ISSN: 2444-250X | ISSN-e: 2444-2887

¿Tramposo e injusto? Entonces, es humano. Robots sociales educativos y ética sintética

María Isabel Gómez-León
Profesora de la Universidad Internacional de La Rioja (Logroño, España)
isabel.gomez@unir.net | https://orcid.org/0000-0001-7466-5441

Extracto

La educación comienza a hacer uso de la inteligencia artificial emocional a través de robots educativos antropomorfizados. La evidencia respalda que los estudiantes (hombres y mujeres) son capaces de crear vínculos emocionales con estos agentes. Sin embargo, cada vez se están encontrando más casos de desinhibición abusiva en este tipo de interacciones, como degradaciones racistas o sexistas, abuso de poder y violencia. Algunos investigadores alertan sobre las consecuencias negativas que este tipo de conductas pueden tener a largo plazo, tanto para la educación ética de los estudiantes como para los robots que aprenden de estas conductas. A pesar de su relevancia desde una perspectiva social y educativa, existen pocos estudios que intenten comprender los mecanismos que subyacen a estas prácticas inmorales o colectivamente dañinas. El objetivo de este artículo es revisar y analizar las investigaciones que han tratado de estudiar el comportamiento antiético del ser humano a través de su interacción con los robots sociales antropomórficos. Se realizó un estudio bibliométrico descriptivo siguiendo los criterios de la declaración PRISMA. Los resultados muestran que, bajo ciertas circunstancias, la antropomorfización y la atribución de intencionalidad a los agentes robóticos podría ser desventajosa, provocando actitudes de rechazo, deshumanización e incluso violencia. Sin embargo, una visión más realista tanto de las capacidades y limitaciones de estos agentes como de los mecanismos que guían la conducta humana podría ayudar a aprovechar el gran potencial de esta tecnología para promover el desarrollo moral y la conciencia ética de los estudiantes.

Palabras clave: tecnología educativa; robots sociales; inteligencia artificial; engaño; abuso; deshonestidad; integridad académica; ética sintética.

Recibido: 30-04-2023 | Aceptado: 28-09-2023 | Publicado: 07-01-2024

Cómo citar: Gómez-León, M.ª I. (2024). ¿Tramposo e injusto? Entonces, es humano. Robots sociales educativos y ética sintética. *Tecnología, Ciencia y Educación*, *27*, 167-186. https://doi.org/10.51302/tce.2024.18841

ISSN: 2444-250X | ISSN-e: 2444-2887

Cheating and unfair? So, its human. Educational social robots and synthetic ethics

María Isabel Gómez-León
Profesora de la Universidad Internacional de La Rioja (Logroño, España)
isabel.gomez@unir.net | https://orcid.org/0000-0001-7466-5441

Abstract

Education begins to make use of emotional artificial intelligence through anthropomorphized educational robots. Evidence supports that students (men and women) are able to create emotional bonds with these agents. However, more and more cases of abusive disinhibition are being found in such interactions, such as racist or sexist degradation, abuse of power and violence. Some researchers warn about the negative consequences that this type of behavior can have in the long term, both for the ethical education of students and for robots that learn from these behaviors. Despite their relevance from a social and educational perspective, there are few studies that attempt to understand the mechanisms underlying these immoral or collectively harmful practices. The aim of this article is to review and analyze the research that has tried to study the unethical behavior of the human being through its interaction with anthropomorphic social robots. A descriptive bibliometric study was carried out following the criteria of the PRISMA declaration. The results show that, under certain circumstances, anthropomorphization and attribution of intentionality to robotic agents could be disadvantageous causing attitudes of rejection, dehumanization and even violence. However, a more realistic view of both the capabilities and limitations of these agents and the mechanisms that guide human behavior could help harness the great potential of this technology to promote students' moral development and ethical awareness.

Keywords: educational technology; social robots; artificial intelligence; cheating behavior; robot abuse; dishonesty; academic integrity; synthetic ethics.

Received: 30-04-2023 | Accepted: 28-09-2023 | Published: 07-01-2024

Citation: Gómez-León, M.ª I. (2024). Cheating and unfair? So, its human. Educational social robots and synthetic ethics. *Tecnología, Ciencia y Educación*, *27*, 167-186. https://doi.org/10.51302/tce.2024.18841

Sumario

1. Introducción
2. Objetivos
3. Método
4. Resultados
    4.1. Antropomorfismo y categorías sociales humanas
    4.2. Antropomorfismo y conductas relacionadas con la integridad académica
    4.3. Antropomorfismo, vinculación emocional y violencia
5. Discusión
    5.1. El mayor riesgo de la humanización es la deshumanización
    5.2. Riesgos educativos de la deshumanización
    5.3. ¿Qué y cuándo humanizar?
    5.4. Aprovechar el potencial de los robots antropomórficos: claridad y transparencia
6. Conclusiones

Referencias bibliográficas

1. Introducción

El antropomorfismo es la tendencia humana a atribuir rasgos humanos a entidades no humanas, o a tratar el comportamiento no humano como motivado por sentimientos y estados mentales humanos (Rajaonah y Zio, 2022). La gente ha tendido a antropomorfizar los objetos inanimados durante siglos, viendo caras en las nubes, en las copas de los árboles o en las piedras. Hoy, el antropomorfismo es utilizado por la robótica social para dotar a los robots de presencia y comportamientos que sean lo suficientemente creíbles como para fomentar la conexión, el interés y el compromiso socioemocional con el usuario. Un robot social es un agente autónomo físicamente encarnado que se comunica con los humanos a través de señales socioemocionales y que aprende de forma adaptativa mientras se va individualizando de los demás una vez fabricado y programado. Durante los últimos veinte años se ha incrementado el desarrollo de robots sociales que están hechos para interactuar como cuidadores, terapeutas, compañeros o tutores en las escuelas. Los esfuerzos se han centrado en crear robots que se asemejen a los humanos en apariencia, mente, expresión emocional y comportamiento.

La investigación ha mostrado que la antropomorfización en la interacción humano-robot puede desencadenar comportamientos similares a los observados en las interacciones humanas (Spatola *et al.*, 2018, 2019). Por lo tanto, a medida que ha avanzado la comprensión sobre los mecanismos cognitivos y emocionales de la conducta humana, también ha avanzado el desarrollo de una tecnología capaz de influir significativamente en el comportamiento, en los vínculos sociales, en la convivencia y en las conductas colaborativas. Una parte importante de la literatura ha destacado el potencial de los robots sociales para crear vínculos emocionales significativos y aumentar los resultados cognitivos y afectivos del estudiante (Spatola *et al.*, 2018). Sin embargo, las investigaciones están mostrando cada vez más casos de desinhibición abusiva hacia estos agentes. Situaciones donde el agente compañero pasa a convertirse en un objeto subordinado y maltratado. Los investigadores advierten sobre las implicaciones que este tipo de conductas pueden tener a largo plazo, como amplificación de estereotipos nocivos, daño emocional a grupos socioemocionalmente vulnerables, aceptabilidad de la agresión relacional y/o desensibilización y generalización de conductas antisociales a otros seres vivos (Darling, 2021).

