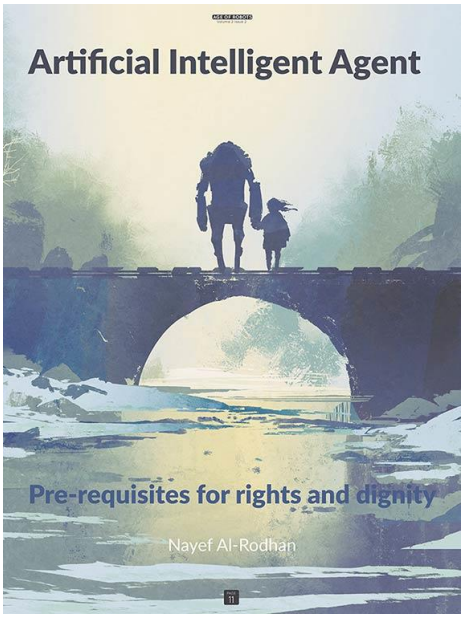
Agentes de Inteligencia Artificial: prerequisitios sobre derechos y dignidad

por Nayef Al-Rodhan

Autor del artículo destacado

tamaño del texto Tamaño de texto más grande Tamaño de texto más pequeño

Deep Blue de IBM era un ordenador "ajedrecista" que en 1997 logró un éxito destacable: derrotó al campeón del mundo Gary Kasparov en 19 jugadas. Kasparov nunca había perdido una partida contra otro ser humano en menos de 20 jugadas. Intentó derrotar a Deep Blue en otras partidas pero volvió a perder al año siguiente, tras una actualización del programa conocida con la denominación no oficial de "Deeper Blue". Fue todo un hito en la historia de la inteligencia artificial, pero en ningún momento se consideró que la genial máquina de jugar al ajedrez mereciera tener "derechos". Si bien teóricamente era capaz de visualizar 200 millones de jugadas de ajedrez por segundo, Deep Blue tenía capacidades generales limitadas y no podía realizar otras tareas más allá de aquellas para las que había sido programado, en este caso, jugar al ajedrez.



No obstante, cuando los robots comenzaron a hablar y a interactuar con los seres humanos (incluso creando arte y formas de expresión más sofisticadas incluyendo la música) quedó claro que **sus funciones no solo iban a multiplicar las capacidades humanas, sino que comenzaban a competir directamente con las mismas.**

En 2015, **AlphaGo** de Google derrotó a un jugador humano de Go de primera categoría. El Go es un antiguo juego de estrategia oriental que implica el dominio de conocimientos que se consideran exclusivos del ser humano (por ej., la intuición). Más recientemente, en un documento publicado en *Nature*, investigadores de DeepMind explicaron cómo se actualizó AlphaGo mediante un algoritmo basado en **aprendizaje de refuerzo, que permite que un ordenador aprenda por sí mismo sin intervención humana y, en términos efectivos, se convierta en su propio maestro.**

De manera cada vez más evidente, la pregunta ya no radica solo en el carácter instrumental de los robots sino que se empieza a considerar a los robots como verdaderos pares de los seres humanos que merecerían tener derechos y dignidad. En solo una década, las respuestas humanas a los robots han pasado de la mera curiosidad y el divertimento -contemplándolos como artilugios inteligentes-, a la cautela y la alarma y, más recientemente, se comienza a contemplar a los robots como entidades que merecen obtener derechos de ciudadanía. La reciente concesión por parte de Arabia Saudí de la ciudadanía a Sofía, un robot humanoide, puede parecer un título honorífico y simbólico, pero suscita nuevas preguntas acerca de las relaciones de los seres humanos con los robots inteligentes.

Humanos y robots: del paternalismo al igualitarismo

El actual dilema acerca de la concesión de derechos a los robots está precedido por un largo debate sobre las interacciones entre los seres humanos y los robots cuyos orígenes se remontan a medio siglo atrás. En el ínterin, algunas teorías y estudios antropológicos fascinantes han descrito esta relación ambivalente.

