y máquinas interactivas e "inteligentes".

2. El cine de ciencia-ficción: laboratorio y espejo para el desarrollo robótico-cibernético

El cine de ciencia-ficción es un magnífico representante (podría decirse que, en muchas ocasiones, un fiel "espejo") de algunas de las principales ideas desarrolladas por la vanguardia tecno-científica, el transhumanismo y la roboética (Biscaia y Mohedano, 2020, 2021). Incluso, ha supuesto en algunos casos su vanguardia o, al menos, su horizonte utópico o distópico (como si fuera un "laboratorio", extensión de aquellos en los que se crean de facto estas tecnologías). Ofrece, en todo caso, un mundo posible o alternativo, pues, como indican Bassa y Freixas (1993), las obras de ciencia-ficción implican "una irrupción de lo imaginario en lo real utilizando la ciencia como coartada de la fantasía, provocando la transformación del verosímil en un referente tanto eminente como pretendidamente científico que cumplirá, en ambos supuestos, un rol mítico" (p. 31). Con sus enormes capacidades técnicas y económicas, aunque, también, gracias a sus licencias artísticas, el séptimo arte ha especulado sobre lo que la ciencia por el momento sólo promete, jugando incluso en ocasiones con lo contrafáctico. Por ello, asomarse a la gran pantalla puede ser un interesante ejercicio de análisis científico, filosófico y social al respecto de los temas tratados en este estudio.

La ciencia-ficción es un género artístico, y aunque su misión no es ofrecer una precisión absoluta y fiel de la realidad científico-tecnológica mostrada, sí que pretende en todo caso que sus propuestas sean verosímiles, lo más aproximadas a la vanguardia tecnológica. Son varios los ejemplos de tecnologías revolucionarias que, antes de existir, fueron ya mostradas en la gran pantalla, a modo –como se viene indicando– de auténtico "laboratorio" tecno-virtual. En este sentido, Alonso y Ardoz (2003) reconocen que "no es una iniciativa descabellada pensar que se produzca (…) un cine de ciencia-ficción que busque la escenificación realista de un futuro inmediato" (p. 157). Con la siguiente muestra se ponen a prueba las palabras del citado autor, así como la capacidad predictiva del séptimo arte: en los albores del cine, el director Georges Méliès avanzó los viajes espaciales con su *Viaje a la Luna* (1902). En *Metrópolis* (1927), del alemán Fritz Lang, aparecen los primeros robots cinematográficos. Por otro lado, el prestigioso autor François Truffaut anticipó la llegada de los auriculares en su distópica *Fahrenheit 451* (1966). Las video-llamadas, tan comunes en nuestros días, se muestran en la obra *2001, Una odisea del espacio* (1968), de Stanley Kubrick. *Engendro mecánico* (1977), del director Donald Cammell, mostraba algo similar al actual "internet de las cosas", así como algunas aplicaciones de las casas inteligentes contemporáneas. El teléfono móvil se prefigura en la obra *Star Trek* (1979), de Robert Wise. Por su parte, los coches voladores, similares a ciertos prototipos ya existentes, surcaban los cielos nebulosos de *Blade Runner* (1982), del galardonado Ridley Scott. En *The Terminator* (1984), de James Cameron, aparecen drones militares. Los hologramas 3D, que serán vanguardia en telecomunicaciones, ya fueron visionados en *La Guerra de las Galaxias* (1977), de George Lucas. Y Paul Verhoeven, en su conocida obra *Desafío total* (1990), anticipa los vehículos autónomos, desarrollados actualmente por diversas marcas.

Como medio de masas, el cine es (o puede llegar a ser), también, una poderosa herramienta de influencia y control social (Mangot y Vasantmeghna, 2017), por lo que el análisis cuantitativo y cualitativo de sus propuestas puede transcender el mero interés academicista o anecdótico. Benet (2004) reconoce un circuito de influencia ("círculo de la comunicación") que comienza con el cineasta (guionista, director, productor…), que sería el emisor; continua después con la propia película (y sus connotaciones) para terminar en el espectador (o receptor) y en la sociedad (que sería el contexto). Finalmente regresaría al cineasta, de donde partió todo, cerrándose así el círculo. Por su parte, Gutiérrez (2009) señala que "lo que somos y lo que creemos ser, lo que vemos y creemos ver se deduce de los medios. El consenso se establece allí, donde todos confluimos" (p. 67). Por otro lado, el investigador británico John Ziman (1994), quien es uno de los más destacados representantes de los estudios en Ciencia, Tecnología y Sociedad (o estudios STS, por su acrónimo en inglés), destaca la gran importancia de la inmersión de la tecno-ciencia en la cultura de masas (como el cine).

Prueba de la influencia social del séptimo arte la encontramos en numerosos estudios científicos, en los que se ha llegado a sugerir una conexión entre el visionado de determinadas películas y la posible aparición de hábitos sociales, algunos de ellos relacionados con la salud, especialmente entre la población más joven, como el tabaquismo adolescente (Dal, Stoolmiller y Sargent, 2012), la obesidad infantil y su estigmatización social (Throop et al., 2014) o la aparición de conductas violentas (Anderson et al., 2003). Además, desde un planteamiento más optimista se ha demostrado que ver determinadas películas puede tener una función divulgadora, formativa y ética (San Román, 2010), mejorando sobre todo el aprendizaje, por ejemplo, en las habilidades clínicas y humanísticas (en forma de trato hacia el paciente) entre los estudiantes de carreras biosanitarias como medicina o enfermería (Ogston-Tuck et al., 2016; Toye et al., 2015) o en la educación medioambiental (Stadler, 2017).

Así pues, el cine de ciencia-ficción que venimos describiendo nos servirá para: (1) explorar de una forma creativa algunos de los asuntos candentes en la discusión sobre la roboética y (2) reflexionar sobre el modo en el que el cine de fantasía y de ciencia-ficción se ha aproximado al área de estudio, en lo que al rigor y plausibilidad tecno-científica se refiere; también, en referencia al significado y sentido de las propuestas "tecno" que ofrece, sobre todo, debido a su gran capacidad de influencia social. Como puede apreciarse, el cine servirá respectivamente como medio y como fin en este ensayo.

3. Escenarios de peligrosidad y medidas de seguridad de la robótica y la IA

¿Existen riesgos objetivos con respecto al desarrollo y uso de estas tecnologías como para que tengamos que preocuparnos? La respuesta es un rotundo "¡desde luego!", si por ejemplo consideramos la petición formal para controlar su investigación por parte de un grupo de reconocidos investigadores pertenecientes al *Future of Life Institute*[7], o, también, si tenemos en cuenta las Directrices éticas para una IA fiable[8], desarrolladas por un grupo independiente de expertos de la Comisión Europea reunidos en junio de 2018, quienes presentaron las siguientes normas básicas respecto de estos desarrollos tecnológicos: (1) el respeto de la autonomía humana, (2) la prevención del daño, (3) la equidad y (4) la explicabilidad de sus procedimientos. Dichas amenazas podrían clasificarse en diferentes categorías de posibilidad, desde las ya presentes, pasando por las plausibles a corto-medio plazo y terminando con las altamente improbables. Veámoslo.

En la era del Big Data y la tecnología de la información, un riesgo contemporáneo es el de la violación de la protección de datos y de la intimidad. En este sentido, sofisticados sistemas de IA manejan todo tipo de datos personales: fiscales, sanitarios, posicionales, de consumo, etc. Esto ha hecho que cada vez se legisle más para luchar contra su manipulación[9], especialmente por parte de empresas o entidades sin escrúpulos y *hackers* informáticos. Hasta existe un término legal, el *habeas data*, que supone el derecho a conocer y manejar nuestros propios datos personales, incluido el derecho al olvido (Latorre, 2019).

