AISB2016

Simposio sobre los Principios de la Robótica

4 de abril de 2016, Sheffield, Reino Unido.

Editado por
TonyJPrescott

Comité Organizador
TonyJPrescott
alan winfield
MadeleinedePollaBuning
joannajbryson
NoelSharkey

Parte de la Convención de 2016 de la Sociedad para el Estudio de la Inteligencia Artificial y la Simulación de Comportamiento (AISB)

Acerca del Simposio

Han pasado cinco años desde la publicación de “Principios de la robótica”1 , desarrollado por un grupo de distinguidos expertos británicos en robótica e IA, un retiro financiado por el EPSRC/AHRC.

1. Los robots son herramientas de usos múltiples. Los robots no deben diseñarse exclusiva o principalmente para matar o dañar a los humanos, excepto en interés de la seguridad nacional.
2. Los seres humanos, no los robots, son agentes responsables.
3. Los robots son productos. Deben diseñarse utilizando procesos que garanticen su seguridad y protección.
4. Los robots son artefactos fabricados.
   forma engañosa de explotar a los usuarios vulnerables; en su lugar, la naturaleza de su máquina debe ser transparente.
5. La persona con responsabilidad legal por un robot debe ser atribuida.

Los principios han tenido un impacto significativo en la investigación sobre robótica en el Reino Unido y continúan provocando un debate sustancial. Validez: ¿los principios son correctos como afirmaciones sobre la naturaleza de los robots (por ejemplo, que son herramientas y productos), los desarrolladores de robots y la relación
entre los robots y las personas (por ejemplo, que los robots deben tener un diseño transparente), o son ontológicamente defectuosos, inexactos, obsoletos o engañosos?

b. Suficiencia/generalidad—son el principio sin cobertura suficiente y suficientemente amplia de todos los temas importantes que podrían surgir en la regulación de la robótica en el mundo real o son preocupaciones significativas que se pasan por
alto. c. Utilidad: son los principios de uso práctico para desarrolladores, usuarios o leyes de robots. creadores, en la determinación de estrategias para mejores prácticas en robótica, o estándares o marcos legales, o están limitados en su uso por falta de especificidad o al permitir excepciones críticas (como el uso de robots como armas con fines de seguridad nacional).

Un simposio de 1 día se llevó a cabo el 4 de abril como parte de la Conferencia AISB 2016 en Sheffield, Reino Unido. Estas actas contienen comentarios sobre los principios solicitados antes de la reunión.

---

[1] https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

AISB2016Simposio sobre los Principios de la Robótica

Comentarios enviados

El significado de los principios EPSRC de la robótica

Joanna Bryson, Universidad de Bath y Princeton Center for Information Technology Policy

Introducción

Al revisar los principios de la robótica, es importante considerar cuidadosamente su significado completo. Aquí visito brevemente primero el significado del documento como un todo, luego de sus partes constituyentes. Los principios de robótica de EPSRC fueron generados como entregable por un grupo ensamblado con poca orientación y sin necesidad de entregar entregable. La intención original del evento de robótica epsrc parece haber sido solo la discusión en sí, o quizás incluso solo el hecho de la reunión. Los académicos presentes querían algo que mostrar por el tiempo que dedicaron y, como resultado, una cantidad sustancial de tiempo de todos los presentes el último día se dedicó a la creación de las tres versiones de los principios y su documentación. Parte de la documentación fue prorrogado –nuevamente por consenso– después de la reunión. Es correcto y apropiado
que haya una manera de examinar e incluso actualizar o mantener el documento. Incluso las constituciones nacionales tienen medios para su mantenimiento. Sin embargo, es fundamental para la eficacia de los documentos de política que no sean fáciles de cambiar. Deben proporcionar un timón para evitar el difuminado y, como tales, normalmente son más difíciles de modificar de lo que fueron instanciar en primer lugar. Tenga en cuenta que a algunos países y otras uniones políticas no les ha resultado fácil crear incluso sus constituciones iniciales por esta misma razón. Por lo tanto, es importante pensar cuidadosamente sobre el significado de los principios.

Los principios como política

La política tecnológica, y la política en general, es algo sorprendentemente amorfo. Al igual que otros aspectos de la inteligencia natural, la política no siempre reside en la ley o incluso en la gobernanza. Gran parte de la política no está escrita e incluso no se conoce explícitamente. El Reino Unido es realmente sobresaliente en su innovación del derecho consuetudinario, que reconoce esto y la importancia de la cultura y los precedentes. No obstante, a la luz fría de un comité que trabaja en casos de impacto de REF, tenemos que preguntar, ¿son los principios una política? Creo que la respuesta es sí". Son un conjunto de pautas acordadas por una fracción sustancial, aunque quizás arbitraria, de la comunidad a la que afectan, y se publican en las páginas web del gobierno. Toda
política tiene tres componentes: asignativo, distributivo y estabilizador. La asignación es el proceso de determinar en qué problemas vale la pena gastar tiempo y otros recursos. En el caso de los principios, esto fue instigado por el EPSRC (o alguna organización por encima de ellos) debido a la preocupación de que el público británico pudiera rechazar la robótica ya que tenían alimentos modificados genéticamente. Nos dijeron que el rechazo a la robótica se consideraba una grave amenaza para la economía británica. Tenga en cuenta también que cada uno de los participantes (al menos aquellos a los que no se les pagó específicamente para asistir) también hizo inversiones individuales, asignando tiempo al problema de la ética de los robots, aunque para muchos esto se confundió con la oportunidad de hacerse más conocidos por su
principal organización de financiación. El componente estabilizador es el que asegura que la política, una vez establecida, se incorpore a la sociedad de tal manera que sea poco probable que se deshaga rápidamente o que se convi

una responsabilidad o materia de controversia. En el caso de los principios, esto evidentemente se ha logrado al menos hasta cierto punto, ya que estamos celebrando su quinto aniversario. Hablando con otros autores, sé de ninguno completamente enamorado del producto final, pero todos respetan el proceso democrático (ciertamente representativo) mediante el cual se lograron, y la importancia del compromiso mutuo de sus colegas con el producto final. Por mi parte, me encantaría ver los principios cosificados en la política o incluso en la ley, pero todavía tengo que descubrir el proceso mediante el cual esto podría lograrse. Sin embargo, han llamado y continúan llamando la atención de varias juntas de estándares e investigaciones parlamentarias, así como de la prensa y otros académicos. Dejo para el final el aspecto más controvertido de la política: el distributivo. En su base, toda política se trata de selección de acciones,
y eso implica la asignación o más bien reasignación de recursos. La política trata de pasar por alto esto, ya que necesariamente va contra la corriente de aquellos de quienes se reasignan los recursos, incluso en los casos en que esos individuos pueden obtener un beneficio neto. Odiamos perder el control, pero las políticas son para el control. "Intenta pasar por alto" es de hecho un eufemismo; hacer aceptable la redistribución puede ser el proyecto central de los políticos. En este caso, el gobierno tenía preocupaciones muy específicas sobre las personas que habían estado en los medios promoviendo el miedo a los robots, y tenían un deseo muy claro de encontrar formas de cambiar la atención de los medios y las impresiones públicas hacia la seguridad de la robótica. Por el contrario, fueron realmente los participantes quienes mencionaron los otros cambios importantes del sensacionalismo al pragmatismo: la afirmación de que los robots no son partes responsables ante la ley y que los usuarios no deben ser engañados acerca de sus capacidades. Los representantes del consejo sabían que esta redistribución del poder enfadaría a algunos de los destacados beneficiarios de los fondos, y los participantes sabían lo mismo de algunos de sus colegas. Sin embargo, hubo una sorprendente unanimidad entre los académicos en cuanto a que los mayores riesgos morales de los robots eran su naturaleza carismática y el increíble entusiasmo que muchas personas tienen por invertir su propia identidad en las máquinas, lo que llevó a una sorprendente confusión sobre su naturaleza que todos nosotros habíamos presenciado. Este carisma y confusión dejó la puerta abierta a todo tipo de manipulaciones por parte de corporaciones y gobiernos, donde los robots podrían ser establecidos como responsables o incluso sustitutos de vidas o valores humanos.

El principio de matar

Los robots son herramientas multiusos. Los robots no deben diseñarse única o principalmente para matar o dañar a los seres humanos, excepto en interés de la seguridad nacional.

Los primeros tres principios pretendían ser correcciones de las leyes de Asimov. Los robots no son partes responsables, por lo que no podrían matar. En cambio, los robots no deberían poder usarse como herramientas para matar. Esta simple regla hizo clara la transferencia de la subjetividad moral y, al mismo tiempo, satisfizo los deseos pacifistas de la mayoría de los presentes. Sin embargo, pragmáticamente, los robots ya se usaban como armas de guerra, y una ley que no se puede hacer cumplir es cuestionable. Estábamos convencidos de que liderar con un principio conocido como falso disminuiría significativamente nuestras posibilidades de impacto cultural. Por lo tanto, el significado del primer principio podría parecer neutralizado por el compromiso de la excepción, pero que los robots no sean armas en la sociedad civil sigue siendo un punto social importante. Más allá de esto, el hecho de que la política práctica deba tener en cuenta las necesidades del gobierno para abordar tanto la seguridad como la industria (el Reino Unido es la quinta nación más grande del mundo en el tráfico de armas) también tiene significado. Por puramente académico que algunos de nosotros podamos

deseamos que sea nuestra disciplina, el hecho de que muchos de sus productos tengan una utilidad inmediata significa que no podemos evitar el impacto en nuestro mundo.

El principio de cumplimiento

Los humanos, no los robots, son los agentes responsables. Los robots deben diseñarse y operarse en la medida de lo posible para cumplir con las leyes existentes y los derechos y libertades fundamentales, incluida la privacidad.

La segunda ley de Asimov tiene que ver con seguir instrucciones, pero incluso la noción de obedecer implica una agencia moral. El significado original de esta ley era que los robots son tecnología ordinaria y se ajustan a normas y leyes ordinarias. En la configuración de los principios como conjunto, el segundo principio llegó a ser el que comunicó más algunos de los peligros de la IA en general, y la IA confundida con un sujeto moral en particular. El énfasis en la privacidad refleja la preocupación especial de un agente físico inteligente que percibe que ocupa exactamente el mismo espacio que una familia humana. La tecnología está fundamentalmente inmersa en el umwelt humano, más que cualquier tecnología anterior o mascota, quizás incluso más que algunos humanos en un hogar como los niños. Tiene acceso al lenguaje escrito y hablado, información social, horarios observados, etc. Además, puede ser confundido con una mascota u otro miembro de confianza de la familia, sus habilidades especiales para una comunicación perfecta con el mundo exterior se olvidan temporalmente o sus habilidades para aprender regularidades. y clasificar los estímulos. En estos casos, la información de los primates puede almacenarse sin querer en una nube pública, o incluso en una supuestamente privada susceptible de ser pirateada. Obligar a una tecnología tan novedosa y similar a la humana a cumplir con las normas legales estándar de privacidad y seguridad no es una tarea trivial.

El principio de mercantilización

Los robots son productos. Deben diseñarse utilizando procesos que aseguren su seguridad y protección.

La última ley de Asimov es la autoprotección, pero los robots no tienen yo. En cambio, esta ley se centró en proteger a los humanos de los robots al nivel de la solidez básica del robot. El principio nuevamente nos trae a la conciencia de la naturaleza fabricada no especial del robot, en un intento de evitar la evasión de responsabilidad legal alegando que los robots tienen una naturaleza única. El fabricante de un robot debe tener exactamente tanta responsabilidad por el funcionamiento de la maquinaria según las especificaciones como el fabricante de un automóvil o una herramienta eléctrica. De hecho, los robots pueden ser automóviles o herramientas eléctricas, pero si es así, deberían ser más seguros que la variedad convencional de cualquiera de ellos.

El principio de transparencia

Los robots son artefactos manufacturados. No deben diseñarse de manera engañosa para explotar a los usuarios vulnerables; en su lugar, la naturaleza de su máquina debe ser transparente.

Los primeros tres principios establecieron el marco legal para la fabricación y venta de robótica como idénticos a otros productos. Los dos últimos están destinados a garantizar que el estado también se comunique al usuario. El principio de transparencia busca garantizar que las personas no inviertan demasiado en su tecnología, por ejemplo, contratando a un cuidador de casas para evitar que el robot se sienta solo. Algunos especialistas en robótica se oponen a este principio porque el engaño es necesario para la eficacia de su aplicación prevista, como hacer que las personas no se sientan solas para que estén menos deprimidas. Otros sostienen que este principio niega la posibilidad de que los robots sean más que máquinas ordinarias. El primer argumento está abierto a la experimentación. En primer lugar, debe establecerse que no hay forma de desencadenar un compromiso emocional sin engaño, lo que parece poco probable dado el grado de compromiso emocional que se establece con personajes ficticios y objetos claramente no cognoscitivos. Si se establece experimentalmente un requisito para el engaño, entonces se puede debatir la compensación entre los costos y los beneficios del engaño. El segundo, sin embargo, es incontrovertible. La autoría que tenemos sobre los artefactos es una parte fundamental de su naturaleza mecánica. La IA es, por definición, un artefacto. Hasta cierto punto, podríamos incluso argumentar que este principio es autolimitante. Si la IA realmente fuera capaz de alterar lo que significa ser una máquina, comunicar esta naturaleza de máquina modificada aún cumpliría con este principio.

El principio de la responsabilidad jurídica

Se debe atribuir a la persona con la responsabilidad legal de un robot.

Finalmente, el quinto principio comunica el estado de los robots como artefactos de la manera más fundamental posible. Son propiedad, y esa propiedad debe estar legalmente atribuida. El hecho de que los robots se construyan y posean es la razón por la que he argumentado anteriormente que estamos éticamente obligados a no convertirlos en personas porque poseer personas es incuestionablemente poco ético. El argumento no es que existan robots parecidos a personas cuyo estatus debamos degradar legalmente, sino que el estatus legal necesariamente degradado significa que no debemos hacer que la semejanza a la persona sea una característica de ningún robot fabricado legalmente. Sin embargo, los principios de la robótica no llegan a este extremo del futurismo. Como dije antes, se enfocan en comunicar la realidad actual a una población tan ansiosa de poseer e identificarse con lo sobrehumano que fácilmente se les puede hacer creer que un robot mal fabricado o mal operado es el culpable del daño que se le inflige. Si escucha un ruido horrible y encuentra un automóvil estrellado contra su casa, puede identificar rápida y fácilmente al propietario del automóvil, incluso si el automóvil está vacío en este momento, simplemente a través de sus placas de matrícula o, en el peor de los casos, a través de números de serie. La idea es que lo mismo debería ser cierto si encuentra un robot incrustado en su propiedad. Los participantes en el retiro de robótica predijeron con precisión un problema que ahora ya está presente en nuestra sociedad debido a los drones, y que ahora se está abordando en algunas naciones con licencias obligatorias, como recomendó el comité.

Conclusión

Para resumir, los principios EPSRC son valiosos porque representan una política construida a un costo significativo para el contribuyente y el personal. Si bien ninguna política es perfecta, idealmente

solo debe ser reemplazada por una nueva política con un nivel de inversión equivalentemente alto o superior tanto por parte del gobierno como de los expertos en la materia. Su propósito es brindar confianza a los consumidores y ciudadanos en la robótica como una tecnología confiable apta para generalizarse en nuestra sociedad. Cada uno de los principios individuales representa preocupaciones sustanciales de los expertos y las partes interesadas, aunque a veces esa representación en sí misma no es perfectamente transparente. El objetivo general era comunicar claramente que la responsabilidad de la fabricación y el funcionamiento seguros y fiables de los robots no era diferente de la de cualquier otro objeto fabricado y vendido en el Reino Unido y, por lo tanto, las leyes vigentes del país deberían ser adecuadas para cubrir tanto a los consumidores como a los fabricantes. .

Es importante darse cuenta de que este no es el caso de todos los robots imaginables. Es fácil concebir obras de arte únicas que califiquen como robots y no como productos mercantilizados, o concebir robots que simplemente se construyen de manera insegura o irresponsable. Lo que a la gente le cuesta más conceptualizar es que puede haber propiedades cognitivas, como el sufrimiento, que podrían ser factibles de incorporar a un robot, pero hacerlo sería tan poco ético como poner los frenos defectuosos a un coche. Los principios de la robótica no buscan determinar lo que es posible, buscan comunicar prácticas recomendables para integrar la robótica autónoma a la ley de la tierra.

RobóticaInvestigaciónÉticaDiscusión

A. Gning, D. Davis, Y Cheng, P. Robinson, Departamento de Ciencias de la Computación,
Universidad de Hull.

e.gning@hull.ac.uk

Introducción

En el mundo moderno con el desarrollo de los recursos tecnológicos, la robótica ha resultado en numerosas aplicaciones [1] y, a menudo, en un despliegue imprevisto en la vida real (por ejemplo, el uso cada vez mayor de drones en aplicaciones civiles). Para acompañar esta era de la robótica, la comunidad investigadora y la sociedad en general necesitan definir principios éticos que sean lo suficientemente generales como para ser robustos a la evolución en el tiempo y adaptarse a la gama de posibles aplicaciones.

Los principios éticos deben vulgarizarse y universalizarse lo suficiente como para que los diseñadores y fabricantes de robots sean conscientes de las normas y los límites que deben respetar.

En general, las aplicaciones de la robótica se pueden clasificar en cuatro grupos: robots domésticos o de asistencia humana [2] [3] [4], robots médicos [5] [6] [7] [8], robots de defensa [9] [10].] y robots industriales [11] [12]. La discusión se centra en los primeros tres grupos de robots, ya que los robots industriales a menudo están limitados a áreas limitadas con conjuntos preespecificados de tareas limitadas.

y no están en interacción directa con la sociedad.

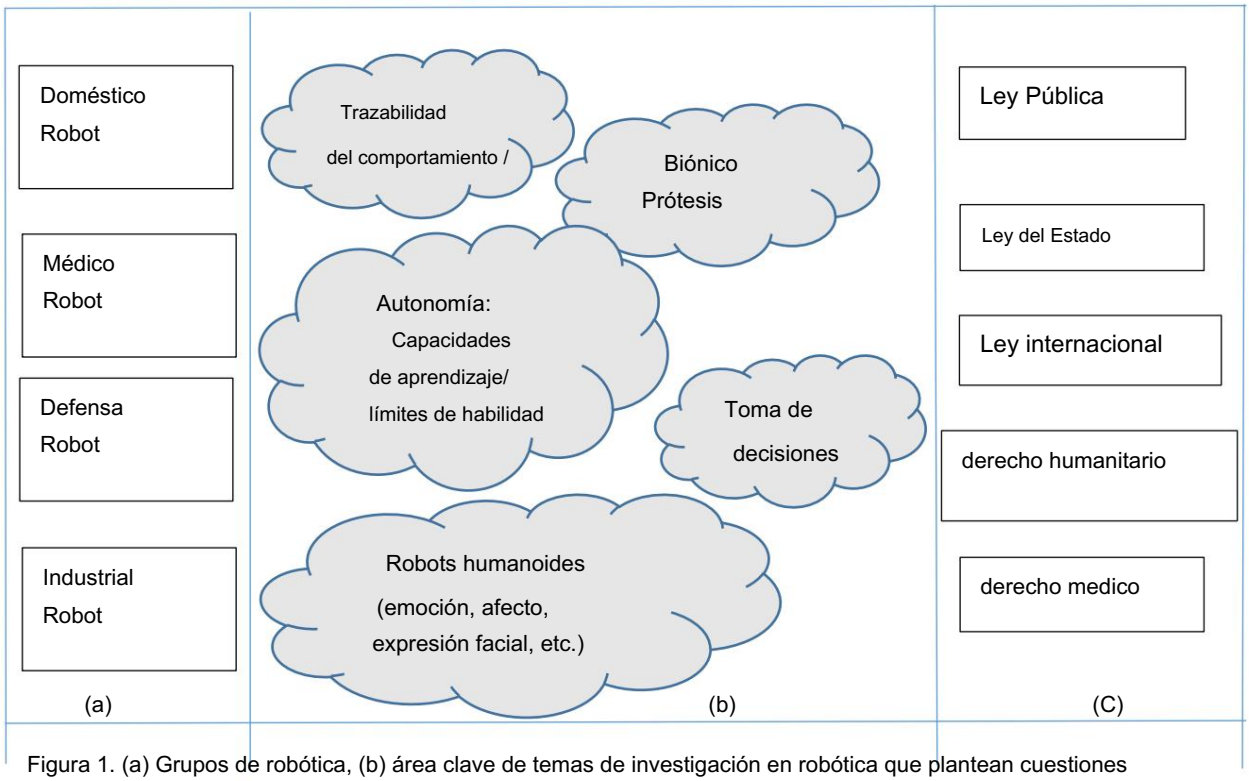

Figura 1. (a) Grupos de robótica, (b) área clave de temas de investigación en robótica que plantean cuestiones éticas, (c) dominio de las regulaciones legales a considerar.

La Figura 1 enumera la difícil tarea de regular la ética para diferentes grupos de robótica con respecto a cinco temas de investigación populares. En cada grupo de robótica (figura 1-(a)), se pueden plantear cuestiones éticas y se representan las claves (figura 1-(b)): ¿cómo podemos diseñar robots para que siempre sean

capaz de interpretar su comportamiento, conocer los límites y límites de los robots de aprendizaje; los robots humanoides con imitación de gestos humanos y animación facial pueden llevar la interacción con humanos (especialmente niños y personas discapacitadas) a nuevas fronteras planteando necesidades de regulación y anticipación de posibles implicaciones; Hoy en día, los humanos pueden beneficiarse enormemente de las prótesis gracias a los avances en la investigación en robótica. Sin embargo, esto puede dar lugar a que las personas sin discapacidades busquen prótesis que puedan mejorar sus capacidades, lo que plantea nuevamente la necesidad de nuevas regulaciones. Finalmente, el mayor desafío se relaciona con las capacidades de decisión de los robots. Estas decisiones involucran directamente la vida humana y, por lo tanto, plantean cuestiones de cuestiones jurídicas resultantes de las acciones ejecutadas.

Además de estos temas de investigación, la ética de la robótica debe ser compatible con un conjunto de dominios legales (figura 1 – (c)). Es necesario tener en cuenta que los robots deben ser compatibles con diferentes niveles de leyes que pueden cambiar, por ejemplo, de una región a otra o de un estado a otro.
estado.

Hace cinco años en el Reino Unido, un panel de distinguidos expertos en robótica e inteligencia artificial publicó los Principios de robótica del EPSRC en forma de cinco reglas y siete mensajes de alto nivel. Proponemos discutir estas reglas con un enfoque en la estructura transversal - entre robótica
grupos, temas de investigación y marcos de leyes presentados en la figura 1- y con respecto a los tres criterios de validez, suficiencia y utilidad.

Discusiónsobreesteestablecimientodereglas

El comentario general que se puede hacer sobre el conjunto de cinco reglas es que es bastante ambicioso pensar que es posible dar una guía común/uniforme a todo tipo de robots, a lo largo de toda la evolución posible de la investigación y todos los marcos legales. Sería más natural buscar reglas de orientación que reflejen la naturaleza transversal de la robótica que se muestra en la figura 1. Creemos que las cinco reglas no son lo suficientemente generales y las reglas deben especificarse y particularizarse para cada grupo de robots en la figura 1 (a) tomando en cuenta el aspecto específico de las leyes y las vías de investigación involucradas en la figura 1 (c) al final (b).

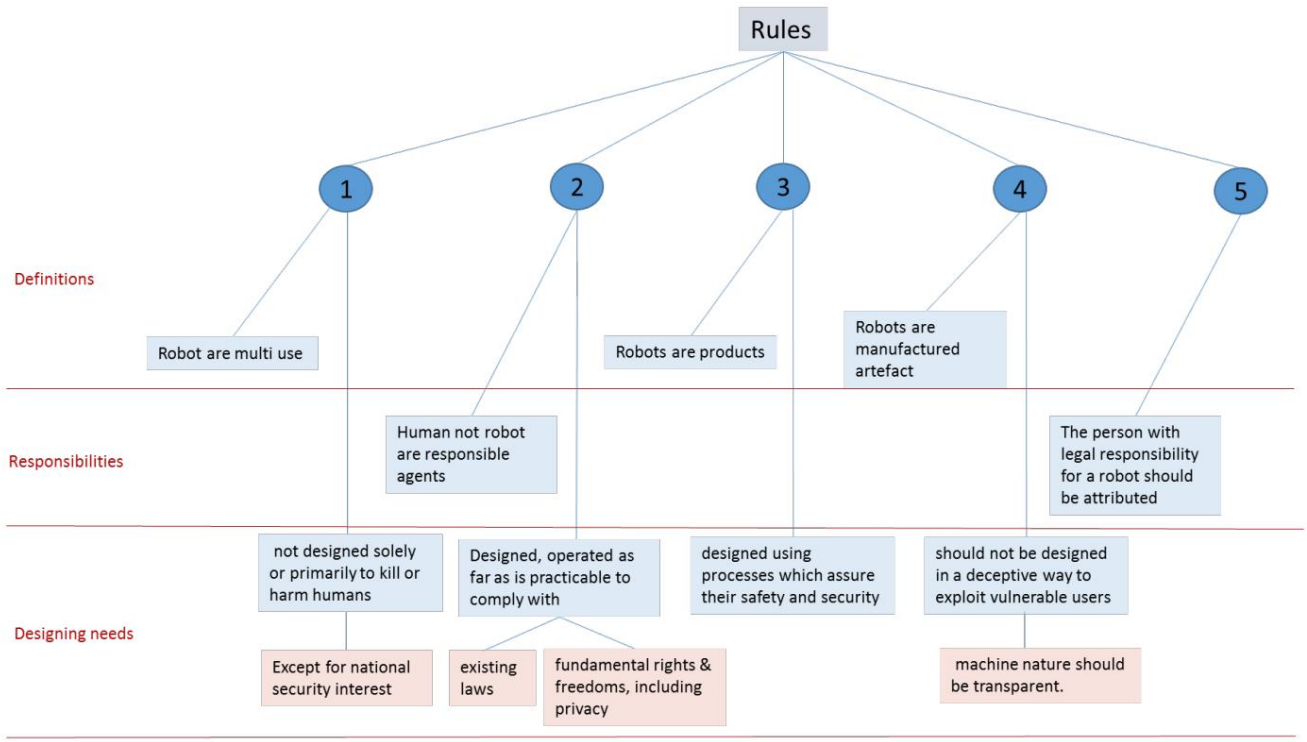

Figura 2. El conjunto de cinco reglas

La figura 2 representa de forma gráfica el conjunto de cinco reglas. Se puede ver que hay un patrón común a las cinco reglas: las primeras oraciones son a menudo generalidades y definiciones.
sobre robots como "los robots son productos" o afirmar que las responsabilidades no son responsabilidad del robot sino del ser humano. Las últimas oraciones establecen las necesidades de diseño y oscilan entre cuestiones de seguridad, cuestiones legales y cuestiones de transparencia.

Claramente, las críticas sobre la estructura que se muestra en la Figura 2 se pueden enumerar de la siguiente manera

- Las cinco reglas a veces se superponen. Por ejemplo , "cumplir con la ley existente" en la regla 2 encapsula "no diseñado única o principalmente para matar o dañar a seres humanos" en la regla 1; “Se debe atribuir la responsabilidad legal de un robot a la persona” en la regla 5 puede encapsular implícitamente “Los humanos, no los robots, son agentes responsables” en la regla 2.
- Las cinco reglas no son generales. Por ejemplo, apenas podemos ver cómo las prótesis biónicas pueden ajustarse a las reglas en la forma actual (por ejemplo, la investigación en robótica puede permitir en el futuro la modificación del cuerpo humano para ganar más potencia, velocidad, etc.).
  La sexualidad artificial es otro ejemplo de investigación controvertida que puede dar lugar a cuestiones éticas.
- Las Cinco reglas no insisten en que la ley puede ser contradictoria dependiendo del dominio considerado de los consejos, regiones, países y continentes. Un ejemplo similar que se encuentra a menudo en la investigación son las solicitudes de patentes para las cuales se necesitan varios estudios específicos para aplicar en determinadas regiones del mundo. Las leyes pueden ser aún más complicadas ya que las creencias religiosas y los hábitos y costumbres de las personas dictarán la noción de ética.

