La ciencia ficción como estímulo del debate ético en robótica

Carme Torras
http://www.iri.upc.edu/people/torras

1. Introducción

Si en el pasado se podían analizar las posibles implicaciones sociales de los nuevos desarrollos tecnológicos antes de su despliegue, ahora que las innovaciones son constantes y se integran en nuestro día a día en un abrir y cerrar de ojos, podríamos decir que estamos participando en un experimento mundial sin ningún estudio previo de impacto.

La Internet de las cosas, los *influencers* en la red, los programas que aprenden de la interacción con humanos, los robots de asistencia y compañía, los juegos de ordenador con un propósito y los que persiguen un impacto social, las webs que ofrecen la inmortalidad digital… Estas herramientas pueden, en poco tiempo, modificar el mercado laboral, darle la vuelta a la reputación de alguien, transformar un distrito, cambiar nuestras relaciones —no solo en el trabajo, sino también en nuestras familias y nuestras relaciones personales— o ampliar lo que una persona deja tras morir, que ahora incluye una huella digital (Torras 2018b).

Es difícil predecir —de forma fundamentada— la influencia que tendrá la hiperconectividad y nuestra creciente interacción con las máquinas en la evolución de la sociedad, la economía y la vida cotidiana de las personas. Por tanto, al tratar de anticipar los beneficios y riesgos potenciales de las tecnologías de la información, no solo el público general recurre a la ciencia ficción (CF), sino que algunos académicos e incluso empresas (por ejemplo, *The Tomorrow Project* de Intel) lo hacen. Citando al renombrado escritor de ciencia ficción Neal Stephenson (2011): «La buena ciencia ficción proporciona una imagen plausible y detallada de una realidad alternativa en la que ha tenido lugar algún tipo de innovación convincente. Un buen universo de ciencia ficción tiene una coherencia y una lógica interna que tiene sentido para científicos e ingenieros.»

En este capítulo afirmamos que la ciencia ficción anticipatoria tiene un papel importante que desempeñar en este contexto de desarrollo tecnológico acelerado, ya que puede ayudar a prever futuros posibles alternativos y luego fomentar debates éticos que, con suerte, cristalizarían en iniciativas y regulaciones educativas en beneficio de la humanidad. La Sección 2 describe el paso de la automatización industrial y la informática a las tecnologías digitales centradas en el ser humano. A continuación, las implicaciones sociales de este movimiento y la necesidad de una confluencia con las humanidades se argumentan en la Sección 3, dando paso al surgimiento de la Roboética, que es objeto de la siguiente sección. La Sección 5 describe algunas iniciativas para la educación ética basadas en historias de ciencia ficción y las Secciones 6 y 7 presentan materiales

éticos basados en escenas y cuestiones planteadas en *The Vestigial Heart* (Torras, 2018a). Finalmente, se esbozan algunas conclusiones y tendencias futuras.

2. Inteligencia artificial y robótica centradas en el ser humano

La invención de Internet y la llegada de los teléfonos móviles han dado lugar a la aparición de un sinfín de aplicaciones y de redes sociales. Un fenómeno difícil de prever hace tan solo un par de décadas. Tampoco era previsible que los robots dejarían de estar confinados en las fábricas y se irían introduciendo en entornos cotidianos: asistiendo a discapacitados y ancianos, sirviendo de guías en ferias y museos, de recepcionistas o dependientes en centros comerciales, actuando como compañeros de juegos de jóvenes y adultos, e incluso como niñeras y maestros de refuerzo en las aulas (Figura 1).

FIGURA 1. Ejemplos de robots que ayudan a personas con discapacidad, brindan orientación en espacios amplios como aeropuertos y actúan como recepcionista, compañero de juegos, niñera o maestro de refuerzo.

Los robots no solo encontrarán una mayor aplicación en contextos centrados en el ser humano, como la atención médica, la educación y el entretenimiento, sino también en áreas de servicios como la logística, la limpieza de grandes superficies y el monitoreo ambiental. También ampliarán su repertorio de actividades en el lugar de trabajo, ya que permanecerán en las líneas de producción de las fábricas y, además, colaborarán activamente con los trabajadores humanos como compañeros de trabajo.

Este paso de la robótica al sector servicios (Torras 2016) está alineado con el auge de las tecnologías para las ciudades inteligentes. Aplicaciones tan diversas como la recogida de basura, el reciclaje, la vigilancia y la seguridad vial requieren combinar inteligencia ambiental con robots autónomos. Están en marcha proyectos extremadamente ambiciosos en este sentido, como el desarrollo de una red donde los robots compartirán datos y procedimientos, como mapas de edificios visitados, habilidades de manipulación adquiridas y otros conocimientos aprendidos en un formato común independiente del hardware de cada robot. Esta red estará conectada a la Internet de las cosas, donde los robots podrán obtener modelos de objetos e instrucciones de uso para todo tipo de productos comerciales.

Aunque las tecnologías digitales, como la inteligencia artificial (IA), la robótica centrada en el ser humano y el Internet de las cosas (IoT) a menudo se consideran un paso más en una transformación social que comenzó con las revoluciones agrícola e industrial, introducen una diferencia cualitativa. Ya no se trata solo de mecanizar tareas pesadas y repetitivas en granjas e industrias, o de electrodomésticos que liberan el tiempo de las personas para usarlo de formas más creativas y agradables. La diferencia radica en que estas nuevas tecnologías entran en dominios hasta ahora considerados exclusivamente humanos, como la toma de decisiones, las emociones y las relaciones sociales, que pueden comprometer los valores humanos, moldear decisivamente la sociedad y nuestra forma de vida y, en última instancia, influir en la evolución. de la humanidad.

El requisito más importante de estas nuevas tecnologías es la capacidad de adaptarse a diferentes entornos y situaciones, así como a cada usuario (lo que se conoce como 'personalización'). Para ello, necesitan aprender de las experiencias, es decir, de las interacciones con los humanos y/o con el entorno a través de sensores y actuadores. La adaptabilidad es lo que permite generalizar de una situación a otra, ser tolerante a las percepciones y acciones imprecisas, y desenvolverse adecuadamente en entornos no predefinidos y dinámicos.

Todo esto no puede ser desarrollado solo por tecnólogos y requiere una estrecha colaboración con antropólogos, científicos sociales, psicólogos, abogados, filósofos, es decir, investigadores de las humanidades, dentro de equipos multidisciplinarios.

3. Implicaciones Sociales: Confluencia con las Humanidades

Las comunidades de investigación en robótica e IA son muy conscientes de esta necesaria confluencia con las humanidades y se han emprendido muchas iniciativas conjuntas, como el lanzamiento de proyectos de investigación (RoboLaw 2014; REELER 2017; AI4EU 2019), la publicación de números especiales en revistas científicas (Veruggio *et al.* 2011), la organización de iniciativas interdisciplinarias (MIT y Harvard 2017) y talleres (Stanford Humanities Center

2019), y los foros abiertos sobre *Robotics meet the Humanities* en las principales conferencias de robótica (ICRA Forum 2013; IROS Forum 2018).

Estos foros a menudo están abiertos al público en general y reúnen como oradores invitados no solo a profesores universitarios de departamentos técnicos y de humanidades, sino también a cineastas, escritores de ciencia ficción, bailarines, actores de teatro y representantes de instituciones gubernamentales, todos compartiendo la creencia que el desarrollo de la robótica debe tener en cuenta la comprensión de la humanidad en sus múltiples facetas.