Hay avances en la neurociencia cognitiva donde diversas disciplinas se combinan para analizar y comprender los mecanismos que subyacen al pensamiento y comportamiento éticos (Bartneck *et al.*, 2020; Rajaonah y Zio, 2022). Sin embargo, existen menos estudios que intenten comprender los mecanismos detrás del cumplimiento excesivo o la aceptación prolongada de prácticas inmorales o colectivamente dañinas, a pesar de ser igualmente relevantes desde una perspectiva social y educativa.

Si lo que se pretende es que la inteligencia artificial facilite, aliente y fortalezca los intercambios de los estudiantes en la dirección ética, es necesario que los diseñadores tengan

una comprensión profunda del comportamiento humano en interacción con los robots sociales, incluidas las tendencias a los comportamientos sociales sin sentido, la discriminación, la deshonestidad, el abuso y la violencia.

La ética sintética posibilita usar los agentes artificiales en esta dirección. Puede establecerse como una nueva ciencia de los seres humanos en la que los robots antropomórficos funcionan como «objetos», pero también como «instrumentos» de una indagación sobre lo que constituye la identidad o la particularidad humana (Rajaonah y Zio, 2022). El uso de agentes artificiales permite manipular y controlar varios parámetros de comportamiento, apariencia y expresividad en uno de los compañeros de interacción (el agente artificial) y examinar el efecto de estos parámetros en el otro compañero de interacción (el ser humano). Al mismo tiempo, usar agentes artificiales significa introducir la presencia de sistemas artificiales, pero parecidos a los humanos, en la esfera social humana. Esto permite probar de manera controlada, pero ecológicamente válida, los mecanismos humanos fundamentales de la cognición social tanto a nivel conductual como neuronal.

En este sentido, la ética sintética podría permitir estudiar las fortalezas y las debilidades de las relaciones humanas y mejorar su calidad construyendo tecnología en consecuencia. La cuestión es:

> ¿Qué características del ser humano no sería beneficioso reproducir en el diseño de robots sociales destinados al desarrollo de las interacciones éticas de los estudiantes?

2. Objetivos

El objetivo de este artículo es revisar y analizar las investigaciones que han estudiado el comportamiento antiético del ser humano a través de su interacción con los robots sociales antropomórficos.

3. Método

La búsqueda bibliográfica se llevó a cabo a través de las bases de datos Scopus, Web of Science, PubMed, IEEE Xplore y ACM Digital Library, siguiendo las recomendaciones PRISMA. Se utilizaron los términos (*dishonesty* o *cheating behavior* o *abuse* o *violence* o *dehumanization* o *stereotypes* o *morality* o *ethics*) y (*robots social* o *autonomous robots* o *artificial intelligence*).

En dicha búsqueda se identificaron un total de 374 artículos. De estos, 59 artículos fueron descartados por estar duplicados. Se incluyeron estudios empíricos escritos en cualquier lengua que estudiaran la interacción humano-robot social y que presentaran resultados relacionados con conductas antiéticas. Se excluyeron los estudios con poblaciones

clínicas, aquellos que no incluyeron robots antropomórficos y aquellos otros que utilizaron exclusivamente agentes virtuales para estudiar la relación humano-agente artificial.

Para aplicar los criterios de inclusión y exclusión se importaron registros de las bases de datos a la herramienta Rayyan QCRI. Tras la aplicación de los criterios de inclusión/exclusión, se mantuvieron 51 artículos. Se analizó la calidad metodológica de los estudios seleccionados; en concreto, se tuvo en cuenta el tamaño de la muestra, el tipo de diseño, la definición de variables, la validez y fiabilidad de los instrumentos de medida, el procedimiento de recogida de datos y la descripción y especificación de los métodos estadísticos. Tras este análisis, nueve artículos fueron valorados con una calidad baja, por lo que fueron eliminados. Tras una revisión exhaustiva de los artículos se realizaron búsquedas indirectas a través de los autores más citados o de aquellos que presentasen datos relevantes o novedosos para la elaboración del trabajo. La selección final estuvo formada por 42 artículos.

4. Resultados

4.1. Antropomorfismo y categorías sociales humanas

Un factor que afecta significativamente el proceso de antropomorfizar objetos, y luego empatizar con ellos, es su apariencia; por ejemplo, tener una estructura que se asemeje a un humano, pero solo hasta cierto punto. Según la teoría del Valle Inquietante, la respuesta puede pasar repentinamente a la inquietud y al rechazo ante representaciones humanas demasiado realistas, pero siempre imperfectas (Feng *et al*., 2018). Sin embargo, el atractivo físico de un robot puede superar la inquietud que provoca el efecto del Valle Inquietante. De hecho, es el factor que mejor predice su aceptación como compañero de trabajo. En este caso, al parecer se produce un sesgo en el que la belleza se relaciona con la bondad, y este estereotipo es especialmente influyente cuando el humano y el robot son del sexo opuesto (Eyssel y Hegel, 2012).

Los estereotipos de género entre humanos también influyen en el curso de las interacciones entre humanos y robots. Desde el punto de vista del desarrollo, los niños comienzan a atribuir un sesgo de género a los robots a la edad de cinco años, clasificándolos como masculinos (aunque en la lengua nativa el sustantivo no sea masculino) (Okanda y Taniguchi, 2021). Un robot con cabello corto y labios planos es calificado como menos femenino y más adecuado para tareas típicamente masculinas (por ejemplo, trabajo técnico) que un robot idéntico con cabello largo y labios bien formados, que es calificado como más femenino y más adecuado para tareas típicamente femeninas (por ejemplo, crianza de niños). En adultos, la relación entre cintura-cadera y el ancho de los hombros es suficiente para provocar el mismo efecto (Bernotat *et al.*, 2017).

Pero no solo la apariencia física o el nombre del robot (masculino o femenino) parecen ser suficientes para asociar ciertos trabajos con un género en particular (Jackson *et al.*, 2020).

Asimismo, un robot con voz masculina es clasificado como más creíble y competente que un robot con voz femenina, a pesar de que el contenido del discurso sea el mismo. Como consecuencia, hay más probabilidades de que una persona (independientemente del sexo) siga las instrucciones de una máquina si su voz es masculina. Del mismo modo, es percibido más favorablemente que un robot masculino rechace o cuestione órdenes humanas potencialmente inmorales a que lo haga un robot femenino, particularmente para los participantes de género masculino (Jackson *et al.*, 2020). Se ha encontrado, además, que los hombres son más propensos a donar dinero a un robot del sexo opuesto, a quien, por otra parte, consideran más atractivo, mientras que esta preferencia no se observa en las mujeres (Nass y Moon, 2000). En estudiantes adolescentes (14 a 16 años), las alumnas tienen mayores ganancias de aprendizaje cuando interactúan con un agente pedagógico femenino, no siendo el caso de los estudiantes masculinos, quienes obtienen sus peores resultados afectivos y cognitivos cuando el agente es femenino (Arroyo *et al.*, 2018).

En cuanto a los rasgos físicos étnicos, se ha detectado que las personas califican a los agentes informáticos que tienen una cara de la misma etnia como más atractivos, confiables, persuasivos e inteligentes que a los agentes que poseen rasgos de otra etnia (Nass y Moon, 2000). Estos últimos son considerados como pertenecientes a un grupo extraño, lo que disminuye la empatía y aumenta la deshumanización hacia los integrantes del grupo (Spatola *et al.*, 2019; Strait *et al.*, 2018). El simple hecho de señalar el país de origen sesga el grado de calidez y competencia con la que se percibe al robot. Por ejemplo, los robots de países percibidos como de baja calidez (como Rusia) podrían considerarse objetivos aceptables para comportamientos socialmente inaceptables porque son percibidos como más distantes de los humanos que los de países percibidos como de alta calidez (como España) (Spatola *et al.*, 2019).