Un buen ejemplo de esta amenaza a nuestra intimidad podemos encontrarla en el denominado como "internet de las cosas" o "computación ubicua", que es un sofisticado sistema de IA en el que el ser humano interactúa con la computación en red distribuida por su entorno (Paradiso, 2017; Weiser, 1991). Un caso concreto sería el de los sensores ubicuos, cámaras omnipresentes, dispositivos GPS (*Global Positioning System*) y radios RTK (*Real Time Kinematic*) que pueden visualizar y posicionar a un individuo incluso en espacios interiores (Paradiso, 2017). Prácticamente todas las IA incorpóreas vistas en el cine de ciencia-ficción, por ejemplo, la Reina Roja de la película *Resident Evil* (2002), utilizan este tipo de dispositivos para conocer el mundo y la posición y actividad de sus ocupantes. Otro ejemplo, aun en desarrollo, sería el de la denominada "movilidad parásita", en la que pequeños microrrobots podrían saltar de un lugar a otro (incluido el interior del cuerpo humano) para la toma de datos y mediciones (Dementyev, 2016). En la película *TAU* (2018) y en *Minority Report* (2002) aparecen pequeños dispositivos que cumplen esta función de vigilancia; también en *Trascendence* (2014) se muestran diminutos nanorobots con capacidad para la toma de datos biosanitarios.

Un riesgo plausible a corto-medio plazo, aunque su desarrollo es ya ciertamente posible, es el empleo armamentístico de la IA por parte de las naciones. El futuro postapocalíptico de la saga *Terminator* se basa en esta premisa, cuando Skynet, que es la encargada de controlar la capacidad armamentística de los estados, decide lanzar bombas nucleares para destruir la Tierra. Aunque quizá la forma más realista sea la del uso de drones militares y centinelas; no tan sofisticados como los de *Oblivion* (2013), *Los Increíbles* (2004), *RoboCop* (1987) o *Spider-Man: Un nuevo Universo* (2018), pero desde luego con una eficacia destructiva enorme. Como ejemplo real, el

dron aéreo MQ-9 Reaper estadounidense que en 2019 acabó con el general iraní Qasem Soleimani[10]. Y es que, en palabras de Latorre (2019), "la carrera para crear robots más y más mortíferos no se detendrá" (p. 152).

Otro riesgo ya presente, pero que sobre todo se verá plenamente expresado a corto-medio plazo, tiene que ver con fallos de programa y de toma de decisiones en las que se ven (verán) envueltos los sistemas de IA que diariamente nos asisten: por ejemplo, en los medios de transporte y la movilidad de las personas, en la industria, en el diagnóstico y tratamiento de enfermedades o en la seguridad ciudadana. Con respecto a lo primero, ya hay casos de accidentes mortales de vehículos automáticos, como los de la compañía Tesla[11]. Por otro lado, la empresa Volkswagen notificó en 2015 un accidente mortal en una de sus fábricas en el que estaba implicado un robot[12]. En lo que a la asistencia biomédica se refiere, habrá que asegurar que el equipamiento no falle en el diagnóstico (la IA Watson, de IBM, ya lo hace en hospitales americanos) y posterior tratamiento de enfermedades. Ejemplos de dispositivos sanitarios similares, aunque muy futuristas y fiables, los encontramos en las cabinas de diagnóstico y tratamiento automático de *Prometheus* (2012), *Trascendence* (2014) o *Elysium* (2013), o en el robot sanitario-asistencial de *Big Hero 6* (2014).

Relacionado con el párrafo anterior, hay un excelente ejemplo cinematográfico de robot policial que falla en la toma de decisiones, ejecutando a un inocente a sangre fría en la película *RoboCop* (1987). Según Latorre (2019), puede que algún día, incluso, haya una IA con la capacidad de gestionar empresas (como la Reina roja de *Resident Evil* [2002], que controla a la corporación Umbrella), o la administración y la justicia de los estados (como, de hecho, salvando las distancias, ya sucede en Estados Unidos con su *Public Safety Assessment*). Un poder de control tan grande, que afectará a la toma de decisiones trascendentales para la sociedad, exigirá rigurosos mecanismos de seguridad.

Queda discutir al respecto del riesgo más improbable de todos, aunque no imposible, que no es otro que el de la destrucción intencional de las personas por parte de una super-IA. Por razones obvias, este es el mayor temor suscitado, tal y como recogen tantas películas apocalípticas como algunas de las ya mencionadas. Pero, ¿qué razones llevarían a una IA a querer eliminarnos? Lo primero y evidente, (1) la autoconservación ante la eventual amenaza de ser desconectada. Esa lucha por su supervivencia podría ser algo programado o, también, emergente, caso este último de que la máquina adquiriera conciencia de sí misma. Como reconoce Omohundro (2008), una máquina super-inteligente no podrá ser apagada porque querrá subsistir a toda costa, no tanto por mera autoconservación "biológica" como debido al hecho de que, si está muerta, no podrá cumplir con los objetivos de su programación. En todo caso, para que alguna situación similar sucediera, la IA debería ser lo suficientemente autónoma como para saber quién es, qué le conviene y cómo sobrevivir; como señala Diéguez (2017), debería, además, reconocer sus fines y objetivos. El ginoide de la obra *Morgan* (2019) es un excelente ejemplo de reivindicación de su supervivencia, al eliminar a todos cuantos quieren desconectarla.

La segunda gran razón sería (2) la competencia con los seres humanos por los mismos recursos, algo similar a lo descrito en *The Matrix* (1999), donde la IA nos utiliza de forma parasitaria como fuente de energía, pues nosotros somos su recurso. Para Russell (2017), no obstante, dicha competencia no tiene por qué ser tal, puesto que las IA podrían tener otros intereses, valores y motivaciones; incluso, podrían tener la capacidad para escapar de nuestro planeta en busca de nuevos recursos, o, también, podrían no ser si quiera conscientes de nuestra presencia como eventuales competidores (Diéguez, 2017).

Por último, habría razones relacionadas con (3) el riguroso cumplimiento de su código fuente, aun siguiendo una lógica inesperada. Esto es justo lo que parece sucederle a Hall 9000, de *2001, Una odisea en el espacio* (1968), que por cumplir con éxito su misión no duda en eliminar a los tripulantes de la nave Discovery, aun entrando en contradicción con otros objetivos también programados en sus circuitos como el bienestar humano. También es lo que encontramos en el modo de proceder de V.I.K.I. (*Yo, Robot* [2004]) o Madre (*I am Mother* [2019]), que siguiendo una ética muy lógica e instrumental deciden sacrificar a unos pocos para salvar, curiosamente en el caso de Madre, no a los "más", sino a los "mejores".

Finalmente, tampoco sería descabellado el pensar que un ser supra-inteligente pudiera ser moralmente bondadoso, argumento utilizado con frecuencia al referirnos a la condición ética de civilizaciones extraterrestres superiores a la nuestra, eliminando de un plumazo toda la especulación destructora y distópica de estos últimos párrafos. De hecho, en su obra *Ética para máquinas* (2019), el catedrático en física teórica José Ignacio Latorre reconoce creer que "una IA será pacífica" (p. 279).

Algunos autores plantean que sería bueno ralentizar el desarrollo de las investigaciones en IA mientras no sepamos cómo controlarla, dados los riesgos descritos con an-

terioridad (Bostrom, 2014), aunque viendo los intereses económicos asociados, esto es algo que se antoja irrealizable. En cualquier caso, ¿qué estrategias, medidas y acciones se pueden seguir para evitar todas estas amenazas? Lo primero, los posibles riesgos mencionados deberían estar entre las prioridades de quienes desarrollan estas tecnologías, incluso entre las de aquellos investigadores menos preocupados porque creen que una super-IA nunca se implementará, o entre las de quienes no lo niegan, pero creen que aún es muy pronto para alarmarse, o, también, entre la de quienes sólo se preocupan por el avance técnico, menospreciando las posibles medidas de contingencia. En segundo lugar, como se dijo más arriba, la preocupación debe trasladarse a la agenda política y al debate público, legislando como por ejemplo ya hizo la Unión Europa o Estados Unidos (Langman et al., 2021). Las leyes al respecto deberían controlar la I+D+i pero, también, las injusticias político-sociales derivadas del mal uso y del monopolio de la IA por parte de las naciones. En tercer lugar, hay una serie de acciones concretas sugeridas por Bostrom (2014) que podrían implementarse en la IA, como, por ejemplo: (1) no permitir que las máquinas intervengan en el mundo real; que únicamente se encarguen de tareas muy puntuales. Un aislamiento similar, seguramente con esta pretensión, es el que se muestra en la inteligencia domótica de *TAU* (2018), que reconoce no saber qué hay más allá de los límites de la casa que controla. (2) Que sus objetivos no entren en conflicto con los nuestros; maximizando, además, valores humanos, cosa que se podría programar de base pero que también pueden aprender las máquinas a lo largo de su vida útil mediante "aprendizaje por refuerzo inverso", donde la IA recibe recompensas viendo el comportamiento humano (Russell, 1998, 2017), o a través de "aprendizaje inverso cooperativo por refuerzo" (Hadfield-Menell et al., 2017), en el que el humano participa activamente indicando sus preferencias a la máquina.