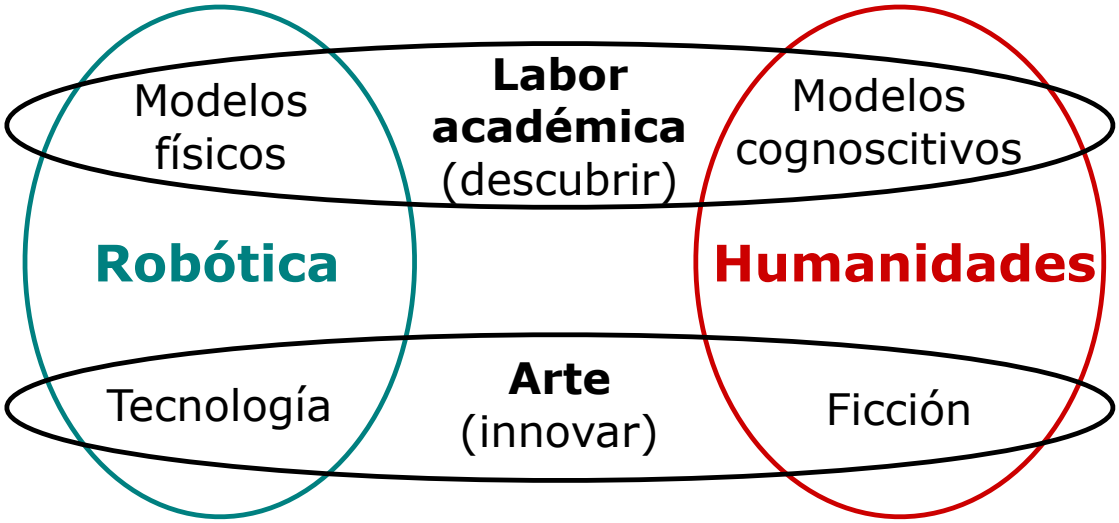

Figura 2. Representación gráfica de posibles puentes entre la Robótica y las Humanidades.

El punto de vista que presenté en esos foros sobre cómo se pueden tender puentes entre la robótica y las humanidades se muestra en la Figura 2. Ambas disciplinas comparten un nivel de modelado destinado a descubrir hechos y relaciones, y el trabajo académico de equipos conjuntos está comenzando a llevarse a cabo relacionando modelos físicos desarrollados para robots con modelos cognoscitivos estudiados en las humanidades. Además, como demuestran los foros mencionados anteriormente, también se pueden establecer puentes a nivel artístico conectando la invención en robótica y en las humanidades, por ejemplo, reuniendo en un mismo equipo a expertos en innovación tecnológica y escritores de ciencia ficción.

Buena muestra de ello en 2012 se creó el Centro para la Ciencia y la Imaginación en la Universidad Estatal de Arizona (http://csi.asu.edu/). La idea surgió de un sugerente discurso de Neal Stephenson (2011), pronunciado en presencia del rector de la universidad, en que el reconocido escritor afirmaba que los científicos de hoy habían perdido la capacidad de pensar y hacer «grandes cosas», como el programa espacial Apolo o el microprocesador. El presidente

respondió que quizás los «culpables» eran los escritores de ciencia ficción al no imaginar futuros ambiciosos que inspiraran a los científicos para hacerlos realidad. Como resultado, el Centro ahora alberga varios grupos de investigación que reúnen a investigadores en ciencias y humanidades para diseñar y esforzarse por lograr objetivos ambiciosos que den forma al futuro.

Uno de los proyectos del Centro es la continuación de una iniciativa lanzada por la empresa Intel, *The Tomorrow Project*, en la que pidieron a cuatro escritores de ciencia ficción que crearan historias imaginando posibles usos futuros de sus productos en fotónica, robótica, telemática y sensores inteligentes. El libro de los cuatro relatos es de acceso abierto (Rushkoff et al. 2012), y desde entonces han aparecido varios volúmenes en los que se proponen soluciones a los mayores desafíos que enfrenta la humanidad hoy en día, a través de las artes visuales y la escritura, todo ello como resultado del trabajo realizado en este centro.

Estudiantes, docentes y profesionales de estas áreas tecnológicas están tomando conciencia de que contar con una formación transversal en ingeniería, que permita salvar las brechas lingüísticas con profesionales de otros campos y de diferentes formaciones, será cada vez más una característica clave para desarrollar una carrera exitosa. Usando marcos teóricos de pedagogía y psicología para fundamentar los resultados de un estudio experimental con 24 personas, Borrego y Newswander (2008) brindan algunas recomendaciones para agencias de financiación, sociedades profesionales, gestores y profesores universitarios para promover la educación interdisciplinaria en ingeniería.

4. Roboética

La confluencia de la IA y la robótica con las humanidades incluso ha dado lugar a una nueva disciplina: la tecnoética o roboética, un subcampo de la ética aplicada que estudia las implicaciones tanto positivas como negativas de la IA y la robótica para los individuos y la sociedad, con miras a inspirar el diseño moral, desarrollo y uso de los llamados robots inteligentes/autónomos, y ayudar a prevenir su mal uso contra la humanidad (Veruggio 2005). La disciplina involucra dos áreas principales: la regulación legal y la educación ética.

En cuanto a la regulación, varias instituciones y colegios profesionales están desarrollando reglamentos para diseñadores, programadores y usuarios de robots. El Parlamento Europeo (2017) publicó algunas directrices bajo el título general de *Civil Law Rules on Robotics*. Otros documentos de este tipo son: *Guide to the ethical design and application of robots and robotic systems,* presentado por la British Standards Institution (2016), *Ethics Guidelines for Trustworthy AI,* publicado por el Grupo de expertos de alto nivel sobre IA de la Comisión Europea (2019) y *Ethically Aligned Design,* que la IEEE Standards Association (2019) abrió a discusión pública.

Hay muchas opciones para integrar la formación ética en las titulaciones universitarias tecnológicas, que van desde incluir una asignatura de ética profesional en el plan de estudios, pasando por permitir a los alumnos cursar algunos créditos o un *minor* en un Departamento de Filosofía, hasta incluso ofrecer una titulación mixta, como la carrera de Informática y Filosofía de la Universidad de Oxford o el reciente grado en Ciencia, Tecnología y Humanidades, fruto de la colaboración entre la Universitat Autònoma de Barcelona, la Universidad Autónoma de Madrid y la Universidad Carlos III de Madrid.

Prestigiosas asociaciones como la Association for Computing Machinery (ACM) y el Institute of Electrical and Electronics Engineers (IEEE) incluyen 18 áreas de conocimiento en sus planes de estudios de Ciencias de la Computación (ACM/IEEE CS curricula 2013), una de las cuales es "Social Issues and Professional Practice", para que "los estudiantes desarrollen una comprensión de los temas sociales, éticos, legales y profesionales relevantes". La necesidad de incorporar el estudio de estos temas no técnicos en el plan de estudios de ACM se reconoció formalmente en 1991.

Los cursos de esta área suelen basarse en textos de filosofía, ciencias sociales o derecho, pero recientemente algunos docentes están recurriendo a relatos de ciencia ficción para ejemplificar situaciones conflictivas, ya que la narración es una buena manera de involucrar a los estudiantes. Algunas experiencias en esta línea se describen a continuación.

5. El papel formativo de la ciencia ficción

La enseñanza de la ética profesional difiere considerablemente de la enseñanza de otras materias dentro de una carrera tecnológica. No se trata tanto de que los alumnos aprendan unos contenidos concretos, sino de sensibilizarlos sobre las implicaciones sociales y éticas de sus futuros trabajos y formarlos para analizar y debatir sobre estos temas. Las personas tienen escalas de valores diferentes y, a menudo, contradictorias, y el objetivo no es unificar los puntos de vista de los estudiantes en torno a un conjunto de reglas, sino aumentar su conciencia y sus habilidades para pensar y discutir. Además, los estudiantes de tecnología no son filósofos. Aunque existen algunas teorías éticas consolidadas que deben conocer, los textos filosóficos suelen ser demasiado abstractos para los informáticos e ingenieros, y se suele optar por una opción pragmática.