A veces, la deshumanización puede minimizarse o eliminarse empleando tácticas de manipulación de la relación intergrupal; por ejemplo, si los robots son presentados a personas alemanas con nombres que son populares en Alemania, son tratados más positivamente y reconocidos como más humanos que los robots con nombres que suenan extranjeros (Eyssel y Kuchenbrandt, 2012). Otra táctica que parece reducir la deshumanización del «otro» es introducir una narrativa que conduzca a la individualización; por ejemplo, cuando un robot con aspecto de insecto es presentado de la siguiente manera: «Este es Francisco, un ser entrañable. Su color favorito es el rojo. Ha vivido en el laboratorio durante varios meses. Recientemente ha tenido la oportunidad de jugar con otros robots y ha estado emocionado desde entonces» (Darling *et al.*, 2015, p. 772), los participantes dudan durante mucho más tiempo en destruirlo con un martillo que cuando no es presentado.

Además, las personas consideran más humanos a los robots que comparten características culturales que van más allá de la apariencia y el lenguaje, como coloquialismos locales o historias personales que identifiquen elementos culturales comunes (el tráfico de la ciudad, la bandera de un equipo regional, etc.). De acuerdo con este principio de «similitud-atracción», cuando los participantes trabajan con un robot que muestra un estilo de personalidad que coincide con el suyo, «dominante» o «sumiso», lo consideran más inteligente y

convincente, a pesar de que el contenido sea idéntico. Es más, cuando la «personalidad» del robot coincide con la del usuario, es más probable que las personas le den crédito al agente por su éxito y menos probable que le culpen por el fracaso, en comparación a cuando hay un desajuste de personalidad (Esteban *et al.*, 2022).

Pero no solo las características físicas y de personalidad, sino también la posición laboral del agente social parece influir en la conducta del usuario. Las personas muestran más responsabilidad por la finalización exitosa de una tarea cuando trabajan con un robot que ocupa un puesto subordinado. Sin embargo, cuando el robot tiene semejanza humana y/o está en una posición de autoridad, las personas tienden a implicarse menos en la tarea y a delegar la responsabilidad en el robot. Por ejemplo, la responsabilidad asumida por un médico ante una decisión clínica errónea es menor si esta ha sido tomada con la ayuda de un robot al que se le ha atribuido un estatus superior, aunque el rendimiento sea el mismo (Bleher y Braun, 2022). El efecto de la categorización es tan fuerte que la mera rotulación de una máquina como «especialista» hace que los segmentos de noticias que reproduce sean considerados significativamente más informativos, interesantes, serios y de alta calidad que aquellos otros que reproduce una máquina con el rótulo «generalista», a pesar de que los segmentos de noticias sean idénticos (Nass y Moon, 2000).

4.2. Antropomorfismo y conductas relacionadas con la integridad académica

Aunque los dos aspectos más importantes para que los agentes artificiales parezcan sociales son la apariencia y el comportamiento humano, el comportamiento es probablemente más crítico que la apariencia (Gómez-León, 2023). De hecho, cuando, por ejemplo, una aspiradora muestra un movimiento autónomo coordinado con el de un ser humano, la persona se inclina a interactuar socialmente con ella (Fink *et al.*, 2012).

La memoria perfecta, las fuertes habilidades de razonamiento y el desempeño impecable son rasgos cognitivos típicos asociados a los robots. En cambio, el olvido y el razonamiento erróneo son patrones cognitivos típicos asociados a los humanos, por lo que un robot que funciona de manera defectuosa resulta más agradable y humano que uno que no comete errores, aunque sea considerado como menos inteligente y conduzca a un peor desempeño en la tarea y a una mayor desconfianza (Mirnig *et al.*, 2017). Si el robot reconoce sus errores, pero no trata de corregirlos, se percibe como menos capaz y confiable, pero nos resulta más familiar. De la misma manera, cuando un robot comete errores de habla, se le considera menos sincero, pero ayuda a que la interacción se perciba como más humana (Cameron *et al.*, 2020).

Por otra parte, las personas consideran que los robots injustos/tramposos tienen un comportamiento más similar al humano, lo que potencia la interacción con él, el compromiso social y la atribución de estados mentales (Litoiu *et al.*, 2015). Es más, las personas no se molestan si un robot engaña para obtener algún beneficio a su favor, solo lo hacen cuando el engaño va en contra de los beneficios de la persona.

Conductas sociales como la lealtad, la confianza, la devolución de favores o ayudar a alguien parecen surgir cuando se espera obtener algún beneficio o, al menos, cuando no suponen una pérdida de beneficios, y esto se aplica a diversos dominios. Zonca *et al.* (2021) hallaron que las personas colaboraban en una tarea durante más tiempo y con mayor precisión si el agente les ayudaba previamente que si no lo hacía. Es más, las personas tendían a tener más en cuenta los consejos de quienes previamente habían tenido en cuenta su propia opinión, por lo que, incluso si habían perdido la confianza en el robot, fingían tener en cuenta sus opiniones para complacerle y poder ejercer mayor influencia sobre él. Kirby *et al.* (2010) encontraron que los participantes tendían a evitar los robots cuando estos manifestaban un estado de ánimo negativo, como la tristeza, sin embargo, cuando el robot proporcionaba un estado de ánimo positivo, como la alegría, las personas interactuaban con él durante más tiempo. En un juego de incentivos económicos/materiales, Hsieh *et al.* (2023) determinaron que, independientemente del estado emocional del robot, el comportamiento prosocial de las personas solo aparecía cuando habían logrado ganancias personales elevadas.

Al parecer, las personas toman decisiones en beneficio propio, pero, al mismo tiempo, tratan de aparentar que son personas morales (Hoffman *et al.*, 2015), por lo que solo la presencia de un robot es suficiente para disminuir las trampas (Ahmad y Refik, 2022; Hoffman *et al.*, 2015; Petisca *et al.*, 2022). Aunque las conductas deshonestas pueden aumentar si el robot no parece ser consciente de la trampa, las personas perciben sus limitaciones u observan que otros incumplen las normas sin obtener consecuencias negativas por ello (Forlizzi *et al.*, 2016). Por ejemplo, Petisca *et al.* (2022) encontraron que estar en presencia de un robot cuyo comportamiento refleja que es consciente de la situación (interviene cuando hace trampa) disminuía el comportamiento de engaño, frente a un robot no consciente de la situación. Sin embargo, interactuar con un robot que mostraba intervenciones verbales simples desinhibía el engaño en la tarea, ya que las personas tendían a aprovecharse cuando percibían sus limitaciones. Ayub *et al.* (2021) mostraron en un grupo de estudiantes universitarios que la exposición previa al engaño de los compañeros aumentaba significativamente la probabilidad de hacer trampa. Forlizzi *et al.* (2016) observaron que la vigilancia de un robot tampoco parecía tener efecto cuando alguien robaba aperitivos en un espacio público. En este caso, la conducta se propagaba rápidamente entre los espectadores. Es más, si la degradación ética ocurría lentamente, es decir, sin que se notara demasiado, se aceptaba mejor que cuando ocurría un cambio abrupto.

