Publicado en "Revista de Filosofía".
and Psychology 11, páginas 881–897
(2020)", que debe citarse para hacer
referencia a este trabajo.

Comprender la IA: ¿podemos y debemos sentir empatía? con robots?

Susanne Schmetkamp1

Published online: 28 April 2020

Abstracto
Ampliando el debate sobre la empatía con seres humanos, animales o personajes de ficción para incluir las relaciones entre humanos y robots, este artículo propone dos perspectivas diferentes desde las que evaluar el alcance y los límites de la empatía con los robots: la primera es epistemológico, mientras que el segundo es normativo. El enfoque epistemológico nos ayuda aclarar si podemos empatizar con la inteligencia artificial o, más precisamente, con robots sociales. El principal enigma aquí se refiere, entre otras cosas, a qué es exactamente con los que empatizamos si los robots no tienen emociones ni creencias, ya que no tienen una conciencia en un sentido elaborado. Sin embargo, al comparar robots con ficticios personajes, el artículo muestra que todavía podemos empatizar con los robots y que muchos de los Las explicaciones existentes sobre la empatía y la lectura de la mente son compatibles con tal punto de vista. por eso Al hacerlo, el artículo se centra en la importancia de tomar perspectiva y afirma que También atribuimos a los robots algo así como una experiencia en perspectiva. El enfoque normativo examina el impacto moral de empatizar con los robots. En este sentido, el documento Analiza críticamente tres posibles respuestas: estratégica, antibarbarización y pragmatista. Esta última posición se defiende subrayando que cada vez nos vemos más obligados interactuar con robots en un mundo compartido y llevar a los robots a nuestra moralidad.
La consideración debe verse como una parte integral de nuestra comprensión de nosotros mismos y de los demás.

Palabras clave Empatía. inteligencia artificial. Robots humanoides. Interacción. Tomando perspectiva. Personajes de ficción. Ética

1. Introducción

Los debates sobre la empatía o, más ampliamente, la comprensión interpersonal han sido un pilar de la erudición en una amplia gama de disciplinas. Sin embargo, si bien mucho ha fue escrito sobre la capacidad humana de empatizar con personas reales o personajes de ficción

* Susanne Schmetkamp
Susanne.schmetkamp@icloud.com

[1] Departamento de Filosofía, Universidad de Friburgo, Friburgo, Suiza

(para resúmenes recientes, ver Coplan y Goldie 2011; Maibom 2017), hasta hace poco los filósofos han descuidado un poco el papel de la empatía en las interacciones entre humanos y robots (HRI) (cf. Brinck y Balkenius 2018; Lin et al. 2017). Sin embargo, en línea con el creciente número de estudios sobre las emociones u otras características de los sistemas de inteligencia artificial,1 ha habido mucho interés filosófico en la posibilidad y necesidad de interactuar y empatizar con diferentes formas de inteligencia artificial, especialmente con las llamadas redes sociales. robots.2 Este interés también ha dado lugar a debates sobre el valor de la empatía para la sociedad en general o para la atención sanitaria y la terapia en particular (Coeckelbergh 2018; Darling 2016; Engelen 2018; Loh 2019; Misselhorn en prensa; Vallor 2011). Cada vez está más claro que, en el futuro, los robots y los androides (es decir, robots que parecen humanos) se convertirán en actores más o menos independientes y con habilidades sociales. Como tales, están llamados a convertirse en compañeros importantes y cada vez más capaces de establecer relaciones con los seres humanos (Benford y Malartre 2007; Breazeal 2002; Dumouchel y Damiano 2017). Además, los sistemas de aprendizaje profundo (Kasparov 2017) se implementarán en muchas (hasta ahora) profesiones humanas, lo que no solo mejorará o facilitará algunas tareas o desafíos (en la investigación médica, por ejemplo); también podrían obligarnos a reconsiderar algunos conceptos clave como inteligencia, agencia, conciencia, autonomía, emociones o perspectivas (Schneider en prensa).

Como han demostrado los estudios (Leite et al. 2013), la forma y el éxito de las relaciones entre humanos y robots a menudo dependen de características humanas, como la capacidad de los robots para expresar emociones, interactuar y ejecutar decisiones (más o menos) autónomas. . Estas capacidades también son importantes para la comprensión empática recíproca.3 Si bien los seres humanos también reconocen y atribuyen emociones en relación con formas virtuales abstractas o incluso ante dispositivos técnicos (los mejores ejemplos son los teléfonos inteligentes y las computadoras), para nuestra interacción cooperativa y colaborativa con los robots. – particularmente en el contexto médico o de atención sanitaria– una fuerte semejanza humana podría ser crucial para que estas interacciones tengan éxito. A medida que crece la presencia de robots sociales humanoides en la sociedad, también crece la necesidad de examinar y dar forma a nuestras interacciones con ellos. La generación actual de robots ya es capaz de expresar una variedad de emociones: la IA humanoide "Sophia", por ejemplo, conoce 60 expresiones faciales diferentes e incluso parece capaz de comunicarse con sentido del humor e ironía. Sin embargo, los robots no tienen conciencia en el sentido de experiencia subjetiva4 y no poseen humor ni emociones en un sentido elaborado (Boden 2016; MacLennan 2014; Scheutz 2011). Sin embargo, es posible que tengan algo que pueda considerarse análogo a las emociones humanas y a algunos procesos mentales. Además, a la luz de las ideas recientes de la filosofía de la cognición encarnada, podría darse el caso de que el cuerpo y el comportamiento humanos y la cognición "extendida" sean los que ayuden a los humanos a reconocer a los androides como compañeros y similares a ellos en algunos aspectos, mientras que permaneciendo totalmente distinto de

[1] Un grupo del MIT Media Lab y de la asociación de estándares IEEE aboga por el concepto de inteligencia "extendida" en lugar de "artificial". Con esta nueva narrativa "extendida" se quiere garantizar que los robots no sustituyan a los seres humanos, sino que los apoyen y cooperen con ellos. Juntos establecieron el Consejo de Inteligencia Extendida CXI, ver https://globalcxi.org (consultado por última vez el 12.12.2019).

[2] Un proyecto financiado por el ERC, ubicado en la Universidad de Glasgow y dirigido por Emily Cross examina en particular la socialización de los seres humanos con inteligencia artificial y la importancia de la interacción y las relaciones con los robots para la cognición social. Uno de los focos está en la capacidad de los robots para ser compañeros, http://www.so-bots.com (consultado por última vez el 20.12.2019).

[3] Sobre el fenómeno del “valle inquietante” ver más abajo.

[4] Al menos cuando seguimos una posición antifisicalista.

ellos en los demás (Benford y Malartre 2007, 181; Hoffmann y Pfeifer 2018; Newen et al. 2018). 5 La

empatía se considera en términos generales como una forma crucial de aprehender y reexperimentar los estados mentales de los demás mediante la lectura de la mente, el intercambio emocional y/o o coexperiencia experiencial (ver, por ejemplo, Engelen y Röttger-Rössler 2012; Goldman 2006; Stueber 2018; Zahavi 2014).6 En filosofía, la empatía generalmente se distingue del contagio afectivo y de la simpatía o compasión moral.7 Mientras que esta última apunta al bien ser de los demás y quiere promoverlo (o al menos no impedirlo) (Darwall 1998), la empatía, en primera instancia, conduce a la comprensión de los procesos mentales de los demás, como las emociones o las creencias. En contraste con el mero contagio emocional, debe existir una diferenciación entre el yo y el otro (De Vignemont y Jacob 2012). Sigue habiendo un importante debate sobre este punto y se han propuesto una variedad de definiciones y enfoques que buscan abordar preguntas como: ¿Cómo percibimos y accedemos a los estados y experiencias de los demás? ¿Cómo se debe caracterizar ¿Cuál es el resultado de este proceso? A grandes rasgos, las teorías predominantes – provenientes de la filosofía de la mente o la fenomenología– son las Neuronas Espejo o Teoría de la Resonancia (MNT) (Gallese 2001), la Teoría Visual (TT) (Fodor 1987; Gopnik y Wellman 1994), la Teoría de la Simulación (ST))) (De Vignemont y Jacob 2012; Goldman 2006, 2011; Stueber 2006), Teoría de la percepción directa (DPT) (Zahavi 2011) con sus variaciones de la Teoría de la interacción (IT) (Gallagher 2008, 2017) y la Teoría de la narratividad (NT) (Gallagher y Hutto 2008). Además, existen teorías híbridas y pluralistas que combinan dos o más enfoques, como la percepción directa y la imaginación8 (Schmetkamp 2017, 2019; Dullstein 2013; para excelentes resúmenes, ver Newen 2015; Stueber 2018; Zahavi 2014; 2018). Dado este conjunto diverso de enfoques, hagamos algunas distinciones adicionales entre empatía cognitiva (como TT) o empatía afectiva (como ST o MNT).9 Al preguntarnos si podemos

empatizar con los robots, la sección 2 se centrará en los muchos aspectos epistémicos Dimensiones de las interrelaciones empáticas con los robots: ¿Qué percibimos y entendemos si no hay realmente emociones, experiencias subjetivas o

---

[5] El artículo se centra principalmente en robots humanoides. Una razón para esto es que ayuda a limitar el alcance del artículo; Otra razón es la suposición de que las características humanas facilitan nuestra interacción social con la inteligencia artificial y hacen más plausible que tratemos a los robots como socios sociales. Sin embargo, también podemos empatizar con formas más abstractas de IA atribuyéndoles estados y motivos emocionales (ver Isik, Koldeewyn, Beeler y Kanwisher 2017). Estoy muy agradecido a un crítico por este comentario.