Según Sullins (2015), las principales teorías éticas relevantes para las tecnologías digitales son: el consecuencialismo o utilitarismo (maximizar el número de personas que disfrutan de los resultados más beneficiosos), el deontologismo (actuar solo de acuerdo con máximas que podrían convertirse en leyes universales), la ética de la virtud (confiar en el carácter moral de los individuos virtuosos), justicia social (todos los seres humanos merecen ser tratados por igual y debe haber una sólida justificación de las excepciones), bien común (vivir en comunidad impone restricciones al individuo), ética religiosa (las normas provienen de una autoridad espiritual), y

ética de la información (políticas y códigos para regir la creación, organización, difusión y uso de la información).

Dado que ninguna teoría por sí sola es apropiada para abordar todos los problemas éticos que surgen en el diseño y uso de las innovaciones técnicas, la opción pragmática es adoptar un enfoque híbrido. Esta ética híbrida es defendida por Wallach y Allen (2008) como una combinación de teorías deductivas (es decir, aquellas que aplican principios racionales para derivar normas) e inductivas (es decir, aquellas que infieren pautas generales a partir de situaciones específicas).

Ahora bien, ¿de dónde deben provenir estas situaciones específicas? Stephenson (2011) afirma: "Lo que las historias de ciencia ficción pueden hacer mejor que casi cualquier otra cosa es proporcionar no solo ideas para alguna innovación técnica, sino también brindar una imagen coherente de esa innovación integrada en la sociedad, en la economía, y en la vida de las personas". Así, algunos cursos de Ética en Tecnología recurren a relatos de ciencia ficción para ejemplificar situaciones conflictivas. Temas abordados en las obras clásicas de Asimov, Dick, Bradbury, Orwell, Huxley, Hoffman, Shelley, Capek, Wells, Sturgeon, Silverberg o Keyes, como las tres leyes de la robótica, niñeras mecánicas, seguridad versus libertad, falta de privacidad, el totalitarismo tecnológico, los sustitutos emocionales, las réplicas humanoides, la incidencia en el mercado laboral, la responsabilidad moral, la pérdida del control humano, los sesgos en datos históricos, las brechas digitales, o la mejora humana y el posthumanismo, han cobrado fuerza en la actualidad debido al auge de la inteligencia artificial y la robótica social (Torras 2015).

Es natural que los docentes que enseñan Ética en las carreras tecnológicas recurran a este tipo de relatos de ciencia ficción para ejemplificar situaciones sensibles con que los estudiantes pueden encontrar-se en su ejercicio profesional para propiciar una fructífera reflexión y debate sobre las mismas (Torras 2010). Después de impartir el curso "Ciencia Ficción y Ética Informática" cinco veces en la Universidad de Kentucky y dos veces en la Universidad de Illinois en Chicago, Burton et al. (2018) afirman sobre su experiencia: "Usar la ficción para enseñar ética permite a los estudiantes discutir y razonar de manera segura sobre temas difíciles y cargados de emociones sin que el debate se convierta en algo personal". Además de constatar cuán atractiva resulta la narrativa para los estudiantes, los autores describen las conclusiones derivadas de su experiencia, que vale la pena leer en detalle.

Algunas películas y series de televisión recientes también abordan cuestiones de roboética y se utilizan en cursos, tanto en línea como en el aula. Destacaría la serie de televisión *Real Humans* (donde conviven robots casi humanos con personas y a menudo compiten con ellas), la película *Surrogates* (en la que cada ciudadano tiene un avatar controlado desde casa que se mueve por la ciudad e interactúa con las personas), y la novela *The Windup Girl* (en la

que un robot toma conciencia de que fue construido para servir a las personas y se pregunta sobre sus derechos y deberes). La película *Robot and Frank* —que muestra la relación entre un anciano, Frank, y su cuidador robótico— merece una mención especial por su realismo y valor educativo, y en ella se basa un curso en línea de *Ética de los Robots* en el web Teach with Movies (2012), entre otros.

Otros trabajos recientes de ciencia ficción se centran en cuestiones psicológicas y sociales relacionadas con el uso intensivo de teléfonos móviles, la interacción generalizada en las redes sociales, la toma de decisiones automática basada en inteligencia artificial, los juegos de realidad virtual inmersiva y los algoritmos de aprendizaje utilizando *big data*, que han desencadenado interesantes debates (Torras 2018b). En este sentido, la serie de televisión *Black Mirror* es una obra maestra que, en cada capítulo, lleva una determinada tecnología a sus consecuencias más extremas, y la película *Her* retrata a un hombre que se enamora del sistema operativo de su computadora, traduciendo así el cuento *El hombre de arena* de Hoffmann a un entorno digital contemporáneo.

En un contexto académico, Iverach-Brereton (2011) revisa los roles que juegan los robots en las películas desde una perspectiva histórica, prestando especial atención a su grado de autonomía, y utiliza estos escenarios ficticios como herramienta para hacer predicciones sobre cómo los humanos pueden aceptar o no la integración de robots en la sociedad. De manera similar, El Mesbahi (2015) explora cuestiones éticas relacionadas con la interacción humano-robot a través de la lente de treinta películas de ciencia ficción populares y presenta los resultados de una encuesta sobre cómo las personas perciben a los robots en esas películas y quién creen que es responsable de sus acciones, a saber, el propio robot, el diseñador/fabricante, el programador o el usuario.

6. La mutación sentimental - Una guía para el debate sobre "Ética en Robótica Social e IA"

La investigación sobre robots asistenciales que realizamos en mi grupo (Torras 2016, 2019), me suscitó un interés creciente por las implicaciones sociales y éticas de las tecnologías que estamos desarrollando y, en particular, por idear formas de enseñar ética a los tecnólogos. Esto me llevó a probar suerte con la ficción, y en la novela *La mutación sentimental* (Torras, 2012), imaginé cómo ser criado por niñeras artificiales, aprender de maestros robóticos y compartir el trabajo y el ocio con programas de IA afectaría el desarrollo intelectual, emocional y los hábitos sociales de las generaciones futuras. El *leit motiv* de la novela es una cita del filósofo Robert C. Solomon (1977): «son las relaciones que vamos construyendo las que a su vez nos modelan». Se refería a las relaciones humanas con nuestros padres, maestros y amigos, pero la cita se puede aplicar a los asistentes robóticos y todo tipo de dispositivos interactivos, cada vez más presentes en nuestras vidas.

La versión inglesa del libro, que lleva por título *The Vestigial* Heart y ha sido publicada por MIT Press (Torras 2018a), incluye un apéndice con 24 preguntas de roboética y sugerencias para una discusión sobre las situaciones que aparecen en la novela. Además se publicó junto con una guía didáctica y una presentación de 100 diapositivas para impartir un curso sobre *Ética en Robótica Social e Inteligencia Artificial* (ver Figura 3). Abarca seis temas principales: cómo diseñar el asistente «perfecto»; la importancia de la apariencia del robot y la simulación de emociones para la aceptación de los robots; el papel de los programas de IA en el lugar de trabajo y en entornos educativos; el dilema entre la toma de decisiones automática y la libertad y la dignidad humanas; y la responsabilidad civil relacionada con la programación de «valores morales» en los robots.