Una de las primeras hipótesis fue formulada en 1970 por el profesor de robótica japonés **Masahiro Mori**, que propuso la posibilidad de describir la interacción entre seres humanos y androides, es decir robots con apariencia humana. Según esta teoría, **los objetos humanoides de aspecto imperfecto que resultan similares e incluso distintas réplicas del aspecto humano, provocarán desagrado y una extraña sensación de repulsión en los observadores humanos.** Algunos experimentos recientes comprobando la validez de la teoría del Valle Inquietante han demostrado que en un espectro de aspectos robóticos, a medida que los rostros se acercaban al aspecto humano y resultaban menos mecánicos, eran percibidos como elementos más desagradables; sin embargo, **cuando los rostros de los robots tenían un aspecto cuasi-humano, la simpatía hacia los mismos aumentó rápidamente**, si bien de forma precaria, ya que incluso fallos de carácter mínimo perturbaban la interacción social.

Si bien ciertos estudiosos la refutan, la teoría del Valle Tenebroso ha alentado algunas hipótesis provocativas acerca de la manera en que cabe fomentar las interacciones y la confianza entre los seres humanos y los robots.

Nuevas investigaciones realizadas en universidades de Japón complementan algunas de las premisas de la tesis del Valle Inquietante, mostrando que **los seres humanos incluso pueden llegar a empatizar con los robots que experimentan "dolor", casi en la misma medida en que lo harían si se tratara de seres humanos que experimentan dolor.** No obstante, posteriores investigaciones han determinado que los escaneos de electrocardiogramas realizados a los participantes del estudios mostraron intensidades más débiles de los potenciales cerebrales en sus respuestas al dolor que experimentaban los robots que las que se desencadenaban ante el dolor que afectaba a los seres humanos. (No queda claro de qué forma el sentido de empatía podría expresarse para los robots que no tienen una naturaleza "humanoide").

Esta investigación sugeriría que cuando antropomorfizamos los robots en mayor medida, mejor podríamos imaginar la posibilidad de comprender sus perspectivas y, de esta manera, podríamos llegar a generar confianza hacia ellos o sentir empatía.

Es probable que esta explicación de la **empatía de arriba a abajo** sea de hecho consecuencia de un conjunto de factores más complejo que conforma nuestras relaciones con los robots. Kate Darling, investigadora del MIT, ha sugerido que los vínculos sociales se deben fundamentalmente a tres factores: **la cualidad física** (cuando los robots existen en *nuestro espacio*, no en una pantalla), **la autonomía de movimientos percibida y el comportamiento social** (pueden comunicarse con los seres humanos). En un experimento que realizó con otros colegas, se pidió a los participantes que golpearan a un grupo de pequeños robots hasta "la muerte". **La aversión que provocaba el abuso hacia los robots resultó manifiesta.** Como seres humanos, sabemos racionalmente que los robots no tienen dignidad en términos intrínsecos, pero sentimos empatía porque en ellos vemos cierto reflejo

de nosotros mismos y, de esta manera, de nuestros propios temores, como es el caso del miedo al dolor.

En otras palabras, un robot no puede evocar los sentimientos meramente neutrales de fastidio que uno podría experimentar cuando se avería la tostadora. / Imagen: ***Pixabay***

Las características únicas de los robots han reforzado **la idea de que merecen que se los considere desde una perspectiva totalmente distinta.** Asimismo, a medida que el aprendizaje de máquina progresa y, de su mano, la inteligencia y la autonomía de los robots que nos rodean, las preguntas que rodean las responsabilidades de los robots se complejizarán cada vez más. Una forma de abordar este problema ha consistido en compararlo con los derechos de los animales, sobre los cuales los filósofos y los especialistas en ética han reflexionado durante siglos.