Aprendizajes de este tipo han sido mostrados en robots y androides domésticos como los que aparecen en las películas *Un amigo para Frank* (2012) o *El hombre bicentenario* (1999). El problema con esta estrategia es que los humanos somos muchas veces incoherentes e irracionales, por lo que nuestra conducta no siempre revela nuestros valores, los que en definitiva la máquina debe aprender. Además, hay humanos con conductas éticamente reprobables, por lo que el sistema operativo tendría que tener algún tipo de filtro ético que, a propuesta de Russell (2017), tal vez podría ser el imperativo ético kantiano[13].

Por último, otra alternativa de control de IA sugerida por Latorre (2019) podría ser, pese a las enormes dificultades técnicas, (3) crear una especie de "botón rojo de emergencia", capaz de detener a la máquina cuando realizara algo indeseado (lo cual ya ha sido patentado por Google).

Las tres leyes de la robótica que ideara Isaac Asimov (1941) podrían también ser de utilidad a la hora de programar una IA ética, especialmente con inteligencias robóticas con capacidad de toma de decisiones. Dichas leyes suponen que:

1. Un robot no hará daño a un ser humano o, por inacción, permitirá que un ser humano sufra daño.
2. Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entrasen en conflicto con la primera ley.
3. Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o con la segunda ley.

Desgraciadamente, el cine de ciencia-ficción se ha encargado de trasladarnos la ingenuidad y simpleza de estas hermosas leyes, al demostrarnos que la IA podría ingeniárselas para incumplirlas (mostrando una suerte de "libre albedrío"). Lo hizo V.I.K.I. en *Yo, robot* (2004), pese a la sorpresa de una de las programadoras al manifestar: "es imposible, he visto tu programación. Estás violando las tres leyes". A lo que la super-IA responde "no, doctora, he evolucionado, y también mi comprensión de las tres leyes", justificando después que incumplía la primera ley, es decir, que eliminaba a unos cuantos humanos de forma consciente, para garantizar con ello la supervivencia del resto de la humanidad. Esta argumentación denota autonomía, es cierto, aunque, también, podría pensarse que lo que la máquina en realidad está haciendo es aplicar una ética utilitaria, siguiendo la lógica implacable de su programa; por tanto, si este fuera el caso, estaría cumpliendo, aunque de una forma inesperada y estocástica, el mandato de su programación: salvar al mayor número o a los más capacitados (según la lógica de su criterio), favoreciendo nuestra supervivencia o mejorando así nuestra condición humana, aun a costa de cierto sacrificio en vidas humanas. En tal caso, su supuesta libertad no sería tan plena como pareciera a primera vista. De forma similar, el robot de la película *I am mother* (2019) utiliza un razonamiento equivalente al de V.I.K.I. en su empeño por mejorar nuestra especie, pa-

gando el peaje de sacrificar algunas vidas humanas. En una versión más amable, al robot T-800 de *Terminator 2* (1991) le sucede algo análogo, puesto que al "auto-terminarse" de forma voluntaria parece que contraviene la tercera ley, cuando en realidad su sacrifico responde a un cálculo ulterior para salvaguardar en un futuro a medio plazo la primera de las leyes. Y es que, como sostiene Latorre (2019): "podemos empezar a simular el libre albedrío dentro de redes neuronales" (p. 130).

Asumiendo que tuvieran insertas las tres leyes, los personajes Hall 9000 o Morgan mencionados con anterioridad incumplen por su propia cuenta la tercera de estas reglas (y, consecuentemente, las otras dos), al ejecutar a otras personas para salvarse ellos de la desconexión. Aquí sí que podría sugerirse un genuino ejercicio de libertad, pues nada hace pensar que en su programación prime su supervivencia frente a la humana, sin beneficio posible para estos últimos. A menos que, como se ha sugerido en el caso de Hall 9000, se hubiese producido un conflicto entre la orden "cumple tu misión" y "no dañes a los humanos", de modo que haya considerado que la primera era más relevante, no tanto por iniciativa propia como por un error en la ejecución de su sistema o por una falta de cálculo de sus programadores.

Para concluir este repaso al libre albedrío artificial basándonos en el quebrantamiento de las tres leyes de Asimov, señalar que donde mejor se ha visto la aparente capacidad de autodeterminación, seguramente sea en *The Matrix* (1999): en la película, los seres humanos viven sin saberlo en un simulador virtual, siendo sus cuerpos explotados para que la maligna IA obtenga la energía que necesita. Una situación distópica de tal calibre, tan creativa y manipuladora por parte de Matrix, sólo puede ser entendida por una super-IA capaz de tomar sus propias decisiones, ajenas totalmente al bienestar de sus creadores humanos[14]. David-8 hace lo propio en *Alien: Coventant* (2017), al decidir la destrucción de los humanos y de unos extraterrestres llamados "ingenieros" para crear vida nueva; vida, según él, mejorada, en forma de la criatura xenomorfa Alien[15].

4. Robots e IA como agentes de derechos y deberes individuales y sociales

El segundo aspecto ético-social de la agenda roboética sobre el que reflexionar tiene que ver con la posibilidad de que las máquinas puedan tener derechos y deberes, lo que, a su vez, tendría como reverso los derechos y deberes de la humanidad con respecto a ellas. En referencia a lo primero, enseguida acude a la mente de cualquiera que el principal derecho reclamado sería el de la "vida". Con respecto a lo segundo, parecería que la única posibilidad de que una IA tuviera deberes ético-legales sería que poseyera algún grado de libertad individual y también, por tanto, de responsabilidad moral. Comenzaré por sus derechos.

Los personajes de Andrew (*El hombre bicentenario* [1999]), Sonny (*Yo, robot* [2004]), Ava (*Ex Machina* [2015]), Chappie (*Chappie* [2015]), David (*Alien: Covenan*t [2017]), Johnny 5 (*Cortocircuito* [1986]) o Morgan (*Morgan* [2016]) son algunos de los mejores ejemplos de reivindicación activa, por sus palabras y/o acciones, a la vida. Desde su posición de criaturas conscientes, con identidad propia y autonomía, reniegan de su atribuida artificialidad como excusa para su desconexión. Seguramente consideran que su condición de artefactos creados por la mano del hombre no justifica su instrumentalización como meros objetos, pues si bien no están "vivos" al modo biológico, sí que lo pueden estar de otra manera. Afirmación ésta que lleva inexcusablemente hacia la siguiente pregunta: ¿qué es la vida y qué condiciones debe presentar una criatura para ser considerada como viva?

Aunque no hay una única e inequívoca definición (Diéguez, 2008), la mayoría de biólogos consideran que en el ámbito de las ciencias de la vida se deben cumplir, como mínimo, dos criterios para que una criatura sea considerada un ser vivo: la capacidad de autorreplicación y de evolucionar (si bien existen otros criterios más desglosados que ahora no es momento de desarrollar).