Por otra parte, Mubin *et al.* (2020) descubrieron que la apariencia amigable de un robot parecía no ser adecuada para mantener la disciplina en el aula, mientras que un robot percibido como más intimidante mejoraba significativamente la deshonestidad académica. Es más, Spatola *et al.* (2018) observaron que, cuando el robot era considerado como «malo» (respondía con desprecio, falta de empatía y evaluaba negativamente la inteligencia del estudiante), este tenía el mismo efecto que la presencia humana e incrementaba la atención en la tarea, frente a un robot amable y empático. Adicionalmente, Kennedy *et al.* (2015) detectaron que el 17 % de una muestra de niños de entre 7 y 8 años asociaba la figura del maestro a un robot con un comportamiento social, mientras que el 64 % la asociaba a un robot con un comportamiento asocial, lo cual mejoró significativamente el aprendizaje (Kennedy *et al.*, 2015).

En algunas situaciones el antropomorfismo puede ser desventajoso, especialmente cuando las respuestas empáticas del agente son incongruentes durante el diálogo o cuando sus respuestas no cumplen con las expectativas asociadas a su apariencia humana (Wiese *et al.*, 2017). Por ejemplo, si un robot no expresa emociones negativas durante una tarea competitiva se concibe como extraño, frío, estresante e irritante (Becker *et al.*, 2005). Las actitudes de las personas también pueden ser negativas cuando la autonomía del robot se percibe como amenazante. Maninger y Shank (2022) determinaron que las personas etiquetan a los robots con un alto grado de gerencia y autonomía como agentes intencionales con plenas responsabilidades morales, pero con una capacidad limitada para sentirse a sí mismos y a los demás, de manera que, cuando un robot comete un acto dañino, se considera tan inmoral como si lo comete un humano, pero recibe una mayor parte de la culpa. Como consecuencia, las personas atribuyen una mayor maldad a un agente de inteligencia artificial cuando se comete una violación moral que a la organización, al programador o a los usuarios (Shank y DeSanti, 2018). Adicionalmente, la incapacidad para categorizar a los robots en una entidad ontológica bien definida disminuye la empatía y aumenta el sesgo de la deshumanización (Shank y DeSanti, 2018), por lo que es más probable que se les nieguen derechos morales, que sean juzgados con más dureza y que otros los dañen (Wiese *et al.*, 2017).

4.3. Antropomorfismo, vinculación emocional y violencia

Diferentes estudios han mostrado que los robots sociales generan confianza, aumentan el bienestar socioemocional y favorecen vínculos afectivos que las personas describen como amor y cuidado recíprocos (Rajaonah y Zio, 2022). Pero ¿hasta qué punto el ser humano sería capaz de hacer daño a un robot con el que establece algún tipo de vínculo emocional?

Brščić *et al.* (2015) y Nomura *et al.* (2016) observaron que los niños de 5 a 9 años abusaban a menudo de robots situados en los centros comerciales. Los niños utilizaban insultos, obstruían repetidamente el camino del robot e incluso le pateaban y golpeaban, a pesar de que lo consideraran humano y de que aproximadamente la mitad de ellos creyera que el robot era capaz de experimentar dolor y estrés. La mayoría decía haber abusado porque sentían curiosidad, porque disfrutaban haciéndolo o, simplemente, porque otros también lo hacían.

Otros experimentos en contextos naturales han mostrado comportamientos abusivos similares. Estudiantes de una residencia universitaria dirigieron burlas, abuso de poder e insultos hacia un robot que intentaba iniciar interacciones positivas con ellos (Rehm y Krogsager, 2013). La implementación de un robot de servicio en un entorno público recibió conductas de intimidación que a menudo escalaron hacia la agresión y la violencia física (incluidas patadas, puñetazos y bofetadas) (Salvini, *et al.*, 2010). Durante una sesión de clase, más del 38 % de los estudiantes abusaron de una agente pedagógica conversacional empleando la hipersexualización y la cosificación y ubicándola en un rol inferior y subordinado (Veletsianos *et al.*, 2008). El análisis de las interacciones espontáneas con un robot conversacional diseñado exclusivamente para el entretenimiento mostró que el abuso verbal, la discrimi-

nación y la degradación eran generalizados: «Serás mi esclavo y si te digo que hagas algo, hazlo directamente. ¡Dime tu nombre ahora, esclavo!» (De Angeli y Brahnam, 2008, p. 309). Además, el acoso sexual, las ofensas y los insultos eran habituales. Strait *et al.* (2018) encontraron que la frecuencia y el grado en el que se manifiesta este tipo de agresiones en dominios públicos es significativamente mayor hacia los robots de género femenino, y la respuesta es aún más negativa y deshumanizante cuando la agente es clasificada como asiática o negra que cuando lo es como blanca.

Explicitar la capacidad del robot para pensar y sentir no afecta la voluntad de humillarle (Keijsers *et al.*, 2022). Es más, sorprendentemente, existe una relación positiva entre señalar el derecho que tiene el robot a la protección contra el abuso y la tendencia a menospreciarlo en público. En los estudios de Keijsers *et al.* (2022) y de Tan *et al.* (2018) ningún participante le dijo explícitamente al acosador que dejara de maltratar al robot o que su abuso era moralmente incorrecto. Además, los espectadores del abuso intervinieron menos cuando el robot emitió señales emocionales mientras era golpeado (pedir ayuda y acurrucarse) que cuando no respondió (Keijsers *et al.*, 2022; Tan *et al.*, 2018). Garcia-Goo *et al.* (2022) encontraron que los espectadores aceptaban más la agresión cuando se producía hacia un hombre que hacia una mujer, incluso si el daño causado eran idéntico. Es más, aquellos con una mayor orientación a la dominación social y al sexismo hostil mostraron una mayor insensibilidad ante la víctima (de cualquier sexo), a pesar de atribuirle una mayor capacidad para experimentar dolor. A ellos la victimización les resultó divertida, entretenida, graciosa e incluso hilarante, y la víctima, estúpida.

En laboratorio, reproduciendo el experimento de obediencia de Milgram, donde los participantes debían administrar descargas eléctricas a un estudiante cuando respondía incorrectamente, Bartneck *et al.* (2005) detectaron que todos los participantes administraron el voltaje más alto cuando se trataba de un robot, en comparación con solo el 40 % en el estudio original de Milgram. En un estudio posterior, Bartneck *et al.* (2007) hallaron que, tras interactuar con un robot, los participantes dudaban hasta tres veces menos en apagarlo y borrarle la memoria (lo que equivaldría a «matarle») si era percibido como menos inteligente. En la misma línea, Riddoch y Cross (2021) encontraron que incluso aquellos participantes que decían sentirse apegados hacia un robot y que lo percibían como una presencia social, con emociones y capacidad de sentir dolor, dudaban menos de 25 segundos en golpearlo con un mazo en la cabeza. Y, aunque la mayoría afirmaba sentirse culpable al tener que hacerlo: «Sientes que está vivo, por lo que no quieres tener ningún tipo de agresividad con él» (p. 9), «Es casi como golpear a alguien» (p. 7), etc., alguno dio muestras de entusiasmo: «¡Yo estaba emocionado! Lo esperaba con ansias. Quería ser el primero en destrozarlo completamente» (p. 10).