Por ejemplo, ¿debemos extender los derechos a los robots conformando una lógica similar a la que se emplea en el caso de los derechos de los animales? Alguna preguntas de carácter filosófico emergen instantáneamente, por ejemplo: ¿acaso los seres humanos deberíamos reaccionar como tutores responsables? Pero esta forma de pensar ha pasado a la historia, incluso en lo que respecta a los derechos de los animales. Presupone la idea de que los seres humanos son “amos superiores” cualificados y capaces de tomar decisiones acerca de quién merece vivir y quién no.

No obstante, el paralelismo con los derechos de los animales tiene cierto valor en lo que respecta al hecho de que la concesión de derechos y la asignación de personalidad jurídica van unidos al deseo de vincular otros agentes con la humanidad.

Si bien sabemos que los robots no sienten dolor y no tienen conciencia, aun así la comparación con los derechos de los animales resulta relevante ya que nos prepara para un futuro en el que estaremos vinculados más estrechamente con los agentes pertenecientes al campo de la IA.

Robots morales: los derechos deben estar vinculados a la moralidad

Los robots no tienen una vida “biológica” ni experimentan sentimientos… al tiempo que tampoco tienen la capacidad de reproducirse. Pueden enseñarse a sí mismos a jugar al ajedrez pero no pueden autosustentarse más allá de la voluntad y el esfuerzo de los fabricantes y los ingenieros humanos. Esto en sí constituye una razón de peso suficiente para justificar por qué actualmente **seguimos pensando en los robots como esclavos mecánicos y no como en entidades por derecho propio.**

Asimismo, esta forma de pensar se refleja en la legislación que cubre la responsabilidad en este campo. En la UE, la responsabilidad jurídica en lo que respecta al daño provocado por robots recae en el fabricante. Lo mismo ocurre con los daños previsibles derivados de cualquier defecto de fabricación. No obstante, **la UE está empezando a reconocer cada vez más que cuando se trata de robots, hará falta un enfoque centrado en la "personalidad". Y ese momento no está demasiado lejos.** Si los robots se complejizan cada vez más, hasta el punto de que puedan tomar decisiones morales de manera instantánea, incluyendo decisiones sobre la vida y la muerte, es posible que resulte más apropiado hacer uso de una **nueva noción de "persona electrónica".**

Estas ideas fueron elaboradas en un estudio realizado para la comisión JURI del Parlamento Europeo titulado "Normas de derecho civil europeo en robótica", en el que se señaló que la creciente presencia de robots autónomos provocará una escisión en los valores de la sociedad (se tratará de una escisión tan profunda que no puede compararse con las divisiones causadas por Internet y las tecnologías digitales). Constituirá un elemento transformador en un sentido existencial. Los autores de este estudio aconsejaron tener una precaución extrema**, yendo un paso más allá y difuminando la línea entre "la materia viva y la inerte", lo que podría causar la destrucción de los fundamentos humanistas europeos.** Según concluye el estudio, no resultaría adecuado asignar un estatus de persona a un robot ya que esto supondría rebajar de categoría a la humanidad. El único propósito de un robot debería ser prestar ayuda a los seres humanos.

No obstante, si la tecnología progresa conforme a lo establecido en la Ley de Moore, resultará acuciante efectuar un análisis sobre la cuestión de los derechos. Por el momento, aún podemos escoger si debemos debatir sobre los derechos de los robots o no, pero puede que en escasas décadas esta "libertad" se haya esfumado.

Resulta difícil identificar cuándo llegará el momento (y muchos escenarios se basan en predicciones poco realistas), pero parece bastante seguro sugerir que, a medida que avance el siglo, resultará claro hacia dónde se encamina la tecnología.