[6] Es muy controvertido si la empatía presupone o implica un reflejo afectivo, una lectura teórica de la mente, una toma de perspectiva simulada, una comprensión emocional y/o una comprensión experiencial, y actualmente este debate no tiene fin a la vista (ver, por ejemplo, Zahavi 2018). Muchos filósofos enfatizan que la lectura de la mente es algo distinto de la empatía y que la empatía es "algo extra". Aquí, sin embargo, he intentado aplicar todos los diferentes enfoques. Sin embargo, mi propia posición es fenomenológica.

[7] Sin embargo, un problema de todo el debate es que no existe un consenso conceptual sobre lo que es e implica la empatía. El proyecto sobre robots sociales, financiado por el CEI, por ejemplo, define la empatía como algo que implica tanto la correspondencia emocional como el comportamiento prosocial. Sin embargo, en filosofía, la empatía no suele verse como una emoción o actitud moral (ver Cross et al. 2018; Zahavi 2018).

[8] Por ejemplo, haciendo referencia a las posiciones clásicas de Stein o Dilthey y combinando la percepción directa con la representación imaginativa (“Vergegenwärtigung”) (ver también Gallagher 2019).

[9] Kanske (2018) distingue entre empatía afectiva propiamente dicha y teoría cognitiva de la mente. Mientras que la primera capacidad nos permitiría sentir lo que otros sienten, la otra nos ayudaría a comprender lo que otros piensan o creen. Aunque reconozco las diferencias, no distinguiré aquí la mentalización de la empatía, sino que examinaré diferentes formas de comprender otras mentes bajo el término general de empatía, ya que éste es el término central en el debate filosófico actual.

perspectivas en un sentido rico? ¿O los robots tienen algo parecido a emociones, creencias y experiencias? ¿Tienen una visión individual del mundo (Schmetkamp 2017) o una narrativa (Gallagher 2012), ya que al menos están encarnados y contextualizados? Al comparar los robots con personajes de ficción, la respuesta será afirmativa: Sí, hasta cierto punto podemos empatizar con los robots de forma cognitiva, afectiva e incluso experimental, ya sea infiriendo, sintiendo, interactuando o imaginando cómo perciben y se mueven a través de ellos. su mundo, así como comprendemos de manera plural (Vaage 2010) cómo un personaje ficticio (por ejemplo, en una película) percibe su mundo, actúa y siente. El aspecto crucial aquí será que atribuimos una perspectiva individual al otro. Lo comprendemos independientemente de si esta perspectiva sólo está narrada, proyectada o programada.10 La segunda pregunta, que se discutirá en la sección 3, es si debemos empatizar con los robots. Esta pregunta tiene dos lados: podemos preguntarnos si la empatía con los robots tiene una mera función estratégica con respecto a la mejora de la comprensión recíproca dentro de la interrelación humano-robot, o podemos preguntarnos si la empatía tiene un impacto ético tal que tienen el deber de empatizar con los robots (para obtener una descripción general sobre el tema de la ética y la IA, consulte Boddington et al. 2017). Si, por ejemplo, podemos comprender epistémicamente lo que los robots perciben, pretenden o incluso podrían “sentir”, también podemos predecir lo que harán a continuación. En general, esto podría ser útil en términos de nuestras interacciones con ellos.11 Obviamente, esto se refiere a un "debería" estratégico o racional. El segundo significado de la pregunta da lugar a una respuesta normativa: ¿debemos empatía hacia los demás en un sentido moral? ¿Y qué ganamos ellos o nosotros, como empatizadores, con esto desde un punto de vista moral? Considerando esta pregunta, a primera vista podría parecer obvia una respuesta kantiana, que sigue el precedente sentado por la visión de Kant sobre los animales y que puede modificarse para aplicarse a la inteligencia artificial: es decir, debemos empatizar, según el argumento, para evitar. ""barbarización moral". Al final, el artículo no tomará ni el camino estratégico ni el kantiano, sino que propondrá una respuesta pragmática y relacional. Esta respuesta está relacionada con las otras dos. Sin embargo, enfatiza el impacto de la interacción y la autocomprensión.

2 ¿Podemos empatizar con los robots?

Por razones de espacio, me concentraré en robots que tienen cara y cuerpo, muestran expresiones y comportamientos similares a los humanos, están destinados a interactuar con seres humanos y, por lo tanto, están encarnados e integrados en nuestra vida cotidiana y, como tales, están sujetos a las condiciones sociales. valoración por parte de los humanos. Una segunda razón para este enfoque es la suposición de que los robots con características y expresiones similares a las humanas son probablemente incluso más capaces de generar confianza y provocar respuestas emocionales similares a las de los humanos reales (Brinck y Balkenius 2018; Mori 2005) y, en este sentido, son más probables . .ser reconocidos y aceptados como socios en la interacción social. Aunque los estudios en psicología cognitiva han demostrado que también podemos sentir empatía o leer la mente con sistemas que tienen poco parecido físico (Bretan et al. 2015), una apariencia humana

[10] El artículo se centra en la cuestión epistemológica. No responderá a la pregunta metafísica de si Los robots o la IA tienen conciencia.

[11] En cuanto al despliegue de sistemas de aprendizaje profundo en medicina, entre otras características, es necesario confiar en la máquina inteligente y comprender lo que va a hacer, por ejemplo en una interacción médico-paciente.

importante para el uso de robots como cuidadores o colegas en la atención sanitaria (Vallor 2011).12 Pero, ¿qué tipo de empatía está en juego aquí? ¿Reflejamos las expresiones de los robots? ¿Interpretamos y predecimos su comportamiento? ¿O empatizamos de una manera más fenomenológica e interactiva?

Como mínimo, la empatía puede definirse como la capacidad humana de comprender los estados mentales de los demás y reexperimentarlos de una forma u otra, aunque sigue siendo un tema de debate si el sujeto empático necesita sentir lo mismo que el otro. Algunas teorías restringen los objetos de la empatía a las emociones de las personas y sus expresiones como indicadores de estados afectivos. Otros son más amplios e incluyen otros procesos cognitivos como objetos de empatía, como creencias, deseos y sus respectivas razones (para una descripción general, consulte Batson 2009; Slote 2017). Una definición destacada implica una condición de isomorfismo: el empatizador y el objetivo están en el mismo estado afectivo o al menos en uno similar (De Vignemont y Singer 2006). Sin embargo, como han argumentado algunos críticos, la empatía no implica necesariamente que repliquemos los estados mentales de los demás (Zahavi y Michael 2018). Tampoco tenemos que preocuparnos por el otro en un sentido más elaborado.