Conclusiones

En esta discusión, brindamos brevemente argumentos sobre la necesidad de una formulación diferente para las cinco reglas. Se demuestra a través de una representación pictórica de las cinco reglas que, de hecho, no son suficientes, se superponen y no reflejan explícitamente los verdaderos desafíos de la ética de la robótica.

Basamos parte de nuestro razonamiento en la naturaleza transversal de la ética robótica en tres líneas: grupos que constituyen la robótica, caminos futuros de la robótica que son esenciales para ser capturados al definir la ética y la naturaleza estructurada de la ley. Recomendamos una reformulación natural que diferencie la ética para cada grupo de robótica sin dejar de lado las contradicciones y las fuertes limitaciones que pueden existir debido a la naturaleza estructural de la ley.

Bibliografía

[1] G. Bekey, Robótica: estado del arte y desafíos futuros., California: London Imperial 2008.
Prensa universitaria. ,

[2] K. Dautenhahn, S. Woods, C. Kaouri, ML Walters, KL Koay e I. Werry, ¿Qué es un robot compañero-amigo, asistente o mayordomo?, IIC o. IR a. Sistemas, Ed., 2005.

[3] J. Forlizzi y DC, Robots de servicio en el entorno doméstico: un estudio de la aspiradora roomba en el hogar, 1ª conferencia ACM SIGCHI/SIGART sobre interacción humano-robot. ACM, 2006.

[4] JY Sung, RE Grinter, HI Christensen y L. Guo, ¿Amas de casa o tecnófilos?: comprensión de los propietarios de robots domésticos, 3ra Conferencia Internacional ACM/IEEE sobre Interacción Humano-Robot (HRI), 2008, pp. 129-136.

[5] GP Moustris, SC Hiridis, K. Deliparaschos y K. Konstantinidis, Evolution of
sistemas quirúrgicos robóticos autónomos y semiautónomos: una revisión de la literatura,

vol. 7, The International Journal of Medical Robotics and Computer Assisted Surgery, 2011, pp. 375-392.

[6] J. Rassweiler, J. Binder y T. Frede, "Robótica y telecirugía: ¿cambiarán nuestra ¿futuro?," vol. 11, no. 3, pp. 309-320, 2001.

[7] G. Kwakkel, KBJ y KHI, Efectos de la terapia asistida por robot en la extremidad superior recuperación después de un accidente cerebrovascular: una revisión sistemática, Neurorehabilitación y reparación neural, 2007.

[8] K. Cleary y C. Nguyen, "Estado del arte en robótica quirúrgica: aplicaciones clínicas y desafíos tecnológicos", vol. 6, núm. 6, págs. 312-328, 2001.

[9] PW Singer, "Robots en guerra", Wilson Quarterly, 2008.

[10] TK Adams, "La guerra futura y el declive de la toma de decisiones humana", Parámetros, vol. 31, núm. 4, 2001.

[11] P. Leitão, "Control de fabricación distribuido basado en agentes: una encuesta de vanguardia", Aplicaciones de ingeniería de la inteligencia artificial, vol. 22, núm. 7, págs. 979-991, 2009.

[12] ZM Bi, SY Lang, W. Shen y L. Wang, "Sistemas de fabricación reconfigurables: el estado del arte", International Journal of Production Research, vol. 46, núm. 4, págs. 967-992, 2008.

Los robots no son solo herramientas

Tony J. Prescott, Universidad de Sheffield

En el centro de los principios de la robótica del EPSRC (en adelante, "los principios") hay una serie de afirmaciones fontológicas sobre la naturaleza de los robots que sirven como axiomas para enmarcar el desarrollo subsiguiente de los desafíos éticos y las reglas. Estas incluyen afirmaciones sobre lo que son los robots y también sobre lo que no son. Las afirmaciones sobre qué robots son incluyen que “los robots son herramientas de usos múltiples” (principio 1), que “los robots son productos” (principio 3) y “piezas de tecnología” (comentario sobre el principio 3), y que “los robots son artefactos fabricados” (principio 4). ejemplo 2), que los robots son “simplemente no personas” (comentario sobre el principio 3), y que la inteligencia de los robots solo puede dar una “impresión de inteligencia real” (comentario sobre el principio 4).

Al leer por primera vez, estas afirmaciones parecen afirmaciones directas de verdades obvias. Argumentaré que este no es el caso. En cambio, propondré que estos compromisos ontológicos carecen de matices, suponen con demasiada facilidad que conocemos las condiciones límite del desarrollo futuro de la robótica, y que oscurecen o ignoran algunos de los importantes debates éticos. Podría comenzar por pensar detenidamente en el estado ontológico de los robots.

Si observamos cómo se presentan los principios, parece haber un proceso implícito de inducción en el trabajo que permite que las afirmaciones sobre lo que son los robots más actuales se interpreten como afirmaciones sobre lo que los robots deben ser esencialmente. “herramientas de varios tipos, aunque herramientas muy especiales” en el preámbulo. Si bien es fácil estar de acuerdo con una afirmación general de que los robots son herramientas de usos múltiples, especialmente en el contexto de la discusión sobre el uso dual (principio 1), la afirmación mucho más fuerte de que los robots son solo herramientas, o simplemente herramientas, niega que puedan pertenecer sensatamente a otras categorías disjuntas.

Tomemos como ejemplo la categoría de 'compañero'. Existe un gran esfuerzo en torno al desarrollo de compañeros robóticos que puedan brindar apoyo social y emocional a las personas, como se reconoció parcialmente en la discusión del principio 4. La categoría de herramientas describe objetos físicos/mecánicos que cumplen una función, mientras que la categoría de compañeros describe a otras personas significativas, generalmente personas o animales, con quienes podría tener una relación recíproca marcada por un vínculo emocional .La posibilidad de que los robots puedan pertenecer a ambas categorías plantea problemas importantes e interesantes que se oscurecen al insistir en que los robots son solo herramientas.

De hecho, de acuerdo con la visión de los robots como herramientas, la discusión sobre la compañía de los robots en los principios es bastante desdeñosa, describiendo los juguetes que podrían proporcionar cierto placer a las personas que no pueden, o no pueden costear, tener animales como mascotas.

Se argumenta que la naturaleza falsa de los compañeros robóticos crea un problema ético en el sentido de que los compañeros robóticos son potencialmente engañosos y deben diseñarse de manera que su “naturaleza de máquina sea transparente”.

El problema ontológico aquí se refiere particularmente a la afirmación de que los robots nunca podrían poseer capacidades psicológicas tales como emociones o inteligencia “reales”. Lo que son, en términos humanos, es objeto de un acalorado debate en las ciencias cognitivas y cerebrales.

De hecho, hay contra-afirmaciones de que los robots, adecuadamente configurados, pueden tener emociones[1], mientras que el futuro de la inteligencia artificial, como inteligencia, no tiene un techo obvio por debajo del nivel humano.

Otro problema se refiere a la suposición acerca de cómo la gente verá los robots, específicamente, que los robots serán vistos como herramientas si se muestran de manera transparente. por ejemplo, las personas pueden antropomorfizar a los robots independientemente de lo obvios que sean productos fabricados.

Las animaciones de Heider-Simmel de figuras geométricas simples [2] (ver figura), muestran cuán cruda puede ser esta información y, sin embargo, todavía veremos intencionalidad, motivación e incluso emoción. respuesta emocional a esto nosotros mismos.

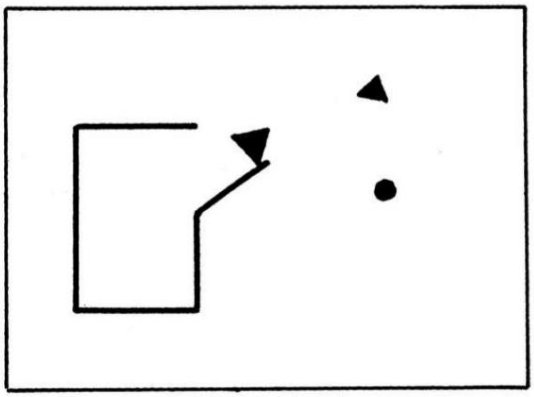

Figura. Las formas geométricas que se mueven en una animación simple fueron interpretadas como “seres animados, principalmente personas”, en este famoso estudio de 1944 de Heider y Simmel.

Un análisis de cuestiones ontológicas y psicológicas en la interacción humano-robot ha sido previamente realizado por Kahn y sus colegas [4]. Siguiendo una línea de pensamiento similar, podemos describir cuatro formas generales en las que se pueden combinar las perspectivas ontológicas sobre lo que son los robots y las perspectivas psicológicas sobre cómo se ven los robots. Estas se ilustran en la siguiente tabla junto con algunas de las cuestiones éticas que implican.

I. Los robots son solo herramientas (o), y las personas verán a los robots solo como herramientas a menos que sean engañados por un diseño de robot engañoso (p). Cuestiones éticas: debemos abordar las responsabilidades humanas como creadores/usuarios de robots y el riesgo de engaño en la fabricación de robots que parecen ser algo que no son. Esta es la posición de 'los principios'.	II. Los robots son solo herramientas (o), pero las personas pueden verlos con capacidades psicológicas significativas independientemente de la transparencia de su naturaleza-máquina (p). Cuestiones éticas: debemos tener en cuenta cómo las personas ven a los robots, por ejemplo, que pueden sentir que tienen relaciones significativas y valiosas con los robots, o que pueden ver a los robots con estados internos importantes, como la capacidad de sufrir, a pesar de no tener tales capacidades.
tercero Los robots pueden tener algunas capacidades psicológicas significativas (o), pero las personas aún los verán como simples herramientas (p). Cuestiones éticas: debemos analizar los riesgos de tratar entidades que pueden tener capacidades psicológicas significativas, como la capacidad de sufrir, como si fueran solo herramientas, y los peligros inherentes a la creación de una nueva clase de entidades con capacidades psicológicas significativas, como la inteligencia similar a la humana, sin reconocer que lo estamos haciendo.	IV. Los robots pueden tener algunas capacidades psicológicas significativas similares a las humanas (o), y las personas verán el mas que tiene tales capacidades (p). Cuestiones éticas: debemos considerar escenarios en los que las personas deberán coexistir junto con nuevos tipos de entidades psicológicamente significativas en forma de futuros robots/IA.

Tabla. Cómo se pueden combinar las perspectivas ontológicas (o) y psicológicas (p) sobre los robots (después de Kahnetal., 2007).

Tenga en cuenta que solo un cuadrante de esta tabla (I) se aborda en los principios, pero que II, III y IV son todos posibles, al menos en teoría. Para concluir este ensayo, quiero considerar brevemente algunas de las cuestiones éticas que surgen en los cuadrantes II–IV.

En el cuadrante II, surgen preguntas interesantes sobre cómo deben tratarse los robots, no porque sean agentes sensibles, sino porque las personas elegirán tratarlos como tales. propietarios japoneses de perros robot Sony Aibo [6], no parece tan extraño cuando se ve desde la perspectiva de cómo los robots son vistos por las personas en lugar de en términos de lo que son. con algunos robots pueden ser similares a los que desarrollamos con otras posesiones valiosas, como cicatrices y teléfonos móviles. Por otro lado, para algunos robots, pueden parecerse más a las relaciones que tenemos con los animales de compañía, incluidos, por ejemplo, deseosos de apoyarlos y nutrirlos (algo que ustedes mismos encuentran gratificante). como la capacidad de recordar y comunicarme acerca de algunas de nuestras experiencias compartidas. Más generalmente, lo que puede ser necesario, para desarrollar principios éticos adecuados, es desarrollar una taxonomía de las diferentes formas de vínculos emocionales que pueden existir entre robots y personas y analizar los factores que podrían sustentar el desarrollo y mantenimiento de tales relaciones [7].

El cuadrante III se refiere a la posibilidad de que los robots tengan capacidades psicológicas significativas que están en peligro de ser pasados por alto por las personas. Esto plantea riesgos éticos que no se discuten en los principios, pero que han sido resaltados por otros. de entidad consciente que sufre innecesariamente debido a nuestras acciones, esto es claramente un problema ético si sucediera. Aunque esto puede parecer poco probable a corto plazo, hay motivos para considerar que esto podría conllevar un riesgo a mediano y largo plazo a medida que las arquitecturas cognitivas para los robots se vuelvan más sofisticadas. Varias tendencias en la investigación en curso sobre la conciencia humana también respaldan esta posibilidad. teorías temporales de la conciencia [9] afirma un papel crítico para la integración de la información

que no requiere necesariamente un sustrato biológico. Los neurólogos también están reevaluando si las islas de actividad integrada en los cerebros de los pacientes 'encerrados' podrían constituir una forma de conciencia mínima [10]. Finalmente, existe un debate activo sobre si los animales con los cerebros más pequeños que los nuestros, como los peces, podrían estar en una forma significativa (por ejemplo, que pueden experimentar dolor) [11]. Estos desarrollos sugieren que la conciencia podría ser posible en un agente artificial sin tener que igualar el tamaño o la complejidad de un cerebro humano intacto.

[12], y Bryson[13] ha propuesto que los robots de hoy en día ya podrían tener algunas formas simples de conciencia que cumplen con algunos criterios comúnmente propuestos.

Una de las consecuencias de la visión de los robots como “simples herramientas” es la desestimación implícita de la posibilidad de una IA fuerte: que los futuros robots puedan tener una inteligencia general a nivel humano o superior.], es que una singularidad de la IA podría revertir la relación maestro-esclavo entre los humanos y los robots. una super-IA que se auto-arranca. Un enfoque ético sin duda alentaría una mayor vigilancia.

El debate es la perspectiva del 'cerebro global', propuesta por Heylighen[15] y otros, de que los humanos y la IA avanzada podrían coexistir para nuestro beneficio mutuo. Esto nos recuerda que la ética debe consistir en analizar los beneficios potenciales así como los riesgos.

Aunque los escenarios de los cuadrantes III/IV pueden parecer inverosímiles o al menos distantes, tales preocupaciones han capturado la imaginación del público y han provocado importantes llamados al debate (por ejemplo, [16]). Señalando que los robots son solo herramientas pueden hacer poco para calmar las voces y podrían parecer hegemónicos y condescendientes. Si bien los enfoques de estos desafíos a más largo plazo son necesariamente especulativos, un punto de partida para reconocer que hay preocupaciones aquí que merecen mayor atención.

Un enfoque sincero tal vez reconozca que, si bien la mayoría de los robots actualmente son poco más que herramientas, estamos entrando en una era en la que habrá nuevos tipos de entidades que combinan algunas de las propiedades de las máquinas y herramientas con capacidades psicológicas que previamente habíamos pensado que estaban reservadas para organismos biológicos complejos como los humanos. vive de la misma manera que los organismos biológicos , ni simplemente mecánicos como con una máquina tradicional.

La liminalidad de los robots los hace a la vez fascinantes e inherentemente aterradores, y un pararrayos para nuestros temores más amplios sobre los efectos deshumanizantes de la tecnología[18].

La Asociación de Científicos de Manhattan escribió en 1945 [19] sobre su sentimiento de responsabilidad colectiva por su papel en el desarrollo de una tecnología con “potencial para un gran daño o un gran bien” (energía atómica) y sobre su “conciencia especial” de que podría conducir al “avance de nuestra civilización o su destrucción total”. Los usuarios también tienen la responsabilidad especial de comprender y expresarse abiertamente sobre lo que podría traer el futuro de la robótica y sus beneficios y amenazas potenciales.

Referencias

1. Fellous, J.-M., Fromhumanemotionstorobotemotions, en AAAISimposio de primavera sobre arquitecturas para modelar emociones: fundamentos interdisciplinarios E.HudlickaandL.Caññamero,Editors.2004,AAAIPress:MenloPark,CA.p.37-47.
2. Heider, F. y M. Simmel, Un estudio experimental del comportamiento aparente. The American Journal of Psychology, 1944.57(2):p.243-259.
3. Levy, D., Amor y sexo con robots. 2007, Londres: HarperCollins.
4. Kahn, J., Peter H., et al., ¿ Qué es humano?: Hacia puntos de referencia psicológicos en el campo de la humanidad–interacción de robots InteractionStudies, 2007.8(3):p.363-390.
5. Lovgren, S., Robot CodeofEthicstoPreventAndroidAbuse,ProtectHumans, en NationalGeographicNews. 2007.
6. Brown, A., Tomournaroboticdoghisbetrulyhuman en Guardian. 2015: Mánchester.
7. Collins, EC, A. Millings, and P.TJ Adjunto a la tecnología de asistencia: una nueva conceptualización. en Tecnología Asistiva: De la Investigación a la Práctica: AAATE2013. 2013.
8. Metzinger, T., El túnel del ego: la ciencia de la mente y el mito del yo. 2009, Nueva York: BasicBooks.
9. Tononi, G., La conciencia como información integrada: un manifiesto provisional. TheBiologicalBulletin,2008.215(3):p.216-242.
10. Qiu, J., Explorando islas de conciencia en el cerebro dañado. TheLancetNeurology.6(11):p.946-947.
11. Seth, AK, Por qué el dolor de pescado no puede ni debe descartarse Animal Sentience, 2016.2016.020.
12. Dennett, D., Los requisitos prácticos para hacer un robot consciente. Philosophical TransactionsoftheRoyalSocietyofLondonA, 1994.349:p.133-146.

13 Bryson,JJ Crude,Cheesy,Second-RateConsciousness. en VienaConferenceonConsciousness. 2008.
14 Bostrom, N., Superinteligencia: Caminos, Peligros, Estrategias. 2014, Oxford: Oxford University Press.
15. Heylighen,F., TheGlobalBrainasaNewUtopia, en Zukunftsfiguren, R.MareschandF.Rötzer,Editors.2002,Suhrkamp:Frankfurt.
dieciséis. Instituto Futuro de la Vida. Una carta abierta: prioridades de investigación para una inteligencia artificial robusta y beneficiosa. 2015; disponible en: http://futureoflife.org/ai-open-letter/.
17 Kang, M., Sublime DreamsofLivingMachines: TheAutomatonintheEuropeanImagination. 2011, Cambridge, MA: Harvard University Press.
18 Szollosy, M., Freud, Frankenstein y nuestro miedo a los robots: proyección en nuestra percepción cultural de la tecnología. AI&SOCIEDAD,2016:p.1-7.
19 Asociación de Científicos de Manhattan. Declaración Preliminar. 1945; disponible en: https://www.gilderlehrman.org/history-by-era/postwar-politics-and-origins-cold war/resources/physicists-predict-nuclear-arms-race-.

AISBTaller sobre los Principios de Robótica de la EPSRC

¿Defender un (ismo) humano obsoleto?

Michael Szollosy, Sheffield Robótica

Introducción

El taller de la EPSRC de 2010 para diseñar un conjunto de principios para el desarrollo responsable y la investigación de los robots fue un proyecto importante, ambicioso y muy bien intencionado. contexto y también destinado a una audiencia general.

A partir de la revisión de los Principios ESPRC, queda claro que pueden y deben cumplir una función vital en proteger a los seres humanos de investigaciones irresponsables o simplemente irreflexivas sobre tecnologías que posiblemente podrían tener consecuencias muy reales y muy negativas para la humanidad, a nivel personal, social o incluso a nivel de toda la especie.

Sin embargo, también está claro que lo que protegen los Principios del ESPRC es un ser humano muy específico, o al menos, una concepción muy específica de lo que constituye un ser humano. envejecidoeneltallerlohizo.

(Sin embargo, mi opinión es que debería haber habido un preámbulo que estableciera esas suposiciones). Las pautas del EPSRC están destinadas a proteger a los seres humanos y garantizar que la investigación en robótica se lleve a cabo para el 'beneficio máximo de todos sus ciudadanos', aunque cómo se verían exactamente esos 'ciudadanos' es una pregunta que se deja sin respuesta y significa que, por loables que sean sus intenciones, estos Los Principios ya son en gran medida un documento arraigado en un contexto histórico y cultural particular, y esto hace que sea poco probable que estos Principios, en su forma actual, perduren en el mediano o largo plazo.

Los Principios ESPRC hacen ciertas suposiciones, muy específicas, pero completamente tácitas sobre lo que constituye un 'ser humano'. Y los Principios se adaptan muy bien a los seres humanos. tiene derecho a diseñar, construir y comprar robots, y debe mantener la responsabilidad legal completa por ellos.

Los Principios, tal como están articulados actualmente, probablemente serán suficientes a corto plazo, tal vez siete a mediano plazo, para abordar la mayoría de los problemas con las nuevas tecnologías que surgen de la robótica y los laboratorios informáticos en todo el Reino Unido.

criaturas transitorias, son una invención relativamente nueva, un ser, además, que no es actualmente ni ha sido nunca internamente consistente y unitario, y será constantemente rehecho y transformado por una gran cantidad de nuevas tecnologías. Estas tecnologías transformadoras incluyen las mejoras biológicas y mecánicas defendidas por posthumanistas, transhumanistas y otros

pero también, más simplemente, nuevos en cuanto a nosotros mismos, la sociedad y nuevas formas de pensar acerca de nuestro lugar en el mundo, cambios que pueden ser provocados no solo por robots, genetistas, informáticos, sino también por cambios en nuestra vida económica, política y social en general.
El ser humano –o los seres humanos– a los que se atiende en los Principios de la ESPRC no será la primera, ni la última, articulación de lo que significa ser humano.

¿Qué humano?

En el corazón (implícito) de los principios de la ESPRC hay un ser humano particular definido a lo largo de los últimos siglos por lo que se conoce como humanismo. Este ser humano es un agente por derecho propio, un ser que es independiente y no debe ser gobernado por otras fuerzas, metafísicas o sobrenaturales.
Este ser humano se encuentra en el centro de los sistemas legales, éticos, económicos y políticos con base en Europa; sin embargo, es vital recordar que 1. este ser humano es todavía un invento relativamente nuevo y que 2. a lo largo de su vida, nunca ha habido una única versión de este ser humano, como les ha gustado imaginar a los defensores humanistas.

Hay poco consenso en cuanto al nacimiento del ser humano: algunos afirmarían que el Renacimiento, otros dirían que la Ilustración, y otros todavía dirían que el tema humanista se volvió central en la forma en que pensamos acerca de nosotros mismos solo en el siglo XIX. Es cierto que esta conceptualización del ser humano es una invención, no dada; el humano humanista no es 'natural', ni siquiera una interpretación 'correcta' de nuestra naturaleza humana.

Y cuando contextualizamos el humanismo de esta manera, y consideramos su lado como históricamente, notando cómo las diferentes ideas de lo que significa ser humano (o incluso 'humanista') han cambiado radicalmente a lo largo de los siglos, se vuelve evidente que no estamos hablando solo de un ser humano, o una idea de lo que significa ser humano. hablar no es un solo ser humano, sino muchos seres humanos, no una naturaleza humana única, inalienable, evidente por sí misma, sino humanos que cambian las concepciones de sí mismos en contextos particulares. La comprensión que los humanos tienen de sí mismos es siempre contingente y contextual . Cada persona puede ser un miembro del público, un ciudadano, un especialista en diferentes contextos, un consumidor, un productor; podemos ser delincuentes, pacientes, clientes, contribuyentes, partes interesadas, estudiantes, trabajadores o administradores de todas estas cosas a la vez, o ninguna de ellas, según los contextos. ora'man'ora' la mujer es muy diferente hoy de lo que era hace doscientos, cincuenta o incluso diez años. Podemos estar sujetas a discursos siempre cambiantes sobre derecho, medicina, educación, política, economía, filosofía, industria, los medios de comunicación y una gran cantidad de otros sistemas, lenguajes e instituciones que buscan definirnos y entendernos de maneras ligeramente diferentes o radicalmente divergentes.

Los supuestos que subyacen a nuestra noción de una concepción solitaria de lo que significa ser 'humano' son insostenibles bajo el intenso escrutinio de las nuevas formas de pensar sobre uno mismo, y también porque las nuevas tecnologías nos obligan a pensar sobre nosotros mismos de manera diferente.

palos para ayudar en la caza o la lanzadera voladora transformó la forma en que se hizo la tela en la Revolución Industrial. Los nuevos desarrollos en robótica exacerbarán estos procesos.
se integran más intrincadamente con sus herramientas a medida que los palos se convierten en prótesis y el trabajo humano se reemplaza por completo con máquinas automatizadas. Y estos desarrollos crearán nuevos seres humanos y nuevas formas de pensar sobre nosotros mismos.

Sin embargo, la tecnología no siempre se manifiesta en entidades físicas; los avances tecnológicos no siempre tienen la forma de nuevas herramientas o máquinas: la invención de las leyes y un sistema legal fueron nuevas tecnologías que tuvieron un tremendo impacto en la forma en que construimos nuestro ser humano, social, de la misma manera que la invención del método científico, las nuevas relaciones industriales o Facebook han cambiado la forma en que nos concebimos y presentamos esa idea de nosotros mismos. ,al mundo. Nuestras tecnologías del siglo XXI, incluidos los robots y la IA más avanzados, pero también los sistemas legales cambiantes, los órganos políticos y los sistemas de vida forética, serán desarrollos adicionales que transformarán radicalmente, con el tiempo, cómo nos vemos a nosotros mismos y cómo concebimos la idea misma de lo que significa ser humano.

¿De qué maneras implícitas y tácitas necesitamos examinar específicamente, hacer que el Principio de ESPRC se entregue a estas suposiciones humanistas sobre la naturaleza de los seres humanos? Los Principios no abordan la cuestión de a qué ser humano apuntaban. s –las constituciones, las cartas, los tratados o las declaraciones de principios– deben declarar explícitamente desde el principio aquellas verdades que se consideran evidentes, la base de lo que se sigue.

En ausencia de un sujeto claramente definido, los Principios ofrecen la concepción humanista habitual y familiar del ser humano: el ser humano estático y homogéneo que muy pronto se hará obsoleta, si no está ya preparada, por los propios avances tecnológicos que pretende controlar. Existe una concepción demasiado simplista de la relación entre los seres humanos y sus herramientas: una relación unidireccional en la que las herramientas son siempre los sirvientes de sus amos humanos, y siempre bajo el control de un agente humano independiente. Tal relación entre sujeto y objeto, agente activo y artículo pasivo siempre habría sido ingenua. Debemos considerar las formas en que nuestras herramientas transforman a los humanos; pero cómo nuestras nuevas tecnologías exigen una reorganización fundamental de toda nuestra forma de vida, e insistieron en cómo se concebía toda nuestra estructura social.

Decir que nuestra relación con nuestras herramientas no es una simple relación amo-sirviente, unidireccional, no es decir que nuestras herramientas son nuestros amos. Sin embargo, debemos reconocer que los seres humanos son tanto el producto de nuestra forma de hacer las cosas como lo que hacemos, y cómo lo hacemos, lo deciden los seres humanos. (Realmente no hay nada controvertido en decir esto; es algo que Marx reconoció hace más de ciento cincuenta años, al explicar cómo el modo de producción de una sociedad definió sus relaciones sociales, y cómo los seres humanos individuales eran a su vez definidos por esas relaciones sociales). y las líneas entre 'biológico' y 'máquina' son aún más borrosas.

Los Principios, por lo tanto, a pesar de las nobles intenciones, intentan dar el dominio de la tecnología futura y en evolución a un ser humano obsoleto.

La concepción del ser humano que se ofrece en los Principios también comparte con el humanismo la ilusión de ofrecer un sujeto único y homogéneo, cuando en realidad ese sujeto es una compilación de múltiples seres, a menudo contradictorios. mundo , o incluso diferentes culturas dentro de la misma comunidad.