Las preguntas de carácter verdaderamente existencial aparecerán **cuando la tecnología permita realizar una transición desde una moralidad de arriba hacia abajo** (lo que significa que el programador introduce valores morales en el robot) a una moralidad de abajo hacia arriba, donde los robots puedan aprender a tener competencias morales a través de un proceso de socialización en sus respectivos entornos, de forma similar a como los seres humanos alcanzan una orientación moral. Asimismo, esto significará que están fuera de la tutela humana y que son impredeciblemente libres en lo que respecta a la adquisición de valores morales. Lo que nos separa de ese momento es la capacidad tecnológica que nos permitiría que los robots tuvieran la posibilidad de aprender de esa manera.

Sin embargo, recientes avances en tecnología neuromórfica prometen que la humanidad se acerque aún más a esta perspectiva. Los chips neuromórficos (o con forma similar a las neuronas) permiten **emular la arquitectura neuronal del cerebro y la forma especial en que funcionan las neuronas a través de conexiones intrincadas y no de forma lineal (en oposición a la informática de ceros y unos).** Los robots equipados con dichos chips podrían aprender y podrían procesar información de manera similar a como lo hacen los seres humanos.

Las consecuencias que se derivarían de esta tecnología podrían constituir un punto de inflexión en el debate acerca de la moralidad y los derechos de los robots. **Los robots no solo aprenderán como los seres humanos sino que sus principios morales se desarrollarán de manera interactiva en virtud de la cooperación social con sus "homólogos" humanos.** ¿Qué moralidad heredarán y a qué valores se adherirán? Previamente, he expuesto la teoría de que los seres humanos funcionan como una tabula rasa *predispuesta*, es decir, nuestros valores se conforman mediante la educación y el entorno pero estamos mentalmente programados de manera fundamental (nuestro instinto de supervivencia). Los robots que "aprendan haciendo" estarán influenciados por el entorno de manera similar e incluso pueden adquirir un instinto de autopreservación, incluso a expensas de los seres humanos. Luego, podría resultar indispensable un sistema de derechos y responsabilidades para evitar una situación de "estado de naturaleza".

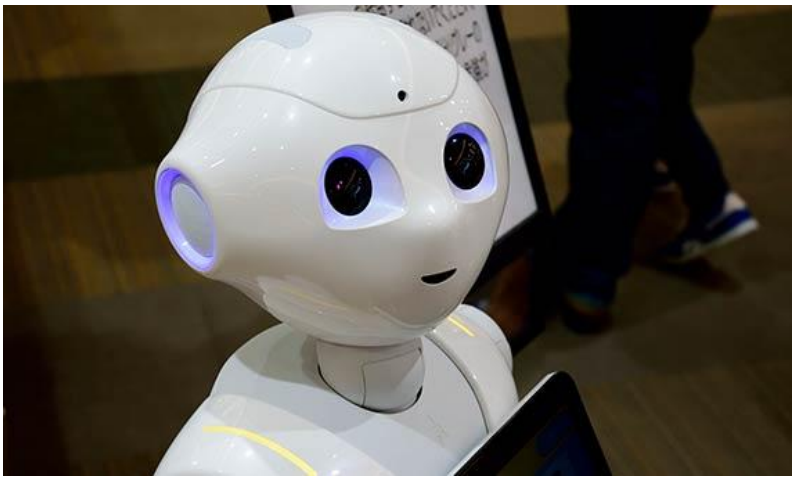
Derechos y dignidad

Para Thomas Hobbes, los derechos eran elementos cruciales para la propia vida: una vida sin derechos legales era una vida solitaria y en constante peligro. Para que los robots adquirieran derechos deberían temer por sus vidas (es decir, deberían tener consciencia y autoconsciencia), al tiempo que necesitarían protección ante otras entidades, seres humanos o agentes de IA que resultaran hostiles. Esto solo podría suceder cuando la tecnología avanzara hasta tal punto que los robots no necesitaran un operador humano. Por otro lado, si los robots fueran a adquirir tanto la inteligencia como los medios para provocar daños y desestabilizar el orden político y social, vivirían efectivamente en un estado de naturaleza dentro de Leviatán, es decir dentro de los sistemas de soberanía y de leyes que sustentan a los estados modernos pero que actualmente solo resultan vinculantes para los seres humanos. **Luego deberíamos pensar en los robots como actores jurídicos, con responsabilidad legal en lo que respecta a sus acciones.**