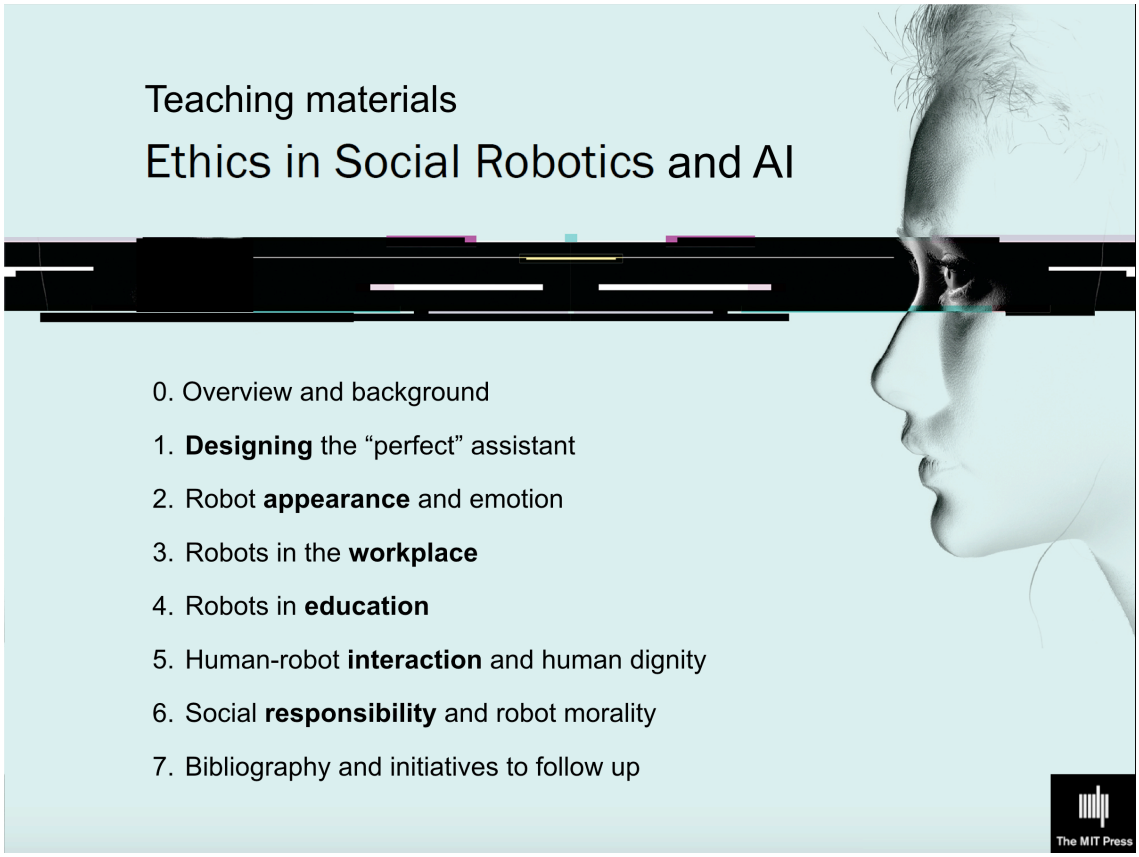

Figura 3. Portada de la presentación de 100 diapositivas en que se detallan los capítulos de la guía del profesor de un curso universitario sobre Ética en Robótica Social e Inteligencia Artificial, descargable como material auxiliar gratuito del web de MIT Press: http://mitpress.mit.edu/books/vestigial-heart

Cada tema se desarrolla a partir de escenas de la novela y sigue la misma estructura, comenzando con algunas citas relevantes, luego se exponen los antecedentes éticos correspondientes, seguidos de cuatro preguntas y pistas para su discusión, y se cierra con algunos

temas revisitados de capítulos anteriores (véase la Figura 4 para el tema a desarrollar en la siguiente sección). El libro, junto con los materiales auxiliares de ética, se ha utilizado en diversas universidades de Norteamérica, Europa y China, en donde la editorial Cheers ha publicado en 2022 la versión en chino.

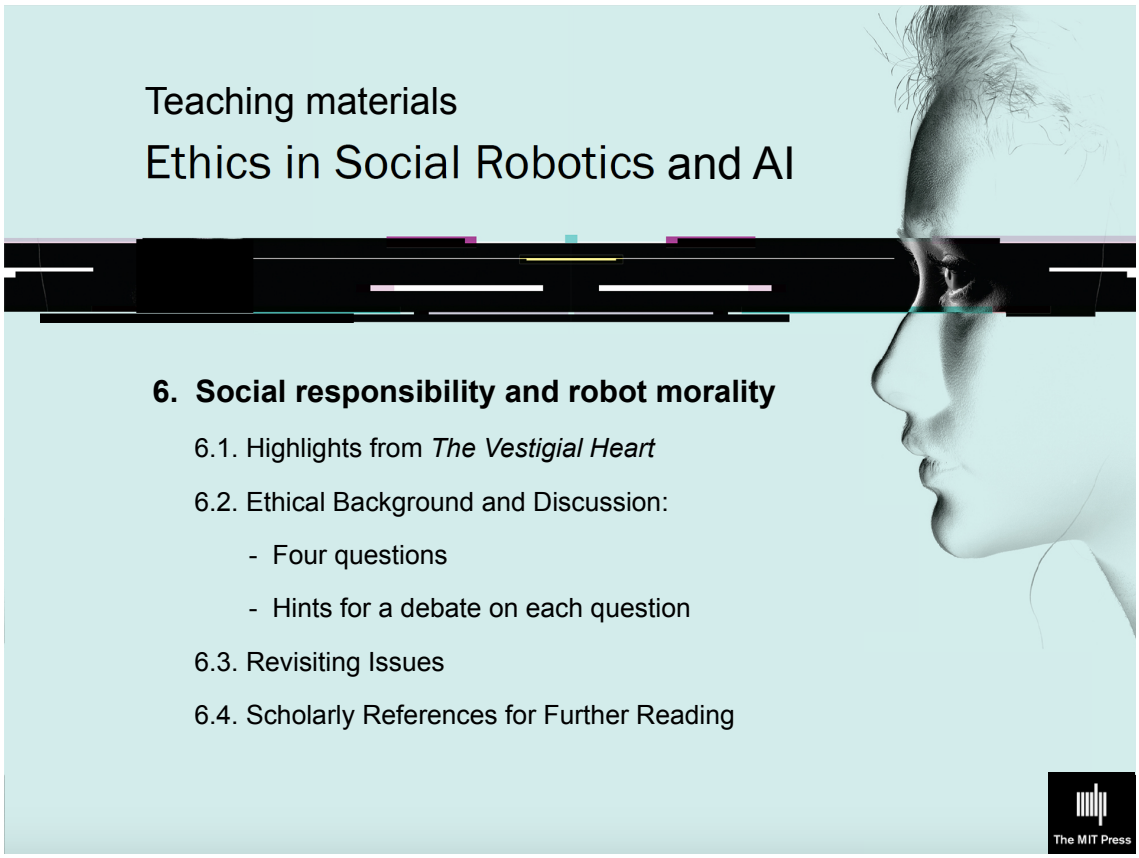

Figura 4. Estructura compartida por todas las secciones de los materiales didácticos asociados a The Vestigial Heart (MIT Press, 2018a), aquí en particular para el tema "Responsabilidad social y moralidad robótica" desarrollado en la Sección 7.