Como es bien sabido, la "exageración" actual en torno al tema de la empatía puede atribuirse en gran medida al descubrimiento de las llamadas "neuronas espejo" (Iacoboni et al. 1999; 2011). A grandes rasgos, las neuronas espejo son aquellas neuronas situadas en una zona del cerebro que se descargan tanto para la observación como para la ejecución de acciones similares. Este proceso de imitación se ha aplicado a la comprensión de las emociones humanas: al observar la expresión afectiva de otra persona –como una cara triste– se activarían las mismas neuronas que si nosotros –como observadores– hubiéramos puesto una cara triste y sintiéramos tristeza. Mientras que esta teoría ha sido ampliamente criticada (Hickok 2014) y rechazada como teoría de la empatía, otros la han invocado en su enfoque más elaborado de la empatía. En su explicación de la Teoría de la Simulación (ST), Alvin Goldman, por ejemplo, distingue entre una forma de lectura mental de bajo y alto nivel o una "ruta del espejo" y una "ruta reconstructiva", aunque la "resonancia" emocional "" se implementa a través de ambas rutas (Goldman 2006, 2011). Las neuronas espejo son la parte principal de los procesos de bajo nivel mediante los cuales comprendemos los estados mentales de otra persona de forma inmediata y automática. En un nivel más complejo y superior, simulamos el estado del otro en nuestra propia mente y luego llegamos al conocimiento de cómo se siente el otro, no implementando una teoría, sino imitando el comportamiento de los demás en nuestra mente y luego proyectando nuestra propio proceso mental sobre el otro. Según ST, simulamos, a través de una perspectiva en primera persona, estar en la situación del otro y utilizamos nuestros propios mecanismos mentales para generar pensamientos, creencias, deseos y emociones. Durante las últimas dos décadas, la ST –junto con su oponente la Teoría Teoría (TT)– ha dominado el debate sobre la lectura de la mente. TT afirma que nuestra comprensión de otras mentes se basa esencialmente en la psicología popular, que es innata o adquirida durante la primera infancia (Baron-Cohen 1995). La TT supone que hacemos inferencias basadas en teorías para comprender a los demás ... y el deseo. La TT ha sido criticada por ser demasiado teórica y demasiado general (Zahavi 2014; pero cf. Fodor 1987). Sus detractores afirman que TT no tiene en cuenta lo otro concreto, ni tampoco

[12] Sin embargo, empíricamente sigue siendo incierto si los robots deben ser realmente humanos en HRI (Brinck y Balkenius 2018).

[13] Un problema, por supuesto, es cómo entendemos el término "comprensión". Monika Dullstein (2012) tiene
Se ha demostrado que los relatos de la Teoría de la Mente utilizan una noción bastante diferente de los relatos fenomenológicos.

reconocer la encarnación y el arraigo de los demás. Además, se considera que tanto TT como ST están engañados por una visión oclusiva cartesiana falsa de la mente, como si no pudiéramos percibir lo que sucede en la mente de otro (Zahavi 2011, 2014) . Por el contrario, las explicaciones fenomenológicas enfatizan la encarnación y el arraigo de los seres humanos y argumentan que somos capaces de ver directamente en el rostro y las expresiones corporales del otro lo que él está experimentando: desde este punto de vista, no tenemos que inferir o imaginar lo que está sintiendo. ; sólo necesitamos percibirlo. Además, lo hacemos en un contexto situacional compartido y a través de la interacción. Por esta razón, este enfoque se denomina Teoría de la Percepción Directa (DPT).

(Zahavi 2011, 2014) o Teoría de la Interacción (TI) (Gallagher 2001; 2012). A diferencia de TT, DPT e IT sostienen que no adoptamos una postura de tercera persona hacia los demás ni los observamos. Además, DPT e IT también argumentan que no tenemos un acceso indirecto imaginativo a los demás. En cambio, interactuamos socialmente de una manera segunda personal, mediante la cual dos "tú" se reconocen de forma complementaria y recíproca (Dullstein 2012; Engelen 2018; Zahavi y Michael 2018). Los límites de la DPT surgen obviamente en situaciones en las que el otro no está presente para nosotros: por ejemplo, cuando alguien nos cuenta una historia sobre otra persona, o si leemos una novela, miramos una película o vemos una obra de teatro donde las experiencias de otros están presentes. de alguna manera mediados por alguien más (por ejemplo, un narrador), no tenemos encuentros directos. Por lo tanto, todos estos casos son casos en los que el otro viene dado por la narración, a veces incluso dentro de un marco ficticio. Por eso algunos filósofos añaden que una narrativa es esencial para comprender otras mentes o suscitar empatía en algo más que el sentido más básico.

Daniel D. Hutto (2008) formulando la Hipótesis de la Práctica Narrativa (NPH). Según esta tesis, entendemos las razones de los demás para actuar, sus creencias y deseos sólo cuando también tomamos en cuenta las circunstancias individuales, la historia del sujeto, su situación actual, sus esperanzas y experiencias, sus rasgos de carácter, etc. En otras palabras, según NPH, para comprender la situación de alguien, tenemos que confiar en la "historia" de esa persona (Gallagher 2012). Esta visión también permite empatizar con "monstruos o extraterrestres de otros planetas, tal como se retratan en las películas" (Gallagher 2012). Sin embargo, aquí se necesita una especie de imaginación: hay muchísimos casos -no sólo sino especialmente en nuestro trato con la ficción- en los que confiamos en nuestra imaginación como una forma de hacer disponible algo que no está presente para nosotros. Incluso una de las primeras pioneras de los enfoques fenomenológicos de la empatía, Edith Stein (1989), afirmó que la imaginación o "representación" es la etapa del proceso de comprensión empática. Esta es la razón por la que algunas teorías de la empatía combinan un enfoque de segunda persona con una forma de representación imaginativa de la situación, narrativa y/o perspectiva concreta del otro (Schmetkamp 2019; Gallagher y Gallagher 2019).15 [14] juega un papel crucial dentro de una multiplicidad Independientemente de si debemos o no Si consideramos todas estas diferentes explicaciones como teorías de la empatía o, más ampliamente, como teorías de la comprensión interpersonal, para cada explicación podemos preguntar lo siguiente desde una perspectiva descriptiva y epistemológica: ¿Cómo podemos

[14] Es difícil dar una traducción exacta del concepto de Stein de "Vergegenwärtigung". La traducción al inglés (Stein 1989) utiliza "representación" o "acto representacional" (Stein 1989: 8) como un "dato" representado no primordial de las experiencias indirectas o de los demás (análogas a la memoria, la expectativa y la fantasía) (ibid.). En el debate a menudo se pasa por alto que Stein propone un modelo escalonado de empatía, según el cual el primer nivel es la percepción directa de la experiencia del otro, siendo el segundo nivel una especie de reflexión y toma de perspectiva (Stein 1989:10) .

[15] Gallagher definió recientemente la empatía de la siguiente manera: "La empatía podría [...] no sólo [contar] como algo que sucede, sino como un método; y eso […] implica ponerse en la perspectiva o situación del otro". (2018). Al hacerlo, Gallagher amplió su enfoque narrativo hacia un enfoque perspectivo (combinando la narrativa con la perspectiva subjetiva).

¿Empatizar con la IA, por ejemplo con robots con apariencia humana, si la explicación respectiva fuera la más plausible? Por ejemplo, si observamos las expresiones y/o acciones de un robot, se podría argumentar que automáticamente resonamos e imitamos el comportamiento expresivo. Si queremos predecir qué hará el robot a continuación, también podríamos implementar una teoría psicológica popular e inferir sus razones para actuar. Podríamos simular lo que haríamos si estuviéramos en su situación y luego proyectar nuestra experiencia en ellos. O, en encuentros directos, podríamos percibir interactivamente sus acciones. Podríamos considerar su inserción dentro de un contexto narrativo y comprender la estructura intencional de sus emociones sin replicar al mismo tiempo su contenido "cualitativo". Empíricamente hablando, estas formas interactivas de comprensión ciertamente ocurren.

Sin embargo, se pueden plantear algunas objeciones metafísicas y epistemológicas obvias. El principal problema es que los robots en realidad no sienten ni experimentan nada. Tampoco tienen realmente estados mentales como deseos o creencias, porque no tienen conciencia. Dicho esto, también parece extraño hablar de la perspectiva individual o la narrativa personal de un robot. En la medida en que la empatía se dirige hacia los estados mentales y el "estar en el mundo" de alguien, la respuesta sería: no podemos empatizar con los robots.