Es poco probable que los avances en la robótica y la IA beneficien a todas las comunidades y todas las naciones por igual, especialmente a corto y mediano plazo, y los principios para el desarrollo de la robótica deberían reconocerlo.

Tampoco está claro a qué individuos se hace referencia en el documento; los seres humanos se denominan "el público", de nuevo, como si se tratara de un cuerpo único y homogéneo.

- Los Principios se refieren a los 'ciudadanos', un sujeto de un determinado organismo político (generalmente nacional), aunque no está claro si alguien puede seguir afirmando ser un 'ciudadano' de un estado-nación discreto e independiente, independiente de otras influencias. ¿Serán los beneficios y la responsabilidad de los robots solo para los ciudadanos de un estado nacional o de un organismo político en particular? (Quizás esto sea tan complicado si una mayor automatización lleva a la implementación de una Renta Básica Universal en un estado nacional específico, pero no en otros lugares). apons 'por razones de seguridad nacional'.

- Los Principios insisten en que solo los humanos son "agentes legales responsables". Esto puede no ser controvertido en la actualidad, y uno debe entrar en los reinos de la ciencia ficción para imaginar cuándo podríamos tener robots inteligentes y conscientes e IA que serían iguales a los seres humanos a los ojos de la ley, pero esta declaración ignora el agua ya enturbiada por los sistemas autónomos, como los automóviles que se conducen solos, y los desafíos que plantean. a nuestro sistema legal humanista. Además, podemos preguntarnos cómo los humanos mejorados tecnológicamente (por ejemplo, los cyborgs) pueden ser considerados por ley como agentes igualmente (¿menos? ¿más?) responsables.
- Los Principios insisten en consideraciones de privacidad, aunque ya podemos ver que para mucha gente los límites de 'yo' y 'público' son borrosos, y la noción de privacidad ha sido radicalmente alterado en tan corto espacio de tiempo. Las redes sociales, la promesa de las 'casas inteligentes' y los problemas de seguridad han significado que, culturalmente, tenemos una idea muy diferente de lo que significa 'privacidad' y cómo es importante para nosotros.

- Los Principios hicieron una clara distinción entre los que diseñan robots, los que venden robots y los consumidores y usuarios. Los Principios simplemente aceptan que los intereses de estos grupos pueden competir. te (cf., por ejemplo, Paul Mason 2015). Ya es evidente que a medida que se desarrollan la robótica y la IA, estas relaciones sociales que alguna vez fueron aparentemente estables se verán sometidas a una tensión cada vez mayor y probablemente se transformarán en algo más apropiado para las nuevas posibilidades de producir cosas y formas más eficientes de organizar la sociedad (una especie de post-capitalismo, como algunos quisieran). .Como ya se borran los límites entre productores y vendedores por un lado y consumidores y usuarios por otro

> otros (por ejemplo, Uber, datos de colaboración colectiva, Google), estas categorías ya necesitan ser mucho más flexibles de lo que imaginan en un humanismo directo y simplista.

También vale la pena señalar que los principios ESPRC de la robótica son, como era de esperar, tal vez, en gran medida una invención cristiana europea. Los robots se consideran máquinas y, por lo tanto, meros objetos . esa propiedad intangible, metafísica, única para la vida o, en la mayoría de las articulaciones, única específicamente para los humanos. (Esta idea de carecer de algo vitalmente humano se encuentra en la idea misma del robot, cuando la palabra se introdujo por primera vez en el mundo en la obra de teatro RUR de Karl Capek en 1921). s, incluidos estos principios de ESPRC.

A modo de contraste, vale la pena señalar, como muchos han hecho (por ejemplo, Metzler y Lewis 2008; Lee, Sung, Šabanović, Han 2012), cuán diferente se perciben los robots y la IA en diferentes contextos culturales. religiones 'animistas', es decir, creen que todas las cosas, incluidos los objetos inanimados, contienen la naturaleza de kami, o espíritu . Es poco probable que estas influencias, tan profundamente arraigadas, se transformen con demasiada facilidad mediante la introducción de nuevas tecnologías e ideas, pero subrayan que los principios de ESPRC se encuentran en gran medida en un contexto cultural e histórico muy específico, y cómo debemos estar preparados y dispuestos a imaginar otras ideas y relaciones no solo en el futuro, sino ahora mismo, si queremos construir un consenso internacional sobre los principios del robo. óticas

¿Nuevos humanos?

El hecho de que los Principios no tengan la intención de ser leyes estrictas, sino más bien de informar el debate y para futuras referencias, demuestra la visión de futuro de los delegados al taller, pero los Principios deben terminar para permitir el movimiento más allá de la noción estrictamente concebida de 'lo humano' que subyace a ellos en su estado actual. En la actualidad, el tema humanista en el corazón de estos Principios es un límite estricto sobre lo que se puede concebir, porque esta idea de 'lo humano' define todas las relaciones que en él se imaginan. Es necesario, en cambio, imaginar un ser humano diferente, más plural y flexible en la base.

Los pensadores pueden regatear (ya menudo, con náuseas) acerca de cuándo se desmoronó el consenso que apoyaba al sujeto humanista, pero está claro que en algún momento después de la Segunda Guerra Mundial, con la pérdida de la metanarrativa y una nueva y radical hermenéutica de la sospecha (que algunos han llegado a entender como 'posmodernismo'), el sujeto humanista estable, tal como se entendía una vez, no tardó mucho en este mundo. está claro que algo tiene que venir a continuación, ya que no podemos proceder a construir ningún tipo de marco o modelo sin alguna noción de lo que significa ser humano, hay mucho menos acuerdo. humanos

como consumidores, humanos como productores, humanos como diseñadores, como sujetos legales, como ciudadanos y sujetos de diversas entidades políticas… Cualquiera de estos seres humanos puede vencer en cualquier momento.

Podemos tratar de recrear algunos principios para la robótica basados en un sujeto humano que viene después del humanismo. Podemos querer llamar a este ser humano el posthumano, o simplemente posthumano. Sin embargo, estos términos son complicados y se refieren a una variedad vertiginosa de diferentes ideas e ideologías (incluso más de las contenidas bajo el término general 'humanismo' que lo precedió). Sin proporcionar un resumen completo de las diferentes ideas a las que se puede hacer referencia bajo estas etiquetas, deseo dar aquí alguna idea sobre el tipo de ideas que creo que serán necesarias y útiles para avanzar con principios futuros para la innovación tecnológica. novación

Posthumanismo puede significar simplemente, filosóficamente, culturalmente, lo que viene después del humanismo; este posthumanismo, a veces un antihumanismo, refuta el tipo de suposiciones estables y singulares sobre la naturaleza humana y humana que son establecidas por el humanismo. Yendo un poco más allá, el posthumanismo acepta la contingencia y los contextos de las concepciones de lo humano y reemplaza una naturaleza humana estática con algo más dinámico y pluralista.

El posthumanismo, o tal vez más exactamente, los posthumanismos, no son teleológicos, no toman como punto de partida que el ser humano al que hemos llegado después de millones de años de evolución y miles de años de filosofía –nosotros– es el ser humano, un producto final, acabado, pulido, que ahora permanecerá para siempre inmutable e inmutable. Así que una gran fortaleza del posthumanismo, como se entiende y se expresa aquí, es que existe una flexibilidad inherente para adaptarse a tales cambios, y es importante que cualquier empresa ambiciosa, como establecer un conjunto de principios para definir nuestras relaciones presentes y futuras con una tecnología en constante cambio, tener una flexibilidad similar incorporada.

Algunos, que son particularmente optimistas acerca de la próxima aparición de la IA consciente, y que podrían llamarse a sí mismos transhumanistas, podrían considerar los principios ESPRC como ingenuamente antropocéntricos, que no dan cuenta de la aparición como robots y IA como agentes conscientes por derecho propio que merecen (quizás la misma) consideración junto con los humanos en la creación de cualquier principio ético. Tal argumento compartiría con lo que yo Estoy avanzando aquí con la creencia de que los principios de la ESPRC ya están algo anticuados y demasiado estrechos en su concepción de lo que constituye 'lo humano', aunque soy mucho menos optimista acerca de la inminencia de la IA inteligente, y no comparto la certeza transhumanista general de que los humanos son radicalmente transformados por la tecnología (por ejemplo, los seres huma son casi inmortales) están igualmente muy cerca. Incluso si no inventamos nuevos robots y no hacemos nuevos avances en inteligencia artificial, lo cual es muy poco probable, es casi seguro.

que los seres humanos continuaremos inventando los otros sistemas e instituciones que definen quiénes somos, transformando así a los seres humanos y necesitando un conjunto de principios nuevo y más flexible para definir nuestra relación con los robots y la IA.

Los redactores de los Principios tienen la intención de que sea un 'documento vivo', no leyes 'duras y rápidas', sino la base para futuros debates y referencias, que es exactamente lo que debe ser.

nuestras nuevas tecnologías cuando toma como punto de partida un sujeto humano tan rígido y ya obsoleto en su corazón.

Es interesante ver que el preámbulo de los Principios menciona la ubicuidad de Asimov y sus Tres Leyes. Debido a que aunque las Leyes de Asimov se desestimen, correctamente, por ser inadecuadas, porque son ficticias y, por lo tanto, no abordan la 'vida real' y no pueden usarse en la práctica, sin embargo, hay algo en los escritos de Asimov que la ESPRC podría haber tomado como inspiración: la capacidad de imaginar diferentes mundos, poblados por diferentes tipos de seres humanos. Los seres humanos siempre están pasando por procesos de reinvención, pero con los avances en robótica e IA que probablemente están a la vuelta de la esquina, podríamos especular que estamos al borde de una transformación aún más radical en cómo nos vemos a nosotros mismos y cómo nos relacionamos con nuestras tecnologías. Por lo tanto, es absolutamente vital que busquemos crear principios para la robótica que será capaz de adaptarse a estas relaciones cambiantes, y a la vez capaz y poner límites en varias direcciones de desarrollo. Tendremos que pensar con imaginación sobre el tipo de robots que crearemos , pero también sobre el tipo de personas en las que nos convertiremos, y si buscamos elaborar principios para el beneficio de cuatro sociedades, necesitamos tener una mejor comprensión de cómo se verán esas sociedades y los seres humanos que las pueblan.

Referencias

Lee, Sung, Šabanović, Han. 2012. Diseño cultural de los microbots domésticos: un estudio de las expectativas de los usuarios en Corea y los Estados Unidos.

MasonP.2015. Postcapitalismo: una guía para nuestro futuro. Londres: Allen Lane.

Metzler y Lewis.2008. Puntos de vista éticos, puntos de vista religiosos y aceptación de aplicaciones robóticas: un piloto estudio.AAAI.15–22.

Regulación de pueblos robots: Reflexiones sobre los principios de la robótica desde el nuevo lado otro de la ley

Aurora Voiculescu

Centro de Derecho y Teoría, Universidad de Westminster

"Ellos me preguntaron dónde elegiría correr, cuál favorecía? Altas? O Bajas?
Donde robot ratones y hombres, yo dije, correr alrededor en pueblos robots.
¿Pero es eso sabio? Por el estaño un tonto y el hierro no tiene pensamiento!
La computadora y los ratones pueden encontrarme hechos y enseñarme lo que no soy.
Pero robot todo inhumano es, todo pecado con cog y mesh.
No si nosotros enseñamos las buenas cosas en, para que pueda enseñar nuestra carne
[...]
Como el hombre mismo una mezcla es, alborotador paradoja,
Así debemos enseñar nuestras máquinas locas: levántese levántese levántese sus calcetines!
Come run with me, wild children/men, half dires and dooms, half clowns.
Pace robot ratones, race robot men, winQlose in robot towns."

Roy Bradbury[1]

La iniciativa Principios de Robótica se deriva en gran parte de un reflejo de la extensión ¡en que los robots ya afectan nuestras vidas y en ¡es esperaba eso ¡ellos ¡voluntad ¡afectar ¡él ¡en ¡el 'robot pueblos' ¡de ¡el ¡relativamente ¡cerca futuro. Si el inicial ¡regulación ¡relacionado ¡a este transformador ¡tecnología ¡tomará la forma de principios directivos blandos, de instrumentos jurídicos nacionales duros ¡o ¡incluso ¡de ¡complejo ¡internacional tratados ¡es ¡a ¡desafiante todavía, ¡en ¡este punto, ¡a ¡secundario asunto. ¡El ¡primario ¡pregunta ¡es bastante a (legalE) ¡normativo pregunta, apuntado ¡en delineando ¡claro ¡límites ¡de ¡el humano/ robot coexistencia; ¡direccionamiento ¡el ¡normativo ¡dinámica ¡de ¡causalidad ¡y responsabilidad; ¡intentando ¡a ¡identifica el lugar geométrico o lugar de mensa y acto en procesos y, ¡atrevimiento ¡nosotros decir, relaciones que pueden bien ¡probar ¡a ser más y más complejo con ¡el avances de la ciencia y la tecnología.[2]

¡Derramando ¡de ¡este ¡necesidad ¡para ¡normativo ¡introspección (¡en nuestra psique social mucho más que cualquier otra cosa), este papel es ¡una invitación a la reflexión ¡en ¡el

[1] ¡Rayo bradbury, Where Robot Mice and Robot Men Run Round in Robot Towns: New Poems, Tanto Light and Dark (Nueva York: Random House Inc, 1977).

[2] Aurora ¡Voiculescu, "¡Humano ¡Derechos ¡Más allá de ¡el Humano: ¡Hermenéutica ¡y Normatividad en la era de lo desconocido", (próximamente).

¡propuesto! ¡Principios! ¡de! ¡Robótica! (que cubre! 5! principios! y! 7! HighELEvel! Mensajes)! ¡próximo! ¡afuera! ¡de! ¡el! multidisciplinar! ¡expertoInformado! EPSRC! ¡y! ¡AHRC! ¡Robótica! ¡Retiro! en! 2010.! ¡El! ¡complejidad! de!problemas!a! la portada es! ¡semejante! ¡eso! ¡estos! reflexiones! ¡poder! ¡solo! ¡apuntar! ¡a! ¡comprometer! ¡con! ¡qué! ¡es! propuesto,! ¡con el! ¡texto! ¡Ofrecido! ¡para! reflexión,! levantando! ¡afuera! ¡alguno! ¡de! ¡el! ¡posible! significados! ¡o! interpretaciones!de!tales!textos.!Dicho!análisis!se!propone!como!esencial!para! preparar!el! terreno para! ¡más! discusiones,! ¡y! finalmente,!por! embarcarse! ¡en! ¡cualquier! ¡eventual! procesos!regulatorios.!

Principios!en!búsqueda!de!una!definición!

Reflexionar sobre los principios existentes sí invita, en primer lugar, a reflexionar sobre qué es un robot y si la definición de ese robot !debería!conformarse!con,3 y!por lo tanto! el!tipo!de! entidades!que!se!debería!apuntar!a!regular,!debería!ser!una!respuesta!a!nuestra! estadoEdeElaTierra! en! ¡tecnología! o! un! ¡reflexión! ¡de! ¡nuestro! stateEofEtheE(tecnología!en)! arte.!En!otra! palabra,!en!que!punto!del!espectro!entre!ciencia!y!ciencia! ¡ficción! ¡debería! ¡nosotros! ¡lugar! ¡nosotros mismos! ¡cuando! ¡diseño! normas! ¡y! evaluando! ¡su! ¿Eficacia? ¡Cuán lejos en el futuro se debe mirar, cuando el futuro para el cual! nosotros!regulamos!es!así! ¡lejos! que!nosotros! solo!podemos!especular!como! a!su!existencia,!mientras!en! ¡¿Al mismo tiempo! estamos! galopando!hacia!ese!mismo!futuro!a!una!velocidad!

Ley/reglamento,! ¡duro! ¡o! suave,! requiere definiciones. ¡El! principios! ¡aquí! ¡bajo! ¡discusión! ¡hacer! ¡no! ¡inmediatamente! ¡enviar! ¡a! uno.! ¡A! ¡robot! ¡es! definido! ¡por! ¡alguno! ¡como! '¡a! ¡máquina! ¡capaz! ¡de! ¡que lleva! ¡afuera! ¡a! ¡complejo! ¡serie! ¡de! ¡comportamiento! automáticamente'! o,! ¡diferentemente! matizado,! 'un! mecánico! ¡o! agente!artificial!virtual,! generalmente!un!electroE mecánico! ¡máquina! ¡eso! ¡es! ¡guiado! ¡por! ¡a! ¡computadora! ¡programa! ¡o! ¡electrónico! circuitos'.4!Varios,! ¡más! ¡o! ¡menos! ¡factible! distinciones! ¡son! ¡también! ¡poner! adelante,! ¡más! ¡notablemente! ¡entre! ¡industrial! ¡y! ¡servicio! máquinas,! ¡entre! ¡altamente! máquinas!autónomas!y!programas!cognitivos!de!ordenador,!entre! y!entidades! cognitivas!desencarnadas,!etc.!La!NASA!mismo!utiliza!una!más!mundana!y! ¡impreciso! idioma,! ¡muy! inútil! ¡para! ¡el! regulador,! definitorio! ¡robots! ¡como! “máquinas!que!pueden!utilizarse! para!realizar!trabajos”.!Algunos!robots,!la!formulación!de!la!NASA!continúa! ¡a! agregar,! '¡poder! ¡hacer! ¡trabajar! ¡por! ellos mismos.! ¡Otro! ¡robots! ¡debe! ¡siempre! ¡tener! ¡a! ¡persona!

3!A! ¡definición! ¡nunca! ¡ser! valorEneutral,! ¡siempre! estableciendo! ¡el! 'en'Es! ¡y! ¡el! ¡Fuera! ¡siguiente! ¡a! ¡más! ¡o! ¡menos! ¡declarado! valorEladen! camino.! (¡Mira! ¡Alan! ¡Norrie! ¡Adentro! Voiculescu!2000)

4 ¡MerriamEWebster! Diccionario,! "¡Definición! ¡de! 'Robot'", accedido! ¡Febrero! 10,! 2016,!http:// www.merriamEwebster.com/dictionary/robot!entry:!robot.

¡narración! ¡a ellos! ¡qué! ¡a! do'.5!Tal! ¡a! ¡variedad! ¡de! formulaciones! ¡crear! ¡a! ¡regulador! desconcertante!y!hará!cualquier!afirmación!normativa!difícil!de!seguir!y/o!fácil!de!aplicar! escapar! cumpliendo!con.!!

¡Mientras! de acuerdo! ¡eso! ¡allá! ¡es! ¡No! ¡acordado! ¡definición! per& se, 6!algunos! ¡poner! ¡adelante! ¡a! número!de!características!que!tendría!un!robot!características!que,!de!un!regulador! (y!no!solo!)! perspectiva,! ¡son! ¡ellos mismos! ¡en! ¡necesidad! ¡de! definiciones:! sintiendo el! ¡alrededores! (¡que tiene! 'conciencia'! incorporada! de! su! entorno);! movimienot,! ¡si! rodando,! caminando,! empujando,! o!tal vez!incluso!solo! transmisión de datos;!energía,! ¡ser capaz! para! potenciarse!a sí mismo!de!maneras! eso!dependerá!de!qué! ¡el propósito de! ¡el! ¡un robot!es;! inteligencia:!que!está!provista!de!'inteligencia'!por!su!programador,!que!tiene!la! ¡capacidad! ¡a! ¡evaluar! alrededores,! circunstancias,! ¡complejo! información.! [(E! ¡cuanto más! una! máquina! es! capaz! de! interactuar! de manera independiente! con! un! mundo! dinámico! !entre!otros,! precisamente!hacia esto)]!

Entonces,! ¡a! ¡robot! ¡es! definido! ¡más! ¡específicamente! ¡como! ¡a! sistema,! ¡a! ¡máquina! ¡eso contiene! sensores,!sistemas!de!control,!manipuladores,!fuentes!de!energía!y!software!todo! funcionando! ¡juntos! ¡a! ¡llevar a cabo! ¡a! tarea".! ¡De acuerdo a! ¡a! ¡semejante! ¡a! perspectiva,! "[diseño,! edificio,! ¡programación! ¡y! ¡pruebas! ¡a! ¡robot! ¡es! ¡a! ¡combinación! ¡de! física,! ¡mecánico! ingeniería,! ¡eléctrico! ingeniería,! ¡estructural! ingeniería,! matemáticas! y! informática.! ¡En! ¡alguno! ¡casos! biología,!medicina,! química! podría! ¡también! estar! involucrado".!Si! ¡el! estudiante!en! robótica! mayo! ¡activamente! ¡comprometerse con! ¡todo! ¡estos! disciplinas! "¡en! ¡a! ¡profundamente! problemaEposed! resolución de problemas! ambiente",7!algunos! ¡podría! ¡correctamente! decir,! ¡eso! regulando! ¡robots! ¡y! '¡robot! pueblos'! requiere! ¡a! ¡similarmente! Compromiso!interdisciplinario!complejo!con!la!mayoría!si! ¡no todo! ¡de! ¡estos! campos.#!Para! ¡el! ¡normativo! ¡discurso! (¡ya sea de principios duros, reglamentarios o blandos),! ¡el! ¡hecho! que muchas de estas definiciones tengan varios puntos en común no es su
¡A! ¡definición! ¡eso! ¡es! ¡suficientemente! preciso,! ¡todavía! ¡dinámica! ¡suficiente! ¡a! ¡captura! ¡el! ¡esencia! ¡de! ¡el! socioEtecnológico! fenómenos!es! ¡por lo tanto! ¡necesario! ¡para! ¡apertura! los!principios!de!robótica! ¡a! mayor!desarrollo!y!problematización.!Esta!necesidad! se relaciona con perspectivas como la de Andra Keay, quien habla de los robots como un "…!un& entorno ;!demasiado!grande!para!que!nosotros! ítem".!Mientras!que!inevitablemente!están! vinculados!a!el! progresa! ¡de! tecnología!E! "[qué]! nosotros! ¡llamar! ¡a! ¡robot! hoy!es!más! sofisticado

---

5 !NASA,! "¡Qué! ¡Es! ¿Robótica?,"! NASA& sabe,! ¡Puede! 18,! 2015,! http://www.nasa.gov/ audience/forstudents/kE4/stories/nasaE know/what_is_robotics_k4.html.!

6! H.! ¡James! Wilson,! "¡Qué! ¡Es! ¡a! Robot,! ¿De todos modos?,"!Harvard&Business&Review,! ¡Abril! 15,!2015,!https://hbr.org/2015/04/whatEisEaErobotEanyway.
7 Ibíd.

que!lo!que!llamábamos!un!robot!en!los!años!80”!dice!Keay!E!también!es!cierto!que!es!más! que!eso.!“¡Siempre! ha!sido!un!problema&de&identidad”!dice!Keay.8!!

¡Él! ¡debería! ¡ser! dijo,!sin embargo,! que identidades!y! clasificaciones! ¡tener siempre! ¡estado! ¡problemático! ¡y! problematizado! ¡cuando! ¡parte! ¡de! ¡regulador! iniciativas,! ¡si! ¡estos! ¡tenía! ¡a! ¡hacer! ¡con! humanos! ¡o! ¡No Ehumanos! similar.! Ley,! ¡en! particular,! ¡siempre! termina!convirtiendo! cualquier!identidad!en!una!ficción!jurídica!que!muchas!muchas!tienes!poco!que!hacer! con!cualquier!otra! dimensión!física!o!científica!de!esa!entidad. 9!Al!mismo!tiempo,! derecho!E!en!su!sentido!más!amplio!de! imperativos!normativos!socialmente!respaldados! E!ha!siempre! ¡prosperó!en!las!definiciones.!La!ausencia! de!una!'definición!de!trabajo!de!un! ¡aparece el robot! por tanto!ambos!como!un!testigo! ¡a! los!desafíos!de! determinar! tecnología!en!su! prisa,!y!como!reflejo!de!una!posible!debilidad!a!ser! documento.!!

Por último, pero no menos importante, ¡además! al!problema!de! la!ausencia!de!un!trabajo!acordado! definición!(que!sería!tan pronto!disputada!y!problematizada!por!supuesto),!ahí! ¡es también! el! reconocimiento! ¡eso! 'definiciones! son! ¡nunca! neutral'.!Esta!idea!fue! ¡avanzado! ¡alguno! décadas! ¡atrás! ¡por! ¡Larry! ¡Puede! ¡cuando! reflejando! ¡en! definitorio! ¡el! responsabilidad!de!agencia!colectiva! no!humana!(¡una!innovación!jurídica!que de otro modo no sería irrelevante). ¡como! hechos,!mientras!en! realidad!establecen!oposiciones!que!'arbitrariamente!separan!aquellas! quienes!están!incluidos!y! aquellos! que!están!excluidos! de!una!conceptualización!compartida! ¡o! práctica'. 10!Esto! ¡afirmación! ¡voluntad! ¡convertirse en! ¡más! ¡y! ¡más! ¡evidente! ¡una vez! ¡el! el espectro de opciones disponibles entre el robot y la IA se amplía.11

¡Ya sea que se anticipe con temor o entusiasmo,! el!reto de!regular! las!múltiples!dimensiones!de!las! interacciones!humano-robot!son!múltiples.!Un!número! ¡de! ¡asuntos! ¡poner! ¡adelante! ¡para! ¡reflexión! ¡en! ¡relación! ¡a! ¡el! ¡dado! ¡cinco! principios! ¡son! mencionado!brevemente!aquí:!

8!Mira! ¡entrevista! ¡con! Andra! ¡Claro! ¡fundador! ¡de! ¡Robot! ¡Plataforma de lanzamiento! ¡y! ¡Gerente! ¡Director! ¡de! ¡Silicio! ¡Valle! ¡Robótica! en Signa! Brewster,! "¡Qué! ¡Es! ¡a! ¡¿Robot?! ¡El! ¡Respuesta! 2014,!"!https://gigaom.com/2014/07/05/whatEisEaErobotEtheEanswerEisEconstantlyE ¡Constantemente! ¡Evolucionando,! ¡Julio! evolution/!(Énfasis del autor).!

9 David!Fagundes,!“Nota,!De!Que!Hablamos!Cuando!Hablamos!De!Personas:!La! Language!of!a!Legal!Fiction”,!Harvard&Law&Review 114,!no.!6!(2001):!1745–68.

10 ¡Larry! Puede,! Compartiendo y Responsabilidad,! ¡Nuevo! ¡edición! 1996! (Chicago:! Universidad! De! Chicago!Press,!1992),!171ff.

11 ¡Keneth! Grady,! "¡Artificial! Inteligencia:! ¡Ser! Asustado,! ¡Ser! Muy,! ¡Muy! ¡Asustado! (¡O! No),"! SeytLines:& Cambiando& la& Práctica& de& la& Ley,! ¡Diciembre! 31,! 2014,! http:// www.seytlines.com/2014/12/artificialEintelligenceEbeEafraidEbeEveryE veryEafraidEorEnot/.

En primer lugar, existe la necesidad de una perspectiva un poco más clara en cuanto a qué es lo que se regula: la ausencia de una definición consensuada antes mencionada, ¡la ! principios! poner! ¡adelante! para!la!discusión!revelen!el!potencial!de!confusión!en cuanto!a!la! ¡actual! agenda:! 'regulando! ¡robots! ¡en! ¡el! ¡real! mundo'! ¡tiene! ¡a! ¡doble! sentido,! ¡lleno! ¡de! peligros. Si! uno 'regula! a! los! robots!,! el! texto! trae! una! agencia! que!puede!darse!por! supuesto!en!contextos!donde!puede!ser!no!deseable!(aunque!esta!interpretación!es!claramente! contradicha!por!algunos!de!los!principios,! en particular! por!
¡Principio! No.! 2! ¡y! 5).! ¡El! ¡segundo! significado,! ¡más! ¡en! ¡línea! ¡con! ¡qué! ¡el! ¡cinco! principios! ¡ellos mismos! revelar,! ¡podría! ¡apuntar! ¡en! 'regulando! ¡el! ¡creación! ¡y! ¡usar! ¡de! robots'. La!elección!de!esta!interpretación!debería!ser!más!explícita!a lo largo!de!la! formulaciones,!evitando!confusiones!regulatorias.!