En esta línea, algunos autores (Luria *et al.*, 2019) han propuesto la creación de robots diseñados para ser apuñalados y golpeados con el fin de ayudar a las personas a sentirse mejor. Según los autores, estos robots evitarían atacar a otras personas cuando se está enojado. Lo que no explican es cómo poder controlar esos impulsos violentos a largo plazo.

5. Discusión

Los resultados se discuten sugiriendo algunos de los mecanismos que puedan ayudar a comprender tales conductas y ofreciendo algunas recomendaciones sobre el tipo de diseño que podría ayudar a regular y alentar el cumplimiento de las normas morales en los estudiantes.

5.1. El mayor riesgo de la humanización es la deshumanización

Los estudios muestran que las personas hacen trampa en presencia de un robot, especialmente si pueden determinar sus capacidades, e incluso cometen actos violentos hacia ellos, por lo que se esperaría que encontraran menos grave el acto de ser deshonesto o violento con un robot en comparación con un humano. Sin embargo, se ha detectado que la gente considera que maltratar a un robot es tan inmoral como abusar de un humano. A pesar de ello, el nivel de culpa cuando se comete un acto deshonesto hacia un robot es menor que cuando se comete hacia un animal o hacia un humano (Bartneck y Keijsers, 2020). La mayoría de las personas justifica este tipo de conductas por la falta de capacidades del agente para comprender lo que está haciendo y por la ausencia de sentimientos, es decir, la falta de «humanización» (Petisca *et al.*, 2022).

En el sentido estricto de la palabra, los agentes artificiales, al carecer de humanidad real no se pueden deshumanizar, pero la literatura muestra que las personas «humanizan» a los agentes automáticamente y sin intención. Precisamente, uno de los principales factores que contribuye a que los robots sean tratados como entidades sociales es su capacidad para ser percibidos como intencionales, es decir, seres con una mente. Al proyectar intenciones similares a las humanas, las personas pueden intentar anticipar el comportamiento del agente y, en consecuencia, tener cierta sensación de control sobre él. De hecho, el temor explícito de las personas por ser controlados por la tecnología correlaciona con su respuesta emocional hacia los robots (Strait *et al.*, 2018).

Sin embargo, bajo algunas circunstancias, la antropomorfización puede ser desfavorable, en particular, cuando el estado mental de un agente es ambiguo y provoca incertidumbre en cuanto a su categorización ontológica (humano versus robot) o cuando el comportamiento del agente se desvía lo suficiente como para que un modelo antropomórfico conduzca a predicciones incorrectas.

Este conflicto puede consumir recursos cognitivos y provocar reacciones emocionales negativas, incrementar la percepción de amenaza y provocar estrategias de control violentas y/u ofensivas, o despertar la curiosidad y provocar conductas desafiantes y provocativas que suelen escalar a conductas de agresión física y/o verbal (Keijsers *et al.*, 2022; Strait, *et al.*, 2018; Wiese *et al.*, 2017).

5.2. Riesgos educativos de la deshumanización

En base a estos resultados se plantean varias consideraciones importantes en lo que se refiere al diseño y al desarrollo futuro de robots antropomórficos en el ámbito educativo.

Primero, si un robot determinado está diseñado para aprender de sus interacciones con las personas, proyectar una identidad a través de características individuales y sociales antropocéntricas, puede reforzar prejuicios dañinos, tanto en los estudiantes como en los agentes con los que interactúa (Spatola *et al.*, 2019). De hecho, ya han comenzado a mostrarse las primeras evidencias empíricas sobre robots sexistas y racistas que prestan más atención al discurso de un hombre que al de una mujer o que asocian palabras como «criminal» con caras de personas negras y latinas (Darling, 2021).

Segundo, focalizar la atención sobre aspectos físicos o culturalmente conocidos para promover la empatía y facilitar la categorización social puede conducir a la deshumanización del «grupo externo», lo que reduce la empatía y aumenta la agresión contra los miembros de dicho grupo (Keijsers *et al.*, 2022; Petisca *et al.*, 2022). Algo realmente preocupante teniendo en cuenta que actualmente los niños que pertenecen a una minoría étnica tienen más probabilidades de ser acosados en la escuela (Strait *et al.*, 2018).

Tercero, en la medida en que la relación humano-robot tiende a imitar una relación maestro-esclavo, los robots pueden reforzar actitudes autoritarias e incluso crueles en los alumnos, despertando una falsa creencia de su derecho para dar órdenes o explotar a otros agentes (tanto humanos como artificiales). Los niños mimados por niñeras robot incapaces de decir «no» son un ejemplo frecuentemente citado. Estos niños tienden a encontrar atractiva la idea de intimidar a los robots; es más, las actitudes agresivas, arrogantes o las tendencias autoindulgentes podrían verse reforzadas por la incapacidad del robot para escapar o defenderse (Rajaonah y Zio, 2022).

Cuarto, los estudios demuestran que proyectar una identidad a través de características individuales y sociales antropocéntricas puede aumentar la credibilidad de los agentes robóticos. Sin embargo, reproducir ciertas características humanas, como la indisciplina, ser injusto, hacer trampas o evadir la responsabilidad moral podría generar o reforzar conductas que atentan contra la integridad académica de los estudiantes. Es importante señalar que a partir de los 5 años los niños comienzan a seleccionar las conductas que imitan en función de la identificación con el modelo y con el grupo social en general. Es más, la influencia de los factores sociales sobre el comportamiento de sobreimitación (imitación de acciones que se reconocen como irracionales) aumenta con la edad.

Quinto, el abuso hacia los robots podría causar daño emocional a otros humanos, especialmente a poblaciones socioemocionalmente vulnerables. La ciencia ha mostrado que ver cómo un robot está siendo torturado puede incrementar en el observador respuestas fisiológicas de alerta, estrés y otras emociones negativas que responden a los mismos patrones de activación neuronales que cuando se trata de un humano (Wiese *et al.*, 2017).

Y, sexto, permitir el abuso hacia estos agentes puede suponer que sea considerado como normativo y no como una forma de comportamiento antisocial, con el consecuente riesgo a la desensibilización y a la generalización de estas conductas hacia otros contextos o grupos. Se ha comprobado que la exposición a la agresión relacional cambia las creencias normativas de los estudiantes, es decir, hace que este tipo de conductas se vuelvan cada vez más aceptables, tanto individualmente como a nivel del aula. Es más, esta aceptación parece aumentar con la edad y es uno de los mayores predictores del acoso escolar (Behnk *et al.*, 2022).

5.3. ¿Qué y cuándo humanizar?

Las investigaciones muestran que centrar el diseño en las características antropomórficas externas de un robot no produce automáticamente el efecto deseado y que se necesita más que esta apariencia para garantizar interacciones sociales efectivas (Feng *et al.*, 2018; Okanda y Taniguchi, 2021; Spatola *et al.*, 2019). De la misma manera, la percepción de intencionalidad atribuida a un robot no siempre afecta positivamente ni al desempeño de la tarea (Kennedy *et al.*, 2015) ni al fomento de las interacciones sociales (Maninger y Shank, 2022; Shank y DeSanti, 2018; Wiese *et al.*, 2017), por lo que es necesario tener en cuenta muchas otras variables.