Sin embargo, se podrían dar dos posibles respuestas: primero, los "estados mentales" de los robots a menudo se describen como "estados computacionales" que se considera que tienen una estructura análoga a los estados mentales humanos. Entonces, si asumimos que los robots tienen algo comparable a los estados mentales humanos, ¿también tienen algo como emociones o experimentan sentimientos con los que empatizamos? Según algunas explicaciones filosóficas actuales sobre las emociones, los estados o procesos emocionales exhiben una estructura compleja que consta de componentes cognitivos y afectivos (De Sousa 1987; Nussbaum 2001): cuando sentimos ira, nuestra ira está dirigida hacia un objeto que evaluamos como molesto. . A esto también se le llama teoría de la evaluación, que implica que hacemos juicios sobre los objetos en nuestro entorno con respecto a su psicología de relevancia para nuestros objetivos. Si las emociones sólo consistieran en este mero componente cognitivo, podríamos suponer que los robots tienen emociones en un sentido mínimo.
Podríamos argumentar que los robots actúan según un conjunto de razones que se basan en un conjunto de creencias sobre el mundo. Sin embargo, las emociones pueden abarcar más que eso: la ira, por ejemplo, también se siente a nivel corporal y sensacional; se siente, por ejemplo, frustrante y estrechante. Dicho esto, la ira también tiene connotaciones negativas, de las que tomamos conciencia propioceptivamente (Colombetti 2013). Los cuerpos de los robots, aunque no sean puramente virtuales, están hechos de metal o plástico y, lo que es más importante, no están relacionados con un concepto rico de conciencia: en el sentido de que ésta se experimenta a sí misma como un ser emocional. No puede sentir referencialmente lo que es estar en un cuerpo de plástico. Además, como han argumentado los relatos de emociones narrativas, las emociones complejas suelen estar incrustadas dentro de un marco narrativo: podemos contar una historia sobre su activación y desarrollo (Goldie 2000). Y por último, pero no menos importante, los seres humanos son capaces de afrontar creativamente sus sentimientos y emociones: pueden aprender nuevas emociones y son capaces de modificar algunas y cultivar otras.

Sin embargo, esto también podría ser posible para y con los robots. El punto crucial aquí es que nosotros, de manera perfectamente intuitiva, también atribuimos emociones a las máquinas. Al colaborar con robots, podríamos adoptar una "postura intencional". Este concepto, originado en el trabajo de Daniel Dennett, implica que tratamos a un objeto cuyo comportamiento queremos predecir como un agente racional; le atribuimos creencias y deseos y, sobre esa base, predecimos su comportamiento (Dennett 1987). Pero aun así, este enfoque se basa en una teoría de la lectura de la mente y no en la teoría de la interacción fenomenológica que los fenomenólogos tienen en mente. Sin embargo, si consideramos que la conciencia implica una experiencia fenoménica, parece difícil aplicar algo más que eso.

La Teoría de la Mente explica la empatía al HRI. En otras palabras, el problema relacionado con la compatibilidad de las teorías fenomenológicas para la HRI parece ser el aspecto fenoménico de los estados mentales, particularmente el lado emocional y experiencial de las emociones. Mientras que podríamos teorizar (TT) sobre los componentes cognitivos de, o simular, la situación de decisión de un robot (ST) y luego inferir o proyectar a partir de nuestras conclusiones sobre la situación del robot, sería difícil hablar de una comprensión empática de la capacidad afectiva del robot. y estados sensacionales de forma no proyectiva. Si ampliamos el problema al concepto de "experiencia" -el término central del enfoque fenomenológico (DPT)- las cosas se vuelven aún más complicadas. Como se describió anteriormente, según la DPT y sus variaciones, en nuestra interacción social con los demás percibimos empáticamente sus experiencias, y lo hacemos desde un punto de vista recíproco de segunda persona. "Experiencia" es un término fenomenológico de elaboración e implica aspectos existenciales y cualidades fenomenológicas. Experimentamos subjetiva y conscientemente nuestro mundo o cómo es sentir o hacer algo, por ejemplo, percibir una mesa roja como roja y cómo se siente ese enrojecimiento. La DPT supone que experimentamos las experiencias fenoménicas de los demás directa e intersubjetivamente, aunque no replicando el carácter cualitativo exacto de una experiencia, sino más bien atendiendo a la estructura intencional de la perspectiva del otro (Gallagher 2012; Zahavi y Michael 2018) . Para que este proceso funcione, la interacción intercorpórea y cara a cara es importante.16 Ahora bien, mientras que esta última está (al menos básicamente) garantizada cuando cooperamos y colaboramos con robots, faltan algunos criterios cruciales de esta relación intersubjetiva. : así como los robots no sienten emociones, tampoco tienen una experiencia subjetiva con su contenido fenomenal y su impacto existencial ... cómo es para ellos.

No tenemos que emplear inferencias, imitaciones o proyecciones teóricas. Experimentamos que el otro tiene experiencias fenomenales. Dicho esto, desde una perspectiva fenomenológica, parece difícil empatizar con los robots. Sin embargo, al comparar la inteligencia artificial con personajes de ficción, propondré una posible solución y también demostraré que no sólo leemos la mente o reflejamos el comportamiento de los robots, sino que es posible, al menos hasta cierto punto, aplicar un enfoque fenomenológico, es decir, empatizar interactivamente con la "experiencia" de perspectiva de los robots. Y el argumento va incluso más allá de esta analogía: cuando interactuamos con robots en un entorno compartido, desarrollamos una intencionalidad compartida e incluso una historia conjunta, y esto es crucial para nuestra relación con los robots (Coeckelberg 2018) .
Sin embargo, al igual que nuestra comprensión empática de los personajes de ficción, nuestra capacidad de imaginación es crucial aquí.

Hagamos una analogía: comúnmente se supone que la empatía juega un papel esencial en nuestro trato con narrativas y personajes ficticios, ya sea en una novela, una película o una obra de teatro. Desde la década de 1990, ha habido un debate considerable dentro de la filosofía de la literatura y el cine sobre si la "empatía" debería subsumirse bajo el término general de "compromiso emocional" con personajes de ficción en general (por ejemplo,

[16] Sin embargo, la versión narrativista de los enfoques fenomenológicos implica un componente imaginativo que nos permite comprender la estructura intencional mediante la imaginación narrativa, por ejemplo, si no se da una interacción intersubjetiva (Gallagher y Gallagher 2019) .

[17] Es una pregunta similar a la del llamado "experimento mental zombi", que analiza si podemos asumir o atribuir una conciencia en el caso de los zombis, que son como nosotros en todos los aspectos físicos pero no tienen experiencias conscientes en un rico sentido (Chalmers 1996; Dennett 1991).

Plantinga 2009; Smith 1995). Otras formas de compromiso incluyen el contagio emocional y el intercambio emocional –especialmente con respecto a los efectos cambiantes de una ficción–, la simpatía o compasión moral, las emociones negativas como la antipatía y los afectos sinestésicos (Plantinga 2009; Schmetkamp 2017) . Como han señalado muchos estudiosos del cine, la empatía desempeña un papel epistémico crucial al permitir al espectador seguir la narrativa y permanecer apegado a los personajes (Smith 1995).18 Dejando de lado el otro complejo debate sobre la llamada "paradoja de la ficción": que analiza si podemos sentir emociones reales hacia entidades ficticias y si estas emociones son racionales (Yanal 1999) – y suponiendo que realmente sentimos y tenemos que sentir empatía hacia personajes ficticios, todavía tenemos que explicar cuál es la mejor manera de conceptualizar la empatía en el caso de ficción. Si bien en general estoy convencido de que utilizamos diferentes formas de empatía, lectura de la mente y comprensión (es decir, todo el espectro de comprensión de los estados mentales de los demás) cuando vemos una película o leemos una novela, mi suposición es que hay un aspecto que es particularmente importante. Vital: Los personajes de ficción expresan y representan ciertas perspectivas individuales sobre su mundo (ficticio). Estas perspectivas se narran en el mundo diegético de la película o novela; Además, a menudo también están enmarcados por un narrador implícito o explícito. Están incrustados dentro de una narrativa plausible. O, dicho de otra manera: una narrativa es una representación estructurada y moldeada de eventos desde una determinada perspectiva (Goldie 2012: 8) y en la ficción, los personajes encarnan, expresan y representan esas perspectivas arraigadas.

La importancia de las perspectivas para la ficción, y de hecho para nuestro compromiso empático con ella, se debe en parte al hecho de que una ficción generalmente (aunque no siempre) tiene diferentes perspectivas técnicas: una historia generalmente se cuenta en primera o tercera persona. perspectiva. Pero aún más importante es que una perspectiva es una visión del mundo. Dicho esto, una perspectiva significa cómo una persona está incrustada en el mundo, cómo percibe el mundo, cómo lo experimenta. Esta perspectiva está moldeada por y, a su vez, moldea emociones, experiencias, historias, recuerdos; está influenciado por y en sí mismo influye en los rasgos de carácter, los juicios y las creencias (Schmetkamp 2017). Cuando, por ejemplo, estamos en un estado de ánimo depresivo, vemos nuestro mundo desde un punto de vista diferente (es decir, depresivo o melancólico) que si estamos en un estado de felicidad.