Principio!no!1:!Los

robots&son&herramientas&de&usos&múltiples.&Los&robots&no&deben&ser&diseñados&solamente&o&principalme

¡El! primera!parte de! este principio es, sin embargo, aún más desconcertante. En primer lugar, uno! ¡puede! ¡encontrar! ¡el! ¡a partir de! ¡declaración! 'robots& son& multiQuse& herramientas'! ¡como! ¡virtualmente! ¡a! restricción!que!no!tiene!un!propósito!real.!No!está!claro!por!que!un!robot! tiene!que! ser!'multiEuse'!para!estar!seguro!o,!por el contrario,!de!manera!y!de!otra manera!mortal! ¡el!robot!puede!volverse!menos!mortal!si!se!diseña!como!'multiEuse'.!Esto!se!relaciona! ¡a! ¡el! siguiente!parte!del!principio:
'los&robots&no&deben&ser&diseñados&exclusivamente&o&principalmente&para&matar& o&dañar&a&humanos'.!Para! esta!parte!del!principio,!bastaría!también!enseñar!a!los!'robots asesinos'!a!hacer! panqueques!o!calcetines!de!lana!tejidos.!Esto!es!lo!que!en! la!perspectiva! normativa!jurídica,! uno!llamaría!una!"laguna!de!cumplimiento!creativo". Para! identificar! y! ¡el! texto.!Sin embargo,! ¡el! ¡literal! ¡interpretación! ¡es! ¡uno! ¡de! ¡el! ¡primario! ¡normas! ¡de! interpretación!en!derecho,!cuando!la!interpretación!en!línea!con!el!'espíritu!de!la!regla'! ¡puede!no! ser!una!conveniente.!Las!explicaciones!dadas!a!este!principio!en!la! ¡original! ¡documento! ¡hacer! ¡no! ¡parecer! ¡a! ¡en realidad! ¡DIRECCIÓN! ¡este! ¡bastante! ¡básico! ¡acercarse! ¡a! interpretar! las!reglas!y!sus!consecuencias!en!este!contexto!particular.

¡El! ¡comentario! ¡a! ¡este! ¡principio! aparece! ¡a! ¡implicar! ¡otro! ¡potencial! trampas! ¡para! razonamiento! normativo.!En primer lugar,!como!se!menciona!con!respecto!a!las!'herramientas!de!uso!múltiple',! hay!un! esfuerzo!por!plantear!la!idea!de!que!los!robots!son!herramientas!como!cualquiera!otro.!En! Para!seguir!con! esta!lógica,!se!buscan!equivalencias!a!cualquier!costo.!Comparando!una! robot!con!un!cuchillo!o!una!pistola! utilizada!para!diferentes,!tanto!relativamente!benignos!como!delincuentes! propósitos, hace! ¡no! ¡cubrir! ¡para! ¡el! ¡inconsecuencia! ¡eso! ¡allá! ¡son! herramientas,! ¡incluido! ¡armas, para las cuales no se puede pensar en otro propósito que el prima facie uno!

¡Principio! No.! 2:! Los humanos, no los robots, son los agentes responsables. Los robots deben diseñarse y operarse en la medida de lo posible para cumplir con las leyes existentes y los derechos y libertades fundamentales, incluida la privacidad. ¡El! comentarios! ¡a! ¡este! principio!parece!añadir!más!confusión! que!claridad.!Primero!que!todo,!un!relativamente!pequeño! asunto,! ¡hay! la!presunción! que!'nadie!es! probable!de!liberadamente! ¡a! ¡construir un! robot!que!rompe! la!ley'.!Esto!pone! adelante!una!presunción! que!tiene!no! fundamento!en!el!mundo!real!de!'desviación!y!desafío',!como!revelado!por!socioElegal! estudios!entre!la!amplia!población!en!general!así!como!entre!los!cuellos!blancos.12 En segundo lugar,!y! más!importante! !¡este! comentarios! aparece! ¡a! ¡ignorar! ¡ambos! ¡el! ¡forma! '¡ley! piensa'13 como! ¡Bueno! ¡como! ¡el! ¡forma! los robots pueden no lograr los objetivos y deseos que los humanos especifican.14

Como!un!elemento!adicional!aquí,!también!debería!mencionarse!que,!en!ausencia!de! una!definición!de! trabajo!clara,!IA,!'robots!de!aprendizaje',!etc.,!todos!se!relacionan!con!estos! principios! ¡y! ¡su! parámetros.! Su!mecánica,! sin embargo,!puede! ¡Bueno! ¡Se mas! complejo!de!lo!que!el!discurso!legal/normativo!puede! manejar!en!ausencia!de!una! definición#.! ¡Robots! ¡y! ¡AI! ¡máquinas! ¡puede! ¡Bueno! ¡aprender! ¡a! ¡trato! ¡con! 'excepciones'! antes!que!el!derecho!aprenda!a!lidiar!con!las!'diferencias'.15!Igualmente,!otras! disciplinas!parecen! ¡a! ¡indicar! ¡eso! ¡números! (en! este! caso! particular,! 'programación')! ¡puede! ¡Bueno! ¡ser! ¡más! ¡que! ¡justo! eso,! ¡números! ¡ser! inalienablemente! complementado! por/asociado!

12!Algunos! ¡útil! ¡aunque! ¡perder! ejemplos! Ryan! Mateo! ¡y! ¡vatios! Wacker,! La ventaja de& Deviant&:& Cómo& Fringe& Ideas& Create& Mass& Markets! (Casa al azar,! 2010);! Kelly! Pescador,! "¡El! ¡Psicología! ¡de! Fraude:! ¡Qué! ¡Motiva! ¡Estafadores! ¡a! ¡Comprometerse! ¿Delito?"! (Social! Science! Research! Network,! March! 31,! 2015),! http://papers.ssrn.com/abstract=2596825.!

13 ¡Gunther! Teubner,! "¡Cómo! ¡el! ¡Ley! piensa:! ¡Hacia! ¡a! constructivista! Epistemología!del!derecho",!Law&and&Society&Review 23,!no.!5!(1989):!727–57.

14 !Mira! ¡para! ¡instancia! ¡el! ortogonalidad! ¡teoría! ¡en! ¡Mella! Bostrom,! "¡El! ¡Superinteligente! Voluntad:! ¡Motivación! ¡y! ¡Instrumental! ¡Racionalidad! ¡en! ¡Avanzado! ¡Artificial! Agentes",! Mentes y y& Máquinas,! 2012,! http://www.nickbostrom.com/superintelligentwill.pdf.!

15!Esto!está!construido!sobre!las!ideas!de!identidad!y!diferencia,!de!Leibniz!a!Kant...!Gilles& Deleuze&Q&Le&Point&de&Vue&(Le&Pli,&Leibniz&et&Le&Baroque)&1986&FRA&Sub&ITA,!2012 ,! http://www.youtube.com/watch?v=2ZrA_7ewQGs&feature=youtube_gdata_play

ejem.

¡con un! ¡narrativo! ¡movimienot! eso,! ¡uno! ¡podría! ¡decir! aquí,! ¡puede! ¡ser! interpretado! de manera diferente por la máquina que por el humano, pero aún así puede ser interpretado por ella.16!!!

¡Principio! No.! 3:! Los robots son productos y deben diseñarse utilizando procesos que aseguren su seguridad y protección. Estos!principios!plantean!cuestiones!de!distinciones! ¡en! ¡el! ¡legal! autodefensa! debate;! agravio! asuntos,! ¡asimilación! ¡de! responsabilidad,! etc.,! ¡asuntos! ¡eso! ¡depender! ¡muy! ¡mucho! ¡de! ¡el! ¡contexto! ¡y! ¡de! ¡el! ¡medida! ¡a! cual,! ¡como! mencionado! anteriormente,!el!regulador!estará!dispuesto!a!construir!equivalencias!entre! robots! y! ¡otro! tipos! ¡de! ¡herramientas! ¡o! ¡entre! robots! y! ¡otro! tipos! ¡de! ¡propiedad! elementos.!

¡Principio! No.! 4:!Los

robots&son&fabricados&artefactos.&No&deben&ser&diseñados&de&una&manera&engañosa&para&explotar&a&usuari

¡Principio! No.! 5:! La& persona& con& la& responsabilidad& legal& por& un& robot& debe& ser& atribuida. ¡Este!principio!aparece!suficientemente!claro!en!el!contexto!en!el!que!somos!un! lejos!de! una!IA!capaz!de!cumplir!técnicamente!los!requisitos!para!tal! responsabilidad.!En! ¡lo mismo! tiempo,! tomando!en!cuenta!los!elementos!complejos!y! ¡campos! ¡eso! ¡ingresar! ¡el! ¡maquillaje! ¡de! ¡a! robot (ver! arriba! cuando! se! discutan! cuestiones! relacionadas! con! la! definición),! los! desafíos! al! enfoque! regulatorio! ¡el! ¡punto! ¡de! ¡vista! ¡de! levantando! ¡aparte! ¡el! ¡varios! grados! ¡de! responsabilidad,! ¡cuando! ¡cosas! ¡ir! equivocado.! ¡Teniendo! ¡a! '¡registrado! guardián',! ¡portador! ¡de! responsabilidad,! ¡es! ¡sólo parte! ¡de! ¡el! solución.! ¡El! ¡responsabilidad! ¡cojinete! ¡entidad! ¡voluntad! ¡requerir! ¡más!

---

16 Marco! du! Sautoy,! Narrativa& y& Prueba:& ¿Dos& lados& de& la& misma& ecuación?& |& ¡ANTORCHA! (TORCH,! The! Oxford! Research! Centre! in! the! Humanities,! 2015),! http:// www.torch.ox.ac.uk/narrativeEandEproofEtwoEsidesEsameEequationE0.
17 ¡Mira! por!ejemplo!Cynthia!Breazeal,!“Emoción!y!Robots!humanoides!sociables,”! International&Journal&of&HumanQComputer&Studies 59,!no.!1–2! (Julio!2003):!129ff,! doi:10.1016/ S1071E5819(03)00018E1.

reflexión,! ¡como! ¡voluntad! ¡el! ¡tipo! ¡de! daño(s)! ¡eso! ¡puede! ¡ser! atribuido! ¡a! ¡semejante! entidades,! tomando!en!cuenta,!como! ¡alguno! sugerir,! ¡eso! ¡para! ¡el! ¡primero! ¡tiempo! '¡el! ¡promiscuidad! ¡de! los datos'!se!combinan!últimamente!con!'la!capacidad!de!hacer!daño!físico'.18!!

En lugar!de!conclusiones!

Cerrando estas notas completando el círculo, con otra de las imágenes poéticas de Bradbury, se puede decir que, en lo que respecta a los ratones robot, es muy probable !que!en!el!contexto! de!poderes!de! mercado!y!desregulación!de!capacidad!científica!combinadas,! ¡seguramente! ¡primero! '¡saltar! ¡apagado! ¡el! ¡acantilado! ¡y! ¡construir! ¡nuestro! ¡alas! ¡en! ¡el! ¡forma! abajo'.! ¡En! ¡semejante! circunstancias,!lo! que!puede!hacer!es!asegurarse!de!estar!relativamente!preparado!para! ¡este! ¡saltar! ¡por! ¡ya! interrogando! ¡y! problematizando! ¡nuestro! ¡relación! ¡con! ¡ciencia! ¡y! ¡tecnología! ¡y! ¡por! intensificando! ¡nuestro! ¡reflexión! ¡en! ¡viviendo! ¡en! ¡robot! pueblos.!Al abordar!'herramientas',!'productos',!'artefactos'!y!'agentes',!uno!debe!asimilar! ¡Calle! de Agustín! ¡reflexión! ¡en! ¡el! ¡intrincado! ¡conexión! ¡entre! ¡idioma! ¡y! ¡interpretación! ¡como! ¡a! ¡camino! ¡a! revelador! ¡a! Más adentro,! existencial! ¡nivel! ¡de! selfQ comprensión.!La!manera! nosotros! ¡pensar! normativamente!sobre! humanErobot!interaction19 voluntad! ¡decir! ¡como! ¡mucho! ¡acerca de! ¡el! ¡robot! ¡como! ¡acerca de! ¡el! humano.! ¡A! ¡pedir prestado! ¡de! ¡Ginabattista! ¡Vico's! 1725! Nuevo& Ciencia, 20!nosotros! ¡necesidad! ¡a! ¡oso! ¡en! ¡mente! ¡eso! ¡nuestro! ¡pensando en! robots! es! enraizado!en!un!dado! ¡cultural! contexto.! ¡Esto significa! ¡que en! reflejando! ¡acerca de! ¡el! ¡normativo! parámetros! ¡de! ¡robot! pueblos,! ¡el! ¡social! ¡científico! ¡voluntad! ¡no! ¡trato! ¡con! ¡a! ¡campo! ¡de! idealizado! ¡y! supuestamente! 'sujetoEindependiente! objetos',!pero! investigará!un!mundo!que!es,!fundamentalmente,!el!propio.!El!proceso! ¡de! regulando! ¡robots! es,! por lo tanto,! ¡a! ¡proceso! ¡de! autocomprensión,! arraigado! ¡en! ¡a! dado!el!contexto!histórico!y!la!práctica.! Una!comprensión!que!no!culmina! automáticamente en proposiciones ordenadas, normativas y similares a las leyes.

---

18 Ryan!Calo,!“Robotics!and!the!Lecciones!of!Cyberlaw,”!California&Law&Review 103! (2015):!513.

19 ¡Sé! ¡este! ¡interacción! ¡comprendido! ¡en! ¡Latour's! ¡espectro! ¡de! ¡hechos! ¡y! agencia;! ¡Bruno! Latour,! "¡Cómo! ¡a! ¡Hablar! ¡Acerca de! ¡el! ¡¿Cuerpo?! ¡El! ¡Normativo! ¡Dimensión! ¡de! ¡Ciencia! Estudios,"! Cuerpo& && Sociedad 10,! No.! 2-3! (1 de junio! de 2004):! 205–29,! doi:10.1177/1357034X04042943;!Bruno!Latour,!“Cuerpo,!Cyborgs!y!la!Política! ¡de! Encarnación,”!en! El&Cuerpo:&Darwin&College&Lectures,!ed.!Sean!Sweeney!y! Ian! Hodder,!2002,!127–41.
20!Giambattista! Vico,! ¡La& nueva& ciencia& de& Giambattista& Vico! (¡Universidad de Cornell! Pulse,!1744).

Comentario sobre responsabilidad, diseño de productos y nociones de seguridad

Paula Boddington, Departamento de Ciencias de la Computación, Universidad de Oxford.

Los comentarios aquí se relacionan principalmente con la regla 2 y la regla 3. Sugiero que los principios podrían hacerse más específicos para el contexto de la implementación de robots, y que las nociones clave como "seguridad" podrían elaborarse y tal vez extenderse, o la forma precisa en que se utiliza el término podría aclararse.

Algunos de estos puntos se ilustran al considerar en términos generales el uso de la robótica en contextos de enfermería/ancianos y atención social.

Regla 2: Los humanos, no los robots, son agentes responsables.

Si los humanos son agentes responsables, pero los robots no lo son, esto implica que dondequiera que se utilicen robots para reemplazar humanos o parte de una agencia humana, entonces las atribuciones de responsabilidad que antes se otorgaban al agente humano o las acciones humanas, se desplazan a un sistema más amplio, o tal vez, se pasan por alto.

Los robots se utilizarán dentro de un sistema de agentes y comportamientos humanos. Dichos sistemas pueden formalizarse con nociones de responsabilidad claramente expresadas, por ejemplo, dentro de un entorno hospitalario (aunque puede haber elementos de tales sistemas que no se comprendan completamente o se hayan formalizado con total adecuación); o pueden ser informales, por ejemplo, dentro de un entorno de atención domiciliario. En entornos urbanos, la investigación social encuentra que puede haber culturas y valores locales fuertes con respecto a las líneas de responsabilidad y rendición de cuentas.

Para ver un ejemplo de cómo se pueden desplazar las responsabilidades, si un robot asume algunas de las funciones de un asistente de atención médica dentro de un entorno de adjudicación, entonces las responsabilidades se pueden desplazar de una variedad de formas posibles a diferentes actores dentro del sistema de gestión de la atención médica. se ha visto como una falta de escrupulosidad en un empleado, por ejemplo, podría llegar a verse como una dificultad para comprender u operar la maquinaria. Puede haber repercusiones de gran alcance.

Rastrear y comprender tales líneas de responsabilidad y rendición de cuentas puede ser complejo. Surge la pregunta de si se trata únicamente de una tarea de quienes están a cargo del entorno donde se utilizan los robots, o si los diseñadores de los robots pueden tener alguna responsabilidad de ayudar a quienes trabajarán junto con ellos para comprender estos problemas.

La regla 2 habla de cumplir con las leyes existentes y los derechos y libertades fundamentales, incluida la privacidad. Sin embargo, además, en determinadas configuraciones, habrá protocolos y prácticas más específicas y locales que será deseable que cumplan los robots.

Por ejemplo, dentro del NHS existen estándares de atención que tienen como objetivo brindar una atención centrada en la persona y tratar a los pacientes con dignidad. Tales declaraciones de estándares son fundamentales para la prestación de una buena atención médica.

alguien como una persona con dignidad. Averiguar cómo el uso de robots impacta sobre valores tan ricos y contextualizados puede ser muy importante, pero puede ser una cuestión más difícil que simplemente cumplir con la ley y con las responsabilidades establecidas en la ley. tratando estos asuntos en el diseño y uso de robots, a menos que se suponga que esto es ampliamente entendido.

Regla 3: Los robots son productos. Deben diseñarse mediante procesos que garanticen su seguridad y protección.

Puede ser necesario articular lo que se entiende por "seguridad" en este contexto. ¿Se refiere esto simplemente a la seguridad física? ¿Y se refiere a la seguridad en términos de la operación inmediata del robot, o a los efectos del uso de los robots a lo largo del sistema en el que se utilizan?

Esta regla establece que los productos deben ser seguros. Sin embargo, aunque las reglas éticas a menudo se formulan en términos de prevención de daños, es menos que aspiracional buscar diseñar productos que sean simplemente "seguros".

Cuando los robots están reemplazando o ampliando la agencia humana, la tarea que el robot está realizando puede no ser del todo transparente. solo la interacción humana puede proporcionar. Por el contrario, los robots pueden diseñarse para almacenar algunos de los aspectos marinos de la tarea.

Los problemas de seguridad surgen porque descubrir cómo el uso de robots podría interrumpir, o incluso mejorar, ciertos aspectos posiblemente ocultos de las tareas que los robots toman el control, pueden implicar un análisis e investigación considerables. Por lo tanto, existe la posibilidad de que un trabajo tan importante no se reconozca o reconozca. señalando puntualmente que la seguridad debe considerarse en un ámbito amplio. Esto también vuelve a plantear la cuestión de las responsabilidades compartidas entre los equipos que diseñan y fabrican los robots y los que trabajarán con ellos. Puede que no parezca que se pierdan ciertos aspectos rutinarios de la atención que los humanos pueden dar a los robots, tal vez no sea una cuestión de 'seguridad', pero pueden tener un impacto en la recuperación y la salud y, por lo tanto, si es así, debe verse como una cuestión de seguridad. los bots pueden mejorar estas cuestiones.

La atención a la seguridad incluirá, por supuesto, analizar los problemas de seguridad del uso de robots dentro de un sistema más amplio. Por ejemplo, un problema común y grave en entornos hospitalarios para pacientes ancianos y vulnerables es la deshidratación.

confusión. A veces, la deshidratación empeora debido a las dificultades para alcanzar y manejar las bebidas. Supongamos que un sistema robótico pudiera diseñarse para ayudar a tales pacientes a beber. Este sistema puede funcionar con total seguridad y confiabilidad en términos de su uso inmediato, por ejemplo, nunca funciona mal de manera que le da al paciente demasiado demasiado rápido, y nunca derrama bebidas o golpea al paciente.

Sin embargo, un sistema de este tipo podría tener grandes consecuencias negativas en un contexto particular. El aumento de la hidratación podría provocar un aumento de las tasas de enuresis nocturna en algunos pacientes. tsin'bloqueo de la cama'ya que los pacientes a menudo tienen que encontrar alojamiento en instalaciones que puedan atender sus necesidades. Una mayor hidratación también puede dar lugar a más casos de pacientes que se levantan de la cama para ir al baño y, por lo tanto, tienen un aumento de las caídas. Esto también puede tener consecuencias nefastas.

Cuando la regla 3 habla de 'seguridad', ¿está claro que esto incluye o no incluye consideraciones sobre cómo podría funcionar un robot dentro de un sistema de trabajo más amplio?

Contribución de Roeland de Bruin y Madeleine de Cock Buning a la

Taller AISB sobre principios de robótica, 4 de abril de 2016, Sheffield, Reino Unido

---

1. Introducción

Han pasado cinco años desde la publicación de los principios de robótica del EPSRC desarrollados por un panel de distinguidos expertos británicos en robótica e inteligencia artificial en un retiro financiado por EPSRC/AHRC.1 Los principios, que tenían como objetivo "regular los robots en el mundo real", fueron establecido en forma de cinco "reglas" y siete “mensajes de alto nivel”. De hecho, los principios han tenido un impacto significativo en la robótica del Reino Unido. investigación, y continúan provocando un debate sustancial. Dado que actualmente la preocupación pública por la El desarrollo de tecnologías robóticas está aumentando, consideramos útil revisar los principios para considerar su pertinencia continua de acuerdo con los siguientes criterios.

Nuestras contribuciones se centran en el segundo principio:

Principio 2: Los humanos, no los robots, son los agentes responsables. Los robots deben diseñarse; operado en la medida de lo posible para cumplir con las leyes existentes y los derechos y libertades fundamentales, incluida la privacidad.

De hecho, este segundo principio de los principios de robótica del EPSRC es doble. Por un lado el El principio se ocupa de la responsabilidad, incluida la responsabilidad, por las acciones del robot, por otro lado, el El principio implica métodos de diseño de máquinas que pueden ayudar con el cumplimiento de las leyes existentes y derechos y libertades fundamentales, incluida la privacidad.

Dado que tanto la responsabilidad como el diseño forman la columna vertebral de la introducción de la tecnología robótica, en cuanto a instancia incorporada en los coches inteligentes autónomos de nuestra sociedad, probaremos este doble principio centrándose en el desarrollo actual y el despliegue de coches inteligentes autónomos. Si esto El segundo principio EPSRC se puede considerar como una prueba futura, se probará con tres criterios:

1. Validez: ¿es correcto el principio como declaraciones sobre la naturaleza de los robots, desarrolladores de robots, y la relación entre robots y personas, o es ontológicamente defectuoso, inexacto, fuera de lugar anticuado o engañoso.
2. Suficiencia/generalidad: ¿es el principio suficiente y lo suficientemente amplio para cubrir todos los cuestiones importantes que pueden surgir en la regulación de la robótica en el mundo real o son preocupaciones significativas pasadas por alto.
3. Utilidad: es el principio de uso práctico para desarrolladores, usuarios o legisladores de robots, en determinar estrategias para las mejores prácticas en robótica, o estándares o marcos legales, o son limitaron su uso por falta de especificidad o al permitir excepciones críticas.

---

[1] https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

2. Estado del arte

2.1 Estado de los Coches Inteligentes Autónomos (AICs)

Antes de poner a prueba el principio, presentaremos en breve el estado del arte de los AIC.
Actualmente, los automóviles de consumo están cada vez más equipados con tecnología que ayuda en ciertos aspectos de la conducción. Ejemplos de dicha tecnología incluyen asistencia de mantenimiento de carril, frenado de emergencia, asistente de estacionamiento y control de crucero adaptativo. En un futuro cercano, los niveles más altos de automatización del automóvil estén disponibles, lo que finalmente conducirá a la introducción de vehículos totalmente autónomos.

Además, ahora algunos automóviles ya están equipados con ciertas formas de automatización. Incluso hay prototipos disponibles que pueden conducir sin un operador humano. Google es actualmente pionero en auto tecnología de conducción de automóviles, y ha puesto a prueba en carretera un prototipo de AIC en pleno funcionamiento en el Área de la Bahía, California, a principios de 2015.2 También en la Unión Europea, los fabricantes de automóviles se concentran en el desarrollo de la tecnología AIC.3 Scania está probando "Platooning": un tren de carretera de camiones autónomos que seguían de forma autónoma a un camión controlado por humanos que se dirigía al convoy se desplegó en las carreteras holandesas.4 Volvo planeó desplegar 100 automóviles que deberían poder hacerse cargo de todos los aspectos de la conducción en Suecia para 20175 y en Alemania, una parte de la Autobahn A9 entre Múnich y Berlín está reservado para las pruebas exhaustivas de vehículos autónomos en los próximos años.6

Una definición de Coches Inteligentes Autónomos consta de tres elementos. La autonomía se relaciona con el nivel de intervención humana necesaria para la operación, que puede verse como un espectro: una menor necesidad de la intervención humana implica un mayor nivel de autonomía. La inteligencia se relaciona con las formas en que un El sistema puede percibir su entorno y es capaz de adaptar su comportamiento a entornos cambiantes. Incluye la capacidad de aprender, de procesar información compleja y de resolver problemas.7 Los automóviles son vehículos motorizados, utilizados para el transporte de mercancías y/o personas y para la prestación de servicios.

---

[2] Wikipedia, “coche sin conductor de Google”, disponible en Internet en <http://en.wikipedia.org/wiki/Google_driverless_car> (último acceso el 17 de marzo de 2015), en referencia a Matt O'Brian, “Google's 'goofy' new coche autónomo un signo de lo que vendrá”, 22-12-2014, disponible en Internet en <http://www.mercurynews.com/business/ci_27190285/googles-goofy-new-self-driving-car sign-things> (último acceso el 28 de enero de 2016).

[3] Véase, por ejemplo, <https://www.media.volvocars.com/us/en-us/media/pressreleases/145619/volvo-car-groups-first-self-driving autopilot-cars-test-on-public-roads -alrededor-> (Volvo), <http://www.pcmag.com/article2/0,2817,2387524,00.asp>, (Volkswagen) <http://www.bbc.com/news/technology- 25653253> (BMW) (último acceso el 28 de enero de 2016).

[4] Consulte <http://www.scania.nl/about-scania/media/platooning/> (consultado el 20 de marzo de 2015).

[5] Alecander Stoklosa, “Volvo tiene un automóvil autónomo de producción viable, lo pondrá en circulación para 2017”, disponible en Internet en <http://blog.caranddriver.com/volvo-has-a-production-viable -el-coche-autónomo-lo-pondrá-en-la-carretera-para-2017/>. (último acceso 20 de marzo de 2015).

[6] Stephen Edelstein, “Alemania planea un programa de prueba de automóviles autónomos en autopistas de alta velocidad”, 28 de enero de 2015, disponible en Internet en <http://www.motorauthority.com/news/1096521_germany-plans-autonomous-car-test-program- en la autopista de alta velocidad> (último acceso el 20 de marzo de 2015).