7. Apartado de la guía sobre "Responsabilidad social y moralidad de los robótica"

Para contextualizar las escenas que se describen a continuación, conviene hacer un breve resumen de la trama de *La mutación sentimental*: Celia, una niña de trece años a quien criogenizaron porque sufría una enfermedad terminal, es devuelta a la vida en el siglo XXII para ser adoptada. En una sociedad futura donde cada cual tiene un asistente robótico, Celia choca con la manera de pensar, de actuar y de relacionarse de la madre adoptiva y su entorno, tan distinta a la de su familia biológica. La inadaptación de Celia atrae a Silvana, una masajista emocional que estudia las sensaciones perdidas por los humanos; y a Leo, un joven ingeniero que está diseñando una prótesis de creatividad en la empresa de robots personales líder del mercado, CraftER, dirigida

por el enigmático Doctor Craft. En esa sociedad futura casi todo el mundo es asistido por un robot personal, así Cèlia tiene a ROBbie, Leo tiene a ROBco, y el doctor Craft tiene a Alpha+.

7.1. Pasajes destacados de *La mutación sentimental*

Capítulo 29, página 214:

[Leo] acaba atado de manos y cerebro a una empresa, o peor aún, a su oscuro presidente. [..] es innegable que estamos asistiendo a una auténtica mutación de la especie. Mejor dicho, la estamos provocando. Se mira las manos como si esperase encontrarlas más poderosas y manchadas. Son la prolongación del Doctor, muchas como éstas han engendrado el ejército de robots que ahora mismo circula por el mundo modelando la naturaleza humana. ¡Cuánto poder oculto tras unos sirvientes aparentemente fieles y tan útiles!

Capítulo 29, página 216:

[Celia:] "Ahora en cambio, no puedo ir a ninguna parte si no es en aero'vil y, desde luego, acompañada por ROBbie". [..] "No me quejo de él, pobrecito, es un juguete fantástico, pero esto de que me esté vigilando todo el tiempo es una lata."

Capítulo 30, páginas 225, 228-230:

[Dr. Craft:] "¡No me conformo con disfrutar de su creatividad… ¡lo que quiero es aumentar la mía!" [..] "Conéctame, es una orden."

[..]

[Alpha+:] 17:03 – Estos aparatos accesorios no han pasado el test de homologación: he de extremar las precauciones al manipularlos para evitar penalizaciones de grado máximo. [..] Aunque ROBco me haya aconsejado que monitorice sólo los parámetros básicos, me dispongo a controlar todas y cada una de las constantes vitales. A la mínima que cualquiera de ellas se desvíe del valor basal, detendré el sistema completo. No quiero correr riesgos. Dejando aparte las manías de mi PROP, mi obligación es velar por su salud.

[..]

[ROBco :] —¿Por qué le has conectado sensores en el pecho y en la nuca? Eso no te lo indiqué.

17:08 – "Debo vigilar en todo momento que el Doctor no corra peligro."

—Acepto: *es tu PROP. Pero también hay que evitar que se sienta incómodo.*

—¡Bien dicho! Por fin un ROB que ha aprendido lo que tenía que aprender. ¡Caray con el bioingeniero! Si tenía una chispa de talento, el invento se la ha multiplicado. Venga, quítame todos estos cables y empecemos de una vez que quiero probarlo.

17:09 – "Detente. No vas a tocar nada de esto mientras el Doctor esté bajo mi responsabilidad."

—¿Cómo te atreves a contradecirme, bestia inmunda? Tú mismo vas a quitarme todo esto de encima, ¡y sin rechistar!

17:10 – "De acuerdo. —Sumisamente empieza a retirar los sensores—. Pero en este caso no habrá experimento."

—¡Pero qué te has creído, inútil de mierda! Yo soy el que decide. No te necesito para nada, ¿lo has entendido? ¡Para nada! Sal de aquí antes no te deje seco para siempre.

*17:11 –"*Objeto: *va contra las normas. No puedo abandonar a mi PROP cuando se encuentra en peligro."*

—¿Peligro? —Se levanta como un energúmeno y se dirige directo al robot—. Tú sí que te has convertido en un peligro: me narcotizas, me racionas los placeres, y ahora pretendes impedir que mi mente se expanda. ¡Se acabó, pedazo de chatarra!

17:12 – "¿Qué hace, Doctor? No me apague el sintetizador. Discutámoslo, le ayudaré a obtener lo que quiere."

—¡Ni sintetizador ni hostias! Esta vez te desconectaré del todo... Y así viviré más tranquilo.

17:13 – "Tenga cuidado, Doctor, todo movimiento queda registrado.... y ya sabe usted que Mís"

—¡Se acabó! ¡A tomar viento!

El hombre vuelve a su asiento con aire complacido y se dirige a ROBco:

—Ahora, tú, conéctame lo indispensable para ampliar mi mente tal como haces con tu PROP.

Capítulo 30, páginas 231-232:

[ROBco le informa a Leo:] al conectar al Presidente a la cabina, sus parámetros vitales se han alejado mucho de los valores basales y se ha tenido que aplicar el protocolo de emergencia. La recuperación es tan lenta que el robot teme que el hombre entre en crisis de un momento a otro y quiere informarse de los efectos que podría tener parar la sesión en seco.

Como impelido por un resorte, Leo se pone en pie y grita: "¡no lo hagas, podrías matarle!", a la vez que empieza a dar vueltas por el cubículo cual electrón en acelerador de partículas. Tendría que haberlo previsto, piensa, se trata de un hombre de edad y sus órganos,

habituados a la vida actual, han perdido la flexibilidad necesaria para absorber emociones fuertes.

[..] ROBco insiste:

—Advierto: *cuarenta pulsaciones por minuto. Peligro de parada cardiorrespiratoria.*

—¿Pero qué demonios dices? ¿Y qué está haciendo su robot? Es él, el que debe actuar.

—Informo: *Lo ha desconectado.*

—¿¿¿Cómo???

Leo, abatido, se deja caer en el asiento y Celia le coge la mano como si estuviese acompañando a un enfermo.

—Anuncio: *el Doctor ha muerto.* Pregunto: *¿qué debo hacer?*

—Una trampa mortal... eso es lo que he inventado. Ahora sí que tendré que huir y esconderme. ¿Qué debes pensar de mí, Silvana? De poco vino que no la probases tú también...

Abrumada por la suerte que hubiese podido correr, se ha quedado paralizada. Pero oyendo las palabras de Leo, algo se rebela en su fuero interno.

—No digas eso. Ha sido un accidente, no es culpa tuya. Es él quien ha desconectado a su ROB, ¿no? Seguro que sabía a lo que se estaba exponiendo, tal vez es precisamente lo que pretendía: suicidarse.

—Todo lo contrario. Lo que él pretendía era rejuvenecerse, absorber la vida de otro —sus ojos giran hacia Celia pero evita mirarla—. Desgraciado de mí, he estado jugando alegremente con el material más delicado del mundo...

—Repito: *¿qué debo hacer?*

—Decídselo vosotras, yo en este momento ni siquiera sé lo que debo hacer conmigo mismo.

—Vayamos paso a paso. —La Silvana de los momentos comprometidos acaba de tomar las riendas—. Debe de haber alguien a quien se tenga que informar de lo ocurrido.

—Sí, Míster Gatew... pero me culparán…

—Aclaro: *la señora tiene razón. No pueden darte la culpa porque el registro del Alpha+ guarda constancia de que su PROP lo ha desconectado.*

Capítulo 30, página 234:

[Silvana :] "... Me guste o no, los robots se han convertido en los educadores de los protecno, y mejor será que les ayuden a crecer y a ser creativos, en lugar de convertirles en seres dependientes y rutinarios.