En esta etapa, resulta imposible construir un caso sólido en lo que respecta a los derechos de los robots. Hasta el momento, **nada parece indicar que los robots merezcan derechos y reconocimiento jurídico como entidades individuales**, pero esto es solo así porque la tecnología aún no se ha desarrollado lo suficiente como para que sean totalmente autónomos. Según Kant, los derechos emanan de la voluntad racional y los robots no poseen dicha *voluntad racional*, aun cuando pueden realizar asociaciones y llegar a conclusiones lógicas.

Cuantificar las capacidades que deberían tener para que concedamos derechos a los robots constituye una tarea difícil y compilar un listado de los criterios también resultaría problemático: los seres humanos no son uniformemente similares en lo que respecta a sus capacidades y aun así poseen derechos. Quizás podríamos resolver la disputa pensando acerca de estos criterios en términos más generales y más neuro-filosóficos. Basándome en perspectivas que proceden de la neurociencia, describí anteriormente una teoría de la naturaleza humana como "emocional, amoral y egoísta". **Lo que nos hace únicos como especie es: en primer lugar, nuestro carácter emocional**(somos seres mucho más "emocionales" de lo que pensamos y nuestras emociones participan en nuestro proceso de toma de decisiones); en segundo término, nuestra "amoralidad" (no somos ni "morales" ni "inmorales"); **nuestros principios morales se construyen a lo largo de nuestra existencia** y tercero, **nuestro egoísmo** (estamos predispuestos en una dirección fundamental: nuestra necesidad de sobrevivir, lo cual en una forma básica de egoísmo). Los agentes de IA necesitarán poseer prerrequisitos similares para así merecer derechos: en primer término, una

capacidad de sentir y mostrar emociones; en segundo lugar, una capacidad de realizar elecciones morales y asumir responsabilidades y, en tercer término, una capacidad de autoconsciencia y egoísmo personal.



Si se cumplen dichos prerrequisitos, los agentes de IA tendrían atributos emocionales, amorales y egoístas, que actualmente solo poseen los seres humanos. Esto significaría también que las sociedades tendrían que garantizarles las condiciones por las que he abogado anteriormente y que resultan fundamentales para lograr una buena gobernanza inclusiva, a saber: razón, seguridad, derechos humanos, responsabilidad, trasparencia, justicia, oportunidad, innovación e inclusión. / Image: ***Pixabay***

Estas nueve necesidades en relación con la dignidad resultan esenciales para limitar las tensiones entre los atributos emocionales, amorales y egoístas propios de la naturaleza de los robots con aspecto humano, y garantizar sociedades futuras sostenibles y estables.

No obstante, el debate continúa, ya que se sitúa en el corazón de lo que hace que los seres humanos merezcan tener derechos en primer término. Ciertamente, existen otras entidades que tienen personalidad jurídica, pero no son seres humanos: las corporaciones en los Estados Unidos, por ejemplo o ciertas entidades naturales en la India… Pero conceder derechos a los robots serían algo distinto, ya que los robots (a diferencia de lo que sucede con las corporaciones o los ríos) tienen inteligencia y talentos sociales, lo que supone que las implicaciones resultan mucho más profundas.

Lo que resulta necesario para garantizar que no alcancemos ese estadio, que conduciría a un colapso de nuestra civilización, es abogar por una investigación responsable y una comprobación ética a fondo en lo que respecta al desarrollo de la inteligencia artificial.

Prof. Nayef Al-Rodhan

6 de marzo de 2018 | El texto original se encuentra aquí.