Ahora podemos decir que los personajes de ficción tienen (o más bien expresan y actúan) una perspectiva en la medida en que están focalizados y narrados por un narrador que construye y dirige su visión del mundo. Como lectores o espectadores, los atendemos como si tuvieran una perspectiva y podemos imaginar cómo sería tener esa perspectiva. La empatía con los personajes de ficción implica una especie de toma de perspectiva centrada en el otro sin reducir este proceso a uno de mera simulación o proyección egocéntrica.19 Es más , una ventaja de las narrativas de ficción es que imparten las perspectivas de los demás de una manera condensada. . Las ficciones nos brindan la oportunidad de sumergirnos en perspectivas que pueden ser similares o totalmente diferentes a las nuestras, y muchas veces lo hacen de forma intensa, condensada y comprensiva.

---

[18] La empatía como toma de perspectiva es de hecho una capacidad que permite a los espectadores comprender las narrativas y perspectivas de los personajes. Sin embargo, como forma de comprensión sensible de por qué el personaje siente, piensa y actúa como lo hace, también es un resultado. Por lo tanto, Coplan (2011) y Goldie (2000) han argumentado que la empatía es a la vez un proceso y un resultado .

[19] Misselhorn planteó un argumento similar al señalar que "al ver el T-ing de un objeto inanimado imaginamos percibir un T-ing humano" (2009: 353).

Al comparar los robots con personajes de ficción, se destaca una característica congruente central: ambos no tienen realmente emociones ni creencias conscientes, pero pueden expresarlas y representarlas. Y en parte sobre esta base nosotros, como receptores o empatizadores, les atribuimos estados mentales similares a los humanos (Weber 2013). Sin embargo, también los experimentamos como entidades encarnadas de alguna manera con las que interactuamos. Como ha sostenido la filósofa fenomenológica del cine Vivian Sobchack, el cine y sus personajes no son sólo proyectos; tienen un cuerpo y una voz, y permiten experiencias cuasi intersubjetivas entre ellos y sus destinatarios (Sobchack 2004). Incluso podrían permitir impresiones táctiles. Esta característica encarnada también se aplica a los robots, quizás incluso más.

Sin embargo, existen algunas diferencias cruciales. En primer lugar, a diferencia de los robots, los personajes de ficción carecen de una capacidad que es vital para toda explicación intersubjetiva de la empatía: a saber, la interacción recíproca. En nuestras relaciones con personajes de ficción, debemos imaginar que los personajes tienen las emociones, experiencias y perspectivas expresadas, pero no interactuamos recíprocamente con ellos. Además, los personajes de ficción no pueden vetar cualquier cosa que les atribuyamos. Por el contrario, en nuestros encuentros con robots, hay al menos una entidad existente y presente, encarnada e incrustada, que interactúa con la que podemos desarrollar una relación. El robot puede oponerse a algo; por ejemplo, si yo fuera un paciente y no quisiera tomar mi medicamento, se le podría encargar al robot que se asegurara de que lo haga. En segundo lugar, se podría objetar que, a diferencia de los personajes de ficción, los robots no tienen (todavía) una perspectiva experiencial o una narrativa individual, como se mencionó anteriormente. De hecho, las ficciones ofrecen una rica imagen de cómo alguien puede percibir y evaluar su mundo; y a través de estos marcos y prácticas narrativas ampliamos nuestro horizonte y aprendemos nuevas emociones o matices emocionales. Sin embargo, las emociones y experiencias de los personajes de ficción también se narran únicamente dentro de un marco narrativo particular; Su desarrollo depende tanto de lo que un narrador ha diseñado dramatúrgicamente como de cómo lo reciben los lectores o espectadores en su propio trasfondo intelectual y experiencial. Las emociones y experiencias ficticias tienen menos flexibilidad y creatividad que sus contrapartes humanas. Dicho esto, cabe preguntarse si los personajes de ficción todavía pueden contrastarse con los robots. Los personajes de ficción en realidad no experimentan nada; De manera similar, los robots no tienen experiencias en un sentido rico que incluya a los qualia. Sin embargo, los robots al menos perciben su entorno, lo categorizan, lo evalúan e interactúan dentro de él. Tienen una manera de ver y estar en el mundo; están encarnados y contextualizados. Si pensamos en el famoso ejemplo antirreduccionista de Thomas Nagel "¿Cómo es ser un murciélago?" (Nagel 1974) nunca seremos capaces de comprender por completo la perspectiva experiencial de otros seres; un murciélago, o al menos eso dice su argumento, tiene un sistema perceptivo totalmente diferente que no se puede comparar con la percepción humana. Sin embargo, los científicos están descubriendo constantemente nuevos hechos sobre entidades no humanas como peces o plantas (Coeckelberg 2018: 148), y un argumento aquí sostiene que incluso si nunca supiéramos cómo es estar con ellos, al menos podemos experimentar ellos y su perspectiva en nuestra relación con el

Si intentamos comparar la perspectiva del robot con la nuestra, hay algunas similitudes y, por supuesto, también muchas diferencias. Pero éste no es un fenómeno nuevo en nuestra cognición social de otras mentes. En primer lugar, un robot percibe literalmente (p. ej. visualmente) el mundo de una determinada manera (quizás como un humano, quizá no). En segundo lugar, como inteligencia artificial, también tiene perspectiva en el sentido de que percibe y evalúa el mundo que la rodea, cómo resuelve problemas, etc. La perspectiva del robot está lejos de ser una perspectiva en un sentido elaborado, como la de los seres humanos, pero es una perspectiva epistémica y evaluativa: un robot sabe algo y emite juicios sobre el mundo. También podemos afirmar que tiene una perspectiva motivacional, pues un robot actúa sobre el

de sus creencias.20 Aún más importante, los robots están incrustados en un contexto que percibimos o con el que interactuamos. Entonces, mi respuesta a la pregunta de si podemos empatizar con los robots es: sí. Además, todas las cuentas existentes son más o menos aplicables a HRI. Por supuesto, la siguiente pregunta que debemos hacernos es: ¿deberíamos hacerlo?

3 ¿Deberíamos empatizar con los robots?

Dado el análisis anterior, supongamos que podemos empatizar con los robots humanoides de manera plural, es decir, podemos sentir, interactuar o inferir de sus "creencias", "emociones", "experiencias" y "perspectivas". Pero ¿por qué deberíamos sentir empatía por ellos? A la luz del creciente uso de robots en la medicina, la salud y el cuidado de personas mayores, por ejemplo, parece mucho más plausible que los robots empaticen con los pacientes que al revés. De alguna manera deben involucrar algunas sensibilidades hacia las necesidades de los pacientes, mientras que, a su vez, los pacientes humanos pueden necesitar un compañero empático.
Dicho esto, parece que la investigación hasta ahora ha sido principalmente una prueba teórica para revelar cuáles de las diferentes explicaciones de la empatía son compatibles con la HRI. Pero, ¿existe también alguna razón por la que nosotros, como humanos, también deberíamos sentir empatía por los robots? Esta pregunta es relevante, ya que la interrelación entre humanos y robots sólo es exitosa y fructífera si ambos realmente interactúan entre sí, y estas interacciones podrían presuponer, de una forma u otra, un compromiso empático.

Se podrían dar tres argumentos para esta tesis normativa:

1. Un argumento estratégico;
2.Un argumento antibarbarización; 3. Un
argumento pragmático y comunitario compartido.

El primer argumento, el estratégico, no es directamente un argumento normativo moralmente relevante. Retoma la idea de que para interactuar con éxito, de alguna manera debemos ser capaces de inferir y comprender qué está haciendo nuestra contraparte interactiva. Más precisamente, es posible que queramos sentir empatía, adoptar una perspectiva o leer otra mente para lograr mejor nuestros objetivos. Nuestra interacción con los robots y nuestra empatía con ellos, en este sentido, sólo sirve para algo más; es mero instrumental. La noción "debería" se refiere a un imperativo hipotético. En este sentido, los robots son considerados más herramientas que colaboradores. De hecho, aquí no se los ve como agentes morales o pacientes, que tienen un estatus moral (Coeckelberg 2018).