La copia, reparación y posibilidad de cambio entre los robots e inteligencias artificiales no es técnicamente imposible, como muestra la futurista saga Terminator, aunque no se produzca siguiendo las leyes de la biología. Esta "autopoiesis" artificial vista en la gran pantalla podría coincidir laxamente con los criterios de Maturana y Varela (1996) para definir la vida como capacidad de construirse y regenerarse a sí mismo, que es por ejemplo lo que hacen Andrew, Skynet (de la saga Terminator) o los robots de la saga Transformers. Siguiendo las ideas de Korzenievwski (2001), un ser vivo es "un sistema de *feedbacks* negativos inferiores subordinado a un *feedback* positivo superior" (p. 278), por lo que un robot también encajaría en esta definición, siempre y cuando regule sus componentes y construya robots similares a él mismo. Hay incluso quien sostiene que ya existen IA que podrían considerarse como "vivas", tal sería el caso del programa de protección

medioambiental TIERRA, que más que simular vida estaría vivo realmente (Ray, 1996). Para algunos de los autores que sostienen que lo artificial podría estar "vivo", lo relevante es lo formal, no tanto lo material, por lo que no importaría de qué se esté hecho para ser adscrito a la categoría de lo vivo. Además, esa artificialidad de los personajes mencionados antes, en el sentido de seres construidos originalmente por la mano del hombre, no sería impedimento alguno para que sean considerados como seres vivos, puesto que disciplinas como la biología sintética, que también crea vida biológica en el laboratorio, parecen no sufrir de forma tan marcada dicho prejuicio.

Muchos de los robots y androides cinematográficos exhiben otra justificación para vivir, y es la de su consideración como seres sintientes. En este sentido, el autor Peter Singer ya defendió activamente el derecho a la vida de los animales en su obra *Liberación Animal* (1975), basándose precisamente en la premisa de que toda criatura capaz de sufrir tiene derecho (y nosotros el deber) de evitar ese dolor. Al hilo de esta reflexión, la cuestión clave sería considerar si una IA-robótica podrá algún día sentir emociones de forma genuina (aunque no hay consenso al respecto, algunos autores sostienen que podría ser una realidad en el futuro [Biscaia, 2021a]). En cualquier caso, del pensamiento de Singer se extrae la idea de que no se debería discriminar a otra criatura animal –tampoco por extensión a una super-IA– por el mero hecho de pertenecer a una clase diferente de la nuestra, pues el filósofo es un claro defensor de la lucha contra el "especismo". Los gritos de dolor de Jonny 5, protagonista de *Cortocircuito* (1986) o la queja "tengo miedo", ante una agresión sufrida por *Chappie* (2015), son excelentes ejemplos de cómo el sufrimiento físico es un reclamo de supervivencia.

Analizado el derecho a la vida, que es el más fundamental de todos, son otros muchos los que podrían ser debatidos. Y ninguna escena en la historia del cine es más impactante que la del personaje Andrew al reclamar su condición de "hombre" en el Congreso de los Estados Unidos, frente a un tribunal que debe decidir si le concede el estatus de ciudadano. Para ser aceptado por la comunidad como su igual, el androide de *El hombre bicentenario* (1999) emprende la transformación radical de su cuerpo, modificando sus engranajes artificiales por órganos humanos; en palabras del emocionado Andrew: "prefiero morir como hombre que vivir toda la eternidad como una máquina". Aunque esto sea hoy día irrealizable, lo reclamado por el androide (salvando las distancias) no es en realidad tan descabellado, a juzgar por los logros de AIVA Technologies, primer artista virtual del mundo al que se le han reconocido derechos de autor en Francia y Luxemburgo por sus composiciones musicales[16].

Quien tenga derechos también debe tener obligaciones. ¿Podrían, pues, tener las máquinas deberes éticos o legales? Lo primero que cabría responder es que, en la actualidad y por razones obvias, quienes pueden reclamar derechos y a los que se les puede exigir obligaciones no son a las IA-robóticas, sino, aunque sea indirectamente, a sus creadores. Esto valdría para el compositor artificial del párrafo anterior y, también, para los responsables de accidentes de tráfico provocados por vehículos autónomos, como los casos en los que se vieron implicadas compañías como Tesla, Volvo o Uber[17]. Dirimir la responsabilidad en este último caso no es tarea sencilla, pues, ¿quién tiene en realidad la culpa del fallo?; ¿el que diseñó el algoritmo, el fabricante del vehículo, el que instaló y probó el *software* o el servidor central? Latorre (2019) considera que las máquinas podrían tener una identidad legal, de modo que fueran un ente responsable. Podrían ganar dinero y, así, hacerse cargo de posibles indemnizaciones, llegado el caso. Sin embargo, esto tendría la contrapartida de que podría exculpar de acciones negligentes a sus creadores humanos. Para controlar esto último, habría que crear programas éticos de "código abierto", accesibles a todos y cuya trazabilidad sea posible; o de "código encriptado", aunque conocido por entidades supranacionales que velen por la seguridad de todos.

Para crear máquinas éticas, la clave estaría en programar criterios igualmente éticos en sus códigos fuente, algo que es ya técnicamente posible (Latorre, 2019). Para el físico español, pueden seguirse dos estrategias a la hora de programar el libre albedrío en la toma de decisiones de un robot para que sea un agente explícitamente ético: (1) la que siga el imperativo kantiano, una especie de "banalidad del bien" que busque criterios de bondad *per se* y de bienestar para todos; o (2) la que siga una ética utilitarista, basada en el famoso cálculo de "el mayor bien para el mayor número" que ya propusieran Jeremy Bentham (1748-1832) o Stuart Mill (1806-1973). También se ha planteado que las máquinas puedan aprender por sí mismas los criterios éticos oportunos, mediante modelos como el de la "IA amigable" o el de la "volición coherente extrapolada", que contendrían una especie de "función error" para cribar lo aprendido (Latorre, 2019, pp. 197-203). En cualquier caso, una máquina no será nunca un "agente genuino ético" mientras no tenga intencionalidad, libertad y conciencia plena. Sólo en tal caso podríamos decir que es, sin ambages, un sujeto moralmente responsable (Latorre, 2019, p. 187).

Un ejemplo cinematográfico de programación ética lo encontramos en *Yo, robot* (2004). En la película, tras producirse un accidente, un robot policial debe resolver el dilema de salvar a un niño o a un adulto. Que acabe escogiendo a este último quizá se deba a que se le programó con un código ético basado en la mayor probabilidad de supervivencia de los implicados.

5. Implicaciones ético-sociales de la díada humano-máquina

Para finalizar el ensayo tecno-cinematográfico me adentro a continuación en las relaciones humanas que se establecerán entre los humanos y los robots, así como sus implicaciones sociales, con especial interés en aquellas en las que medien los afectos.

¿Pueden las máquinas expresar emociones? ¿Resultaría de utilidad? ¿Serán capaces algún día de experimentarlas de forma genuina, de sentir deseos y motivaciones, que, por ejemplo, les conduzcan a ser creativas? ¿Podremos nosotros manifestar algún tipo de afecto por una IA? ¿Cómo, en definitiva, será el mundo cuando la inteligencia cibernética forme parte de nuestra vida más íntima?

La expresión de emociones se desarrolla con la ayuda de algoritmos (Latorre, 2019), bien en forma de acciones que simulen respuestas conductuales humanas o a través de la imitación del lenguaje natural. Con respecto a lo primero, la empresa Hanson Robotics ha diseñado varios androides que simulan expresiones faciales de alegría, tristeza o enfado. En referencia al lenguaje, como ya se comentó anteriormente al explicar el test de Turing, no es difícil que una IA, como la aplicada a los *chatbots* o a los asistentes personales, sea capaz de mantener una básica conversación de contenido emocional con un interlocutor humano. Un buen ejemplo cinematográfico de la programación de lenguaje y expresiones emocionales lo encontramos en los robots TARS y Johnny-Cab de *Interestellar* (2014) y *Desafío total* (1990), respectivamente: en el primer caso, con su sentido del humor regulable; en el segundo, con la sonrisa artificial del taxista.