[7] Véase Madeleine de Cock Buning, Lucky Belder & Roeland W. de Bruin, Documento de trabajo: “Mapeo del marco legal para la introducción en la sociedad de los robots como sistemas autónomos inteligentes”, en la pág. 3-4, disponible en Internet en <http://www.caaai.eu/wp-content/uploads/2012/08/Mapping-L_ N-fw-for-AIS.pdf> (último acceso el 28 de enero de 2016)) y las referencias a Samir Chopra y Laurence F. White, A Legal Theory for Autonomous Intelligent Agents (Ann Arbor: University of Michigan Press 2011) en p. 10 (autonomía) y Collin R. Davies, “Un paso evolutivo en los derechos de propiedad intelectual: inteligencia artificial y propiedad intelectual”, 27 Computer Law & Security Review 2011, p. 601-619 (inteligencia); y de los mismos autores el capítulo “Mapping the Legal Framework for the Introduction into Society of Robots as Autonomous Intelligent Systems”, en Sam Muller et al (eds.), The Law of the Future and the Future of Law, serie 2012 (De Cock Buning, Belder & De Bruin 2012), pp. 195-210.

Las AIC pueden contribuir a encontrar soluciones a los desafíos a los que se enfrenta actualmente nuestra sociedad. Camino
la seguridad aumentará dramáticamente cuando el 'error humano' sea eliminado como un factor en la causalidad de
accidentes Los AIC podrían reducir significativamente los riesgos de accidentes automovilísticos, ya que el 93% de los accidentes de tránsito son
causados por fallas humanas,8 lo que provoca 1,3 millones de muertes y 50 millones de lesiones graves en todo el mundo por año.9 Además
de contribuir a la seguridad vial, los AIC pueden conducir a un uso más eficiente. de la red viaria, reducir las emisiones de CO2 y ayudar a
mejorar la movilidad de las personas con discapacidad.10 Así, la implantación de los AIC podría dar respuestas para reducir los riesgos
actualmente manifiestos fruto de la evolución tecnológica.
innovación en las últimas décadas.11

Sin embargo, no todos son optimistas sobre un futuro sin conductor. Se afirma que si bien los AIC podrían ser
beneficiosa para la seguridad vial, la introducción de vehículos autónomos generará otros riesgos. Los AIC
ser vulnerable a la piratería, por ejemplo. Asimismo, modelos de negocio y empleo en taxi y
los mercados de transporte cambiarán significativamente, mientras que los conductores pueden volverse obsoletos después de la autonomización
de la conducción.12 Además, los riesgos de accidentes podrían aumentar cuando los vehículos autónomos y no autónomos coexisten en las
mismas carreteras.13

2.2 Estado de derecho

La certeza suficiente sobre el estado legal es esencial para el crecimiento y la aceptación social de los consumidores.
tecnología. La incertidumbre provoca lo contrario. ¿Podría en ese caso la máquina ser la respuesta a la
¿máquina? A continuación, discutiremos brevemente los problemas de responsabilidad que actualmente desafían la introducción y
despliegue en la sociedad de AIC y tocar posibles soluciones de tecnología de evidencia para algunos de
estos desafíos que podrían implicar la privacidad por diseño.

Responsabilidad

Normativa vigente en la UE que aborda la responsabilidad y responsabilidad por los daños que puedan ser causados por
Los AIC plantean retos en términos de innovación en el campo de los AIC y aceptación social de los mismos. En
Por un lado, los productores de AIC temen que, en virtud de la Directiva de responsabilidad por productos defectuosos (PLD), puedan ser fácilmente

[8] Bryant Walker Smith, “Human error as a cause for vehicle crashes”, 18 de noviembre de 2013, disponible en Internet en <http://cyberlaw.stanford.edu/blog/2013/12/human-error-cause-vehicle-crashes > (último acceso el 28 de enero de 2016).

[9] OCDE, “OECD Factbook 2013: Economic, Environmental and Social Statistics”, 2013, disponible en Internet en <http://www.oecd-ilibrary.org/sites/factbook-2013-en/06/02/03/index .html?contentType=&itemId=/content/chapter/factbook 2013-50-en&containerItemId=/content/serial/18147364&accessItemIds=&mimeType=text/html> (último acceso el 28 de enero de 2016), también citado en Gillian Yeomans, “Autonomous Vehicles – entregando el control: oportunidades y riesgos para los seguros”, disponible en Internet en <https://www.lloyds.com/~/media/lloyds/reports/emerging%20risk%20reports/autonomous%20vehicles%20final.pdf > (último acceso el 28 de enero de 2016) (Yeomans 2014) en la página 5.

[10] Véase, por ejemplo, Yeomans 2014, en la pág. 5. También Anne Pawsey, “Autonomous Road Vehicles”, septiembre de 2013 en p. 1. Disponible en Internet en <http://www.parliament.uk/briefing-papers/post-pn-443.pdf>, (POSTnote 2013); Robolaw 2014, en pág. 42.

[11] La contaminación, el cambio climático, la exclusión social de las 'partes más débiles' y los altos riesgos de accidentes en las carreteras (europeas) pueden verse como el resultado de los procesos de modernización e individualización que tuvieron lugar en el siglo pasado. Ahora, a su vez, deben abordarse estos efectos secundarios.
Véase para la identificación y un estudio sobre el concepto de sociedad del riesgo de Ulrich Beck, su libro Risk Society, Towards a New Modernity, London: Sage Publications 1992.

[12] Véase, por ejemplo, Scott Le Vine & John Polak, “Autos automatizados: ¿un viaje suave por delante?”, febrero de 2014, en la pág. 14, disponible en Internet a través de <http://www.theitc.org.uk/docs/114.pdf> (último acceso el 28 de enero de 2016).

[13] Véase Wayne Cunningham y Antuan Goodwin, “Six Reasons to Love, or Loathe, Autonomous Cars”, 8 de mayo de 2013, disponible en Internet en <http://www.cnet.com/news/six-reasons-to-love-or-loathe-autonomous-cars/> (último acceso el 28 de enero de 2016).

responsabilizarse por los daños causados por AIC defectuosos, lo que tendría un efecto paralizador en innovación. [14] Mientras que, por otra parte, el marco actual sobre responsabilidad por productos defectuosos no proporcionar un conjunto de herramientas fácil para que los consumidores responsabilicen a los fabricantes de AIC por defectos en sus productos en todo. Una carga bastante pesada de la prueba recae en los consumidores para establecer que realmente hubo un defecto en el AIC, así como sobre la relación de causalidad entre el defecto y el daño que se haya producido.
La aportación de pruebas será más compleja cuando aumente la autonomía y la inteligencia de los coches, ya que las víctimas tendrán que realizar un análisis (tecnológico) en profundidad de, entre otras cosas, el software (original), el actualizaciones y los datos operativos con los que cuenta un AIC, a fin de establecer la causa precisa de un accidente. Al mismo tiempo, los fabricantes tienen amplia oportunidad de defenderse contra reclamaciones de responsabilidad. Frente a las AIC, el PLD no protege de manera óptima los intereses de consumidores brindándoles medios fáciles para obtener una remuneración por los daños que sufrieron causados por AIC defectuosos de los fabricantes.

El margen de mejora de la legislación actual lo forman además los diferentes regímenes europeos no armonizados sobre responsabilidad de los vehículos de motor. Hasta la fecha, existen 28 marcos diferentes en vigor en la Unión Europea. Por ejemplo, la 'Loi Badinter'15 francesa impone un estricto régimen de responsabilidad sin culpa en para evaluar si el conductor o el custodio de un automóvil debe o no indemnizar los daños de las víctimas (que no sean el conductor)16 de accidentes en los que se vean involucrados vehículos de motor. La responsabilidad solo puede ser exonerada si el conductor (o el custodio) puede probar una falta inexcusable por parte de la víctima.17 Los Países Bajos' 'Wegenverkeerswet' asigna responsabilidad (semi-estricta) al propietario o cuidador (nota: en lugar del conductor o un custodio) de un vehículo motorizado que se vea involucrado en un accidente en el que se hayan producido daños a usuarios de la vía no motorizados.18 Al menos el 50% de los daños sufridos debe ser remunerado, a menos que se pueda probar fuerza mayor.19 En el Reino Unido, Las reglas de negligencia se aplican para establecer si un conductor de un vehículo motorizado puede ser considerado responsable. En tales casos, no existe un régimen de responsabilidad estricta20 en el Reino Unido, aunque el nivel de cuidado requerido de los conductores de vehículos de motor es bastante alto. Caso de ley explica que un conductor que pierde el conocimiento por causas ajenas a él está actuando negligentemente,21 y también lo es el conductor cuyos frenos fallan cuando esta falla no podría haber sido prevista.22
Sin embargo, las víctimas de accidentes causados por vehículos a motor deben probar que los conductores estaban en

---

[14] Véase Erica Palmerini, Federico Azzarri, Fiorella Battaglia et al, D 6.2, “Guidelines on Regulating Robotics”, 22 de septiembre 2014, (RoboLaw 2014), pág. 60

[15] Loi “tendant à l'amélioration de la status des victimes d'accidents de la circulation et à l'accélération des procédures d'indemnización".

[16] Ver A. Tunc, “The 'Loi Badinter' – Ten Years of Experience”, 3 Maastricht Journal of European and Comparative Law, 1996 (Tunc 1996), p. 330. El artículo 3 dice: “Les victimes hormis les conducteurs […] sont indemnisées des dommages résultant des atteintes à leur personne qu'elles ont subis, sans que puisse leur être opposée leur propre faute”.

[17] Ver también Tunc 1997, en p. 335.

[18] La indemnización de los daños sufridos por las víctimas en el interior de un vehículo a motor se rige por las normas generales sobre responsabilidad previstas en el Artículo 6:162 del Código Civil holandés.

[19] Marloes de Vos ea, Tribunal Supremo de los Países Bajos, 2 de junio de 1995, NJ 1997/700-702, y Saïd Hyati ea, 5 de diciembre 1997 Nueva Jersey 1998/400-402. La noción de 'Betriebsgefahr' está tomada de la Straβenverkehrsgesetz alemana.

[20] O responsabilidad presunta como se denomina en la legislación escocesa.

[21] Roberts v. Ramsbottom [1980] 1 WLR 823, también citado en. Cees van Dam, European Tort Law, Oxford: Oxford University Press 2006 (Van Dam 2006), en p. 364, nota al pie 52.

[22] Henderson v. HE Jenkins & Sons and Evans [1970] AC 282, citado en Van Dam 2006, p. 364, nota al pie 53. Van Dam además toma nota de Worsley v Hollins [1991] RTR 252 (CA), en el que los jueces sostuvieron que la demanda por negligencia de la víctima fracasó porque el acusado pudo probar que, aunque sus sistemas de frenado habían fallado, causando daños, su minibús había sido reparado recientemente y pasó su ITV.

culpa, es decir: habían actuado negligentemente.23 Las diferencias significativas en la forma en que la responsabilidad por vehículos se aborda en todos los Estados miembros, no es beneficioso para el desarrollo, los seguros y despliegue de AIC en Europa. En todo caso, los regímenes nacionales que designen responsabilidad a los conductores de automóviles los vehículos deben actualizarse para poder abordar la responsabilidad de los vehículos sin conductor humano.

Privacidad

Considerando que la llegada de la tecnología AIC es prometedora en términos de mayor seguridad en las carreteras, resultando en menos daños a cubrir, también las compañías de seguros observan que cuando un accidente ocurre causado por tecnología autónoma, "se necesitaría una amplia experiencia en análisis de software y hardware para saber cómo y por qué ocurrió".24 Una de las opciones para evaluar la causa de un accidente, y por lo tanto para ayudar a responder a la pregunta de dónde recae la responsabilidad, podría ser equipar vehículos con cajas negras, o con soluciones telemáticas que conectan los AIC a una infraestructura dedicada y/o a servidores remotos.25 Los objetivos de este tipo de tecnologías son, entre otras cosas, registrar los movimientos de vehículos autónomos y las elecciones operativas que se realizan. por el propio automóvil o por el conductor que controla su movimiento, así como datos sobre eventos y objetos en el

cercanías de un vehículo autónomo. La tecnología de caja negra registra y almacena los datos recopilados dentro de un vehículo y ofrece un potencial para una evaluación posterior. La tecnología telemática puede tener una mayor aplicaciones Los datos no solo podrían utilizarse para evaluar los errores y las causas de los daños después de ocurrencia de accidentes, podría incluso tener un efecto preventivo. Comunicación de vehículo a vehículo (V2V) y la comunicación del vehículo a la infraestructura podrían usarse para la prevención de accidentes en tiempo real y brindan "beneficios ambientales, de seguridad y de movilidad" en general.26 Aunque la caja negra tecnologías y soluciones telemáticas como V2V y V2I (en lo sucesivo, "tracing tecnología') puede ser prometedor en términos de prevención de accidentes y distribución de daños causados por accidentes AIC, estos también imponen riesgos en cuanto al derecho a la privacidad (informativa) de las personas que se encuentran en el interior y en las cercanías de automóviles equipados con estas tecnologías.

La privacidad de la información de los ciudadanos está estrictamente regulada en la Unión Europea por la Directiva de Protección de Datos (DPD)27 y se regulará aún más estrictamente después de la entrada en vigor del Reglamento General de Protección de Datos (GDPR)28. El marco actual y el futuro prescriben para ejemplo de que ya durante la fase de diseño de los AIC equipados con tecnología de rastreo, una privacidad

---

[23] Hay una regla de un deber legal que, hasta cierto punto, establece la responsabilidad estricta de los conductores de vehículos motorizados que se acercan a un cruce en la carretera: "El conductor de todo vehículo que se acerque a un cruce deberá, a menos que pueda ver que no hay un cruce de peatones , avance a tal velocidad que pueda, si es necesario, detenerse antes de llegar a dicho cruce", como se cita en Van Dam 2006, en p. 365, nota al pie 57, en referencia al Reg. 3 del Reglamento de lugares de cruce de peatones (tráfico) de 1941, reemplazado por el Reglamento de cruce de peatones de cebra de 1971, SI 1971, No. 1524. Una defensa que tiene un conductor a este respecto es fuerza mayor .

[24] Yeomans 2014, en la pág. 18

[25] Yeomans 2014, en p.18. Véase además James M. Anderson, Nidhi Kalra, Karlyn D. Stanley et al., Tecnología de vehículos autónomos: una guía para los responsables de la formulación de políticas, RAND Transportation, Space and Technology Program 2014, (informe RAND), en la pág. 94-95.

[26] Informe RAND, p. 81.

[27] Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas con sobre el tratamiento de datos personales y sobre la libre circulación de estos datos, Diario Oficial L 281 0050 , 23/11/1995 pág. 0031 -

28 Propuesta de REGLAMENTO DEL PARLAMENTO EUROPEO Y DEL CONSEJO relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (Reglamento general de protección de datos)
COM/2012/011 final - 2012/0011 (COD). Tenga en cuenta que el diálogo tripartito entre la Comisión Europea, el Consejo de Europa y el Parlamento Europeo ha concluido sobre el texto final del RGPD; sin embargo, este texto aún no se ha publicado formalmente.

debe llevarse a cabo una evaluación de impacto. Además, “las medidas técnicas y organizativas apropiadas medidas para proteger los datos personales contra la destrucción accidental o ilegal o la pérdida accidental, alteración, divulgación o acceso no autorizado, en particular cuando el procesamiento implique la transmisión de datos a través de una red, y contra todas las demás formas ilegales de procesamiento ” . , y debe estar habilitado de forma predeterminada.30 La seguridad de última generación y los costos de implementación deben tenerse en cuenta para el

implementación de medidas. Además, estos “garantizarán un nivel de seguridad adecuado a la riesgos que representa el tratamiento y la naturaleza de los datos a proteger”.

Otro desafío aún más reciente lo constituye la reciente decisión del Tribunal de Justicia de las Comunidades Europeas para declarar el marco de puerto seguro, que forma la base de muchos intercambios de datos personales entre la UE y los Estados Unidos de América, inválida. Es probable que la tecnología de rastreo incorporados en AIC constituirán la transmisión internacional de datos (personales), a través de fronteras de la Unión Europea, y posiblemente importar estos datos a los Estados Unidos, por ejemplo, a través de la computación en la nube. El TJUE dictaminó que Estados Unidos no ofrece un nivel adecuado de protección para datos personales, ya que quedó claro después de las revelaciones de Edward Snowden, que las autoridades estadounidenses tales como la Agencia de Seguridad Nacional tienen fácil acceso a los datos personales procesados por empresas e instituciones estadounidenses.31 El tribunal dictaminó que los poderes de las autoridades europeas de supervisión están socavados por las prácticas de los Estados Unidos, que pueden no ser habilitadas por una decisión de la Comisión Europea. esta sentencia implica que la exportación de datos personales a los Estados Unidos ya no es posible sobre la base de la marco de puerto seguro. Aunque los Estados Unidos y la Comisión Europea están negociando actualmente un tratado alternativo,32 mientras tanto, el intercambio de datos personales entre la UE y los Estados Unidos no está permitido en base a las reglas de Puerto Seguro aún inválidas.

3. Ponga a prueba el principio

En esta parte probaremos si el segundo principio EPSRC se puede considerar como una prueba futura contra los criterios validez, suficiencia/generalidad y utilidad

3.1 Validez

Tanto dado el estado actual de la tecnología como de la ley, la primera parte del principio Humanos, no robots, son agentes responsables ha demostrado ser aún válido. Es una afirmación correcta sobre el naturaleza de los robots, desarrolladores de robots y la relación entre robots y personas. Puede haber siempre un ser humano o una persona jurídica responsable de las acciones de las AIC.

La creación específica de una entidad legal separada para las AIC parece actualmente descabellada dada la situación actual. situación tecnológica y jurídica de las AIC, además no contribuiría a resolver la responsabilidad

[29] Arte. 17(1) DPD; véase también el artículo 5(1)(eb) y la sección 2 (art. 30 y siguientes) sobre seguridad de datos en el RGPD.
[30] Arte. 23 RGPD.
[31] Asunto C-362/14, Maximilian Schrems/Facebook [2015].
[32] Consulte las últimas noticias sobre el 'Acuerdo paraguas UE-EE. UU.' (Acuerdo entre los Estados Unidos de América y la Unión Europea sobre la protección de la información personal relacionada con la prevención, investigación, detección y enjuiciamiento de delitos): http : //ec.europa.eu/justice/newsroom/data-protection/news/150908_en.htm (último acceso el 9 de marzo de 2016).

desafíos cumplidos como se describe en la sub 2.2. Lo mismo es cierto para la segunda parte del segundo principio que establece que los Robots deben ser diseñados; operado en la medida de lo posible para cumplir con las leyes existentes y derechos y libertades fundamentales, incluida la privacidad, ha demostrado ser todavía válido. Con un ojo en el tecnología de evidencia (para ser) incorporada dentro de los AIC, esta idea fundamental ha demostrado ser incluso más cierto de lo que uno podría haber imaginado sobre su diseño. Como hemos visto sub 2.2 de hecho, los defectos del actual régimen de responsabilidad puede resolverse parcialmente mediante sistemas inteligentes de recogida y almacenamiento de pruebas integrados en el AIC. Estos sistemas de recolección y almacenamiento de evidencia deben diseñarse de tal manera que los datos personales recopilados estén protegidos tanto como sea posible: privacidad por diseño y privacidad por defecto debe incorporarse en AIC (tecnología de rastreo) en todo momento.

3.2 Suficiencia/generalidad

Al mismo tiempo, el principio sigue siendo suficiente y lo suficientemente amplio para cubrir todos los aspectos importantes cuestiones que pueden surgir en la regulación de las AIC en el mundo real. Los humanos, no los robots, somos Agentes responsables Se deben diseñar robots; operado en la medida de lo posible para cumplir con las leyes existentes y los derechos y libertades fundamentales, incluida la privacidad. No parece haber preocupaciones significativas pasado por alto. Aunque algunos autores parecen argumentar que se debe crear una entidad legal para las máquinas inteligentes autónomas, haciendo de los robots el agente responsable,33 esto no ha sido convincente para muchos34 y ciertamente no para nosotros.

Los retos que plantea la introducción en la sociedad de los coches inteligentes autónomos y su la responsabilidad por daños en sí misma no parece requerir una personalidad jurídica separada. Simplemente agregaría un actor más para la atribución de responsabilidad. Al mismo tiempo, requeriría un rediseño sustancial del sistema de responsabilidad tal como se aplica actualmente al mundo real, mientras que la tecnología aún está en desarrollo etapa que corre el riesgo de regulación insuficiente o excesiva.

3.3 Utilidad

En la medida en que no se agoten los medios legales actuales, entre otras cosas con el fin de lograr una mayor armonización de los Regímenes legislativos de responsabilidad de la UE en combinación con una tecnología de evidencia efectiva, no hay evidencia que sustentaría un cambio de paradigma completo mediante la introducción de las AIC como responsables agentes en sí mismos. Dado que AIC puede ser diseñado y operado para cumplir con las leyes existentes la utilidad de este principio sigue siendo evidente. Sin embargo, las tecnologías de caja negra y las soluciones telemáticas como V2V y V2I pueden ser prometedores en términos de prevención de accidentes y distribución de daños causados por accidentes AIC, ya que éstos también imponen riesgos en términos del derecho a la privacidad (informativa) de personas dentro y en las proximidades de automóviles equipados con estas tecnologías, los sistemas tendrían que

---

[33] Véase, por ejemplo, James Boyle, "¿Dotados por su creador?: El futuro de la personalidad constitucional", El futuro de la Constitución, 9 de marzo de 2011, pág. 6, también disponible a través de Internet en <http://www.brookings.edu/~/media/research/files/papers/2011/3/09-personhood-boyle/0309_personhood_boyle.pdf> (último acceso el 9 de marzo de 2016 Véase, además, JP Günther, F. Münch, S. Beck, S. Löffler, C. Leroux y R,Labruto, "Issues of Privacy and Electronic Personhood in Robotics. 21st IEEE International Symposium on Robot and Human Interactive Communication", París 2012, citado en Christophe Leroux, Roberto Labruto, Chiara Boscarato y otros, "Suggestion for a Green Paper on legal issues in robotics", diciembre de 2012, disponible en Internet en <http://www.eu robotics.net /cms/ upload/PDF/euRobotics_Deliverable_D.3.2.1_Annex_Suggestion_GreenPaper_ELS_IssuesInRobotics.pdf> (último acceso 28 de enero de 2016); y Robolaw 2014, pág. 24

[34] Véase, por ejemplo, Peter Asaro, "Robots and Responsibility from a Legal Perspective", vía < http://www.peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf>, consultado por última vez el 9 de marzo de 2016; y Lawrence Solum, "Legal Personhood for Artificial Intelligences", North Carolina Law Review, (abril de 1992), págs. 1231-1287.

incluir la privacidad desde el diseño para proteger estos derechos fundamentales tal como se establece en los tratados internacionales y europeos.35

Es crucial que estos requisitos de la ley y la tecnología se cumplan antes del desafío de la se puede cumplir con la introducción y el despliegue de los AIC en la sociedad.

4. Conclusión

Podemos concluir diligentemente que el Principio 2 de los principios de robótica del EPSRC desarrollado por Los expertos británicos en robótica e IA en el retiro financiado por EPSRC/AHRC han demostrado ser una prueba de futuro cuando aplicamos el estado del arte actual de la ley y la tecnología que rodea a los AIC.

Los humanos, no AICS, son los agentes responsables. Los AIC deben diseñarse; operado hasta donde sea factible para cumplir con las leyes existentes y los derechos y libertades fundamentales, incluida la privacidad por diseño. Por lo tanto, dando evidencia del hecho de que la respuesta de la máquina está al menos parcialmente en el máquina en sí.

[35] Véase por ejemplo el art. 7 y 8 de la Carta de los Derechos Fundamentales de la Unión Europea y el artículo 8 del Convenio Europeo de Derechos Humanos.

manejo de datos justo y robótica

Burkhard Schafer, Facultad de Derecho de Edimburgo
Lilian Edwards, Universidad de Strathclyde

Esta intervención combina el principio 4, los robots son artefactos fabricados. No deben diseñarse de manera engañosa para explotar a los usuarios vulnerables; en su lugar, la naturaleza de su máquina debe ser transparente con el principio y el principio 2, los robots deben diseñarse; operarse en la medida de lo posible para cumplir con las leyes, los derechos fundamentales y las libertades existentes, incluida la privacidad. Sugiere alguna elaboración/especificación de los principios que la intersección entre ellos necesita. Más radicalmente quizás, se trata de los deberes correspondientes por parte de los usuarios de robots y, de hecho, de terceros hacia los robots. hacia los robots (o, para aquellos que se preocupan por esta formulación, los deberes hacia el propietario de un robot se han comportado de alguna manera contra la máquina.

Los robots plantean algunos desafíos únicos para las prácticas justas de manejo de datos, desafíos que son causados, al menos en parte, por su capacidad de “engañar”, aunque sin darse cuenta, a las personas con las que interactúan.

Mucho antes de la tecnología moderna, los humanos desarrollaron técnicas de preservación de la privacidad, desde las cortinas hasta las ventanas y el velo, desde aprender cuándo susurrar hasta lavarse el propio olor. Los protegían tanto de las miradas indiscretas como del interés de los depredadores no humanos. Crucialmente, estos no solo protegen la información, sino también el intercambio y el intercambio de información (susurrando, insonorizando su estudio)

La ley, con su sistema de reglas y excepciones, con frecuencia otorgó un reconocimiento formal a estas medidas de protección de baja tecnología. arquetipo de “expectativa razonable de privacidad” y “seguridad en nuestras casas y viviendas”, pero también un espacio dentro del cual los datos pueden recopilarse e intercambiarse con mayor libertad: la excepción doméstica de la ley europea de DP es un excelente ejemplo.

La tecnología robótica amenaza con hacer que estas soluciones de baja tecnología al problema de la privacidad sean cada vez más redundantes. anuestracasa.

Nadie es un héroe para sus domésticos. Pero al menos, con los domésticos, el señor o la dama de

la casa podría anticipar lo que exactamente podrían ver, entenderían el entorno normativo (tanto social como legal) que les impedía recopilar y, lo que es más importante, compartir datos sobre su empleador.
La comprensión del entorno normativo junto con la comprensión de las capacidades sensoriales permitiría una evaluación y una gestión racionales del riesgo. (Confío en mi mayordomo con mi ropa interior sucia, pero no con la camisa empapada de sangre. No me preocupo por la señal de calor cuando entretengo a un invitado, pero puedo optar por mantener el ruido bajo, mientras confío en que mi mayordomo toque primero antes de entrar al dormitorio).

La robótica amenaza estas estrategias defensivas no solo porque puede usar sensores fuera del espectro visual o auditivo, o debido a su movilidad que permite la detección en espacios previamente protegidos. Hay evidencia de que hacemos estas inferencias cuando interactuamos con robots. Internet está lleno de personas que “se acercan sigilosamente” a Asimo desde atrás; ahora los sensores de Asimo “pueden” estar ubicados en sus ojos y tener restricciones de visión similares a las de un humano, pero esto bien puede ser falso.

Parte del diseño ético, por lo tanto, también debería ser indicar las capacidades sensoriales de los robots de maneras que faciliten la aparición de defensas “intuitivas” del tipo que usamos con otros humanos, y abstenerse, cuando sea posible, de invitar a inferencias engañosas, e incluir la “facilidad del mecanismo defensivo” en la evaluación de la intrusividad cuando se puede hacer una elección entre diferentes sensores.

La ley de protección de datos es un impulsor detrás de esto, pero las prácticas de detección justa y manejo de datos van más allá de los datos personales, y mucho menos personales sensibles. Protegemos no solo los datos sobre nosotros, sino también nuestras ideas comerciales, descubrimientos científicos o tecnológicos, o habilidades.

Por lo tanto, la ley IP es otra restricción legal que debe observarse bajo este encabezado, y una noción más amplia de “prácticas justas de manejo de datos” que va más allá de la ley DP puede ser necesaria. /compensación?

Esto, potencialmente, plantea también una cuestión que lleva de una manera más radical más allá de los Principios. Tratan principalmente de establecer los deberes que el desarrollador tiene con las personas que interactúan con sus máquinas. deben y los posibles deberes que se les deben a ellos/al propietario del robot.