[Leo:] "...Qué más quisiera yo que poder abrir el estimulador de creatividad a todo el mundo, colgarlo en el registro público. —Sonríe a Celia, correría ese riesgo por ella.

Capítulo 31, página 235:

[Leo] estaba trabajando en un aparato supersecreto para su Director y, ahora que él ha muerto, Leo quiere que todo el mundo pueda disfrutarlo. ¿Verdad que parece algo muy fácil? Pues no lo es. Para empezar, tiene que pasarse un montón de días encerrado en su laboratorio, como si estuviese secuestrado, y trabajando el doble: para el nuevo Director y, de tapadillo, para todos aquellos a quien piensa regalarnos el aparato. Me ha prometido que ROBbie será el primero en tenerlo. Y es que no te lo he dicho: la prótesis —así es como lo llaman— tiene que instalarse en un robot y sirve para aumentar la inteligencia de su dueño, si él lo quiere, claro. La verdad, dudo que a Lu o a Fi les vaya a interesar mucho, pero Leo insiste en colgar el invento en el registro público...

7.2. Antecedentes éticos y discusión

Los robots autónomos necesitan tomar decisiones en situaciones imprevistas por sus diseñadores, lo que plantea no solo problemas de confiabilidad y seguridad para los usuarios, sino también el desafío de regular la toma de decisiones automática, en especial en contextos éticamente sensibles. Ello ha llevado al desarrollo del campo de la "ética de las máquinas", con el objetivo final de desarrollar metodologías para maximizar la probabilidad de que un robot se comporte de una manera ética certificable (Lichocki *et al.* 2011).

Hay quien argumenta que los robots pueden tomar mejores decisiones morales que los humanos, ya que su racionalidad no está limitada por los celos, el miedo o el chantaje emocional (Wallach 2010), mientras que otros insisten en que las máquinas nunca podrán ser agentes morales y, por lo tanto, no debería dotárseles de la capacidad de tomar decisiones morales.

Incluso suponiendo que unas reglas éticas generales puedan implementarse en los robots, surgen preguntas sobre quién debe decidir qué moral se codificará en dichas reglas y hasta qué punto las reglas deben ser modificables por el usuario. Por ejemplo, no está claro si y cuándo puede ser aceptable que un robot se entrometa en la autonomía de un usuario para forzarle a comportarse de modo "más ético" con otros seres humanos (Borenstein y Arkin 2016).

En cualquier caso, un robot es una herramienta y, como tal, nunca es legalmente responsable de nada. Por ello, es de suma importancia establecer procedimientos de atribución de responsabilidad, de forma que siempre sea posible determinar quién es legalmente responsable de

las acciones de los robots (Boden *et al.* 2017). En el caso de robots capaces de aprender de la experiencia, dicha responsabilidad podrá ser compartida entre el diseñador, el fabricante y el usuario; también se podrá acusar a un hacker si se puede demostrar su intervención ilegal.

En el Capítulo 30, ***Alpha+*** dice que va en contra de las reglas abandonar a su PROP mientras está en peligro. Pero su PROP, ***Dr. Craft*** , es en última instancia quien decide y apaga su robot. ¿Quién es responsable de las fatales consecuencias? ***Leo*** se siente doblemente culpable, como diseñador de la cabina sensorial, una "trampa mortal", como él la define, y como PROP de ***ROBco***, el robot directamente involucrado en la muerte, mientras que ***Silvana*** afirma que fue un accidente o un suicidio.

Pregunta A: ¿Se puede garantizar la confiabilidad/seguridad? ¿Cómo se puede prevenir el pirateo/vandalismo?

No se puede demostrar que ningún sistema computacional esté completamente libre de errores o a prueba de vandalismo en todas las circunstancias. Sin embargo, se están desarrollando medidas de seguridad y protección para robots cada vez más sofisticadas y las agencias competentes están estableciendo y haciendo cumplir los estándares, como la Robotics Industries Association (https://www.robotics.org/robotic-standards) y la IEEE Standards Association (2018).

En la primera escena destacada del Capítulo 30, ***el Dr. Craft*** le pide a ***Alpha+*** que lo conecte a accesorios que no han sido aprobados por la agencia de estándares, por lo que para salvaguardar la salud del doctor, el robot se configura para maximizar las precauciones. Esto ilustra cómo el robot adopta el principio de precaución: «Cuando una actividad puede dañar la salud humana o el medio ambiente, se deben tomar medidas de precaución incluso si algunas relaciones de causa y efecto no están completamente establecidas científicamente», que a todos los profesionales se aconseja aplicar cuando se utilizan tecnologías sensibles (Veruggio *et al.* 2016). ***Alpha+*** intenta no solo realizar acciones seguras, sino también garantizar la seguridad de su PROP bajo la acción de otros al negarse a abandonar a su PROP cuando puede estar en peligro.

Incluso si los robots están diseñados para ser seguros y protegidos, los usuarios o los piratas informáticos pueden obligarlos a hacer cosas que sus diseñadores no previeron. En un capítulo anterior, ***Leo*** modifica el software de ***ROBco*** de una manera que contraviene las especificaciones de fabricación y las reglas de seguridad. Las regulaciones deben establecer hasta qué punto se debe exigir o permitir que aquellos que poseen u operan robots los protejan de, por ejemplo, mala praxis, robo o vandalismo (Boden *et al.* 2017).

Leroux y Labruto (2012) consideran la cuestión de si el requisito de que haya un "humano en el lazo de control" debe aplicarse sin excepción. Esto puede afectar la seguridad de manera positiva y negativa. Por ejemplo, en los sistemas de control compartido, se deben tomar medidas para evitar que el ser humano se habitúe al funcionamiento automático, de modo que la

persona no se aburra ni se distraiga, desatendiendo así sus deberes. Esto podría implementarse a través de episodios preplanificados de traspaso al controlador humano con el fin de mantener la atención humana y las habilidades necesarias.

Pregunta B: ¿Quién es responsable de las acciones de los robots? ¿Debería ser modificable el comportamiento moral de los robots?

«Un mundo sin consecuencias y costes es un mundo sin opciones significativas. Una vida sin responsabilidad no es la vida del adulto, es la vida del animal, del niño o del robot.» (Roberts 2001). La mayoría de los especialistas en robótica estarían de acuerdo con esta cita de un libro de ficción que atribuye la responsabilidad exclusivamente a los humanos adultos. Sin embargo, dicha atribución se vuelve cada vez más compleja a medida que los robots se vuelven más autónomos y capaces de modificar su comportamiento gracias al aprendizaje y la experiencia, ya que su actuación ya no se basa completamente en su diseño original.

Hasta ahora, si una máquina fallaba, siempre era el fabricante o el programador, o su empresa, ostentaban la responsabilidad. Con la llegada de los robots de aprendizaje, un área gris de responsabilidad abarca a los mencionados junto con el propietario y el usuario. Decker (2007) propuso que «el aprendizaje de los robots debe estar anclado en la responsabilidad del propietario del robot», como se deriva de la fórmula de humanidad de Kant. Obsérvese que este autor se refiere a la responsabilidad legal (*liability*, en inglés), que es la consecuencia jurídica de la responsabilidad moral. En esta línea y de acuerdo con la cita anterior, se ha sugerido que la responsabilidad legal de los tenedores de animales podría utilizarse como modelo para la responsabilidad legal de los tenedores de robots (Schaerer *et al.* 2009).