Más sustantivo y moralmente normativo es el segundo argumento de la no barbarización o cultivo. Al no sentir empatía por los demás, según el argumento, corremos el riesgo de volvernos insensibles. A su vez, la empatía podría cultivar un comportamiento prosocial y mejorar nuestro carácter moral. Antes de explorar los principales problemas de esta tesis, explicaré dos de sus raíces: un argumento kantiano y uno aristotélico. El argumento kantiano se planteó originalmente con respecto a la relación entre humanos y animales. Eso implica

[20] Una vez más, se podrían presentar argumentos similares para otras formas de IA de agentes no humanos, por ejemplo, formas virtuales abstractas. El objetivo de este artículo son los robots humanoides con los que los seres humanos cooperan y colaboran. Para que esto tenga éxito, los seres humanos deberían atribuir a la IA no sólo estados mentales básicos, sino también una perspectiva y una narrativa. Esto podría ser importante para las intenciones colectivas y la atención colectiva.

que no debemos ser crueles con los animales porque esto dañaría o corrompería nuestro carácter moral en general. Según este argumento, los animales son sólo pacientes morales indirectos, sin tener un estatus moral propio, ya que Kant vincula el estatus moral a la competencia para actuar de forma autónoma según razones y atribuye esta competencia sólo a las personas. El mismo argumento se aplicaría entonces a los robots sociales que no serían per se destinatarios morales: esto se debe a que podrían no tener autonomía en un sentido elaborado. Sin embargo, al no sentir empatía por ellos, faltaríamos el respeto a una condición crucial de la humanidad : nuestra moralidad: si tratamos a los animales de manera inhumana, nos convertimos en personas inhumanas. Esto lógicamente se extiende al tratamiento de los compañeros robóticos. […] También puede prevenir la desensibilización hacia criaturas vivientes reales y proteger la empatía que tenemos unos por otros" (Darling 2016: 19).

El argumento aristotélico va en una dirección similar. Implica que podemos cultivar nuestras emociones tomando una perspectiva, definiendo así la toma de perspectiva como un distanciamiento del propio punto de vista de primera persona, o compartiendo emociones y familiarizándonos así con nuevas emociones (Nussbaum 2011; Rorty 2001) . Mientras que la visión kantiana enfatiza el problema de la barbarización, la visión aristotélica enfatiza el impacto ético de cultivar algo mediante la empatía: nuestras emociones, percepción moral, imaginación y poder de juicio.

Como dije, aquí surgen algunos problemas, que acosan a la visión kantiana en particular: el primero es que el reconocimiento de un estatus meramente indirecto de no personas o seres sin "racionalidad" es insatisfactorio: es contraintuitivo, antropocéntrico y excluye muchas más entidades que los no humanos (Gruen 2017). ¿Pero esto también afecta a las entidades inanimadas? Por lo tanto, la pregunta sigue siendo: ¿Qué es lo que dañamos cuando usamos la violencia contra robots que tal vez no sientan nada de una manera elaborada y subjetiva? ¿Tienen un concepto de respeto y dignidad? ¿Tienen derechos morales? Estas complejas preguntas tendrán que permanecer sin respuesta aquí, ya que requerirían una investigación propia y dedicada. Otra objeción contra la visión kantiana es que el argumento se basa en una explicación específica de la empatía como conducta prosocial. Esto no sólo implica una comprensión de otras mentes, sino que también implica la preocupación por el bienestar de otra entidad. Es decir, el empatizador no sólo está interesado en las experiencias del otro y "siente" ellas; también están motivados para aliviar el sufrimiento del otro o promover su bienestar. Y si fuéramos crueles con ellos y no respetáramos su bienestar -por ejemplo, golpeándolos o violándolos (si pensamos en los robots sexuales)- esto también repercutiría en nuestro comportamiento hacia los humanos. Sin embargo, como se señaló anteriormente, el impacto ético de la preocupación o el cuidado es más bien el impacto de la simpatía o la compasión como emoción moral sui generis y, como tal, es distinta de la empatía (Darwall 1998) . Como han demostrado especialmente los fenomenólogos, la empatía no es necesariamente una actitud positiva hacia los demás, sino que también puede conducir a un comportamien Una persona sádica tiene que ser empática también en este sentido, es decir, comprende el sufrimiento del otro pero no quiere aliviarlo (Breithaupt 2019; Zahavi y Michael

[21] Kant escribe: "Si un hombre mata a su perro porque el animal ya no es capaz de prestarle servicio, no falta a su deber para con el perro, porque el perro no puede juzgar, pero su acto es inhumano y daña en sí mismo la humanidad que tiene. es su deber mostrar hacia la humanidad. Si no quiere reprimir sus sentimientos humanos, debe practicar la bondad hacia los animales, porque quien es cruel con los animales se vuelve duro también en su trato con los hombres" (Kant 1997: 212) .

2018) 22 En otras palabras, un enfoque kantiano combina algunas diferenciaciones conceptuales importantes, a saber, entre empatía y compasión. Aquí se podría plantear otra objeción: empíricamente, no está nada claro por qué alguien que no empatiza con los demás se convierte necesariamente en barbarizado (Brinck y Balkenius 2018).

Sin embargo, desde un ángulo más optimista, algunos sostienen que la comprensión empática frecuente o la toma de perspectiva pueden ayudarnos a aprender cómo podrían sentirse o pensar los demás. Cuanto más implementamos la empatía, más capaces seremos de involucrarnos con los demás, tanto en nuestras interacciones cotidianas como en encuentros más inusuales. Es más, esto podría convertirnos en personas más tolerantes o más virtuosas. Nuevamente, esto se argumenta con respecto a los personajes y narrativas de ficción. Prestar atención a las perspectivas y experiencias de los demás tiene, como famosamente afirmó Richard Rorty, un valor ético, ya que al hacerlo abandonamos nuestra perspectiva egocéntrica (Rorty 2001). Pero, por supuesto, podríamos adoptar este argumento para HRI: empatizar con los robots mejoraría nuestras interacciones cooperativas y colaborativas en la medida en que nos familiarizaríamos más con ellos. Esto lleva al tercer argumento que comparte algunas características tanto con el enfoque estratégico como con el kantiano/aristotélico, pero enfatiza la interacción, la relación y la autocomprensión social de los empatizadores.

Este argumento (que describe mi propia posición) retoma el enfoque de Rorty pero lo modifica hacia una tesis aún más pragmatista y relacional de la cognición social y sus precondiciones. En contraste con el enfoque kantiano y aristotélico, esta visión parte de un punto de vista antiantropocéntrico y enfatiza la relación interactiva entre seres humanos y robots. Esta posición supone que empatizar con los demás -en todas sus variantes, pero especialmente en la tradición fenomenológica interactiva- puede permitirnos familiarizarnos con el "estar en el mundo" de los demás y, por lo tanto, ampliar nuestros horizontes, cambiar nuestras perspectivas y dar forma a nuestras interacciones sociales. .y comportamiento moral hacia otros no humanos.

En vista de las perspectivas, mi suposición aquí es que incluso podemos hablar de (futuros) robots y sistemas de aprendizaje profundo23 como si tuvieran una visión específica de su propio mundo. Esta visión será en algunos aspectos similar y en otros diferente de la perspectiva humana. Películas de ciencia ficción como HER (EE.UU. 2013) han imaginado lo que podría llegar a ser la IA independiente: sistemas superinteligentes que superan con creces las capacidades del pensamiento humano. Empatizar con los robots humanoides con los que interactuamos cada vez más (en el contexto de la atención sanitaria, por ejemplo) podría ayudarnos a prepararnos para futuros desarrollos. Por el momento, sin embargo, lo cierto es que, en la medida en que ya compartimos acciones y entornos con los robots, y en la medida en que la empatía y la cognición social pueden mejorar nuestras interacciones con los demás, también podemos suponer que nuestras interacciones con los robots se beneficiarán de un punto de vista empático, aunque no meramente en un sentido instrumental y estratégico. Esto también podría tener un efecto formativo, argumento que, como se ha señalado, también se ha adelantado en relación con los mundos ficticios. Pero el punto más importante es que tal visión toca la cuestión de cómo queremos entendernos a nosotros mismos: tomar en serio a los robots como compañeros sociales debería implementarse como parte de nuestra autocomprensión como humanos y miembros de sociedades democráticas.

La forma en que interactuamos con los robots depende en gran medida de cómo pensamos en ellos: como ocurre con las herramientas

[22] El fenómeno de que los empáticos pueden volverse aún más crueles cuanto más humanos se llaman robots "valle inquietante" (ver Misselhorn 2009; Mori 2005).