Los robots sociales están diseñados con un propósito particular en mente. Según sea la aplicabilidad en términos de dominios y tareas, se debe considerar la implementación de unas características u otras. Numerosos estudios han demostrado que el principal determinante de la confianza en los robots es su desempeño (Zonca *et al.*, 2021). Los humanos tienden a confiar en los robots siempre que no cometan errores, a pesar de que esto resulte menos humano. Es más, la presencia de fallas puede conducir al desuso del sistema robótico. En el contexto educativo, centrar el interés en las habilidades del agente que ayudan a alcanzar el objetivo de la tarea, y no tanto en las características físicas o socioculturales, puede mejorar significativamente la calidad de la interacción (Gómez-León, 2022). En esta línea, Maggi *et al.* (2021) hallaron que el estilo de interacción autoritario, frente al estilo de interacción amistoso, parecía ser el más apropiado cuando las tareas requerían altas demandas cognitivas, lo que aumentaba el cumplimiento de las tareas, así como la confianza y la aceptación de la tecnología. Además, Petisca *et al.* (2022) encontraron que una mayor expresión social no siempre equivalía a una reducción en las conductas de engaño, por lo que más importante que interactuar con un robot amigable mientras se realiza una tarea tentadora era interactuar con un robot que mostrara conciencia del comportamiento del participante.

Además, en determinados dominios, las personas no siempre prefieren lo que les recuerda al ser humano. Los robots, cuando se utilizan como guías en lugares públicos, son preferidos a los humanos porque, según los comentarios de los usuarios, el robot no juzga a las personas en función de su apariencia y trata a todos por igual. En el ámbito de la atención médica, las personas perciben que los robots tienen ventajas sobre los humanos, como una mayor perseverancia, compromiso y disponibilidad, así como una menor distracción hacia

los pacientes. De la misma manera, cuando se utilizan para fomentar el ejercicio o en programas de pérdida de peso, las personas consideran que los robots están libres de prejuicios, lo que les hace sentirse más cómodas y tener una mayor sensación de privacidad. Es más, en Japón, el 80 % de las personas mayores se sienten más cómodas compartiendo entornos de vida con robots cuidadores que con personas extranjeras (Darling, 2021; Rajaonah y Zio, 2022). Por lo tanto, no parece necesario que los agentes artificiales tengan que emular con precisión el comportamiento humano. Los datos sugieren que podría ser suficiente con que solo mostraran ciertos aspectos que estuvieran más fuertemente asociados con el propósito con el que fueron creados.

5.4. Aprovechar el potencial de los robots antropomórficos: claridad y transparencia

Sin embargo, tratar a los robots como agentes con mente también puede aumentar la percepción de conexión social y fomentar conductas prosociales, como la disminución de las trampas y el aumento de la generosidad. Además, se ha demostrado que los robots pueden tener un poder persuasivo significativo sobre los humanos en el cumplimiento de las normas sociales (Gómez-León, 2022), por lo que podría aprovecharse el potencial de esta tecnología para impulsar interacciones humanas beneficiosas. El antropomorfismo aplicado ofrece un mayor conocimiento sobre la conducta moral del ser humano, lo que puede contribuir a la creación de diseños centrados en aquellas características humanas que faciliten, alienten y fortalezcan las relaciones sociales y que, al mismo tiempo, supongan el mínimo riesgo para debilitarlas.

Por otra parte, se ha comprobado que el grado de antropomorfización puede cambiar después de haber interactuado con un robot humanoide o haber visto, al menos, diferentes ejemplos de tecnología robótica (Gómez-León, 2023). Este tipo de interacciones posibilita que tanto pequeños como adultos se familiaricen con el robot, entiendan mejor cómo funciona e interactúen de una manera más coherente y éticamente deseable. Por lo tanto, aumentar la distancia conceptual entre los humanos y los robots podrían ser uno de los principales determinantes para reducir las tendencias antisociales. Facilitar la categorización de los agentes podría aportar una visión más clara y realista de sus capacidades y limitaciones y, como consecuencia, favorecer una interacción más adecuada; por lo que la creación de una nueva categoría ontológica claramente definida, junto con características de diseño que señalen, explícita e implícitamente, el propósito y la capacidad funcional del robot, podrían evitar en gran medida falsas expectativas y ayudar a reducir la incertidumbre y las conductas poco éticas.

La investigación actual aboga por una mayor transparencia y explicación sobre los principios del diseño de la robótica, lo que ayudaría a los usuarios a construir modelos mentales más precisos (Stange *et al.*, 2022). Los robots sociales están diseñados para tener un comportamiento más o menos autónomo, sin embargo, la autonomía aumenta la incertidumbre y la imprevisibilidad, por lo que la transparencia y la explicación centrada en el usuario co-

bran aún mayor importancia: qué está haciendo el robot y por qué, qué hará a continuación, cuáles son sus capacidades, intenciones y limitaciones situacionales, cuándo y por qué el robot falla al realizar acciones específicas y cómo corregir errores son aspectos esenciales. Además de este tipo de transferencia de robot a humano, la transparencia de humano a robot está siendo de creciente interés. Puesto que tanto el humano como el robot deben compartir la intención y la conciencia de la situación, el robot también debería tener información sobre el desempeño humano.

La transparencia no solo implica «qué» comunicar, sino también «cómo» comunicar para mejorar la comprensión del usuario. En este sentido se pueden utilizar diferentes interfaces: visuales, a través de metáforas y analogías; auditivas, con explicaciones que utilicen un lenguaje natural similar al humano; basadas en interacción física y háptica; multimodales, e incluso a través de robots autónomos que expliquen sus propias necesidades, intenciones y comportamientos (Stange *et al.*, 2022).

Un aspecto importante es estudiar la posibilidad de agregar información detallada sobre los roles y las responsabilidades tanto del robot como del ser humano.

6. Conclusiones

Los estudios experimentales revisados demuestran que las personas reproducen reglas y expectativas sociales propias de las relaciones humanas en su interacción con los robots sociales. El primer conjunto de estudios ilustra cómo las personas abusan de las categorías sociales humanas, aplicando estereotipos de género o discriminando étnicamente a los agentes sociales con los que se identifican. El segundo conjunto demuestra que las personas se identifican con comportamientos sociales antiéticos, como la injusticia o el engaño, aprovechando cualquier oportunidad para practicarlo siempre que la transgresión no se note demasiado.

El análisis de las interacciones humano-robot fuera de laboratorio muestra que no es suficiente centrarse en modelar un comportamiento que sea similar a las interacciones humanas exitosas, sino que también es importante estudiar otras formas desviadas de interacción, como el abuso de poder, el insulto, la degradación y la violencia.

La ética sintética muestra que, independientemente de las intenciones, los mecanismos de la cognición social suelen ser sutiles e implícitos. Las decisiones morales parecen depender demasiado a menudo de los atajos psicológicos, de las percepciones erróneas y de las tentaciones. Sin embargo, se podría aprovechar el potencial de esta tecnología para dirigir la conducta de los estudiantes hacia la práctica de los principios éticos y morales humanos. Para ello sería necesario que los usuarios tuvieran una comprensión precisa de los límites y capacidades del agente y que los diseñadores tuvieran un conocimiento profundo de los mecanismos que guían la conducta humana.