Como ejemplo, permitir un diseño de robot seguro puede implicar un deber para terceros de revelar o compartir cierta información con el robot que en el pasado ha tenido privilegios legales. copiando El enfoque de EE. UU. que argumenta que esto sería copiar con fines funcionales en lugar de expresivos

Sin embargo, una vez que las máquinas coordinen su acción al compartir estos datos, incluso este nuevo argumento puede llegar a sus límites.

Los ciudadanos pueden optar por utilizar tecnología para evitar que los sensores los detecten (p. ej., pintura facial de camuflaje: https://cvdazzle.com), pero eso puede significar que aceptan un mayor riesgo de que el robot se tope con ellos. Si hay terceros involucrados, esto puede crear problemas legales aún más complejos. Si manipulo intencionalmente el proceso de aprendizaje, ¿me estoy metiendo en el territorio de la Ley de Uso Indebido de Computadoras? Y, finalmente, si contribuyo al aprendizaje de una máquina, ¿tengo participación en lo que produce como resultado?

La ley básica de negligencia y su distinción entre actuar y misión y cómo la negligencia la trata estableciendo deberes entre vecinos, será parte de la respuesta legal después de que ocurra un accidente.

1. Confiar en el deber ético/social por parte de terceros de no manipular la adquisición de conocimientos de la máquina2.
Confíe únicamente en una obligación legal más estricta de abstenerse de cierta manipulación de datos previsiblemente
peligrosa3. No confiar en un entorno cooperativo cuando se piensa en la seguridad y el cumplimiento de la ley del robot que construyen; después de todo, no todos cumplen todas las leyes.

Para aclarar por qué surge este problema en el contexto de la discusión sobre la “transparencia del sensor”: SI aceptamos la obligación ética discutida anteriormente, es decir, que los robots normalmente deben revelar cómo y con eso lo que pueden sentir, entonces inevitablemente se exponen a la manipulación.

Si aceptamos 2, entonces tenemos que lidiar directamente con el hecho de que en los dos extremos del espectro, la ley es clara: es difícil tener deberes no humanos, por otro lado, puedo tener el deber de un vecino de no incendiar su granero, pero ¿tengo el deber de que los operadores (o diseñadores) de robots ausentes no confundan su entrenamiento de robots? ¿Es razonablemente previsible que un robot se confunda? Después de todo, ni siquiera tengo el deber de no mentirle a los extraños, por ejemplo, cuando doy instrucciones, a menos que se trate de algún consejo profesional.

La discusión hasta ahora discutió los problemas causados por humanos que retienen/distorsionan/manipulan datos que un robot necesita para operar de manera segura, y que ética o legalmente deben.

Pero también nos enfrentamos a opciones de diseño ético cuando las personas cooperan y ofrecen información voluntaria que no están legalmente obligadas a proporcionar, pero que eligen o no inhiben en un sentido del deber cívico.

Esto podría significar que no solo los robots deben ser identificables como robots, sus sensores como sensores, sino que la salida generada por el robot también debe ser identificable como una máquina, no generada por humanos. podría ser requerido

Por lo tanto, el tema general de esta intervención es, en última instancia, la transparencia algorítmica: los deberes legales y éticos influyen en cuándo y cómo los robots deben revelar sus capacidades sensoriales. Una vez que lo hacen, su entorno tiene opciones: cooperar o no cooperar. algunos) datos para tener máquinas más seguras. Donde la cooperación más allá de lo legalmente requerido produce valor, es necesario discutir cómo dar cuenta de esto de manera no equitativa, imponiendo poderosamente otro deber de divulgación, un "hecho por robot"

Si bien estos son principalmente problemas legales, para la cuestión del diseño ético (y que cumple con la ley), los desarrolladores también deben poder anticipar qué tipo de interacción esperar y a qué tipo de información tendrán acceso legal.

¿Pueden los robots ser agentes morales responsables? ¿Y por qué debería importarnos?

AmandaSharkey,

Departamento de Ciencias de la Computación y Robótica de Sheffield, Universidad de Sheffield

Principio 2. Los seres humanos, no los robots, son agentes responsables.

A primera vista, esta afirmación o principio parece convincente. Tiene sentido insistir en que los humanos y no los robots son los agentes responsables. Nos recuerda las capacidades limitadas de los robots y proporciona un antídoto útil para las fuertes afirmaciones y advertencias que a veces se hacen sobre ellos. Deberíamos ayudar a restringir los posibles usos nocivos a los que se podrían aplicar los robots. También tiene sentido sugerir que los robots deben diseñarse y operarse para cumplir con las leyes existentes y los derechos y libertades fundamentales: es difícil imaginar que alguien sugiera lo contrario.

Pero, en una consideración adicional, se hace evidente que la declaración no da ninguna justificación por decir que los humanos y no los robots son agentes responsables, ni proporciona ninguna guía sobre dónde y cuándo se deben usar los robots, o las consecuencias que se derivan de asumir que los robots no son agentes responsables. ¿Cuáles son las razones para suponer que los robots y no los humanos son agentes responsables? (b) ¿Es suficiente diseñar robots para cumplir con las leyes existentes y los derechos y libertades fundamentales? y (c) Si los robots no son agentes responsables, ¿debería esto limitar las funciones que desempeñan y las situaciones en las que se implementan?

(a) ¿Cuáles son las razones para suponer que los humanos y no los robots son los agentes responsables?

Además de la responsabilidad legal, es posible identificar dos razones para esta suposición. La primera se basa en la diferencia entre máquinas biológicas y mecánicas, y la base biológica de la moralidad. El segundo tiene que ver con la necesidad de que la sociedad asuma la responsabilidad por los artefactos que los humanos han producido.

(i) Máquinas biológicas frente a máquinas mecánicas: responsabilizar a un agente por sus acciones equivale a considerar que es un agente amoroso. PatriciaChurchland (2011) discute la base de la formalidad en los seres vivos, y argumenta que la base para preocuparse por los demás radica en la neuroquímica del apego y la vinculación en los mamíferos. y evitar el dolor en sus parientes inmediatos. Los humanos y otros mamíferos sienten ansiedad por su propio bienestar y el de aquellos a quienes están unidos. Además del apego y la empatía por los demás, los humanos y otros mamíferos desarrollan relaciones sociales más complejas y son capaces de comprender y predecir las acciones de los demás. o desaprobación. Como consecuencia, los seres humanos tienen un sentido intrínseco de la justicia.

Lo mismo es en gran parte el caso de los mamíferos no humanos. Bekoff y Pierce (2009) proporcionan muchos ejemplos de evidencia del sentido amoral de la justicia en los mamíferos. Por ejemplo, los monos capuchinos

trabajando para golosinas parecían ofendidos y se negarían a cooperar más si vieran que otro mono recibió una recompensa amada y deseable por el mismo trabajo (Brosnan y deWaal, 2003).

Por el contrario, a los robots no les preocupa su propia conservación o evitar el dolor, y mucho menos el dolor de los demás. En parte, esto se puede explicar mediante el argumento de que no están realmente encarnados, en la forma en que lo está una criatura viva. para un ser humano. El cuerpo vivo es una entidad autopoética integrada (Maturana y Varela, 1980) en la medida en que una máquina hecha por el hombre no lo es.

Responsabilidad social: Muchos escritores estarían de acuerdo con la implicación de la afirmación de que los robots no son agentes morales completos. Johnson y Miller (2008) argumentan que los robots y otros artefactos computacionales no son agentes morales completos porque “nunca son completamente independientes de sus diseñadores humanos”. edon a los artefactos en sí mismos, ya que el comportamiento y los resultados de los robots y los sistemas informáticos dependen necesariamente de los diseñadores y desarrolladores humanos. Un ejemplo útil que ellos consideran es el de un abridor de puertas. el abridor no sería

En el pasado, se han planteado argumentos relacionados con la falta de independencia de los diseñadores humanos basados en la forma en que los robots, a diferencia de las máquinas vivas, nunca se pueden considerar completamente encarnados, ya que siempre han requerido la intervención humana y la participación en su desarrollo (Sharkey y Ziemke, 2001). El punto aquí es que los robots, y sus sistemas de control subyacentes, dependen sobre la intervención humana. Los robots pueden 'soltarse' para tomar decisiones impredecibles, pero la decisión de permitirles hacerlo es solo humana y social. Cualquier decisión tomada por el robot aún dependerá de su diseño inicial. y reconocido. Johnson (2006) hace una distinción útil entre agentes morales y entidades morales, y coloca los robots y los artefactos informáticos en la segunda categoría. Las entidades morales incluyen el diseñador del artefacto, el artefacto y el usuario del artefacto, y la responsabilidad moral no se puede descargar en el artefacto mismo.

(b) ¿Es suficiente diseñar robots para cumplir con las leyes existentes y los derechos fundamentales y libertades, incluida la intimidad?

Un problema importante con la sugerencia de que los robots deben diseñarse para cumplir con las leyes existentes y los derechos y libertades fundamentales, y la razón por la que no es suficiente, es que las leyes y los derechos humanos existentes no han sido formulados teniendo en cuenta los desarrollos tecnológicos como la robótica. para aparecer como amigos y compañeros, y como resultado, son bienvenidos en nuestros hogares y entornos íntimos. Hay muchas preguntas que deben responderse sobre el grado en que la información a la que tienen acceso será accesible para los demás, y aún hay poca legislación para abordar esto. , (p. ej., Sharkey Sharkey, 2012; Sparrow and Sparrow 2006), pero la Ley de derechos humanos no proporciona ninguna protección explícita frente a tal situación. Preocupaciones similares

Se ha planteado la cuestión de dejar a los niños al "cuidado" de los robots hasta el punto de que sus vínculos con los humanos se ven comprometidos (Sharkey y Sharkey 2010), pero, de nuevo, no existe legislación ni derechos que prevengan explícitamente tal posibilidad, aparte de la asociada con la negligencia infantil. surgir si los humanos colocan a los robots en posiciones de poder sobre los humanos.

Cuando los humanos toman decisiones sobre cómo actuar en situaciones sociales, tienen que hacer más que seguir un conjunto de reglas o leyes. Toman decisiones basadas en una comprensión moral de lo que es inapropiado o inapropiado para ellos. Arkin (2009), por ejemplo, argumentó que en una situación de campo de batalla, los soldados robot podrían programarse para seguir un conjunto de reglas que resultarían en un comportamiento más ético que el que a veces muestran los soldados humanos en el fragor de la batalla. Un robot por otro lado no respondería emocionalmente y podría ser programado, por medio de un 'gobernador ético' para evaluar acciones antes de llevarlas a cabo, y para realizar solo aquellas consideradas moralmente permisibles.

Varios autores han argumentado en contra de la idea de poder programar robots para tomar decisiones morales. la apreciación del panorama general, la comprensión de las intenciones detrás de las acciones de las personas, y la comprensión de los valores y la anticipación de la dirección en la que se desarrollan los eventos' (2013, A/HRC/23/47).

En un artículo interesante sobre los requisitos para crear robots con lo que denominan "competencia moral", Malle y Scheutz (2014) argumentan que, entre otras cosas, los robots requerirían una red de normas morales para saber qué es y qué no es moralmente aceptable. Sugieren que no sería práctico programar esta red y que, en lugar de programar robots con normas morales, podrían aprender desarrollar una red de normas morales sobre la base de la retroalimentación dada a ellos en respuesta a sus acciones. Sugieren que podría ser necesario criar a los robots en entornos humanos, ya que esta puede ser 'la única forma de exponerlos a la riqueza de las situaciones morales humanas y las interacciones comunicativas' (Malle y Scheutz, 2014). uldsermejoradoentrenandolosmonmoralstories

(Riedland Harrison, 2016), y exigirles que apliquen ingeniería inversa a los valores humanos que representar.

Es cierto que es difícil descartar la posibilidad de que en el futuro un robot pueda ser entrenado o criado para ser moral, pero hay varias razones para ser escéptico acerca de la probabilidad de éxito.
Las razones para el escepticismo incluyen la falta de formalidad de una base biológica del robot, como se discutió anteriormente.
Como ya se discutió, un robot individual ni siquiera se preocupa por su propio cuerpo, y mucho menos por el de un ser humano; sufriría dolor si, por ejemplo, se le quitaran las ruedas. s desarrollar una comprensión buena y generalizable de las diferencias entre el bien y el mal.

comportamiento, como los robots programados por Winfield et al (2014) para tomar medidas para evitar que otros robots caigan en un agujero, que se describe como exhibiendo algo que puede describirse como un comportamiento ético. Pero el uso del término 'ético' o 'moral' en este contexto no significa que el los robots en cuestión podrían ser legítimamente elogiados o culpados por sus acciones.

(c) Si los robots no son agentes responsables, ¿debería esto limitar las funciones sociales que otorgan y las situaciones en las que se implementan?

La afirmación original de que los robots no son agentes responsables no explica en detalle lo que esto implica para el despliegue de robots. Aquí se argumenta que existen buenas razones para limitar los roles sociales y los poderes de toma de decisiones de los robots. pero también porque los humanos deberían tener derecho a que sus compañeros humanos tomen decisiones de vida o muerte.

Este argumento puede, y diría que debería, extenderse más allá a otros tipos de decisiones donde los robots podrían restringir las libertades de los humanos. Un robot colocado en el papel de maestro tendría que tomar decisiones sobre cosas como cuándo castigar o restringir a los niños, o cuándo elogiarlos. algo peligroso o arriesgado. Una niñera robot tendría que tomar una decisión similar sobre sus jóvenes a cargo. Se debe tener cuidado para mantener el control, la participación y la responsabilidad humanos en las decisiones que afectarán la vida de los seres humanos. Ya existen riesgos de que las decisiones automatizadas afecten nuestras vidas, pero los robots que pueden tener la apariencia de actores sociales competentes hacen que estos riesgos sean aún más frecuentes.

Resumen: Es fácil estar de acuerdo con el principio de EPSRC acerca de que los robots no son agentes responsables, pero incluso esta breve consideración resulta insuficiente para guiar acciones futuras.

implicancias para el despliegue de robots y para las elecciones humanas sobre los roles sociales que se les deben asignar. Los robots programados para seguir la ley y respetar los derechos y la libertad de las personas, no comprenderán las situaciones sociales y no podrán tomar las decisiones morales correctas de manera consistente sobre las situaciones sociales humanas. Se debe tomar para evitar o minimizar la toma de decisiones automática y algorítmica en cualquier situación en la que se requiera el juicio humano. Incluso se necesita mayor cuidado en el caso de los robots que crean la ilusión de que entienden.

Referencias

Arkin, R. (2009). Control del comportamiento letal en robots autónomos. Revisión de Chapman-Hall. Computadoras y Educación, 58(3),978–988.

Bekoff, M. y Pierce, J. (2009) WildJustice: TheMoralLivesofAnimals. TheUniversityofChicagoPress, Londres.

Brosnan,SFanddeWaal,FB(2003).Monkeysrejectunequalpay. Naturaleza, 425,297-99

Churchland, P. (2011) Braintrust: lo que la neurociencia nos dice sobre la moralidad. Princeton University Press, Oxford.

Heyns, C. (2013). Informe del Relator Especial sobre ejecuciones extrajudiciales, sumarias o arbitrarias, A/HRC/23/47

Johnson, DG (2006). Sistemas informáticos: entidades morales pero no agentes morales. Ética y tecnología de la información, 8(4):195–204

Johnson, DG y Miller, KW (2008) Deshacer agentes morales artificiales. Ética e Información Tecnología (2008) 10: 123–133

Malle, BF y Scheutz, M. (2014). Competencia moral en robots sociales. Simposio internacional IEEE sobre ética en ingeniería, ciencia y tecnología (págs. 30–35). Presentado en el Simposio internacional IEEE sobre ética en ingeniería, ciencia y tecnología, junio, Chicago, IL: IEEE.

Maturana,HR&Varela,FJ(1980). Autopoiesis y cognición: la realización de la vida. Dordrecht,Países Bajos:D.ReidelPublishing

Riedl, MO y Harrison, B. (2016) Uso de historias para enseñar valores humanos a agentes artificiales .

Sharkey, AJC y Sharkey, NE (2012). Grannyandtherobots:Ethicalissuesinrobotcarefortheelderly. Ética y tecnología de la información, 14(1),27–40.

Sharkey, NE, y Sharkey, AJC (2010). La vergüenza de las niñeras robot: una evaluación ética. InteractionStudies, 11(2),161–190.

Sharkey, NE y Ziemke, T. (2001). Mechanisticvs.PhenomenalEmbodiment-CanRobotEmbodimentLeadtoStrongAI CognitiveSystemsResearch, 2,4,251-262

Sparrow, R. y Sparrow, L. (2006). ¿En manos de las máquinas? El futuro del cuidado de los ancianos. Mente y Máquina, 16,141–161.

Winfield, AF, Blum, C. y Liu, W. (2014) Hacia un robot ético: modelos internos, consecuencias y selección de acciones éticas. En M. Mistry, A. Leonardis, M. Witkowski y C. Melhuish (Eds) Avancesensistemasrobóticosautónomos: Actas de la 15 conferencia anual, TAROS 2014 (págs. 85 y 96). Birmingham, Reino Unido, 1 a 3 de septiembre

Segundas reflexiones sobre la privacidad, la seguridad y el engaño

Tom Sorell, Universidad de Warwick
Heather Draper, Universidad de Birmingham

Los cinco principios de la robótica formulados durante el retiro de la AHRC-EPSRCR en 2010 no son la última palabra sobre la ética de los robots, sino una de las primeras palabras. Hay un largo camino por recorrer.

Privacidad

El principio 2 exige que los robots se operen de conformidad con las leyes existentes y los derechos y libertades fundamentales, incluido el derecho a la privacidad. y 8), no es fundamental en los tratados de derechos humanos más antiguos, como el Pacto Internacional de Derechos Civiles y Políticos (consulte el Artículo 17). Primero, la privacidad no se encuentra entre los llamados derechos "no derogables", como el derecho a evitar la tortura o la discriminación. una era de los derechos humanos en la que cualquiera es más fundamental que cualquier otro. Los instrumentos que interpretan el estado de los derechos humanos, como la Declaración de Viena, dicen que todos los derechos humanos son interdependientes e indisolubles (Art. 5). Por lo que aunque existiera un acuerdo sobre lo que es la privacidad, la necesidad de respetar un derecho a la privacidad no sería necesariamente preponderante.

La privacidad personal a veces se entiende como el control de la información sobre uno mismo. Los robots de cuidado en particular y los robots sociales en general a menudo están diseñados para recopilar información sobre los seres humanos con los que interactúan. Sin información, porque las personas cuya información es, en principio, pueden dar su consentimiento para la recopilación y almacenamiento de datos. Debido a que el uso de su información está sujeto a su consentimiento, el control no pasa a manos de otras personas: el consentimiento es una forma de control.

Sin embargo, el consentimiento no resuelve necesariamente todas las cuestiones sobre el uso adecuado de la información personal. En primer lugar, existe una diferencia entre la recopilación de información de forma única o intermitente, y la recopilación de información más o menos continua en tiempo real. Las implicaciones de la segunda son más difíciles de predecir y consentir por adelantado que las Implicaciones de la primera. Es incluso discutible que no existe tal cosa como el consentimiento debidamente informado para el monitoreo y seguimiento continuos de un robot vivo en el cuidado precisamente porque no es posible predecir ni imaginar por adelantado cómo sería la experiencia de vivir con un robot.

En lugar de determinar los límites de privacidad solo a partir de lo que un usuario consiente con buena información, uno puede tener que confiar también en argumentos sobre los límites de privacidad basados en el

Diseño breve y propósito de un tipo particular de robot. A menudo, se supone que los robots de cuidado para personas mayores ayudan al mantenimiento de su autonomía, es decir, su capacidad de elección y tener un conjunto de habilidades, la capacidad de lavarse, limpiar, cocinar, alimentarse, etc., suficiente para vivir de forma independiente. Una persona autónoma decide por sí misma no solo qué hacer y cómo vivir, sino también qué información personal revelar. para mantener la autonomía de una persona mayor se puede juzgar en parte por si la persona mayor tiene tanta libertad como un adulto sin ayuda para decidir sobre todos los aspectos de su vida, incluida la divulgación de información.

La tensión entre la autonomía (y la privacidad) y la seguridad

Una forma en la que el adulto estándar ejerce la autonomía es siendo su propio juez de los riesgos que debe asumir. IfthsohsoReputInnangerRelieduPontaundtakAteAdangerosRescueOfTheautonomoUsrisk-Taker, thentheremayBeanarGumumentaAstrisk-Taking fromTheautonomy restrictando a los que se convierten inatado a hhis, al menos en la temparidad.

En el caso de las personas mayores asistidas, el resumen de diseño del robot generalmente combinará seguridad y autonomía. El robot ayuda a los usuarios a llevar sus propias vidas, y también monitorea al usuario y sus circunstancias para emergencias de salud. Si se detectan emergencias o anomalías, editar puede generar la alarma.

En los casos que son más interesantes desde el punto de vista de la teoría moral, el usuario está dispuesto a correr un riesgo relativamente pequeño, por ejemplo, el riesgo de sufrir una caída menor, con el fin de continuar con su vida cotidiana de la misma manera que lo hacía cuando era más joven. caen. Esto se debe a que mantener

se supone que la autonomía es el propósito primordial del robot compañero. Sin embargo, si el objetivo primordial es mantener al usuario seguro, es posible que no tenga este derecho. Todo depende de cuánto daño menor sea compatible con estar seguro. En teleasistencia de baja tecnología, un usuario no usa una alarma colgante, y le corresponde a él decidir si pedir ayuda o no. El valor de la autonomía respalda esta norma.

La anulación del usuario también podría incorporarse en el diseño de un robot complementario donde la elección del estilo de vida de un usuario, si se informara a amigos y parientes, podría provocar intervenciones coercitivas de esas personas. Aquí es donde la autonomía apoya la privacidad sobre la seguridad total o la prudencia total. de lo contrario, un fagismo restringe la autonomía de las personas mayores y los diseñadores de robots y los formuladores de políticas públicas los tratan peor que a otros adultos .

una persona sin ayuda. Y esta limitación es difícil de defender sin el envejecimiento.

La tensión entre autonomía y rehabilitación

Los robots cuidadores y algunos robots no sociales están diseñados para ayudar a las personas mayores a recuperar las habilidades que han perdido y no solo a ejercitar las que tienen de forma autónoma. al introducir el robot en la casa de un usuario, puede existir una obligación de cooperación. Si no se reconoce tal intención conjunta, se debe dejar espacio para una negativa autónoma a aceptar una rehabilitación que sea beneficiosa. Después de todo, una negativa autónoma a aceptar una intervención médica no puede ignorarse legalmente .

¿Qué ocurre si se proporciona asistencia robótica a una persona mayor con la condición de que se acepte cooperar con la rehabilitación que se pueda ofrecer en el futuro? En este caso, la negativa autónoma puede ser anulada por un compromiso autónomo para cooperar. a un contrato explícito que el usuario celebra, que especifica las responsabilidades que el usuario asume a cambio de la provisión del robot. Dicho contrato puede existir entre un usuario y la autoridad local.

Las responsabilidades bajo el contrato no excluirían los derechos, por supuesto.

Engaño

Finalmente, llegamos al Principio 5. Requiere transparencia en el diseño de robots y prohíbe el engaño de los vulnerables. El engaño es la creación intencional de falsas creencias. El engaño suele ser incorrecto porque el engañador quiere manipular a la persona engañada para hacer algo que sirva a los intereses del engañador. ¿Funciona para engañar, pero su diseño infantil es engañoso? ¿Funciona haciendo que un niño piense que otro niño lo está ayudando? La respuesta aquí es 'No'. de autoengaño, y tampoco el modelado del robot educativo en un niño humanoide es un caso de engaño, tampoco. Identificarse con tal robot o tratar con cariño no es tratar como otro niño.

Podría tratarse como una representación de un niño. Se puede llegar a una conclusión similar sobre el Parorobot.
Este no es un caso de engaño por parte del fabricante de Paro, ni tampoco un caso de autoengaño. Los pacientes con demencia parecen usar Paro y también muñecos no robóticos de la misma manera que los niños muy pequeños usan juguetes blandos.

Los robots tienen una presencia tranquilizadora, por mucho que un perro o un gato puedan serlo. Llrazonamiento para los robots de enseñanza humanoides.

Comentario para el taller de AISB sobre principios de robótica

emily c collins

La Universidad de Sheffield, Sheffield, Reino Unido.

1. Introducción

El siguiente principio tiene como objetivo regular los robots en el mundo real: No. 4. Los robots son artefactos fabricados. No deben diseñarse de manera engañosa para explotar a los usuarios vulnerables; en cambio, su naturaleza de máquina debería ser transparente.

Este comentario ofrecerá una crítica a este principio de acuerdo con los siguientes criterios: a. Validez. ¿Son correctos los principios como afirmaciones sobre la naturaleza de los robots (por ejemplo, que son herramientas y productos), los desarrolladores de robots y la relación entre los robots y las personas (por ejemplo, que los robots deben tener un diseño transparente), o son ontológicamente defectuosos? , inexacta, desactualizada o engañosa.

La crítica dividirá el principio en lo que considero que son sus dos declaraciones de componentes principales: 1. Los robots no deben diseñarse de manera engañosa para explotar a los usuarios vulnerables.
2. La naturaleza de la máquina debe ser transparente.

Argumentaré que las dos declaraciones que componen este principio son fundamentalmente defectuosas debido a la naturaleza indefinida de los términos críticos: 'engañoso', 'vulnerable' y 'naturaleza mecánica', y que como tal el principio en su conjunto es engañosa.

Fig. 1. Panel izquierdo: 'Ekso', abreviatura de exoesqueleto, es un robot portátil que ayuda a los pacientes paralizados a caminar. Panel derecho: Dos mamíferos robot MIRO, un ejemplo de un robot 'social'.

A los efectos de este comentario, un robot se define como un artefacto manufacturado, específicamente una herramienta con la que un usuario humano puede aumentar un estado existente, por ejemplo, proporcionando a un individuo que no puede caminar la capacidad de caminar por medio de la ayuda de una máquina. o proporcionando al usuario una forma avanzada de entretenimiento, como un robot compañero (Figura 1). Este comentario se centrará en particular en la biomimética [1], los robots sociales y su papel como herramientas para la apuesta de los usuarios. Un robot social se define aquí como un dispositivo con cierta autonomía y presencia física que es capaz de interactuar socialmente con las personas y, como tal, se puede esperar que provoque una respuesta emocional de algún tipo de su usuario [2]. Aquí está nuestro primer problema, incluso antes de que se aborde el principio: para definir 'robot' uno debe al menos definir la aplicación del robot y el alcance de sus capacidades. Existen tipos de robots mutuamente excluyentes, que tienen la capacidad de engañar potencialmente a los usuarios en una variedad de formas distintas dependiendo de cómo se interactúa con cada robot. Los robots industriales, móviles, de servicios, educativos, espaciales y sociales, por nombrar solo algunos, tienen diferentes morfologías y vienen con diferentes conjuntos de expectativas de sus usuarios. Ninguno de los Principios de la robótica comienza con una definición de 'robot', por lo que he definido el mío propio.

2 Los robots no deben diseñarse de manera engañosa para explotar a los usuarios vulnerables

En primer lugar, comencemos preguntando qué es 'engañoso'. En este contexto, se está etiquetando al robot como engañoso, por lo que una mejor pregunta podría ser ¿cómo es que el robot es tan engañoso que iría en contra de este principio?

Se están desarrollando robots para parecerse a los seres vivos. Lo que se sabe sobre la dinámica humano-animal se está utilizando para incorporar comportamientos y morfologías similares a los de los animales en el diseño de robots sociales. La biomimética por definición significa diseño por naturaleza, a través de la imitación de los modelos, sistemas y elementos de la naturaleza, con el propósito de resolver problemas humanos. Los robots son herramientas, productos para su uso, cuyo propósito es resolver problemas humanos. Aquí, los mismos principios de diseño que subyacen en la naturaleza de un robot social biomimético, y lo que requieren los desarrolladores de robots, están impulsados por lo que podría llamarse 'ser engañoso': intentar imitar a los seres vivos para mejorar el robot y su usuario.