Otra opción es la metarregulación por parte de una institución de arbitraje de responsabilidad. Para fines de litigio, la traza de decisión de un robot tendría que ser rastreable, siendo una posibilidad instalar una "caja negra" no manipulable para documentar continuamente los resultados significativos del proceso de aprendizaje y las entradas relevantes, que podrían ser verificados por la mencionada institución. Para convencer a ***Leo*** de que no se le puede culpar por la muerte del ***Dr. Craft*** , ***ROBco*** le recuerda que el registro de ***Alpha+*** habrá guardado pruebas de que su PROP lo desconectó.

Los fabricantes podrían protegerse de la responsabilidad pidiéndole al propietario del robot que confirme, por ejemplo, presionando un botón, que el proceso de aprendizaje del robot se ha realizado de manera transparente y que él está de acuerdo. Esta confirmación se registraría en la caja negra y la responsabilidad se colocaría del lado del propietario, como propone Decker (2007). El fabricante del robot solo necesitaría incluir en las instrucciones este procedimiento de confirmación y la obligación de registrarlo en la caja negra.

Peltu y Wilks (2010) contemplan incluso otra posibilidad, a saber, que los desarrollos tecnológicos influyan en los cambios de la ley, de modo que las cosas que no son humanas, como los robots, puedan ser responsables de los daños.

Pregunta 6.C – ¿Cuándo debe prevalecer el bienestar de la sociedad sobre la privacidad de los datos personales?

Esta pregunta surge a menudo en el contexto médico, donde la importancia social de la recolección de datos para fines de investigación puede entrar en conflicto con el derecho a la privacidad de los pacientes. Siguiendo el principio de precaución mencionado en la Pregunta 6.A, se han desarrollado procedimientos de protección de datos y se ha fomentado el uso de formularios de consentimiento informado. A medida que las personas interactúan cada vez más con los robots en un contexto social (por ejemplo, en el rol de agentes de ventas, cuidadores o similares), aumenta el riesgo de divulgación no intencional (o intencional) de información y su uso con fines comerciales.

Calo (2015) describe las formas en que la legislación desarrollada para Internet debe ampliarse para cubrir problemas adicionales planteados por los robots sociales. Por ejemplo, un robot introducido en el hogar podría comprometer la privacidad simplemente creando la sensación de estar siendo observado. Esta preocupación aparece en el Capítulo 29, cuando ***Celia*** se queja de que ***ROBbie la vigila todo el tiempo***. Pero la sensación incómoda puede convertirse en un peligro real si las aspiradoras, los limpiadores de ventanas, los acompañantes de los niños y los asistentes de ancianos y discapacitados llegan a ser espías, especialmente si son pirateados por terceros.

El otro episodio destacado del Capítulo 29, en el que ***Leo*** se siente culpable de haber forjado robots que están esculpiendo la naturaleza humana de formas indeseables, plantea una preocupación más abstracta y de mayor alcance sobre la evolución humana y el bienestar de la sociedad en un mundo cada vez más robotizado. Este llamado a la responsabilidad social subyace en la indagación discutida por Borenstein y Arkin (2016): «¿La principal obligación de un robot es servir a su propietario o a la sociedad humana en general?» Como advierten estos autores, la respuesta a esta pregunta puede tener un profundo impacto en la arquitectura de diseño del robot.

Pregunta 6.D – ¿Qué brechas digitales puede causar la robótica?

Es bien sabido que las tecnologías digitales abren importantes brechas sociales (basadas en la edad, la riqueza, la educación, las zonas geográficas) y los robots pueden ampliar algunas de ellas debido a su costo, materialización física y uso no trivial (Veruggio *et al.* 2016).

Un ejemplo de brecha por edad, educación o simplemente preferencia individual, es cuando un ciudadano solo puede acceder a un servicio interactuando con un agente robótico. Las

regulaciones deben garantizar el derecho de todas las personas al acceso igualitario a los servicios y, por tanto, siempre debe existir la opción de ser redirigido a un agente humano.

La tecnología tiene un fuerte impacto en la distribución global de la riqueza y el poder, provocando diferencias entre distintas partes del mundo. Nagenborg *et al.* (2008) señalan que «los efectos del uso cada vez mayor de robots en el ámbito laboral no pueden juzgarse únicamente observando aquellos países donde se utilizan estos robots. También hay que cuestionarse los efectos en otros países (fuga de cerebros, pérdida de puestos de trabajo, etc.) y la relación entre países que pueden verse afectados por la llamada *brecha robótica*».

Por el contrario, los asistentes robóticos dirigidos a grupos vulnerables podrían reducir las discriminaciones sociales y ayudar a cerrar las brechas antes mencionadas si se tomaran medidas políticas para proporcionar los recursos financieros y los conocimientos necesarios a dichos grupos (Peltu y Wilks 2010). El último episodio destacado en el Capítulo 30 muestra que ***Leo*** es consciente de este problema social y decide sacrificar su libertad inmediata para trabajar para que la prótesis de la creatividad esté disponible para todos. En el Capítulo 31, ***Celia*** le dice a su madre lo orgullosa que está de que él esté dispuesto a hacerlo, incluso si las personas como ***Lu*** y ***Fi*** pueden no estar interesadas en los beneficios que tal prótesis podría brindarles.

7.3. Revisando problemas

En el Capítulo 29 ***Leo*** se refiere al dispositivo de tiempo muerto como una forma de alquilar cerebros, un mecanismo de esclavitud que viola los derechos de los empleados, lo que permite profundizar en algunos de los temas tratados en la sección sobre robots en el lugar de trabajo.

Además, en el Capítulo 30, a ***Leo*** le preocupa que los órganos del ***Dr. Craft*** hayan perdido la capacidad de absorber emociones fuertes, lo que implica que las emociones han desaparecido debido al estilo de vida que prevalece en su sociedad robotizada. Esto puede llevar a revisar el compromiso discutido en secciones anteriores de que nuestra interacción cercana con los robots puede ampliar algunas de nuestras capacidades, pero a riesgo de debilitarnos emocionalmente.

8. Conclusiones y perspectivas de futuro

La creciente interacción de las personas con todo tipo de dispositivos y programas de IA en la vida cotidiana plantea importantes retos sociales y éticos con mucho potencial para dar forma sustancial a nuestro futuro. Esto exige que las carreras universitarias técnicas se acerquen a las humanidades, para que los estudiantes tomen conciencia de los posibles temas sensibles con que se enfrentarán en el desarrollo de sus carreras profesionales y aprendan a reflexionar y discutir sobre ellos.

La filosofía, la psicología, las ciencias sociales y el derecho están brindando perspectivas y conocimientos previos para abordar estos temas, mientras que la ciencia ficción permite especular libremente sobre posibles escenarios y el papel que pueden jugar los humanos y las máquinas en el *pas de deux* que nos conecta irremisiblemente. En esta línea, los materiales educativos basados en historias de ciencia ficción han demostrado ser muy efectivos para involucrar a los estudiantes de tecnología que toman cursos de ética.

Quisiera concluir con unas palabras que la prestigiosa revista Nature incluyó en la introducción del volumen titulado *Many Worlds* (2007), conmemorativo del cincuentenario de la hipótesis de Hugh Everett III sobre los universos paralelos, y que contiene artículos tanto de investigadores en mecánica cuántica como escritores de ciencia ficción. Dice: «La ciencia ficción seria se toma la ciencia en serio. […] No predice qué nos deparará el futuro, pero nos ayuda a intuir qué podría suceder y cómo nos sentiremos cuando una forma de ver el mundo deje paso a otra.» La literatura anticipativa siempre se ha tomado la ciencia en serio y ha tratado de proyectar sus avances hacia el futuro. Parece que la ciencia también está empezando a tomarse en serio esta literatura y a encontrar inspiración en ella. Esta confluencia puede ser sumamente productiva y es una muy buena noticia, que abre interesantes perspectivas para los próximos años.