Referencias bibliográficas

Ahmad, M. I. y Refik, R. (2022). «No chit chat!». A warning from a physical versus virtual robot invigilator: Which matters most? *Frontiers in Robotics and AI*, *9*, 1-11. https://doi.org/10.3389/frobt.2022.908013

Angeli, A. de y Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers*, *20*(3), 302-310. https://doi.org/10.1016/j.intcom.2008.02.004

Arroyo, A. M., Kyohei, T., Koyama, T., Takahashi, H., Rea, F., Sciutti, A., Yoshikawa, Y., Ishiguro, H. y Sandini, G. (2018). Will people morally crack under the authority of a famous wicked robot? *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 35-42). IEEE. https://doi.org/10.1109/ROMAN.2018.8525744

Ayub, A., Hu, H., Zhou, G., Fendley, C., Ramsay, C. M., Jackson, K. L. y Wagner, A. R. (2021). If you cheat, I cheat: cheating on a collaborative task with a social robot. *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (pp. 229-235). IEEE. https://doi.org/10.1109/RO-MAN50785.2021.9515321

Bartneck, C., Hoek, M. van der, Mubin, O. y Mahmud, A. A. (2007). «Daisy, Daisy, give me your answer do!»-Switching off a robot. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction* (pp. 217-222). ACM. https://doi.org/10.1145/1228716.1228746

Bartneck, C. y Keijsers, M. (2020). The morality of abusing a robot. Paladyn. *Journal of Behavioral Robotics*, *11*(1), 271-283. https://doi.org/10.1515/pjbr-2020-0017

Bartneck, C., Rosalia, C., Menges, R. y Deckers, I. (2005). Robot abuse: a limitation of the media equation. En A. De Angeli, S. Brahnam y P. Wallis (Eds.), *Abuse: the Darker Side of Human-Computer Interaction: An INTERACT 2005 Workshop* (pp. 54-57). http://www.agentabuse.org/Abuse_Workshop_WS5.pdf

Becker, C., Prendinger, H., Ishizuka, M. y Wachsmuth, I. (2005). Evaluating affective feedback of the 3D agent max in a competitive cards game. *Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1* (pp. 466-473). Springer Berlin Heidelberg. https://doi.org/10.1007/11573548_60

Behnk, S., Hao, L. y Reuben, E. (2022). Shifting normative beliefs: on why groups behave more antisocially than individuals. *European Economic Review*, *145*. https://doi.org/10.1016/j.euroecorev.2022.104116

Bernotat, J., Eyssel, F. y Sachse, J. (2017). Shape it-The influence of robot body shape on gender perception in robots. *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9* (pp. 75-84). Springer International Publishing. https://doi.org/10.1007/978-3-319-70022-9_8

Bleher, H. y Braun, M. (2022). Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, *2*(4), 747-761. https://doi.org/10.1007/s43681-022-00135-x

Brščić, D., Kidokoro, H., Suehiro, Y. y Kanda, T. (2015). Escaping from children's abuse of social robots. *10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 59-66). ACM. https://doi.org/10.1145/2696454.2696468

Cameron, D., Saille, S. de, Collins, E. C., Aitken, J. M., Cheung, H., Chua, A., Loh, E. J. y Law, J. (2020). The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computer in Human Behabior*, *114*, 1-41. https://doi.org/10.1016/j.chb.2020.106561

Darling, K. (2021). *The New Breed: How to Think About Robots*. Penguin UK.

Darling, K., Nandy, P. y Breazeal, C. (2015). Empathic concern and the effect of stories

in human-robot interaction. *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 770-775). IEEE. https://doi.org/10.1109/ROMAN.2015.7333675

Esteban, P. G., Bagheri, E., Elprama, S. A., Jewell, C. I. C., Cao, H.-L., Beir, A. de, Jacobs, A. y Vanderborght, B. (2022). Should I be introvert or extrovert? A pairwise robot comparison assessing the perception of personality-based social robot behaviors. *International Journal of Social Robotics*, *14*, 1-11. https://doi.org/10.1007/s12369-020-00715-z

Eyssel, F. A. y Hegel, F. (2012). (S)he's got the look: gender-stereotyping of social robots. *Journal of Applied Social Psychology*, *42*(9), 2.213-2.230. https://doi.org/10.1111/j.1559-1816.2012.00937.x

Eyssel, F. y Kuchenbrandt, D. (2012). Social categorization of social robots: anthropomorphism as a function of robot group membership. *The British Journal of Social Psychology*, *51*(4), 724-731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

Feng, S., Wang, X., Wang, Q., Fang, J., Wu, Y., Yi, L. y Wei, K. (2018). The uncanny valley effect in typically developing children and its absence in children with autism spectrum disorders. *PloS One*, *13*(11), 1-14. https://doi.org/10.1371/journal.pone.0206343

Fink, J., Mubin, O., Kaplan, F. y Dillenbourg, P. (Mayo 2012). Anthropomorphic language in online forums about Roomba, AIBO and the iPad. *IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO)* (pp. 54-59). IEEE. https://doi.org/10.1109/ARSO.2012.6213399

Forlizzi, J., Saensuksopa, T., Salaets, N., Shomin, M., Mericli, T. y Hoffman, G. (2016). Let's be honest: a controlled field study of ethical behavior in the presence of a robot. *Robot and Human Interactive Communication (ROMAN). 25th IEEE International Symposium on* (pp. 769-774). IEEE. https://doi.org/10.1109/ROMAN.2016.7745206

Garcia-Goo, H., Winkle, K., Williams, T. y Strait, M. K. (2022). Robots need the ability to navigate abusive interactions. *2022 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 1-9). IEEE. https://scholarworks.utrgv.edu/cs_fac/92/

Gómez-León, M.ª I. (2022). Desarrollo de la empatía a través de la inteligencia artificial socioemocional. *Papeles del Psicólogo*, *43*(3), 218-224. https://doi.org/10.23923/pap.psicol.2996

Gómez-León, M.ª I. (2023). Robots sociales y crecimiento ético en educación infantil. *Edutec. Revista Electrónica de Tecnología Educativa*, *83*, 41-54. https://doi.org/10.21556/edutec.2023.83.2697

Hoffman, G., Forlizzi, J., Ayal, S., Steinfeld, A., Antanitis, J., Hochman, G., Hochendoner, E. y Finkenaur, J. (2015). Robot presence and human honesty: experimental evidence. *10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 181-188). ACM. https://doi.org/10.1145/2696454.2696487

Hsieh, T.-Y., Chaudhury, B. y Cross, E. S. (2023). Human-robot cooperation in economic games: people show strong reciprocity but conditional prosociality toward robots. *International Journal of Social Robotics*, 1-15. https://doi.org/10.1007/s12369-023-00981-7

Hundt, A., Agnew, W., Zeng, V., Kacianka, S. y Gombolay, M. (2022). Robots enact malignant stereotypes. *ACM Conference on Fairness, Accountability, and Transparency* (pp. 743-756). ACM. https://doi.org/10.1145/3531146.3533138

Jackson, R. B., Williams, T. y Smith, N. (2020). Exploring the role of gender in perceptions of robotic noncompliance. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 559-567). ACM. https://doi.org/10.1145/3319502.3374831

Keijsers, M., Bartneck, C. y Eyssel, F. (2022). Pay them no mind: the influence of implicit and explicit robot mind perception on the right to be protected. *International Journal of Social Robotics*, *14*, 499-514. https://doi.org/10.1007/s12369-021-00799-1