[23] O como las llama Susan Schneider: "mentes futuras" (en prensa).

que se supone que interactúan desde un punto de vista meramente instrumental, o como socios que deberíamos tomar en serio por sí mismos. Son, por tanto, la relación y la comunidad compartida las que pasan a primer plano aquí. Semejante posición subraya el impacto pragmático y fenomenológico de las interacciones. Esto también podría tener implicaciones para el estatus de los robots como agentes morales y pacientes morales, como sostiene Mark Coeckelbergh: "La cuestión de la posición moral siempre está relacionada con la cuestión de quién es parte de la comunidad moral y qué juegos morales ya se juegan. ." (Coeckelberg 2018: 149).
En lugar de una implementación de la moralidad de arriba hacia abajo, Coeckelbergh aboga por una perspectiva de abajo hacia arriba. Al considerar a los robots como compañeros en un contexto relacional y al empatizar con su narrativa de perspectiva, desarrollamos una relación con ellos que a su vez tiene efectos en cómo los vemos moralmente (ibid.).24 Sin embargo, discutir el estatus moral iría más allá el alcance de este trabajo. Como se mencionó anteriormente, la empatía no es en sí misma una emoción moral o una actitud de preocupación. Pero podría sembrar las semillas relevantes a este respecto, ya que proporciona la base epistemológica para una moralidad intersubjetiva. Además, tiene mucho que ver con nuestra autocomprensión social y moral: "[La] forma en que tratamos a otras entidades, la forma en que las experimentamos, lo que decimos sobre ellas, la forma en que las tratamos, etc. También dice mucho de mí y dice mucho de nosotros". (Coeckelberg 2018: 150). Pero en lugar de una visión antropocéntrica, se trata más bien de una visión relacional que trata a las entidades no humanas como socios en la interacción.

4. Conclusión

La Inteligencia Artificial en general y los robots humanoides en particular cambiarán nuestras vidas y quizás también a nosotros mismos. Los filósofos tienen mucho que considerar en términos de los impactos epistémicos, éticos, estéticos y políticos de estos nuevos desafíos. La empatía es sólo uno de los muchos temas que HRI cuestiona. Este trabajo ha contribuido a las investigaciones necesarias que ya están en marcha o las que están por venir. Discutí el enigma epistémico de si podemos empatizar con los robots, aplicando las explicaciones contemporáneas dominantes sobre la empatía a este dominio. Luego examiné la cuestión normativa de si deberíamos empatizar con los robots y por qué. El artículo proponía un punto de vista pragmático al demostrar que a) de hecho podemos empatizar con los robots humanoides, no sólo en un nivel básico, sino también, al menos hasta cierto punto, en un nivel de toma de perspectiva imaginativa; Además, se demostró que incluso desde un punto de vista fenomenológico e intersubjetivo, es posible hablar de empatizar con los robots que están incrustados en nuestro mundo, con los que interactuamos y compartimos una narrativa contextual. La atención se centra en la empatía como un proceso de interacción mutua más que como un resultado. Sin embargo, el artículo también argumenta que b) deberíamos empatizar con los robots humanoides porque al hacerlo podemos adquirir nuevos conocimientos de un ser en el mundo muy desconocido, ampliando así nuestros horizontes, capacitándonos para futuros desarrollos de IA y mejorando la HRI en un entorno social compartido. Se consideró que esto no sólo tenía un valor instrumental, sino también valioso para nuestra comprensión de nosotros mismos y de nuestra sociedad, en la que los robots y otras formas de IA pueden considerarse compañeros.

[24] Coeckelbergh propone un enfoque similar al mío, pero se inspira en los conceptos de Wittgenstein sobre una forma de vida y juegos de lenguaje. Sin embargo, su artículo carece de una definición clara de lo que él cree que implica la empatía (por ejemplo, si la empatía realmente implica preocuparse por el bienestar del otro, como parece sugerir su artículo).

Referencias

Baron-Cohen, S. 1995. Ceguera mental. Un ensayo sobre el autismo y la teoría de la mente. Cambridge, MA: MIT Press.

Batson, CD 2009. Estas cosas llamadas empatía: ocho fenómenos relacionados pero distintos. En La neurociencia social de la empatía, ed. J. Decety y W. Ickes, 3-15. Cambridge, MA: MIT Press.

Benford, G. y E. Malartre. 2007. Más allá de lo humano. Tom Doherty Associates: Viviendo con robots y cyborgs. Nueva York.

Boddington, P., P. Millican y M. Wooldridge. 2017. Número especial Mentes y máquinas: Ética e inteligencia artificial. Mentes y máquinas 27 (4): 569–574.

Boden, MA 2016. AI. Su naturaleza y futuro. Oxford: Prensa de la Universidad de Oxford.

Breazeal, CL 2002. Diseño de robots sociales. Cambridge, MA: MIT Press.

Breithaupt, F. 2019. Los lados oscuros de la empatía. Ítaca: Prensa de la Universidad de Cornell.

Bretan, M., G. Hoffman y G. Weinberg. 2015. Comportamientos físicos dinámicos emocionalmente expresivos en robots. Revista internacional de estudios humanos-computadores 78: 1–16.

Brinck, I. y C. Balkenius. 2018. Reconocimiento mutuo en la interacción humano-robot: una cuenta deflacionaria. Filosofía y tecnología: 1–18. https://doi.org/10.1007/s13347-018-0339-x.

Chalmers, DJ 1996. La mente consciente. Oxford: Prensa de la Universidad de Oxford.

Coeckelberg, M. 2018. ¿Por qué preocuparse por los robots? Empatía, posición moral y el lenguaje del sufrimiento. Kairós. Revista de Filosofía y Ciencia 20: 141–158.

Colombetti, G. 2013. El cuerpo sentimiento. La ciencia afectiva se encuentra con la mente activa. Cambridge, MA: MIT Press.

Coplan, A. 2011. Comprender la empatía, 3–18. Sus características y efectos. En Empatía. Filosófico y perspectivas psicológicas. Oxford: Prensa de la Universidad de Oxford.

Coplan, A. y P. Goldie. 2011. Empatía. Perspectivas filosóficas y psicológicas. Oxford: Oxford Prensa universitaria.

Cross, ES, Riddoch, KA, Pratts, J, Titone, S, Chaudhury, B y Hortensius, R. 2018. Una investigación neurocognitiva del impacto de socializar con un robot en la empatía por el dolor. Preimpresión. https://doi. org/10.1101/470534.

Darling, K. 2016. Ampliación de la protección legal a los robots sociales: los efectos del antropomorfismo, la empatía y el comportamiento violento hacia los objetos robóticos. En Ley de robots, ed. M. Froomkin, R. Calo e I. Kerr. Cheltenham: Edward Elgar.

Darwall, S. 1998. Empatía, simpatía, cuidado. Estudios filosóficos 89: 261–282.

De Sousa, R. 1987. La racionalidad de la emoción. Cambridge, MA: MIT Press.

De Vignemont, F. y P. Jacob. 2012. ¿Cómo es sentir el dolor ajeno? Filosofía de la Ciencia 79 (2): 295–316.

De Vignemont, F. y T. Singer. 2006. El cerebro empático: ¿cómo, cuándo y por qué? Tendencias en cognitivo ciencias 10 (10): 435–441.

Dennett, D. 1991. Explicación de la conciencia. Boston: Little, Brown y compañía.

Dullstein, M. 2012. La segunda persona en el debate sobre la teoría de la mente. Repaso de Filosofía y Psicología 3 (2): 231–248.

Dullstein, M. 2013. Percepción directa y simulación: la explicación de Stein sobre la empatía. Revista de Filosofía y Psicología 4: 333–350.

Dumouchel, P. y L. Damiano. 2017. Viviendo con robots. Cambridge, MA: Harvard University Press.

Engelen, EM 2018. ¿Podemos compartir un sentimiento de nosotros con una máquina digital? El intercambio emocional y la reconocimiento de uno como otro. Reseñas de ciencias interdisciplinarias 43 (2): 125–135.

Engelen, EM y B. Röttger-Rössler. 2012. Debates disciplinarios e interdisciplinarios actuales sobre la empatía. Revisión de emociones 4 (1): 3–8.

Fodor, J. 1987. Psicosemántica. El problema del significado en la filosofía de la mente. Cambridge, MA: MIT Prensa.

Gallagher, S. 2008. Percepción directa en el contexto interactivo. Conciencia y Cognición 17 (2): 535–543.

Gallagher, S. 2017. Empatía y teorías de la percepción directa. En El manual de filosofía de Routledge empatía, ed. H. Maibom, 158-168. Nueva York: Routledge.