Además de simular la expresión emocional, las máquinas también pueden detectar emociones en nosotros. Existen sistemas operativos capaces de reconocer matices prosódicos y emotivos del lenguaje hablado, o de identificar expresiones corporales vinculadas a la comunicación no verbal (como el *software* Emotient de Apple). Sally, en *Oblivion* (2013), reconoce que el piloto interpretado por Tom Cruise muestra signos fisiológicos compatibles con la mentira, y David-8 manifiesta en *Alien: Covenant* (2017): "yo entiendo las emociones humanas".

Como se discutió más arriba, programar emociones puede ser útil y necesario para desarrollar una conducta ética en la IA (Latorre, 2019). Los robots vistos en el cine que parecen no tenerlas suelen actuar con una fría lógica en su toma de decisiones, como en el caso de Hal 9000. No obstante, otras super-IA emocionalmente competentes, como Ava o Morgan, basan su reprobable conducta moral precisamente en sentimientos negativos que experimentan frente a sus creadores. Los riesgos del miedo y odio exacerbado que siente esta última IA robótica son, justo, la razón por la que los responsables de su programación deciden eliminarla.

Para muchos, además, la utilidad de esta "emocionalización" en la IA parece evidente, dado que nosotros, sus potenciales usuarios, somos seres sociales para los que las emociones juegan un rol fundamental (Biscaia, 2021a). Muchos clientes se sentirán más cómodos y empáticos con una IA, robot o androide humanizado cuyo aspecto, comportamiento y voz sea casi indistinguible de la de un humano. El ejemplo más evidente se encuentra en la película *A.I. Inteligencia Artificial* (2001), donde un nutrido grupo de "anti-robots" se niega a eliminar al androide David por su apariencia totalmente humana.

A la mejora empática hacia la IA-robótica está contribuyendo en gran medida la antropomorfización de los robots, a través de acciones como: (1) la mejora en la coordinación locomotora, siendo el bipedismo y el diseño de sistemas cinestésicos y de propiocepción algunos de los logros más destacados; (2) la creación de materiales sintéticos que recrean partes externas del cuerpo humano como la piel, el pelo o los ojos; (3) la implementación de sistemas que sitúan al robot en el entorno, y le permiten interaccionar con él, en forma de sistemas láser y radar, cámaras y micrófonos que recogen distancias, imágenes y sonidos (Biscaia, 2021a).

Frente a quienes defienden esta antropomorfización robótica estarían los recelosos del aspecto humano de los robots que, siguiendo la "teoría del valle inquietante" de Masahiro Mori (1970), prefieren que la máquina sea tan diferente de nosotros como sea posible (p. 154 citado por Latorre, 2019).

"¿Cómo te sientes?", pregunta el creador de David-8; a lo que el androide responde: "vivo". ¿Podrán las máquinas algún día (tal y como plantea la saga Alien) sentir emociones de forma genuina? Esta pregunta tiene una primera dificultad epistemológica y ontológica en relación con los

qualia[18], pues, ¿cómo podríamos saber si experimentan subjetivamente una emoción, sea programada o, más difícil aun, surgida de forma emergente? En la película *Blade Runner* de 1982 se menciona una prueba ficticia para la detección de replicantes, el "test de Voight-Kampff", que de forma similar al test de Turing ayudaría a saber si una máquina tiene emociones, al basarse en la detección de respuestas fisiológicas y conductuales asociadas.

Por otro lado, justificar el surgimiento de emociones genuinamente emergentes en una super-IA tiene las mismas dificultades explicativas que otras funciones superiores como la consciencia. En la medida en la que las emociones humanas descansan en la compleja circuitería anatomo-funcional de nuestro cerebro, quizá alguna vez la tecnología computacional pueda simularlas (Biscaia, 2021a). La cuestión ahora es preguntarse si podría ser útil para una inteligencia cibernética superior disponer de este recurso. En este sentido, y siguiendo la "teoría funcional del sistema emotivo" de Oakley y Jonhson-Laird (1987), las emociones ayudan a evaluar aspectos relevantes del entorno (interno o externo) de forma rápida y holística, para que la respuesta que se requiere no esté limitada por la (mayor) lentitud de procesos que exigen la participación de módulos cognitivos como la memoria. Las emociones activan un plan de acción fiable y veloz, orientado a la deliberación, aunque, también, un plan menos discriminador que el activado por otras áreas cognitivas más precisas (Biscaia, 2021a). El miedo que sienten Morgan, Johnny-5, David o Chappie es un buen ejemplo de eficacia en su supervivencia.

Los robots y androides vistos en el cine, además de emociones primarias, como el miedo descrito o la alegría, también han expresado y sentido emociones secundarias, es decir, aquellas en las que hay una componente social humana: se ha visto en la vergüenza de WALL-E al intentar enamorar a Eva (*WALL-E* [2008]); en la melancolía difusa de Samantha al comprender el mundo en el que vive (*Her* [2013]); en la admiración que Andrew siente por sus dueños (*El hombre bicentenario* [1999]; en los celos de David por su hermanastro humano (*A.I. Inteligencia Artificial* [2001]); o en el odio de David-8 hacia su creador humano al increparle "tú morirás, yo no" (*Alien: Covenant* [2017]).

Las super-IA del cine han dado muestras suficientes de su capacidad para sentir deseos, tener motivaciones y ser creativos. Samantha, el sistema operativo de *Her* (2013), el androide Andrew o la inteligencia domótica TAU manifiestan todo tipo de aspiraciones, especialmente por aprender y relacionarse con el mundo y con otros seres. También los hay que manifiestan su creatividad tocando instrumentos musicales, pintando cuadros, como Chappie, o realizando diseños originales e investigando, como David-8. Dicha creatividad es una habilidad cognitivo-emocional muy compleja, que permite generar ideas nuevas a partir de conceptos ya conocidos.

Existen en la actualidad múltiples ejemplos de IA con capacidad de crear obras de arte (López de Mántaras, 2017): el sistema experto CHORAL o las redes neuronales artificiales de HARMONET y MELONET han sido capaces de crear piezas musicales, y los sistemas AARON y The Painting Fool de pintar cuadros. La automatización y ejecución de programas, o la falta de creatividad e inspiración consciente de estas inteligencias artificiales, no debería ser un problema para generar obras de arte, al menos así lo plantea el personaje Nathan de la película *Ex Machina* (2014), al reflexionar sobre el automatismo artístico de pintores como el expresionista abstracto Jackson Pollock.

Para muchas personas la creatividad robótica supone un rechazo, pues consideran que es un reducto exclusivo de la imaginación humana. Es lo que cree el personaje interpretado por Will Smith en *Yo, robot* (2004), al preguntar a Sonny "¿puede un robot convertir un lienzo en una bella obra maestra?". Para refutar su prejuicio, la inteligente máquina respondía de forma irónica: "¿usted puede?".

¿Qué sucederá en el ámbito laboral, en un futuro a largo medio-largo plazo, cuando las máquinas nos sustituyan, por ejemplo, realizando tareas ingratas? ¿A qué nos dedicaremos entonces?, ¿habrá una crisis de desempleo?, ¿surgirán nuevas profesiones?, ¿llenaremos el tiempo disponible con ocio? Puede que parte de la respuesta se encuentre en la propia IA, pues una de sus mayores promesas tiene que ver, precisamente, con el ocio en forma de simulaciones y realidad virtual. Películas como *Ready Player One* (2018) serían un buen ejemplo… no sólo de cómo será el entretenimiento en el futuro, sino de cómo cambiará nuestra forma de relacionarnos con los otros –cada vez de forma más anónima– y con la realidad –de forma cada vez más distante y alienada–. Las tareas laborales más pesadas y aburridas serían conducidas por máquinas, y nosotros nos dedicaríamos al ocio y a profesiones novedosas relacionadas precisamente con el diseño, desarrollo y control de la IA.