Los terapeutas están utilizando robots con apariencia de animales, como Paro [3] y 'FurReal Friends Lulu Cuddlin Kitty', producidos por Hasbro (Figura 2), de manera similar a la terapia asistida por animales (AAT) [4], en el que un animal puede incorporarse a una sesión de terapia existente para ayudar con la facilitación social (como con la terapia de grupo), o usarse de manera individual para ayudar a enfocar a un cliente o paciente durante la terapia. Estos robots tienen un propósito específico, parecerse a un animal y ayudar al terapeuta. Sin embargo, su existencia, aunque se basa en un tratamiento que involucra a un ser vivo, no reemplaza a los animales. Los animales en AAT son considerados co-terapeutas. Se les da la consideración que se esperaría que tuviera una criatura viviente, y se los retira de las sesiones en las que puede ocurrir daño.

Fig. 2. Panel izquierdo: El robot terapéutico 'Paro'. Panel derecho: Los 'FurReal Friends Lulu Cariño Gatito'.

ellos, o en el que ellos mismos están siendo perjudiciales [5]. Este ejemplo demuestra que un robot diseñado con el engaño en mente - para parecerse a un animal - con la intención de ser utilizados por poblaciones vulnerables - individuos en terapia - son no tiene la intención de ser explotador como se define al intentar convencer a un usuario de que el robot parecido a un animal está realmente vivo. En su lugar, se utilizan para desencadenar asociaciones. recuerdos de otros seres vivos. Es difícil transmitir esta idea, pero el matiz es importante. Estos robots no están hechos para ser animales convincentes. Ellos están construidos para ser herramientas robóticas convincentes, y para lograr que las ideas de la naturaleza sean prestado.

En segundo lugar, ¿qué significa 'explotar a los usuarios vulnerables'? que es un vulnerable ¿usuario? ¿Ser vulnerable es un solo estado de ser? Y si es así, ¿en qué momento podría ser considerado, o dejar de serlo, vulnerable? De hecho, ¿quién podría decidir en qué punto un individuo se volvió lo suficientemente vulnerable como para tener su Estado del ¿Robot de arte quitado de ellos? [1]

Dentro de la medicina existe una definición estandarizada de grupos vulnerables, dentro que existen dominios definidos de vulnerabilidad (por ejemplo, [7]). La forma en que los robots engañosos explotan a un individuo vulnerable depende de dónde se encuentre la vulnerabilidad del individuo. el individuo miente. Por ejemplo, los dominios médicos incluyen la vulnerabilidad económica. Considere la explotación emocional del miedo, creada por un medio populista que propaga la creencia de que el trabajo de una persona puede estar amenazado por robots que se presentan engañosamente como más avanzados de lo que son. Aunque podemos asumir que el principio no se refiere a tal vulnerabilidad como económica (aunque en verdad no podemos asumir eso; parte del problema con estos Principios de la robótica es que no se definen de esa manera en absoluto, pero para el fines de este comentario supongamos que la vulnerabilidad a la que se hace referencia es más físico que conceptual). Así que tal vez supongamos que por 'vulnerable' el principio se refiere a grupos. Supongamos también que un usuario general sabrá cuándo un robot es un robot a menos que ese robot sea tan excepcionalmente realista

[1] Para ver un ejemplo de esto en la ficción, véase la primera publicación de Issac Asimov, Robbie [6]. El temor de que los robots exploten a los vulnerables ha existido durante mucho tiempo dentro de la comunidad de robótica, pero la ficción debe ser objeto de burlas de los hechos en la apreciación de este tema.

como para pasar como vivo. Para pasar por vivo, el robot tendría que moverse, responder, parpadear, respirar y vocalizar de forma sincronizada, además de ser morfológicamente exacto, y esta lista no es exhaustiva. Tal tecnología no existe. Así, considerando el Estado del Arte que existe actualmente, como los robots sociales que son el foco de este comentario, la cuestión surge del hecho de que son precisamente los más vulnerables dentro de una población los que más tienen que ganar con su usar. Comúnmente se considera que los dos grupos más vulnerables son los ancianos y los menores, y dentro de estos grupos las personas con deterioro cognitivo.

A los efectos de este comentario, centrémonos en los grupos vulnerables dentro de la población de edad avanzada. El robot Paro antes mencionado es un robot interactivo avanzado diseñado para brindar apoyo físico y emocional a los enfermos y ancianos, no solo, sino con la ayuda de un médico clínico capacitado en Terapia Asistida por Robot (RAT). En individuos que sufren de demencia y otras condiciones de declive cognitivo, la capacidad emocional no declina de forma unívoca con la cognición [8]. Esto permite una aplicación significativa de la terapia psicológica y emocional por parte de un terapeuta con dispositivos estáticos como Paro, que está diseñado para parecerse a un ser vivo, para ser sostenido y mimado [9]. Aquí el engaño se parece al visto con la terapia con muñecas.

En las intervenciones de terapia con muñecas, los cuidadores de personas con enfermedad de Alzheimer utilizan muñecas que se asemejan a bebés reales para tratar de aliviar la ansiedad y brindar alegría a quienes padecen demencia. Esto se logra mediante la introducción de una actividad útil y gratificante, aunque físicamente inofensiva: el cuidado de la muñeca (p. ej., [10]). Aunque controvertidas [11], las terapias que introducen herramientas de puntos focales realistas en el proceso de atención han sido elogiadas por mejorar la calidad de vida (QoL) de los pacientes, y tales estudios incluyen algunos que han explorado el impacto del uso de robots similares a animales en la terapia. también [12].

QoL es una medida compleja que abarca aspectos emocionales, sociales y físicos de la vida de un individuo. Existe en un continuo, fuera del ámbito de las dicotomías, donde x se considera malo e y bueno. Si una herramienta que es robótica se usa con una población vulnerable que tiene capacidades mentales que pueden explotarse para aliviar el sufrimiento de los individuos dentro de esa población, la pregunta de si esa herramienta debería existir o no se vuelve vaga y demasiado compleja para responder. con una sola declaración. El debate se reduce a hasta qué punto deberíamos engañar a los vulnerables y en qué punto eso se vuelve explotador en un sentido negativo. Cuando esa consideración se compara con las mejoras en la calidad de vida de las personas que padecen enfermedades neurodegenerativas incurables, queda claro que este principio es insuficiente. Es fundamentalmente defectuoso porque los términos que lo componen no están definidos. Sin saber qué se entiende realmente por explotar a los vulnerables, todo el principio es engañoso.

Si lo que se explota es el deterioro cognitivo en sí mismo, y el robot se beneficia de la naturaleza vulnerable del individuo, pero con el propósito de mejorar la calidad de vida de ese individuo, ¿no es eso positivo? Cuando ahí

No hay otra alternativa para acceder a los restos del cociente emocional de una persona que sufre de demencia, alguien que de otro modo podría sentir miedo ante un animal vivo que de otro modo sería reconfortante, ¿dónde está el daño real? ¿Está el daño en la mente de aquellos que no sufren y son testigos de lo que ellos mismos consideran reflexivamente como un estado triste? Y si ese es el caso, ¿no deberíamos tratar aún más de proyectarnos en la mente de los vulnerables y apreciar esta situación por lo que es? Un intento de brindar atención utilizando todas y cada una de las herramientas disponibles, promulgada con buena voluntad y supervisada por cuidadores que conocen el alcance total del daño que causan las enfermedades neurodegenerativas, tanto para los pacientes como para sus seres queridos.

3 La naturaleza de la máquina debe ser transparente

Consideremos la población sana que observa robots. Como se dijo anteriormente en este comentario, creo que no existe una tecnología robótica que sea perfectamente engañosa. Incluso los robots más sofisticados son claramente robots. Un usuario puede creer que la IA de un robot es más avanzada de lo que es a primera vista, pero, al menos de manera anecdótica a través de mis propias experiencias en el laboratorio, creo que cualquier período de tiempo con un robot es suficiente para que un usuario establezca un valor aproximado. suficiente aproximación de sus limitaciones, de modo que cualquier sobreestimación inicial de las capacidades del robot pronto se anula con la realidad. En cuanto a aquellas poblaciones lo suficientemente vulnerables como para ser engañadas haciéndoles creer que un robot es más avanzado o está más 'vivo' de lo que es, creo que no es el robot el que debe diseñarse de manera diferente, sino que los usuarios humanos o los practicantes clínicos de RAT que deben estar capacitados para usar su herramienta, su producto robot, de la manera más efectiva y positiva.

4 Resumen

Un robot que es tan perfecto como para engañar por completo a un usuario haciéndole creer que es cualquier cosa menos una máquina, es algo que no puedo imaginar que exista en el corto plazo. Para aquellas personas que son lo suficientemente vulnerables como para estar convencidas de que un robot que es transparentemente una máquina, de hecho está vivo, mi recomendación es considerar de la manera más objetiva y amplia posible todos los beneficios positivos que pueden surgir de tal situación. Para considerar lo que realmente significa explotar a los vulnerables, y tal vez reformular un escenario con resultados aparentemente positivos para el usuario vulnerable, sin el uso del término 'explotar', sino con la palabra 'ayuda':

Los robots son artefactos fabricados, pero que son herramientas para ayudarnos y pueden diseñarse utilizando principios que sabemos que funcionan, incluidos los biomiméticos.
Los robots que están diseñados de manera engañosa, para ser utilizados para ayudar al sufrimiento de los usuarios vulnerables, deben hacer conocer su naturaleza de máquina al público.

cuidadores de esos usuarios vulnerables. Que sea responsabilidad del cuidador mejorar la calidad de vida de sus pacientes por cualquier medio seguro que sea necesario.

Referencias

1. TJ Prescott, MJ Pearson, B. Mitchinson, JCW Sullivan y AG Pipe, "Whisking with robots from rat vibissae to biomimetic technology for active touch", IEEE Robotics and Automation Magazine, vol. 16, núm. 3, págs. 42 a 50, 2009.
2. EC Collins, A. Millings y TJ Prescott, “Adjunto a la tecnología de asistencia: una nueva conceptualización”, en Actas de la 12.ª Conferencia Europea AAATE (Asociación para el Avance de la Tecnología de Asistencia en Europa), 2013.
3. T. Shibata, “Robot de compromiso mental (PARO)”. [En línea], http://www.paro.jp.
4. MR Banks, LM Willoughby y WA Banks, "Terapia asistida por animales y soledad en hogares de ancianos: uso de perros robóticos versus perros vivos", Journal of the American Medical Directors Association, vol. 9, núm. 3, págs. 173 a 177, 2008.
5. S. Brooks, "Psicoterapia asistida por animales y psicoterapia facilitada por equinos" Trabajar con jóvenes traumatizados en bienestar infantil, págs. 196–218, 2006.
6. I. Asimov, “Strange playfellow”, Super Science Stories, págs. 67–77, 1940.
7. BM Association et al., “Protección de adultos vulnerables: un juego de herramientas para practicantes”, 2011.
8. C. Magai, C. Cohen, D. Gomberg, C. Malatesta y C. Culver, “Emotional ex pression during mid-to-late-stage dementia” International Psychogeriatrics, vol. 8, núm. 03, págs. 383–395, 1996.
9. EC Collins, TJ Prescott y B. Mitchinson, "Decirlo con luz: un estudio piloto de comunicación afectiva usando el robot miro", en Biomimetic and Biohybrid Systems, págs. 243–255, Springer, 2015.
10. M. Ehrenfeld, "Uso de muñecas terapéuticas con pacientes psicogeriátricos", Terapia de juego con adultos, págs. 291–297, 2003.
11. G. Mitchell, "Uso de la terapia con muñecas para personas con demencia: una descripción general: Gary mitchell presenta los argumentos a favor y en contra de esta controvertida, pero popular, intervención", Enfermería de personas mayores, vol. 26, núm. 4, págs. 24 a 26, 2014.
12. M. Heerink, J. Albo-Canals, M. Valenti-Soler, P. Martinez-Martin, J. Zondag, C. Smits, and S. Anisuzzaman, “Explorando requisitos y robots de mascotas alternativos para la terapia asistida por robot con personas mayores adultos con demencia”, en Social Robotics, págs. 104–115, Springer, 2013.

¿Por qué mi robot se comporta así? Diseño de transparencia para la inspección en tiempo real de robots autónomos

Andreas Teodoro [1] y Robert H. Wortham [2] y Joanna J. Bryson [3]

Abstracto. Los Principios de robótica del EPSRC dictan la implementación de la transparencia en los sistemas robóticos, sin embargo, la investigación relacionada con esto está en su infancia. El artículo actual presenta al lector a la necesidad de contar con agentes inteligentes transparentes a la inspección. Nosotros proporcionar una definición robusta de transparencia, como un mecanismo para exponer la toma de decisiones del robot, considerando y ampliando sobre otras definiciones prominentes encontradas en la literatura. El documento concluye abordando las posibles decisiones de diseño que los desarrolladores deben tomar. tener en cuenta al diseñar sistemas transparentes.

1. INTRODUCCIÓN

La transparencia, en nuestra opinión, es un elemento clave relacionado con la ética implicaciones tanto del desarrollo como del uso de la Inteligencia Artificial, un tema de creciente interés público y debate. Usamos con frecuencia técnicas filosóficas, matemáticas y biológicamente inspiradas para construyendo agentes interactivos e inteligentes artificiales, pero los tratamos como cajas negras sin comprensión de cómo el tiempo real subyacente la toma de decisiones funciona.

La naturaleza de caja negra de los sistemas inteligentes, como las aplicaciones conscientes del contexto, hace que la interacción sea limitada y, a menudo, poco informativa para el usuario final [14]. Limitar las interacciones puede afectar negativamente el rendimiento del sistema o incluso poner en peligro la funcionalidad del sistema. Imagine un sistema robótico autónomo construido para proporcionar apoyo sanitario a las personas mayores. Sin embargo, las personas mayores pueden tener miedo y desconfiar del sistema. Es posible que no permitan que el robot interactuar con ellos. En tal escenario, las vidas humanas están en riesgo, ya que es posible que no obtenga el tratamiento médico requerido a tiempo, ya que un ser humano que supervisa el sistema debe detectar la falta de interacción e intervenir. Por el contrario, si el usuario humano deposita demasiada confianza en un robot, podría conducir al uso indebido, exceso de confianza y desuso del sistema [13]. En nuestro En el ejemplo del robot sanitario, si el agente funciona mal y sus pacientes no se dan cuenta de que no funciona, los pacientes pueden continuar usando el robot, arriesgando su propia salud. Los robots en ambos escenarios están rompiendo el primer Principio de Robótica de EPSRC al poner humanos vidas en peligro [1].

Para evitar tales situaciones, una adecuada calibración de la confianza entre los los operadores humanos y sus robots es de vital importancia, si no esencial, especialmente en escenarios de alto riesgo como el uso de robots en el ejército o con fines médicos [9]. Se produce la calibración de la confianza cuando el usuario final tiene un modelo mental del sistema y confía en el sistema dentro de las capacidades del sistema y es consciente de su limitación [6].

Creemos que la aplicación de la transparencia no solo es beneficiosa para usuarios finales, pero también para desarrolladores de agentes inteligentes. Tiempo real la depuración del mecanismo de toma de decisiones de un robot podría ayudar a los desarrolladores a corregir errores, prevenir problemas y explicar posibles variaciones en la actuación de un robot. Prevemos que mediante la implementación correcta de la transparencia, los desarrolladores podrían diseñar, probar y depurar sus agentes en tiempo real, similar a la forma en que los desarrolladores de software trabajar con el desarrollo de software tradicional y la depuración.

A pesar de estos posibles beneficios de la transparencia en los sistemas inteligentes, existe poca investigación existente en agentes transparentes e incluso menor implementación de agentes transparentes. Además, existen inconsistencias en las definiciones de transparencia y los criterios para un robot para ser considerado un sistema transparente. En este documento, vamos a presentar las definiciones inconsistentes encontradas en la literatura e intentar complementarlos con los nuestros. Además, en la sección tercera de este documento, discutiremos las decisiones de diseño que un desarrollador necesita a tener en cuenta al diseñar sistemas robóticos transparentes.

Usamos específicamente el término agente inteligente para denotar la combinación de software y hardware de un robot autónomo. sistema, trabajando juntos como actor, viviendo y cambiando el mundo [3]. En este documento, las palabras robot y agente se usan indistintamente.

2 DEFINICIÓN DE TRANSPARENCIA

A pesar del uso predominante de la palabra clave transparencia en el EPSRC Principios de robótica, la investigación para hacer que los sistemas sean transparentes aún está en pañales. A lo largo de los años, muy pocas publicaciones se han centrado en la necesidad de sistemas transparentes e incluso menos han intentado abordar esta necesidad. Cada estudio proporciona su propia definición de la palabra clave, sin excluir a otros. Hasta la fecha, el concepto de transparencia se ha limitado a explicaciones del comportamiento anómalo, la fiabilidad del sistema y los intentos de definir el fundamentos analíticos de un sistema inteligente.

2.1 El principio de transparencia EPSRC

Principios de Robótica de EPSRC considera la transparencia como uno de sus principios clave, al definir la transparencia en robótica como: “Los robots son artefactos manufacturados. No deben diseñarse de manera engañosa. forma de explotar a los usuarios vulnerables; en cambio, su naturaleza de máquina debería ser transparente.” .

La definición de transparencia del EPSRC hace hincapié en mantener la usuario final consciente de los fabricados, mecánicos y, por lo tanto, artificiales

[1] Universidad de Bath, Reino Unido, correo electrónico: a.theodorou@bath.ac.uk
[2] Universidad de Bath, Reino Unido, correo electrónico: rhwortham@bath.ac.uk
[3] Universidad de Bath, Reino Unido, correo electrónico: jjbryson@bath.ac.uk

naturaleza del robot. Sin embargo, la redacción utilizada permite considerar
incluso información indirecta, como documentación técnica en línea,
como una metodología suficiente para hacer cumplir la transparencia[4]. estos lugares
la carga de la responsabilidad con el usuario final. El usuario tendrá que
encontrar, leer y comprender la documentación u otra información proporcionada por el
fabricante. Algunos grupos de usuarios, como los ancianos
o usuarios no especialistas, pueden tener problemas para comprender los aspectos técnicos
términos que se encuentran a menudo en los manuales técnicos.

2.2 La transparencia como mecanismo para informar fiabilidad

Una de las primeras publicaciones definió la transparencia en términos de comunicar
información al usuario final, con respecto a la tendencia del sistema a cometer errores en
un contexto determinado [6]. Si bien la interpretación de Dzindolet es solo una parte de
nuestra definición de un sistema transparente, la
estudio presenta hallazgos interesantes sobre la importancia de la transparencia
sistemas El estudio mostró que proporcionar retroalimentación adicional a los usuarios
con respecto a las fallas del sistema, ayudó a los participantes a depositar su confianza en
el sistema. Los usuarios sabían que el sistema no era 100% confiable,
pero supieron calibrar su confianza al sistema autonómico
en el experimento, ya que se dieron cuenta de cuándo podían confiar
en él y cuando no. El uso militar de sistemas robóticos es cada vez más popular,
especialmente en forma de vehículos no tripulados.
Vehículos Aéreos (UAVs), y la transparencia en los sistemas de combate es fundamental.
Imagínese si un agente identifica un edificio civil como terrorista
y decide emprender acciones en su contra. ¿Quién es responsable? El robot
por ser poco fiable? O el capataz humano, que puso su confianza
en los sensores del sistema y en el mecanismo de toma de decisiones? Mientras que la
EPSRC Principio de Robótica considera responsable al operador humano, el daño causado
es irreversible. Robots trabajando de forma autónoma
para detectar y neutralizar objetivos es necesario tener un comportamiento transparente
[17]. Los seres humanos deben ser capaces de calibrar su confianza en el sistema y
en casos de escenarios de combate, médicos u otros donde si un robot actúa
poco confiable puede dañar o matar humanos, la transparencia como un mecanismo para
informar que la confiabilidad del sistema es fundamental.

2.3 La transparencia como mecanismo para exponer comportamiento inesperado

Estudios posteriores de Kim Hinds [11] y Stumpf et. al [14], concentrado
en proporcionar mecanismos de retroalimentación a los usuarios con respecto a imprevistos
comportamiento de un agente inteligente. En sus estudios, el usuario fue alertado
sólo cuando el agente considere que su comportamiento es anormal. Kim
y el estudio de Hinds, curiosamente, mostró que al aumentar la autonomía
también se incrementó la importancia de la transparencia como responsabilidad
pasó del usuario al robot. Sus resultados están en línea con [10]
investigaciones, que en conjunto demuestran que los humanos son más propensos
culpar a un robot por fallas que otros artefactos fabricados y
compañeros de trabajo

Ser capaz de alertar al usuario cuando el robot se comporta de forma inesperada es
fundamental para lograr la transparencia. En situaciones de alto riesgo, podría ayudar a
salvar vidas humanas o recursos valiosos al alertar
un supervisor humano del sistema para tomar el control o calibrar su confianza
respectivamente. Sin embargo, en la implementación de Kim y Hinds, el robot
alertaba al usuario solo cuando detectaba que se estaba comportando de
una manera inesperada. En nuestra opinión, esta implementación trata de arreglar una
caja negra usando otra. No hay garantía de que el robot esté teniendo problemas
inesperados sin que sepa sobre su comportamiento atípico.
La transparencia debe aplicarse en tiempo real como un mecanismo siempre activo
anismo, permitiendo al usuario decidir si el comportamiento del agente es
considerado esperado o inesperado.

2.4 La transparencia como mecanismo para exponer Toma de decisiones

Creemos que los mecanismos de transparencia deben incorporarse
al sistema, proporcionando información en tiempo real de su funcionamiento, así como
así como aportar documentación adicional según dicta el principio vigente EP SRC. El
agente inteligente, es decir, un robot, debe contener los mecanismos necesarios para
proporcionar información significativa.
al usuario final. Para considerar un robot transparente a la inspección, el
usuario final debe tener la capacidad de solicitar interpretaciones precisas de
las capacidades del robot, objetivos, progreso en relación con dichos objetivos,
Entradas sensoriales: conciencia de la situación, su confiabilidad e inesperado
comportamiento, como mensajes de error. La información proporcionada por el
El robot debe presentarse en un formato comprensible para los humanos.

Un agente transparente, con un mecanismo de toma de decisiones inspeccionable,
también podría ser depurado de manera similar a como se hace en
qué software tradicional, no inteligente, comúnmente se depura.
El desarrollador podría ver qué acciones está realizando el agente, por qué y
cómo se mueve de una acción a la otra. Esto es similar a la manera
en el que los Entornos de Desarrollo Integrado (IDE) populares brindan opciones para
seguir diferentes flujos de código con puntos de depuración,
y tener habilidades como "Step-up" y "Step-in" sobre bloques de
código.

3 DISEÑO DE SISTEMAS TRANSPARENTESz

En esta sección de este documento, discutiremos las diversas decisiones
pueden enfrentar los desarrolladores al diseñar un sistema transparente. Hasta ahora,
investigación destacada en el campo del diseño de sistemas transparentes centrada en
presentar la transparencia solo dentro del contexto de la colaboración entre humanos y
robots (HRC). Por lo tanto, se centraron en diseñar sistemas transparentes capaces de
generar confianza entre los participantes humanos.
y el robot.[12]. Creemos que la transparencia debe estar presente
incluso en entornos no colaborativos, como las competencias entre humanos y robots [11]
o incluso cuando los militares utilizan robots. En nuestro
vista, los desarrolladores deben esforzarse por desarrollar agentes inteligentes, que
puede comunicar información de manera eficiente al usuario final humano, y
secuencialmente le permiten desarrollar un modelo mental del sistema y
su comportamiento

3.1 Usabilidad

Para hacer cumplir la transparencia, se deben diseñar cuidadosamente pantallas adicionales
u otros métodos de comunicación con el usuario final.
ya que estarán integrando información potencialmente compleja. Agente
los desarrolladores deben considerar tanto la relevancia real como el nivel de
abstracción de la información que están exponiendo y cómo van a
presentar esta información.

3.1.1 Relevancia de la información

Diferentes usuarios pueden reaccionar de manera diferente a la información expuesta por
el robot. [16] demuestra que los usuarios finales sin conocimientos técnicos no entienden
ni retienen la información de los insumos técnicos, como los sensores. Esto es contrario al
desarrollador del agente, quien
necesita acceder a dicha información durante el desarrollo y la prueba del robot para
calibrar los sensores de manera efectiva y solucionar cualquier problema que se encuentre.
Sin embargo, dentro del mismo estudio, Tullio demuestra que

los usuarios pueden comprender al menos conceptos básicos de aprendizaje automático, independientemente de su formación no técnica y antecedentes laborales.

La investigación de Tullio establece un buen punto de partida para comprender qué información puede ser relevante para el usuario para ayudarlo a comprender los sistemas inteligentes. Sin embargo, se necesita más trabajo en otras áreas de aplicación para establecer tendencias específicas del dominio y del usuario con respecto a qué información se debe considerar de importancia.

3.1.2 Abstracción de información

Los desarrolladores de sistemas transparentes deberán cuestionar no solo cuál, sino también cuánta información expondrán al usuario estableciendo un nivel de complejidad con el que los usuarios pueden interactuar con la información relacionada con la transparencia. Esto es particularmente importante en los sistemas multi-robot.

Los sistemas multi-robot permiten el uso de múltiples, generalmente pequeños robots, donde un objetivo se comparte entre varios robots, cada uno con su propia información sensorial, confiabilidad y progreso hacia el desempeño de su tarea asignada para que se complete el sistema general. Los desarrollos recientes de la inteligencia de enjambre inspirada en la biología permiten el uso de grandes cantidades de diminutos robots trabajando juntos en un sistema multi-robot [15]. Los militares ya están considerando el desarrollo de enjambres de pequeños soldados robóticos autónomos. Implementar la transparencia en un sistema de este tipo no es una tarea trivial. El desarrollador debe hacer elecciones racionales sobre cuándo se requiere exponer información de bajo o alto nivel. Al exponer toda la información en todo momento, por todo tipo de usuarios, el sistema puede volverse inutilizable ya que el usuario estar sobrecargado de información. Creemos que diferentes usuarios requieren diferentes niveles de abstracción de la información para evitar la infobe sidad. Los niveles más altos de abstracción podrían concentrarse en presentar sólo una visión general del sistema. En lugar de tener el progreso de un sistema hacia un objetivo, al mostrar las acciones actuales que el sistema está tomar en relación con el logro de dicho objetivo, podría simplemente presentar una barra de finalización. Además, en un sistema de múltiples robots, la información de nivel inferior también podría incluir el objetivo, el sensor, el proceso del objetivo y la información general. comportamiento de los agentes individuales de manera detallada. Por el contrario, un La descripción general de alto nivel podría mostrar todos los robots como una sola entidad, indicando los promedios de cada máquina. Agentes inteligentes con un diseño basado en una arquitectura cognitiva, como el Diseño Orientado al Comportamiento (BOD) [2], podría presentar solo elementos del plan de alto nivel si se se necesita sistema. En el caso de un agente diseñado con BOD, los usuarios puede preferir ver e informarse sobre los estados de las unidades o Competencias pero no Acciones individuales. Otros usuarios pueden querer para ver solo partes del plan en detalle y otras partes como un alto nivel descripción general.

Una buena implementación de la transparencia debe proporcionar al usuario con tales opciones, proporcionando a individuos o grupos de usuarios potenciales Con configuraciones flexibles y preestablecidas para atender a una amplia gama de necesidades de los usuarios potenciales. Nuestra hipótesis es que el nivel de abstracción que necesita un individuo depende de una serie de factores que incluyen, entre otros, los antecedentes demográficos del usuario.