Gallagher, S. y J. Gallagher. 2019. Actuar como otro: la empatía de un actor por su personaje. topoi (primero en línea), https://doi.org/https://doi.org/10.1007/s11245-018-96247.

Gallagher, S. y D. Hutto. 2008. Comprender a los demás a través de la interacción primaria y la práctica narrativa. En La mente compartida: perspectivas sobre la intersubjetividad, ed. J. Zlatev, T. Racine, C. Sinha y E. Itkonen, 17–38. Ámsterdam/Filadelfia: John Benjamins Publishing Company.

Gallese, V. 2001. La hipótesis de la 'variedad compartida': de las neuronas espejo a la empatía. Diario de Estudios de conciencia 8: 33–50.

Goldie, P. 2000. Las emociones. Oxford: Prensa de la Universidad de Oxford.

Goldie, P. 2012. El desorden interior. Narrativa, emoción y mente. Oxford: Prensa de la Universidad de Oxford.

Goldman, A. 2006. Simulación de mentes: la filosofía, la psicología y la neurociencia de la lectura de mentes. Oxford: Prensa de la Universidad de Oxford.

Goldman, A. 2011. Dos rutas hacia la empatía: conocimientos de la neurociencia cognitiva. En Empatía: perspectivas filosóficas y psicológicas, ed. A. Coplan y P. Goldie, 31–44. Oxford: Prensa de la Universidad de Oxford.

Gopnik, A. y HM Wellman. 1994. La teoría teórica. En Mapeo de la mente: especificidad de dominio en cognición y cultura, ed. LA Hirschfeld y SA Gelman, 257–293. Cambridge: Prensa de la Universidad de Cambridge.

Gruen, L. 2009. Atendiendo a la naturaleza: compromiso empático con el mundo más que humano. La ética y la Medio ambiente 14 (2): 23–38.

Gruen, L. 2017. El estatus moral de los animales. En La enciclopedia de filosofía de Stanford (edición de otoño de 2017), ed. ES Zalta, https://plato.stanford.edu/archives/fall2017/entries/moral-animal/.

Hickok, G. 2014. El mito de las neuronas espejo: la verdadera neurociencia de la comunicación y la cognición. Nuevo York: WW Norton & Company.

Hoffmann, M. y R. Pfeifer. 2018. Los robots como poderosos aliados para el estudio de la cognición corporal desde abajo hacia arriba. En El manual de Oxford de cognición 4E, ed. A. Newen, L. de Bruin y S. Gallagher. Oxford: Prensa de la Universidad de Oxford.

Hutto, DD 2008. La hipótesis de la práctica narrativa: aclaraciones e implicaciones. Exploraciones filosóficas 11 (3): 175–192.

Iacoboni, M. 2011. Uno dentro del otro: mecanismos neuronales de la empatía en el cerebro de los primates. En Empatía: perspectivas filosóficas y psicológicas, ed. A. Coplan y P. Goldie, 45–57. Oxford: Prensa de la Universidad de Oxford.

Iacoboni, M., R. P. Woods y col. 1999. Mecanismos corticales de imitación humana. Ciencia 286: 2526–2528.

Kanske, P. 2018. La mente social: desenredar los métodos afectivos y cognitivos para comprender a los demás. Reseñas de ciencias interdisciplinarias 43 (2): 115–124.

Kant, I. 1997. Conferencias sobre ética, ed. andtrans. P. Heath y JB Schneewind. Cambridge: Cambridge Prensa universitaria.

Kasparov, G. 2017. Pensamiento profundo: donde termina la inteligencia de las máquinas y comienza la creatividad humana. Nueva York: Asuntos publicos.

Leite, A., A. Pereira, S. Mascarenhas, C. Martinho, R. Prada y A. Paiva. 2013. La influencia de la empatía en las relaciones entre humanos y robots. Revista internacional de estudios humanos-computadores 71(3): 250–260.

Lin, P., R. Jenkins y K. Abney. 2017. Ética de los robots 2.0: de los coches autónomos a la inteligencia artificial. Oxford: Prensa de la Universidad de Oxford.

Loh, J. 2019. Roboterethik. Eine Einführung. Berlín: Suhrkamp.

MacLennan, BJ 2014. El tratamiento ético de los robots y el difícil problema de las emociones de los robots. Revista Internacional de Emociones Sintéticas 5 (1): 9–16.

Maibom, H. 2017. El manual de filosofía de la empatía de Routledge. Londres: Routledge.

Misselhorn, C. 2009. Empatía con los objetos inanimados y el valle inquietante. Mentes y Máquinas 19 (3): 345–359.

Misselhorn, C. En prensa. ¿Es la empatía con los robots moralmente relevante? En Máquinas emocionales: perspectivas desde la informática afectiva y la interacción emocional hombre-máquina, ed. C. Misselhorn y M. Klein. Wiesbaden.

Mori, M. 2005. En el valle inquietante. En Actas del taller Humanoides-2005: Vistas del valle inquietante. Tsukuba: Japón.

Nagel, T. 1974. ¿Cómo es ser un murciélago? La revisión filosófica 83 (4): 435–450.

Newen, A. 2015. Comprender a los demás: la teoría del modelo de persona. En En Open MIND: 26(T), ed. T. Metzinger y JM Windt. Fráncfort del Meno: Grupo MIND.

Newen, A., L. De Bruin y S. Gallagher. 2018. El manual de Oxford de cognición 4E. Oxford: Oxford Prensa universitaria.

Nussbaum, M. 2011. Trastornos del pensamiento: La inteligencia de las emociones. Cambridge: Universidad de Cambridge Prensa.

Plantinga, C. 2009. Espectadores en movimiento: el cine americano y la experiencia del espectador. Berkeley: Universidad de Prensa de California.

Rorty, R. 2001. Redención del egoísmo: James y Proust como ejercicios espirituales. Telos 3 (3): 243–263.

Scheutz, M. 2011. Roles arquitectónicos del afecto y cómo evaluarlos en agentes artificiales. Revista internacional de emociones sintéticas 2 (2): 48–65.

Schmetkamp, S. 2017. Adquirir perspectivas sobre nuestras vidas: estados de ánimo y experiencia estética. Filosofía 45(4): 1681–1695.

Schmetkamp, S. 2019. Theorien der Empathie - Ein Einführung. Hamburgo: Junius Publisher.

Schneider, S. En prensa. Mentes futuras: mejorar y trascender el cerebro.

Slote, M. 2017. Las muchas caras de la empatía. Filosofía 45(3): 843–855.

Smith, M. 1995. Personajes atractivos: ficción, emoción y cine. Oxford: Prensa de Clarendon.

Sobchack, V. 2004. Pensamientos carnales: encarnación y cultura de la imagen en movimiento. Berkeley: Universidad de Prensa de California.

Stein, E. 1989. Sobre el problema de la empatía: las obras completas de Edith Stein. vol. 3 (tercera edición revisada), trans. W. Stein. Washington, DC: Publicaciones ICS.

Stueber, K. 2006. Redescubriendo la empatía: agencia, psicología popular y ciencias humanas. Cambridge, MA: Prensa del MIT.

Stueber, K. 2018. Empatía. En La enciclopedia de filosofía de Stanford (edición de primavera de 2018), ed. ES Zalta, https://plato.stanford.edu/archives/spr2018/entries/empathy/.

Vaage, MB 2010. El cine de ficción y las variedades de compromiso empático. Estudios del Medio Oeste en Filosofía 34: 158–179.

Vallor, S. 2011. Carebots y cuidadores: sostener el ideal ético del cuidado en el siglo XXI. Filosofía y Tecnología 24 (3): 251–268.

Weber, K. 2013. ¿Cómo es encontrarse con un agente artificial autónomo? IA Y SOCIEDAD 28: 483–489.

Yanal, R. J. 1999. Paradojas de la emoción y la ficción. Pensilvania: Penn State University Press.

Zahavi, D. 2011. Empatía y percepción social directa: Una propuesta fenomenológica. Revista de Filosofía y Psicología 2(3): 541–558.

Zahavi, D. 2014. Uno mismo y el otro: explorando la subjetividad, la empatía y la vergüenza. Oxford: Universidad de Oxford Prensa.

Zahavi, D. y J. Michael. 2018. Más allá del reflejo: perspectivas 4E sobre la empatía. En El manual de Oxford de cognición 4E, ed. A. Newen, L. de Bruin y S. Gallagher, 589–606. Oxford: Prensa de la Universidad de Oxford.