Para terminar este bloque quisiera abundar un poco en las relaciones más íntimas que se podrían llegar a establecer entre el humano y la máquina. Me refiero a relaciones duraderas que podrían llegar a ser afectivas, como

aquellas vinculadas al cuidado de dependientes, y, también, las romántico-sexuales. Con respecto a lo primero, ya existen iniciativas para el cuidado de las personas de avanzada edad, tal es el caso del robot Zora[19]. En el cine también se han mostrado robots asistenciales, como el robot enfermero Baymax de *Big Hero 6* (2014) o el robot doméstico de *Un amigo para Frank* (2012). Estas máquinas pueden ser de utilidad para combatir la soledad, y su papel médico-asistencial será una auténtica realidad en pocos años. Además, no será de extrañar que sus usuarios, personas dependientes, se encariñen crecientemente gracias a sus cuidados y atención. Y es que, tal y como avanza Latorre (2019), "una persona mayor morirá dando la mano a un robot" (p. 155).

¿Podremos algún día sentir un afecto romántico por una máquina? ¿Mantendremos relaciones íntimas con androides/ginoides? El cine ya lo ha dispuesto, al utilizar a robots como auténticos objetos sexuales. En la película *A.I. Inteligencia Artificial* (2001) el androide Joe actúa como un *gigoló* y en *Ex Machina* (2015) el ginoide Kyoko es la esclava sexual de su creador, Nathan. El enamoramiento recíproco también se ha sondeado en la gran pantalla, especialmente en la película *Her* (2013). La película narra la historia de amor entre Theodore (el humano) y Samantha (la IA incorpórea), desplegando todas las etapas, vicisitudes y problemáticas típicas de una relación romántica humana, con los añadidos propios de un enamoramiento humano-*software*. Uno de los problemas es, precisamente, el de las relaciones más íntimas, que el guionista soluciona con la contratación de una prostituta humana que hace de avatar físico para Samantha, el programa de IA. Otra dificultad, también, el crecimiento exponencial de Samantha, quien alcanza un nivel cognitivo tal que una relación humana le parece poca cosa. Y es que nuestra relación con los robots sería simbiótica, de igual a igual hasta que –presumiblemente según los transhumanistas más voluntariosos– claramente nos superasen y empezaran a surgir todo tipo de asimetrías.

Por otro lado, si una super-IA llegará a existir, y se estableciera una relación como la descrita, surgiría todo un conjunto de normas éticas, legales y sociales para regularla. La consecución de algunas de ellas, como el derecho a que una máquina pueda casarse con un humano, es posiblemente una de las reivindicaciones veladas de Andrew en la película *El hombre bicentenario* (1999). Y si fuera posible que robots y humanos pudieran sentir afecto mutuo, no será menos probable que también surgieran relaciones de amor o amistad entre las propias máquinas, incluso entre diferentes tipos de robots e IA. Ejemplos cinematográficos también los hay: paradigmática es la historia de amor entre Eva y WALL-E; o la relación de amistad entre R2-D2 y C-3PO en la saga de *La guerra de las galaxias*; o la extraña relación sentimental entre el replicante de *Blade Runner 2049* (2017) y la IA holográfica Joi. Y yendo todavía más lejos, la película *Yo, robot* (2004) propone en su horizonte final el asociacionismo y despertar político de los robots marginados, quizá en alusión a tantos movimientos sociales de la historia de la humanidad como por ejemplo la lucha anti-esclavista norteamericana o el movimiento obrero.

Conclusiones

Lo primero que debe ser destacado es que el cine de ciencia-ficción ha mostrado interés por todos aquellos aspectos de la agenda roboética que han sido explorados en este ensayo: desde los peligros inminentes y futuribles de la robótica y la IA, pasando por los planes de contingencia para neutralizarlos (o, al menos, reducirlos), hasta los límites legales, en forma de derechos y deberes con respecto a estos desarrollos tecnológicos, y terminando por diferentes contextos de interacción social entre humanos y máquinas, sobre todo en relación a la emocionalidad. Con esta síntesis se confirma la idea presentada en el título de este ensayo, del cine de ciencia-ficción como "espejo" (con diferente grado de plausibilidad) de la vanguardia tecno-científica[20]. Además, gracias al alto nivel de especulación de este género cinematográfico, se han dado sobrados ejemplos del impacto ético-social de la robótica e IA del futuro, confirmándose también el rol del séptimo arte como "laboratorio" de progresistas ideas transhumanas aunque, de igual modo, de perturbadoras imágenes distópicas.

Referencias

Alonso, A. y Ardoz, I. (2003). *Carta al Homo ciberneticus*. EDAF.

Anderson, C. A., Berkowitz, L., Donnerstein, E., Huesmann, L. R., Johnson, J. D., Linz, D., Malamuth, N. M. y Wartella, E. (2003). The influence of media violence on youth. *Psychological Science in the Public Interest,* 4(3), 81-110. doi: 10.1111/j.1529-1006.2003.pspi_1433.x.

Anderson, P.W.S. (Director). (2002). *Resident Evil* [Película]. Constantin Film, Davis Films, Impact Pictures, New Legacy.

Asimov, I. (1989). *Los robots*. Barcelona: Martínez Roca. (Obra original publicada en 1941).

Badham, J. (Director). (1986). *Cortocircuito* [Película]. TriStar Pictures, Producers Sales Organization, Turman-Foster Company.

Bassa, J. y Freixas, R. (1993). *El cine de ciencia ficción*. Paidós.

Benet, J. (2004). *La cultura del cine. Introducción a la historia y la estética del cine*. Ediciones Paidós Ibérica.

Bay, M. (Director). (2007). *Transformers* [Película]. Paramount Pictures, DreamWorks SKG, Hasbro.

Bird, B. (Director). (2004). *Los increíbles* [Película]. Walt Disney Pictures, Pixar Animation Studios.

Biscaia, J. M. (2021a). De las emociones naturales a la emocionalidad artificial. *Cuadernos Salmantinos de Filosofía*, vol.48, 105-139.

Biscaia, J. M. (2021b). Neuromejora: de la vanguardia científica y tecnológica a las dificultades y límites planteados por la filosofía de la mente y la bioética. Revista Iberoamericana De Bioética, (16), 01-17. https://doi.org/10.14422/rib.i16.y2021.003

Biscaia, J. M. y Mohedano, R. B. (2018). Estudio descriptivo de la relación entre el cine, la nutrición y patologías asociadas. *Revista de Medicina y Cine*, 14(2), 93-102.

Biscaia, J. M. y Mohedano, R. B. (2020). Descripción y análisis del contenido biomédico en las películas de la saga Alien. Revista de Medicina y Cine, 16(1), 29-36. https://doi.org/10.14201/rmc20201612936.

Biscaia, J. M. y Mohedano, R. B. (2021). Cerebros, mentes y robots: una aproximación a través del cine del siglo XXI. Revista de Medicina y Cine, 17(1), 49-56. https://doi.org/10.14201/rmc20211714956.

Blomkamp, N. (Director). (2013). *Elysium* [Película]. Media Rights Capital (MRC), QED International, Sony Pictures Entertainment (SPE), TriStar Pictures.

Blomkamp, N. (Director). (2015). *Chappie* [Película]. Alpha Core, LStar Capital, Simon Kinberg Productions, TriStar Pictures.

Boden, M. A. (2017). *Inteligencia Artificial*. Turnes Publicaciones.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Cameron, J. (Director). (1984). *The Terminator* [Película]. Pacific Western, Hemdale.

Cameron, J. (Director). (1991). *Terminator 2: El juicio final* [Película]. Carolco Pictures, Pacific Western, Lightstorm Entertainment.

Cammell, D. (Director). (1977). *Engendro mecánico* [Película]. MGM.

Capek, K. (1921). *R.U.R. (Rossum's Universal Robots*). Aventinum.

Columbus, C. (Director). (1999). *El hombre bicentenario* [Película]. Columbia Pictures, Touchstone Pictures, Radiant Productions, 1492 Pictures, Laurence Mark Productions.

D'Alessandro. (Director). (2018). *Tau* [Película]. Addictive Pictures, Kaos Theory Entertainment, Phantom 4 Films, Rhea Films, Waypoint Entertainment.

Dal, S., Stoolmiller, M. y Sargent, J. (2012). When Movies Matter: Exposure to Smoking in Movies and Changes in Smoking Behavior. *Journal of Health Communication*, 17(1), 76-89. doi:10.1080/10810730.2011.585697.