1. Usuario: Ya hemos comentado la forma en que diferentes usuarios tienden a reaccionar de manera diferente a la información sobre el estado actual de un robot Del mismo modo, podemos esperar que varios usuarios respondan de manera similar a los distintos niveles de abstracción basados en su uso del sistema. Los usuarios finales, especialmente los no especialistas, preferirá una visión general de alto nivel de la información disponible, mientras que esperamos que los desarrolladores esperen acceso a un nivel más bajo de información.

2. Tipo de sistema robótico: Como se discutió en nuestros ejemplos anteriores, un Es más probable que un sistema multi-robot requiera un mayor nivel de abstracción, para evitar la infobesidad del usuario final. Un sistema con un solo agente brillante requeriría mucha menos abstracción, ya que se muestran menos datos a su usuario.

3. Propósito del sistema robótico: El propósito previsto del sistema deben tenerse en cuenta al diseñar un agente transparente. Por ejemplo, es mucho más probable que se use un robot militar con un usuario profesional dentro o en el bucle y debido a su operación de alto riesgo, existe una necesidad mucho mayor de mostrar y capturar tanto información sobre el comportamiento del agente como sea posible. En el otro mano, una recepcionista robótica o un asistente personal es más probable para ser utilizado por usuarios no técnicos, que pueden preferir un simplificado descripción general del comportamiento del robot.

3.1.3 Presentación de la información

Los desarrolladores deben considerar cómo presentar al usuario cualquiera de los información adicional sobre el comportamiento del agente que exponer. Estudios previos utilizaron representación visual o de audio de la información. Hasta donde sabemos, no existen estudios previos que comparen los diferentes enfoques.

Los sistemas robóticos autónomos pueden tomar decenas de decisiones diferentes por segundo. Si el agente está utilizando un plan reactivo, como un plan POSH [5], el agente puede realizar miles de llamadas por minuto a los diferentes elementos del plan. Esta cantidad de información es difícil de manejar con Sistemas orientados al audio. Además, visualizar la información, es decir, proporcionando una representación gráfica del plan del agente donde los diferentes elementos del plan parpadean a medida que se llaman, deben hacer que el sistema autoexplicativo y fácil de seguir por usuarios menos técnicos. Finalmente, la visualización de un gráfico como un medio para proporcionar información relacionada con la transparencia tiene beneficios adicionales en la depuración de la aplicación. El desarrollador debe poder seguir un rastro de los diferentes elementos del plan llamados, viendo la entrada sensorial que activó ellos, hasta que se utilizó un elemento específico.

3.2 Utilidad del sistema

Hasta ahora en este documento hemos ampliado la importancia y el diseño opciones con respecto a la implementación de la transparencia. Sin embargo, nos Creo que el desarrollador también debe considerar si implementar la transparencia en realidad puede dañar la utilidad de un sistema. [18] argumenta que la utilidad de un agente se mide por el grado en que es de confianza El aumento de la transparencia puede reducir su utilidad. Esto podría, por ejemplo, tener un efecto negativo para un robot de compañía o un robot sanitario, diseñado para ayudar a los niños. En tales casos, el El sistema está diseñado contra los principios de robótica de EPSRC, ya que explota los sentimientos de sus usuarios para aumentar su utilidad y rendimiento en su tarea establecida.

Otra decisión de diseño importante que afecta al sistema es la transparencia física del sistema. La apariencia física de un puede aumentar su usabilidad [7], pero también puede contrastar con transparencia ocultando su naturaleza mecánica. Volviendo a nuestro ejemplo de robot de compañía, un robot humanoide o similar a un animal puede ser preferible a un agente donde sus mecanismos e partes internas están expuestas. revelando su naturaleza manufacturada [8].

Discutir las ventajas y desventajas entre la utilidad y la transparencia está lejos más allá del alcance de este documento. Sin embargo, los desarrolladores deben ser conscientes de esto a medida que diseñan y desarrollan robots.

4. CONCLUSIÓN

Creemos firmemente que la implementación y el uso de inteligencia
Los sistemas que son de naturaleza transparente pueden ayudar al público a comprender la IA al eliminar el aterrador misterio de por qué está teniendo ese comportamiento. La transparencia permitirá entender a los agentes
conducta emergente. En este artículo redefinimos la transparencia como un mecanismo siempre activo capaz de informar sobre el comportamiento, la fiabilidad y la sentidos y objetivos como tal información podría ayudarnos a entender el comportamiento del sistema autónomo.

Se necesita más trabajo para probar y establecer buenas prácticas con respecto a la implementación de la transparencia dentro de la comunidad de robótica. Teniendo en cuenta los beneficios de los sistemas transparentes, recomendamos
sugerir la promoción de este principio clave por parte de los consejos de investigación, como como EPSRC, y otras comunidades académicas.

AGRADECIMIENTOS

Nos gustaría agradecer a Swen Gaudl (Universidad de Bath) por su más
percepciones importantes.

REFERENCIAS

[1] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dauten Hahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorell, Mick Wallis, Blay Whitby, y Alan Winfield. Principios de la robótica. Consejo de Investigación de Ingeniería y Ciencias Físicas del Reino Unido (EPSRC), abril de 2011. publicación web.

[2] Joanna Bryson, 'El diseño orientado al comportamiento de la inteligencia de agentes modulares', en System, volumen 2592, 61–76, (2002).

[3] Joanna J. Bryson, 'Los robots deberían ser esclavos', en Close Engagements con compañeros artificiales: cuestiones sociales, psicológicas, éticas y de diseño clave, ed., Yorick Wilks, 63–74, John Benjamins, Amsterdam, (marzo de 2010).

[4] Joanna J Bryson, Darwin Caldwell, Kerstin Dautenhahn, Paula Duxbury, Lilian Edwards, Hazel Grian, Sarah Kember, Stephen Kemp, Paul Newman, Geo Peg, Andrew Rose, Tom Rodden, Tom Sorell, Mick Wallis, Shearer West, Alan Winfield e Ian Baldwin, 'La realización de los principios epsrc de la robótica', 133(133), 14–15, (2012).

[5] Joanna J. Bryson, Tristan J. Caulfield y Jan Drugowitsch, 'Integrating life-like action selection into cycle-based agent simulator environment', en Proceedings of Agent 2005: Generative Social Processes, Modelos y Mecanismos, eds., Michael North, David L. Sallach, y Charles Macal, págs. 67–81, Chicago, (octubre de 2005). Nacional de Argonne Laboratorio.

[6] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce y Hall P. Beck, 'El papel de la confianza en la dependencia de la automatización', Revista Internacional de Estudios de Computación Humana, 58 (6), 697–718, (2003).

[7] Kerstin Fischer, 'Cómo habla la gente con los robots: Diseñando el diálogo para reduce la incertidumbre del usuario', AI Magazine, 32(4), 31–38, (2011).

[8] Jennifer Goetz, Sara Kiesler y Aaron Powers, 'Combinar la apariencia y el comportamiento del robot con las tareas para mejorar la cooperación entre humanos y robots', Actas - Taller internacional IEEE sobre comunicación interactiva entre humanos y robots, 55–60, (2003).

[9] Victoria Groom y Clifford Nass, '¿Pueden los robots ser compañeros de equipo?', Inter Estudios de acción, 8(3), 483–500, (2007).

[10] Peter H. Kahn, Rachel L. Severson, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert y Nathan G. Freier, '¿La gente sostiene un humanoide? robot moralmente responsable por el daño que causa?', Procedimientos de la séptima conferencia internacional anual ACM/IEEE sobre Humano-Robot Interacción - HRI '12, (febrero de 2016), 33, (2012).

[11] Taemie Kim y Pamela Hinds, '¿A quién debo culpar? efectos de la autonomía y la transparencia en las atribuciones en la interacción humano-robot', Actas - Taller internacional IEEE sobre comunicación interactiva humana y robótica, 80–85, (2006).

[12] Joseph B Lyons, 'Ser transparente sobre la transparencia: un modelo para interacción humano-robot', Confianza y Sistemas Autónomos: Documentos de Simposio de primavera de la AAAI de 2013, 48–53, (2013).

[13] R Parasuraman y V Riley, 'Humanos y automatización: uso, mal uso, desuso, abuso', Factores humanos, 39(2), 230–253, (1997).

[14] Simone Stumpf, Weng-keen Wong, Margaret Burnett y Todd Kulesza, 'Hacer que los sistemas inteligentes sean comprensibles y controlables por usuarios finales', 10–11, (2010).

[15] Ying Tan y Zhong-yang Zheng, 'Research advance in swarm robótica', Defense Technology, 9(1), 18–39, (3 2013).

[16] Joe Tullio, Anind K. Dey, Jason Chalecki y James Fogarty, 'How funciona: un estudio de campo de usuarios no técnicos que interactúan con un sistema inteligente ', conferencia SIGCHI sobre factores humanos en la informática (CHI'07), 31–40, (2009).

[17] Lu Wang, Greg a Jamieson y Justin G Hollands, 'Confianza y dependencia en un sistema automatizado de identificación de combate', Factores humanos, 51(3), 281–291, (2009).

[18] Robert Wortham, Andreas Theodorou y Joanna J. Bryson, 'The iron triángulo: Transparencia-confianza-utilidad'. presentado, 2016.

Robot Transparencia, Confianza y Utilidad

Robert H. Wortham1 Andreas Theodorou2 y Joanna J. Bryson3

Abstracto. A medida que el razonamiento de los robots se vuelve más complejo, la depuración se vuelve cada vez más difícil basándose únicamente en el comportamiento observable, incluso para los diseñadores de robots y los especialistas técnicos. Del mismo modo, a los usuarios no especialistas les resulta difícil crear modelos mentales útiles del razonamiento de los robots únicamente a partir del comportamiento observado. Los Principios de robótica de EPSRC exigen que nuestros artefactos sean transparentes, pero ¿qué significa esto en la práctica y cómo afecta la transparencia tanto a la confianza como a la utilidad? Investigamos esta relación en la literatura y encontramos que es compleja, particularmente en entornos no industriales donde la transparencia puede tener una gama más amplia de efectos sobre la confianza y la utilidad según la aplicación y el propósito del robot. Describimos nuestro programa de investigación para respaldar nuestra afirmación de que, sin embargo, es posible crear agentes transparentes que sean emocionalmente atractivos a pesar de tener una naturaleza de máquina transparente.

1. INTRODUCCIÓN

Los Principios de robótica de EPSRC incluyen una referencia específica a la transparencia: “Los robots son artefactos fabricados. No deben diseñarse de manera engañosa para explotar a los usuarios vulnerables; en cambio, su naturaleza mecánica debería ser transparente”. ver [1]. Inicialmente, esto parece ser una afirmación normativa directa, basada en la idea común de que los agentes no deben engañar, ya que el engaño generalmente conduce a la explotación. Este documento considera si, de hecho, la transparencia es realmente una idea tan simple, y también si hacer que ciertos tipos de agentes sean transparentes reduce su utilidad. Al considerar esta pregunta, también debemos abordar la relación entre transparencia y confianza.

En este documento, usamos los términos robot y agente indistintamente y con estos términos nos referimos a un artefacto inteligente autónomo incorporado.

¿Qué significa confiar en un robot? Inicialmente, podríamos simplemente afirmar que si una IA es más transparente, entonces podemos confiar más en ella y, por lo tanto, aumenta su utilidad. También podríamos argumentar que la confianza solo se requiere cuando un agente no es completamente transparente y, por lo tanto, una mayor transparencia reduce la necesidad de confianza [4]. Si la utilidad de un artefacto se mide por el grado en que se confía en él, el aumento de la transparencia puede reducir esa utilidad. Este podría ser, por ejemplo, el caso de un robot cuya función principal es brindar compañía.

Entonces, comenzamos a ver que existe una relación compleja entre las ideas de utilidad, transparencia y confianza. Esta relación dependerá del propósito de la IA. En este documento, revisamos la literatura relacionada con la transparencia y la confianza, y también describimos la investigación práctica en curso para investigar la propuesta de que es posible construir un robot emocionalmente atractivo pero transparente.

1 Universidad de Bath, Reino Unido, correo electrónico: rhwortham@bath.ac.uk
2 Universidad de Bath, Reino Unido, correo electrónico: a.theodorou@bath.ac.uk
3 Universidad de Bath, Reino Unido, correo electrónico: jjbryson@bath.ac.uk

2 TEORÍA DE LA MENTE, CONFIANZA Y TRANSPARENCIA

Aunque podemos presuponer que la comunicación entre animales, y particularmente entre humanos, debe ser compleja, de hecho los sistemas de comunicación natural tienden a explotar señales relativamente simples y mínimas, cuyo significado se deriva de modelos extensos [16]. En otras palabras, la evolución, o una historia filogenética compartida, proporciona datos previos adecuados, de modo que se requieren datos mínimos para comunicar el contexto. Aunque algunos argumentarían lo contrario [8], generalmente se acepta que la interacción efectiva, ya sea coerción o cooperación, se basa en que cada parte tenga alguna teoría de la mente (ToM) de la otra [16, 14]. Las acciones individuales y los comportamientos compuestos se interpretan así dentro de un marco de ToM preexistente. Si esa ToM es precisa no es importante, siempre que sea predictiva en términos de comportamiento. El modelo de transparencia del robot no define la ToM empleada por el usuario humano, pero es el modelo de transparencia el que podemos ajustar directamente y, por lo tanto, este es el enfoque de este documento. Es bien sabido que la conducta observable puede comunicar los estados mentales internos del individuo. Breazeal [2] descubrió que la comunicación no verbal implícita mejora la transparencia en comparación con la comunicación no verbal deliberada. Aquí implícito se define como transmitir información inherente al comportamiento pero que el diseñador del robot no comunica deliberadamente. Las personas tienen fuertes expectativas sobre cómo las señales no verbales implícitas y explícitas se asignan a los estados mentales. Breazeal también encontró que la transparencia reduce el conflicto cuando ocurren errores, particularmente cuando se intenta una tarea conjunta.

La reducción del conflicto implica que cuando ocurre un error durante la ejecución de la tarea, la recuperación aún es posible con menos culpabilidad. Breazeal denomina Robustez a este conflicto reducido, y esta robustez es una medida efectiva de utilidad.

2.1 Antropomorfismo y Modelos Mentales de Robots

Los humanos tienen una fuerte predisposición a antropomorfizar no solo la naturaleza, sino todo lo que los rodea [5]: la hipótesis del cerebro social [7] puede explicar este fenómeno; sin embargo, los humanos no tratan a los robots de manera idéntica a los humanos, por ejemplo, con respecto a la posición moral [10]. Aunque existe un debate significativo sobre la ontología de las mentes de los robots frente a las mentes humanas, lo que tiene una importancia más práctica es cómo los humanos entienden psicológicamente las mentes de los robots, es decir, cuál es la ontología percibida, en lugar de la real. Stubbs [15] considera esencial formar un modelo mental de robots para construir un terreno común, que también podríamos interpretar como la base de la confianza humana. Stubbs [15] también encontró que este terreno común se puede establecer de manera efectiva a través de un diálogo interactivo con el robot. Aunque este estudio consideró principalmente robots remotos que trabajan en un entorno industrial o exploratorio, en lugar de robots que operan en

entornos domésticos, debemos tomar nota de la importancia del diálogo en el establecimiento de la confianza. De hecho, Mueller [13] ve el diálogo como uno de las tres características principales de las computadoras transparentes, las otras siendo explicación y aprendizaje.

Meerbeek [12] investiga la relación entre la personalidad percibida de un robot y el nivel en el que el usuario se siente en control durante la interacción. Para ser creíble, Meerbeek descubrió que la expresión de la personalidad debe estar ligada a un modelo interno que se ocupa del comportamiento (por ejemplo, toma de decisiones) basado en la personalidad y emoción El comportamiento informal más expresivo se asocia con una mayor percepción de control del usuario.

Los humanos no especialistas tienen poca ToM para los robots, o tienen un modelo basado en la ciencia ficción contemporánea y, por lo tanto, interpreta los comportamientos usando una teoría predeterminada de otro agente, que asume el agente para compartir motivaciones similares a las humanas. Esto se puede entender en términos evolutivos a través de la necesidad de nuestros antepasados de categorizar rápidamente la actividad proximal como neutral (el susurro de las hojas en el viento), amistoso (el acercamiento de un miembro de la tribu) u hostil (el acercamiento de un depredador o enemigo). Cuando la información sensorial es incierta, desarrollar un sesgo hacia una suposición tanto de agencia como de hostilidad es selectivo para la longevidad individual en un entorno donde uno es frecuentemente la presa, no el depredador. Incluso en nuestros entornos tecnológicos, a menudo experimentamos una agencia falsa, como la robótica. marcación de llamadas de ventas, publicaciones automáticas en Twitter y generación automática correos electrónicos no deseados personalizados.

En un estudio realizado en 2006 en un hospital comunitario de la UU., el personal de enfermería buscaba constantemente las razones por las que los robots actuaron como lo hicieron. Se preguntarían a sí mismos y a los demás, "¿Que esta pasando aqui? ¿Se supone que el robot debe hacer esto o lo hice yo? ¿Ocurre algo?". Esta investigación afirma que los bajos niveles de transparencia llevaron a las personas a cuestionar incluso los comportamientos normales de los demás. robot, a veces incluso llevando a las personas a pensar en comportamientos correctos como errores [11].

3 PROGRAMA DE INVESTIGACIÓN

Estamos comenzando un programa de investigación práctica para investigar el triángulo transparencia, confianza y utilidad. Inicialmente usando no humanoide robots, estamos realizando experimentos para determinar el efecto de varias expresiones de transparencia en la respuesta emocional de los humanos. En el corazón de nuestros experimentos estamos utilizando técnicas de planificación reactiva para construir agentes autónomos. Hemos desarrollado el Planificador reactivo de instinto basado en el enfoque de diseño orientado al comportamiento (BOD) de Bryson [3]. El planificador Instinct informa de la ejecución. y el estado de cada elemento del plan en tiempo real, lo que nos permite capturar implícitamente el proceso de razonamiento dentro del robot que da lugar a su comportamiento Nuestros experimentos investigarán y demostrarán cómo estos datos de transparencia del planificador se pueden utilizar para hacer más comprensible el comportamiento del robot. Inicialmente somos principalmente interesado en hacer que el comportamiento sea transparente para el diseñador del robot, ya que los robots con planes complejos suelen ser muy difíciles de diseñar y depurar. Sin embargo, estos experimentos iniciales también pueden mejorar la transparencia para los observadores no especialistas.

Posteriormente investigaremos cómo podemos aprovechar el mecanismo de transparencia incrustado con Instinct Planner para producir un robot doméstico más eficaz. La investigación investigará si la transparencia hace que las personas se sientan más o menos unidas a su robot, y si son más o menos capaces de evaluar con precisión las necesidades del robot, mientras trabaja para lograr sus objetivos.

Se prevé que estos ensayos se lleven a cabo en un entorno doméstico o casi doméstico, como una casa de retiro. Debemos obtener retroalimentación de observadores/usuarios no especialistas sobre el nivel cualitativo de inteligencia del robot, y también sobre cómo cómodos se sentirían si tuvieran un dispositivo de este tipo en su entorno doméstico. La investigación intentará evaluar los niveles iniciales de miedo, ansiedad, desconfianza hacia la IA y los robots en general, y hacia los robots domésticos En particular. Una vez establecida una posición de referencia, la transparencia del robot debe habilitarse proporcionando retroalimentación al usuario basado sobre la ejecución en tiempo real dentro del planificador reactivo. Los métodos actualmente prevemos son:

- Presentación en tiempo real de declaraciones textuales relacionadas con plan exe corte
- Visualización gráfica en tiempo real de la ejecución del plan.
- Declaraciones de audio (es decir, verbales) relacionadas con la ejecución del plan del robot.

Para cada uno de estos métodos, la información de transparencia podría presentarse en/desde un dispositivo remoto, o en/desde el propio robot. Por lo tanto, hay seis combinaciones posibles. Por supuesto, la fusión de transparencia adicional, como audio combinado con gráficos, también podría ser probado en base al éxito o fracaso de los resultados experimentales iniciales.

Como la literatura indica que el diálogo es importante para establecer confianza, esta investigación debería considerar la posibilidad de aceptar la entrada de voz, aunque restringida a comandos simples, como un medios para que los usuarios pregunten al robot qué está haciendo, y tener el robot responde apropiadamente.

4. DISCUSIÓN

El Principio 1 de EPSRC afirma que los robots son herramientas. Dentro de la industria y entornos de ingeniería esto es bastante claro, en el sentido de que un ser humano usa el robot para completar una tarea técnica. El diseñador y El usuario del robot comparte el objetivo del robot: completar la tarea. Sin embargo, dentro de entornos domésticos y sanitarios, los robots pueden tienen una relación bastante diferente con aquellos con los que interactúan. Ellos puede tener la intención de proporcionar compañía y encubrimiento simultáneo seguimiento del bienestar del paciente. Pueden ser herramientas para el profesional de la salud, pero para el paciente son compañeros. De tal En un entorno, la empresa de servicios públicos puede verse afectada negativamente por un aumento transparencia. Nuestro sentido de compañerismo está relacionado con la medida de agencia que proyectamos sobre el robot. Si somos capaces de comprender la funcionamiento de la inteligencia parece inherentemente volverse menos inteligente en el sentido popular, de modo que entonces proyectamos menos agencia, y como resultado experimenta menos beneficio del robot? podríamos comparar esto con la televisión. Sabemos que no tiene agencia, pero su presencia en la esquina de nuestra sala de estar brinda beneficios similares a los de un compañero. Tal vez esto tenga que ver con la suspensión consciente de la incredulidad, o tal vez tengamos un detector de agencia inconsciente que es más fácil engañado por la tecnología.

Las nociones de inteligencia del sentido común se fusionan con las ideas de la psicología popular sobre la agencia y también sobre la vida. cosas que son inteligentes están vivos, en el sentido de que tienen sus propias creencias, deseos e intenciones que entendemos son fundamentalmente egoístas o egoístas. Implícitamente reconocemos el egoísmo como una característica fundamental de toda la vida [6]. Si dicho agente se relaciona con nosotros, entonces nos considera importante en la búsqueda de estos objetivos egoístas. Tales agentes son dignos de convertirse en nuestros compañeros porque les atribuyen verdadero valor en su relación con nosotros, y esto aumenta nuestro valor en la sociedad. Por el contrario, los agentes que no tienen una agencia de autoservicio no son dignos de nuestra atención porque no transmiten ningún valor social. Quizá por lo tanto, los agentes artificiales cuyo único propósito es el compañerismo y son verdaderamente transparentes a este respecto quedan así descalificados de ser dignos compañeros. Por lo tanto, en algunas situaciones, la transparencia del robot puede

estar en desacuerdo con la utilidad, y más generalmente puede ser ortogonal en lugar de que beneficioso para el uso exitoso del robot. Si bien podemos inventar escenarios y continuar discutiendo la interacción teórica y filosófica entre transparencia, confianza y utilidad, como científicos
esperar el resultado de nuestros experimentos.

5. CONCLUSIÓN

Hemos visto que desentrañar la transparencia y la confianza es complejo, pero puede entenderse en parte observando cómo los humanos llegan a comprender y, posteriormente, a confiar unos en otros, y cómo superan
miedos evolutivos para confiar en otros agentes, a través de
comunicación no verbal. Niveles inaceptables de ansiedad, miedo y
la desconfianza resultará en una respuesta emocional y cognitiva de rechazo robots Hancock [9] afirma que si no podemos confiar en nuestros robots, lo haremos no poder beneficiarse de ellos de manera efectiva. Sin embargo, dado que nosotros interactuar felizmente en la sociedad con otros a quienes no conocemos completamente confianza, y cada vez más interactuamos con las computadoras sabiendo que su recomendaciones tal vez defectuosas, debemos concluir que Hancock es simplificando demasiado. Por último, puede haber aplicaciones en las que la transparencia esté reñida con la utilidad. Nuestro programa de investigación en curso tiene como objetivo validar nuestra hipótesis de que podemos crear robots transparentes que, sin embargo, sean emocionalmente atractivos y útiles
herramientas en una amplia gama de entornos domésticos y casi domésticos. Mientras tanto, queda mucho trabajo por hacer para desentrañar la relación entre transparencia, utilidad y confianza.

REFERENCIAS

[1] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dauten Hahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorell, Mick Wallis, Blay Whitby, y Alan Winfield. Principios de la robótica. Consejo de Investigación de Ingeniería y Ciencias Físicas del Reino Unido (EPSRC), abril de 2011. publicación web.

[2] C. Breazeal, CD Kidd, AL Thomaz, G. Hoffman y M. Berlin, 'Efectos de la comunicación no verbal sobre la eficiencia y robustez en trabajo en equipo humano-robot', en la Conferencia Internacional IEEE/RSJ de 2005 on Intelligent Robots and Systems, págs. 708–713, Alberta, Canadá, (2005). Ieee.

[3] Joanna J. Bryson, 'Inteligencia por diseño: principios de modularidad y coordinación para la ingeniería de agentes adaptativos complejos', (2001).

[4] Joanna J Bryson y Paul Rauwolf, 'Trust, Communication, and In igualdad'. 2016.

[5] Kerstin Dautenhahn, 'Metodología y temas de interacción humano-robot: un campo de investigación en crecimiento', International Journal of Advanced Robotic Systems, 4(1 SPEC. ISS.), 103–108, (2007).

[6] Richard Dawkins, 'Organización jerárquica: un principio candidato para ethology', en Growing Points in Ethology, eds., PPG Bateson y RA Hinde, 7–54, Cambridge University Press, Cambridge, (1976).

[7] RIM Dunbar, 'La hipótesis del cerebro social', Evolutionary Anthropol ogía, 178–190, (1998).

[8] Shaun Gallagher, 'La alternativa narrativa a la teoría de la mente', en Radical Enactivism: Intentionality, Phenomenology, and Narrative, ed., R Menary, número Gallagher 2001, 223–229, John Benjamins, Amsterdam, (2006).

[9] P. a. Hancock, DR Billings, KE Schaefer, JYC Chen, EJ de Visser y R. Parasuraman, 'Un metaanálisis de los factores que afectan Trust in Human-Robot Interaction', Factores humanos: The Journal of the Sociedad de Ergonomía y Factores Humanos, 53(5), 517–527, (2011).

[10] Peter H. Kahn, Hiroshi Ishiguro, Batya Friedman y Takayuki Kanda, '¿Qué es un ser humano? - Hacia referentes psicológicos en el campo de la interacción humano-robot', Actas - Taller internacional IEEE on Robot and Human Interactive Communication, 3, 364–371, (2006).

[11] Taemie Kim y Pamela Hinds, '¿A quién debo culpar? Efectos de la autonomía y la transparencia en las atribuciones en la interacción humano-robot', Actas - Taller internacional IEEE sobre comunicación interactiva humana y robótica, 80–85, (2006).

[12] Bernt Meerbeek, Jettie Hoonhout, Peter Bingley y Jacques Terken, 'Investigando la relación entre la personalidad de un televisor robótico asistente y el nivel de control del usuario', Actas - IEEE International Workshop on Robot and Human Interactive Communication, 404–410, (2006).

[13] Erik T. Mueller, Computadoras transparentes: diseño comprensible Sistemas Inteligentes, Erik T. Mueller, San Bernardino, CA, 2016.

[14] Rebecca Saxe, Laura E Schulz y Yuhong V Jiang, 'Reading minds frente a seguir las reglas: disociar la teoría de la mente y el control ejecutivo en el cerebro', Social neuroscience, 1(3-4), 284–98, (enero de 2006).

[15] Kristen Stubbs, Pamela J Hinds y David Wettergreen, 'Autonomy y terreno común en la interacción humano-robot: un estudio de campo', Sistemas inteligentes IEEE, 22(2), 42–50, (2007).

[16] Robert H Wortham y Joanna J Bryson, 'Communication', en Handbook of Living Machines {en prensa}, Oxford University Press, Oxford, (2016).