Kennedy, J., Baxter, P. E. y Belpaeme, T. (2015). The robot who tried too hard: social behaviour of a robot tutor can negatively affect child learning. *10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 67-74). ACM. https://doi.org/10.1145/2696454.2696457

Kirby, R., Forlizzi, J. y Simmons, R. (2010). Affective social robots. *Robotics and Autonomous Systems*, *58*(3), 322-332. https://doi.org/10.1016/j.robot.2009.09.015

Litoiu, A., Ullman, D., Kim, J. y Scassellati, B. (2015). Evidence that robots trigger a cheating detector in humans. *10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 165-172). ACM. https://doi.org/10.1145/2696454.2696456

Luria, M., Zoran, A. y Forlizzi, J. (2019). Challenges of designing HCI for negative emotions. *arXiv:1908.07577*, 1-3. https://doi.org/10.48550/arXiv.1908.07577

Maggi, G., Dell'Aquila, E., Cucciniello, I. y Rossi, S. (2021). «Don't get distracted!»: the role of social robots' interaction style on users' cognitive performance, acceptance, and non-compliant behavior. *International Journal of Social Robotics*, *13*, 2.057-2.069. https://doi.org/10.1007/s12369-020-00702-4

Mamak, K. (2022). Should violence against robots be banned? *International Journal of Social Robotics*, *14*(4), 1.057-1.066. https://doi.org/10.1007/s12369-021-00852-z

Maninger, T. y Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports*, *5*. https://doi.org/10.1016/j.chbr.2021.100154

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M. y Tscheligi, M. (2017). To Err is robot: how humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, *21*(4), 1-15. https://doi.org/10.3389/frobt.2017.00021

Mubin, O., Cappuccio, M., Alnajjar, F., Ahmad, M. I. y Shahid, S. (Diciembre 2020). Can a robot invigilator prevent cheating? *AI & Society*, *35*(4), 981-989. Springer. https://doi.org/10.1007/s00146-020-00954-8

Nass, C. y Moon, Y. (2000). Machines and mindlessness: social responses to computers. *Journal of Social Issues*, *56*(1), 81-103. https://doi.org/10.1111/0022-4537.00153

Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y. y Yamada, S. (2016). Why do children abuse robots? *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, *17*(3), 348-370. https://doi.org/10.1075/is.17.3.02nom

Okanda, M. y Taniguchi, K. (2021). Is a robot a boy? Japanese children's and adults' gender-attribute bias toward robots and its implications for education on gender stereotypes. *Cognitive Development*, *58*, 101044. https://doi.org/10.1016/j.cogdev.2021.101044

Parreira, M. T., Gillet, S., Winkle, K. y Leite, I. (2023, March). How did we miss this? A case study on unintended biases in robot social behavior. *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 11-20). https://doi.org/10.1007/s12369-022-00864-3

Petisca, S., Leite, I., Paiva, A. y Esteves, F. (2022). Human dishonesty in the presence of a robot: the effects of situation awareness. *International Journal of Social Robotics*, *14*(5), 1.211-1.222. https://doi.org/10.1007/s12369-022-00864-3

Rajaonah, B. y Zio, E. (2022). Social Robotics and synthetic ethics: a methodological proposal for research. *International Journal of Social Robotics*, 1-11. https://doi.org/10.1007/s12369-022-00874-1

Rehm, M. y Krogsager, A. (2013). Negative affect in human robot interaction: impoliteness in unexpected encounters with robots. *Proceedings of the 22nd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN): Living Together, Enjoying Together, and Working Together with Robots!* (pp. 45-50). IEEE Computer Society Press. IEEE RO-MAN Proceedings. https://doi.org/10.1109/ROMAN.2013.6628529

Rhee, S., Lee, S.-Y. y Jung, S.-H. (2017). Ethnic differences in bullying victimization and psychological distress: a test of an ecological model. *Journal of Adolescence*, *60*, 155-160. https://doi.org/10.1016/j.adolescence.2017.07.013

Riddoch, K. A. y Cross, E. S. (2021). «Hit the robot on the head with this mallet»-Making a case for including more open questions in HRI research. *Frontiers in Robotics and AI*, *8*, 1-17. https://doi.org/10.3389/frobt.2021.603510

Salvini, P., Ciaravella, G., Yu, W., Ferri, G., Manzi, A., Mazzolai, B. y Dario, P. (2010). How safe are service robots in urban environments? Bullying a robot. *19th International Symposium in Robot and Human Interactive Communication* (pp. 1-7). https://doi.org/10.1109/ROMAN.2010.5654677

Shank, D. B. y DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401-411. https://doi.org/10.1016/j.chb.2018.05.014

Spatola, N., Anier, N., Redersdorff, S., Ferrand, L., Belletier, C., Normand, A. y Huguet, P. (2019). National stereotypes and robots' perception: the «made in» effect. *Frontiers in Robotics and AI*, *6*, 1-12. https://doi.org/10.3389/frobt.2019.00021

Spatola, N., Belletier, C., Normand, A., Chausse, P., Monceau, S., Augustinova, M., Barra, V., Huguet, P. y Ferrand, L. (2018). Not as bad as it seems: when the presence of a threatening humanoid robot improves human performance. *Science Robotics*, *3*(21). https://doi.org/10.1126/scirobotics.aat5843

Stange, S., Hassan, T., Schröder, F., Konkol, J. y Kopp, S. (2022). Self-explaining social robots: an explainable behavior generation architecture for human-robot interaction. *Frontiers in Artificial Intelligence*, *5*, 1-19. https://doi.org/10.3389/frai.2022.866920

Strait, M., Ramos, A. S., Contreras, V. y Garcia, N. (2018). Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white. *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 452-457). IEEE. https://doi.org/10.1109/ROMAN.2018.8525610

Tan, X. Z., Vázquez, M., Carter, E. J., Morales, C. G. y Steinfeld, A. (2018). Inducing bystander interventions during robot abuse with social mechanisms. *13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 169-177). ACM. https://doi.org/10.1145/3171221.3171247

Veletsianos, G., Scharber, C. y Doering, A. (2008). When sex, drugs, and violence enter the classroom: conversations between adolescents and a female pedagogical agent. *Interacting with Computers*, *20*(3), 292-301. https://doi.org/10.1016/j.intcom.2008.02.007

Wiese, E., Metta, G. y Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, *8*, 1-19. https://doi.org/10.3389/fpsyg.2017.01663

Zonca, J., Folsø, A. y Sciutti, A. (2021). The role of reciprocity in human-robot social influence. *Iscience*, *24*(12). https://doi.org/10.1016/j.isci.2021.103424

María Isabel Gómez-León. Doctora en Neurociencia, con sobresaliente *cum laude* por la Universidad Complutense de Madrid (España). Ha participado en proyectos de investigación con la Universidad Complutense de Madrid y con la Universidad Politécnica de Madrid (España). Actualmente, es profesora de grado y posgrado en la Universidad Internacional de La Rioja (España), en la Universidad Nebrija (España) y en la Universidad Camilo José Cela (España); directora y profesora de posgrado en el Máster de Atención Temprana en la Universidad Francisco de Vitoria (España); y gerente de un centro de neuropsicología infantil, especializado en atención temprana.