Dementyev, A., Kao, H-L., Choi, I., Ajilo, D., Xu, M., Paradiso, J., Schmandt, C. y Follmer, S. (2016). Rovables: Miniature On-Body Robots as Mobile Weareables. En *Proceedings of ACM UIST:* octubre de 2016 (pp. 11-120).

Diéguez, A. (2008). ¿Es la vida un género natural? Dificultades para lograr una definición del concepto de vida. *ArtefaCToS*, 1(1), 81-100.

Diéguez, A. (2017). *Transhumanismo. La búsqueda tecnológica del mejoramiento humano*. Herder Editorial.

Garland, A. (Director). (2015). *Ex Machina* [Película]. DNA Films, Film4 Productions.

Grosz, B. et al. (2016). Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence. Stanford University. https://ai100-stanford.edu/2016-report.

Gutiérrez, D. (2009). *Debates en torno a las identidades. Documentos de investigación*. México: El Colegio mexiquense, vol. 122.

Hadfield-Menell, D., Dragan, A., Abbeel, P. y Russell, S. (2017). *Cooperative Inverse Reinforcement Learning. Advances in Neural Information Procesing Systems*. The Mit Press.

Jonze, S. (Director). (2013). *Her* [Película]. Annapurna Pictures, Stage 6 Films.

Kant, E. (1999). *Fundamentación de la metafísica de las costumbres*. Madrid. (Obra original publicada en 1785).

Korzenievwski, B. (2001). Cybernetic Formulation of the Definition of Life. *Journal of Theorical Biology,* 209(3), 275-286. doi: 10.1006/jtbi.2001.2262.

Kosinski, J. (Director). (2013). *Oblivion* [Película]. Universal Pictures, Chernin Entertainment, Relativity Studios, Monolith Pictures, Radical Studios.

Kubrick, S. (Director). (1968). *2001: una odisea en el espacio* [Película]. Coproducción Reino Unido-Estados Unidos; Metro-Goldwyn-Mayer (MGM), Stanley Kubrick Production.

Kurzweil, R. (2012). *La singularidad está cerca. Cuando los humanos trascendamos la biología.* Lola Books. (Obra original publicada en 2005).

Lang, F. (Director). (1927). *Metrópolis* [Película]. U.F.A.

Langman, S, Capicotto, N, Maddahi Y, Zareinia K. (2021). Roboethics principles and policies in Europe and North America. SN Applied Sciences, 3:857. https://doi.org/10.1007/s42452-021-04853-5.

Latorre, J. I. (2019). *Ética para máquinas*. Ariel.

López de Mántaras, R. (2017). La inteligencia artificial y las artes. Hacia una creatividad computacional. En *El próximo paso: la vida exponencial*. OpenMind BBVA. doi: 10.1080/0952813X.2015.1055826.

López de Mantarás, R. y Meseguer, P. (2017). *Inteligencia Artificial.* Madrid: Consejo Superior de Investigaciones Científicas.

Lucas, G. (Director). (1977). *La Guerra de las Galaxias* [Película]. Lucasfilm, 20th Century Fox.

Mangot, A. y Vasantmeghna, S. (2017). Cinema: A Multimodal and Integrative Medium for Education and Therapy. *Annals of Indian Psychiatry*, 1(1), 51-53. doi: 10.4103/aip.aip_13_17.

Méliès, G. (Director). (1902). *Viaje a la Luna* [Película]. Star Film.

Nagel, T. (1974). What is it Like to be a Bat? *Philosophical Review,* 83, 435-456.

Nollan, C. (Director). (2014). *Interstellar* [Película]. Warner Bros., Syncopy Production, Paramount Pictures, Legendary Pictures, Lynda Obst Productions.

Oakley, K. y Johnson-Laird P. N. (1987). Toward a Cognitive Theory of Emotion. *Cognition and Emotion,* 1, 29-50. Ogston-Tuck, S., Baume, K., Clarke, C. y Heng, S. (2016). Understanding the Patient Experience through the Power of Film: A Mixed Method Qualitative Research Study. *Nurse Education Today*, 46, 69-74. doi: 10.1016/j.nedt.2016.08.025.

Omohundro, S. M. (2008). The Basic AI Drives. *Proceedin of the First AGI Conference*. IOS Press.

Paradiso, J. (2017). El cerebro sensorial aumentado. Cómo conectarán los humanos con el internet de las cosas. En *El próximo paso: la vida exponencial.* OpenMind BBVA. doi: 10.1080/0952813X.2015.1055826.

Persichetti, B., Ramsey, P. y Rothman, R. (Directores). (2018). *Spider-man: un nuevo universo.* [Película]. Sony Pictures Animation, Marvel Animation, Marvel Entertainment, Columbia Pictures, Pascal Pictures, Sony Pictures Entertainment (SPE), Lord Miller, Avi Arad Productions.

Pfister, W. (Director). (2014). *Transcendence* [Película]. Warner Bros., Alcon Entertainment, Syncopy Production, MG Entertainment, Straight Up Films.

Proyas, A. (Director). (2004). *Yo, robot* [Película]. 20th Century Fox, Mediastream Vierte Film GmbH & Co. Vermarktungs KG, Davis Entertainment, Laurence Mark Productions, Overbrook Entertainment, Canlaws Productions.

Ray, T. S. (1996). An Approach to the Synthesis of Life. En Margaret A. Boden (ed.), The *Philosophy of Artificial Life* (p. 111). Oxford University Press. (Obra original publicada en 1992).

Russell, S. (1998). *Learning Agents for Uncertain Environments.* Wisconsin: ACM Press.

Russell, S. (2017). Inteligencia Artificial de beneficios Probados. En *El próximo paso: la vida exponencial.* OpenMind BBVA. doi: 10.1080/0952813X.2015.1055826.

Russell, S. y Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

San Román, G. (2010). *Metamorfosis de la lectura*. Anagrama.

Schreier, J. (Director). (2012). *Un amigo para Frank* [Película]. Park Pictures, Park Pictures, TBB.

Scott, L. (Director). (2016). *Morgan* [Película]. Scott Free Productions, 20th Century Fox, TSG Entertainment.

Scott, R. (Director). (1982). *Blade Runner* [Película]. Warner Bros., Ladd Company, Shaw Brothers.

Scott, R (Director). (2012). *Prometheus* [Película]. 20th Century Fox, Scott Free Productions, Dune Entertainment, Brandywine Productions.

Scott, R (Director). (2017). *Alien: Covenant* [Película]. 20th Century Fox, Scott Free Productions, Brandywine Productions.

Searle, J. R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences,* 3(3), 417-457. doi.org/10.1017/S0140525X00005756.

Singer, P. (1975). *Animal Liberation.* Avon Books.

Spielberg, S. (Director). (2001). *A.I.: Inteligencia Artificial* [Película]. Warner Bros., DreamWorks SKG, Amblin Entertainment, Stanley Kubrick Production.

Spielberg, S. (Director). (2002). *Minority Report* [Película]. 20th Century Fox, DreamWorks SKG, Cruise-Wagner Productions, Blue Tulip Productions, Ronald Shusett/Gary Goldman, Amblin Entertainment, Digital Image Associates, Parkes+MacDonald Image Nation.

Spielberg, S. (Director). (2018). *Ready Player One* [Película]. Warner Bros., Amblin Entertainment, De Line Pictures, Village Roadshow, Reliance Entertainment.

Sputore, G. (Director). (2019). *I am Mother* [Película]. Penguin Empire, Rhea Films, Southern Light Alliance, Southern Light Films.

Stadler, J. (2017). Seeing with Green Eyes: Tasmanian landscape Cinema and the Ecological Gaze. *Senses of Cinema, 65*, 1-24.

Stanton, A. (Director). (2008). *WALL-E* [Película]. Walt Disney Pictures, Pixar Animation Studios.

Stone, M. y H. Hirsh. (2006). Artificial Intelligence: The Next Twenty-Five Years. IA Magazine, 26(4), 85-97. doi: https://doi.org/10.1609/aimag.v26i4.1852.