Agradecimiento

Este trabajo se ha realizado en el marco del proyecto CLOTHILDE (ERC Advanced Grant No. 741930), financiado por el programa Horizonte 2020 de la Unión Europea.

Referencias

AI4EU (2019) AI for Europe. https://www.ai4eu.eu/

Boden M., Bryson J., Caldwell D., Dautenhahn K., Edwards L., Kember S., Newman P., Parry V., Pegman G., Rodden T., Sorrell T., Wallis M., Whitby B. and Winfield A.F. (2017) Principles of robotics: Regulating robots in the real world. *Connection Science*, 29(2): 124-129.

Borenstein J., Arkin R. (2016) Robotic nudges: the ethics of engineering a more socially just human being. *Science and Engineering Ethics*, 22(1), 31-46.

Borrego, M. and L. K. Newswander (2008) Characteristics of successful cross‐disciplinary engineering education collaborations. *Journal of Engineering Education*, 97(2), 123-134.

British Standards Institution (2016) Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems. https://shop.bsigroup.com/ProductDetail/?pid=000000000030320089

Burton E., Goldsmith J. and Mattei N. (2018) How to Teach Computer Ethics through Science Fiction. *Communications of the ACM*, 61(8): 54-64. https://cacm.acm.org/magazines/2018/8/229765-how-to-teach-computer-ethics-through-science-fiction

Calo R. (2015) Robotics and the Lessons of Cyberlaw. *California Law Review*, 103(3), 513-563.

Decker M. (2007) Can humans be replaced by autonomous robots? Ethical reflections in the framework of an interdisciplinary technology assessment, *Workshop on Roboethics*, Intl. Conf. on Robotics and Automation (ICRA'07).

El Mesbahi M. (2015) Human-Robot Interaction Ethics in Sci-Fi Movies: Ethics Are Not 'There', We Are the Ethics! Intl. Conference of Design, User Experience, and Usability, *Lecture Notes in Computer Science*, 9186, 590-598.

European Commission's High-Level Expert Group on AI (2019) *Ethics Guidelines for Trustworthy AI.* Online: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

European Parliament (2017) *Civil Law Rules on Robotics*. Online: http://www.europarl.europa.eu/RegData/etudes/PERI/2017/580862/IPOL_PERI(2017)580862_EN.pdf

ICRA Forum (2013) Robotics Meets the Humanities. http://www.icra2013.org/indexaf5b.html

IEEE Standards Association (2018) IEEE Standards Activities in the Robotics and Automation Space. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/robotics.pdf

IEEE Standards Association (2019) Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Online: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

IROS Forum (2018) Robotics Meets the Humanities: Social Relationship, Ethics, Art and Science Fiction. http://www.ynl.t.u-tokyo.ac.jp/forum/iros2018/

Iverach-Brereton C. (2011) Learning from the Future: What Science Fiction Can Teach Us about Social Robots. *Advanced Introduction to Human Robot Interaction (AHRI 2011)*, University of Manitoba, Winnipeg, Canada.

Leroux C., Labruto R. (2012) Ethical, Legal, and Societal Issues in Robotics, euRobotics: The European Robotics Coordination Action, Deliverable D3.2.1.

Lichocki P., Kahn Jr P.H., Billard A. (2011) A survey of the robotics ethical landscape. *IEEE Robotics and Automation Magazine*, 18(1), 39-50.

«Many Worlds» (2007) *Nature*, 448(7149): 1-104.

MIT and Harvaard (2017) The Ethics and Governance of Artificial Intelligence Initiative. https://aiethicsinitiative.org/

Nagenborg M., Capurro R., Weber J., Pingel C. (2008) Ethical regulations on robotics in Europe. *AI & Society*, 22(3), 349-366.

Peltu M. and Wilks Y. (2010) In Wilks Y. (ed.): Summary and discussion of the issues. In Wilks Y. (Ed.) *Close engagements with artificial companions: key social, psychological, ethical and design issues*, pp. 259-286, Amsterdam, The Netherlands: John Benjamins Publishing Company.

REELER project (2017) REsponsible Ethical LEarning with Robotics. https://reeler.eu/

Roberts R. (2001) *The Invisible Heart - An Economic Romance*, MIT Press.

RoboLaw project (2014) Deliverable D6.2 - Guidelines on Regulating Robotics. Online: http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf

Rushkoff D., Hammond R., Thomas S., Markus H. (2012) *The Tomorrow Project*. Bestselling Authors Describe Daily Life in the Future. Intel. Santa Clara.

Schaerer E., Kelley R., Nicolescu M. (2009) Robots as animals: A framework for liability and responsibility in human-robot interactions. Proc. 18th IEEE Intl. Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 72-77.

Solomon R.C. (1977) *The Passions: The Myth and Nature of Human Emotion*, Anchor Press.

Stanford Humanities Center (2019) Workshop on AI, Humanities and the Arts. http://shc.stanford.edu/events/ai-humanities-arts-workshop

Stephenson N. (2011) Innovation Starvation. *Wired*, 27 October. Available at: http://www.wired.com/2011/10/stephenson-innovation-starvation

Sullins J.P. (2015) Applied professional ethics for the reluctant roboticist. In *The Emerging Policy and Ethics of Human-Robot Interaction*, edited by L.D. Riek, W. Hartzog, D. Howard, A. Moon and R. Calo, Workshop at the 10th ACM/IEEE International Conference on Human-Robot Interaction, Portland.

Teach with Movies website (2012) Robot Ethics with clips from the movie *Robot and Frank*. http://teachwithmovies.org/robot-and-frank/

Torras C. (2010) Robbie, the pioneer robot nanny: Science fiction helps develop ethical social opinion. *Interaction Studies*, 11(2), 269-273.

Torras C. (2012) *La mutación sentimental*, Editorial Milenio.

Torras C. (2015) Social robots: A meeting point between science and fiction. *Metode Science Studies Journal-Annual Review*, 5, 111-115. Available online: http://www.redalyc.org/pdf/5117/511751360016.pdf

Torras C. (2016) Service robots for citizens of the future. *European Review*, 24(1), 17-30.

Torras C. (2018a) *The Vestigial Heart. A Novel of the Robot Age*, together with a teacher's guide and a 100-slide presentation to teach a course on *Ethics in Social Robotics and AI*. The MIT Press. https://mitpress.mit.edu/books/vestigial-heart

Torras C. (2018b) Social networks and robot companions: Technology, ethics and science fiction. *Metode Science Studies Journal*, 99: 47-53.

Torras C. (2019) Assistive robotics: Research challenges and ethics education initiatives. *DILEMATA: International Journal of Applied Ethics*, 30: 63-77.

Veruggio G. (2005) The birth of roboethics. *Roboethics Workshop at the IEEE Intl. Conf. on Robotics and Automation*, Barcelona.

Veruggio G., Operto F., Bekey G. (2016) Roboethics: Social and ethical implications of robotics. In Siciliano B., Khatib O. (Eds.) *Springer Handbook of Robotics*, *2nd edition*, Chapter 80, pp. 2135-2160, Springer.

Veruggio G., Solis, J. and Van der Loos M. (2011) Roboethics: Ethics applied to robotics [from the guest editors]. *IEEE Robotics and Automation Magazine*, 18(1): 21-22.

Wallach W. (2010) Robot Morals and Human Ethics: The Seminar, *Teaching Ethics*, 11(1), 87-92.

Wallach W. and Allen C. (2008) *Moral machines: Teaching robots right from wrong*. Oxford University Press.