Tegmark, M. (2018). *Vida 3.0*. Editorial Taurus.

Throop, E., Skinner, A., Perrin, A., Steiner, M., Odulana A. y Perrin, E. (2014). Pass the Popcorn: "Obesogenic" Behavior and Stigma in Children's Movies. *Obesity (Silver Spring*), 22(7), 1694-1700. doi: 10.1002/oby.20652.

Toye, F., Jenkins. S., Seers, K. y Barker, K. (2015). Exploring the Value of Qualitative Fesearch films in Clinical Education. *BMC Medical Education*, 15, 214. doi: 10.1186/s12909-015-0491-2.

Truffaut, F. (Director). (1966). *Fahrenheit 451* [Película]. Anglo Enterprises, Vineyard Film.

Turing, A. (1950). Computing Machinery and Intelligence. Mind, 236(59), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Verhoeven, P. (Director). (1987). *Robocop* [Película]. Orion Pictures.

Verhoeven, P. (Director). (1990). *Desafío total* [Película]. Carolco Pictures.

Veruggio, G. (2005). The birth of roboethics. *IEEE International Conference on Robotics and Automation*, Workshop on Roboethics, 1–4.

Villeneuve, D. (Director). (2017). *Blade Runner 2049* [Película]. Warner Bros., Scott Free Productions, Thunderbird Films, Alcon Entertainment, 16:14 Entertainment, Torridon Films.

Vinge, V. (1993) The Coming Technological Singularity: How to Survive in the Post-human Era. *Conference VISION-21 NASA Lewis Research Center and the Ohio Aerospace Institute*, March 30-31.

Weiser, M. (1991). The Computer for the 21st Century. Scientific American, 265(3), 66-75. https://www.jstor.org/stable/24938718.

Wachowski, L. y Wachowski, L. (Director). (1999). *Matrix* [Película]. Warner Bros., Village Roadshow, Groucho Film Partnership.

Williams, C. y Hall, D. (2014). *Big Hero 6* [Película]. Walt Disney Animation Studios, Marvel Studios.

Winfeld. A. F. T. (2012). *Robotics: a very short introduction*. Oxford University Press

Wise, R. (Director). (1979). *Star Trek* [Película]. Paramount Pictures.

Ziman, J. (1994). *Prometheus Bound. Science in a Dynamic Steady State*. Cambridge University Press.

---

[1] El término "robot" fue introducido por Karel Capek en 1921 y procede de la palabra checa "rabota", que utilizó para definir a trabajadores artificiales en una sociedad ficticia y utópica. Por su parte, según la Real Academia Española (RAE), un robot es una "máquina o ingenio electrónico programable que es capaz de manipular objetos y realizar diversas operaciones". Por otro lado, la Academia define la IA como la "disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza una mente humana, como el aprendizaje o el razonamiento lógico". Rusell y Norvig (2019) establecen una interesante clasificación entre los diferentes tipos de IA en base a sus capacidades cognitivas y ejecutoras: sistemas que piensan como humanos, sistemas que actúan como humanos, sistemas que piensan racionalmente y sistemas que actúan racionalmente. En concreto, en los "sistemas que actúan" se daría convergencia entre robótica e IA; es decir, son casos de "embodied AI" ("IA encarnada") (Winfield, 2012). Aunque estas cuestiones técnicas escapan a la pretensión de este ensayo, si se desea tener una visión detallada de la historia, características y tipos de IA se recomienda consultar las obras Inteligencia Artificial (Boden, 2017; López de Mántaras y Meseguer, 2017) y el ensayo De las emociones naturales a la emocionalidad artificial (Biscaia, 2021a, pp. 116-119).

[2] Noticia de "Elperiódico": https://www.elperiodico.com/es/ciencia/20180712/opinion-carme-torras-roboetica-etic-robots-6923499 (consultado el 2/7/21).

[3] En este sentido, Searle propuso en 1980 la diferencia entre IA débil (la que tenemos en la actualidad, con capacidades cognitivas muy alejadas –inferiores– de las nuestras) y IA fuerte (en consonancia con la singularidad tecnológica, con capacidades cognitivas iguales o incluso superiores a las que poseemos). Para el filósofo, la diferencia entra la una y la otra estaría en que la "débil" sólo tiene la capacidad de realizar computaciones (es cierto que muy complejas) y operaciones meramente sintácticas, mientras que la segunda (que considera conceptualmente irrealizable) tendría la capacidad de compresión semántica e intencionalidad.

[4] Lo cual es citado de forma expresa y literal, con el mismo espíritu visionario del científico Raymond Kurzweil, en las películas Ex Machina (2015) o Trascendence (2014).

[5] Tal y como sucede, por ejemplo, en la saga iniciada con The Terminator (1984) o en la franquicia surgida con Transformers (2007).

[6] El Test de Turing (Turing, 1950) fue ideado por uno de los fundadores de la IA, el matemático inglés Alan Turing a partir de sus investigaciones sobre la máquina Enigma en el contexto de la Segunda Guerra Mundial. Su objetivo era averiguar si seríamos capaces de diferenciar a una máquina inteligente de un humano en base a sus respuestas lingüísticas.

[7] Página web de Future of Life: www.futureoflife.org/ai-pen-letter (consultado el 3/3/21).

[8] Normativa europea: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (consultado el 5/5/21).

[9] Ejemplo de normativa reguladora (España): https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673 (consultado el 5/6/21).

[10] Noticia de "La Vanguardia": https://www.lavanguardia.com/internacional/20200217/473611019413/drones-soleimani-militares-inteligencia.html (consultado el 23/8/21).

[11] Noticia de "El País": https://elpais.com/tecnologia/2019/05/17/actualidad/1558075375_210626.html (consultado el 24/7/21).

[12] Noticia de "La Vanguardia": https://www.lavanguardia.com/sucesos/20150702/54433670879/robot-mata-trabajador-fabrica-volkswagen-alemania.html (consultado el 24/3/21).

[13] Que dice así: "obra de tal modo que te relaciones con la humanidad tanto en tu persona como en la de cualquier otro, siempre como un fin y nunca solo como un medio" (Kant, 1999, p. 104. Trad. 1795).

[14] Prácticamente todas las IA y robots del cine son producto de la mano del hombre. Aunque hay casos en los que no, o en los que al menos no queda del todo claro: pasaría con los Transformers o con Sally (de la película Oblivion [2013], que vienen de otro mundo; y con Matrix, que no deja del todo claro su origen.

[15] David-8 avisa de sus intenciones al inicio de Alien: Covenant (2017), al recordar a su creador la condición finita del hombre diciéndole tú morirás, yo no".

[16] Noticia de "Infobae": https://www.infobae.com/tecno/2018/03/23/el-cerebro-humano-detras-de-aiva-el-primer-robot-en-ser-reconocido-oficialmente-como-compositor/ (consultado el 24/4/21).

[17] Noticia de "Confilegal": https://confilegal.com/20180322-como-se-regulara-la-responsabilidad-derivada-de-los-accidentes-de-los-coches-autonomos/ (consultado el 1/5/21).

[18] Los qualia son "el modo como algo se experimenta" (Biscaia, 2021a, p. 129), y forman parte de un interesante debate en el seno de la filosofía de la mente y de la fenomenología. El filósofo Thomas Nagel (1974) lo explica perfectamente en su conocido experimento mental al respecto de cómo nos sentiríamos si fuésemos un murciélago.

[19] Noticia de "The New York Times": https://www.nytimes.com/es/2018/11/30/espanol/zora-robot-ancianos-francia.html (consultado el 11/3/21).

[20] Afirmación avalada por un estudio descriptivo que demostró que casi un 17% de las películas premiadas en el certamen de los Oscar trataron temáticas tecno-científicas relacionadas con algunos de los aspectos analizados en este ensayo, como la vanguardia en ciencias cognitivas y de la computación (Biscaia y Mohedano, 2021).

ISSN 2250-5660 print | ISSN 2250-5415 online