EL PROBLEMA DE IDENTIDAD PERSONAL: CASTIGOS Y RECOMPENSAS.

Seguridad, Privacidad y Aspectos Legales.

Curso 2021/2022

Facultad de Ciencias.

Máster en Ciencia de Datos.

Autor: Jesús Octavio Raboso.

Santander, 11 de abril de 2022.

Works of art make rules; rules do not make works of art.
Claude Debussy (1862-1918)

The unpredictable and the predetermined unfold together to make everything the way it is.
Tom Stoppard (1937-)

It is not a human move. I have never seen a human play this move. So beautiful.
Fan Hui (1981-)

Índice general

Introducción.

Cuando vemos una cosa en un instante y espacio determinados, tenemos la certeza, sea cual sea la cosa, de que es la misma cosa que vemos y no otra que, al mismo tiempo, exista en otro lugar por semejante e indistinguible que pueda ser en todos los demás aspectos.

Precisamente en eso consiste la identidad, en aquellas ideas que atribuimos a las cosas y que no varían en nada desde el momento en el que consideramos la existencia previa de dicha cosa.

La discusión filosófica sobre la identidad personal se refiere, fundamentalmente, a la identidad de una persona a través del tiempo y el espacio. Pretende dilucidar qué criterio ha de aplicarse para afirmar que una persona en un instante y espacio determinados se corresponde con la misma persona en otro instante y espacio. Consideremos la siguiente situación:

Cuando llega la noche, para descansar de los agobios del día, me siento en la poltrona y fijo la mirada en la vieja foto en blanco y negro que cuelga de la pared del salón. En ella, veo a un niño con su madre, posando ante un paisaje marítimo. De vez en cuando me hundo en la perplejidad: ese niño soy yo. ¿Cómo es posible? Lo sé porque me lo llevan diciendo desde hace años, pero ¿qué tiene que ver ese niño conmigo? No nos parecemos físicamente y, con toda seguridad, nuestros pensamientos y actitudes ante la vida no tienen nada en común. ¿No podía yo ser, o mejor aún, haber sido otro niño? ¿Qué nos hace a ese niño de diez 10 años y a mí, de más de 20, la misma persona?

Cabe preguntarse: ¿existen tests que puedan aplicarse para determinar la identidad personal a través del tiempo? ¿No los usan desde hace mucho tiempo las autoridades policiales o judiciales? ¿Cuándo y por qué podremos decir que el veinteñero que descansa es, o no, la misma persona que el niño de la foto?

Pero, el empeño filosófico en desentrañar el concepto de *identidad personal* no nace sólo por el propio interés de aprender algo importante sobre nosotros mismos sino por la vertiente práctica: importa mucho saber en qué consiste ser la misma persona -y no el mismo ser humano u objeto físico- ya que, desde los puntos de vista ético y jurídico, resulta decisivo poder determinar las condiciones que una persona ha de cumplir para ser considerada agente responsable y, por tanto, ser objeto de la atribución de recompensas y castigos.

En el actual mundo tecnológico cabe preguntarse si sólo pueden atribuirse recompensas y castigos a las personas, ¿no podríamos hacer lo propio con los

algoritmos? La teoría de la Inteligencia Artificial fuerte plantea un mundo de posibilidades que no podemos dejar de considerar.

Capítulo 1

La identidad personal como problema filosófico.

1.1. La continuidad de la sustancia material.

Algunos criterios sobre la persistencia de la identidad personal a través del tiempo se basan en el hecho de tener una existencia material continua.

En primer lugar, abordaremos estos criterios para objetos físicos e inanimados.

Teseo ha navegado por los mares del mundo durante un año. Al concluir su travesía, observa que su nave se ha deteriorado gravemente. Por ello, la sitúa en un dique seco para repararla. Pero la reparación, que le lleva un año completo, es más seria de lo que parecía a priori. Teseo ha sustituido todas las piezas del buque por otras exactamente iguales a las anteriores. Concluida la reparación, Teseo vuelve al mar. Pero, lo hace en compañía, pues, mientras reparaba su nave e iba desechando sus piezas, un rival las restauraba una a una y ha construido un buque exactamente igual al de Teseo con el que se lanza al mar.

Esta paradoja, propuesta por Plutarco (s. I - s. II), plantea varias cuestiones: ¿cuál de las dos naves, la que dirige Teseo o la que dirige el rival, se corresponde con la nave antigua?, ¿ninguna, ambas, la que dirige Teseo, o la que dirige su rival?

Los interrogantes aumentan al tratar con personas. Quizá, el sentido común parezca indicar que los juicios de identidad se basan en lo que llamaremos *criterio corporal*: el mismo cuerpo es la misma persona. Si bien, no podemos esperar mucho de este criterio. ¿Puede servir el cuerpo físico para dar cuenta de la problemática del veinteañero y el niño de la fotografía? En general, se reconoce que el cuerpo de las personas sirve como criterio de identidad personal, pero más bien lo hace como criterio de evidencia y no como criterio constitutivo. Es decir, que el cuerpo sea el mismo, normalmente, es indicio de que la persona es la misma. Si bien, como ya se ha dejado entrever, los criterios de identidad personal no tratan de establecer qué va a aceptarse como evidencia de la identidad personal sino en qué consiste la identidad personal. Las huellas dactilares parecen ser una evidencia de identidad personal, pero nadie zanjaría la cuestión alegando que

las huellas dactilares son en sí la identidad de una persona. Continuaremos con el ejemplo de las huellas dactilares y las manos en adelante.

Derek Parfit (1942-2017) plantea el siguiente experimento mental para sacar a la luz las intuiciones sobre la continuidad corporal.

He estado en Marte, pero sólo por el método clásico, un viaje espacial que dura varias horas mientras estás acomodado en un teletransportador. Esta máquina me envía hasta allí a la velocidad de la luz tras pulsar un botón verde. Como es habitual, estoy muy nervioso. ¿Funcionará? Recapitulo los consejos e instrucciones que me han indicado. Cuando apriete el botón verde, perderé la conciencia y, tiempo después, me despertaré con la sensación de que ha pasado sólo un instante. Sin embargo, habré estado inconsciente un par de horas. El escáner situado en la Tierra, destruirá mi cuerpo y mi cerebro a la vez que graba los estados de todas mis células. Acto seguido, transmitirá esta información al Replicador de Marte a la velocidad de la luz. Este Replicador creará, a partir de materia nueva, un cerebro y un cuerpo exactamente iguales a los míos. Aprieto el botón. Según lo previsto, pierdo la conciencia y parece que la recupero en seguida, pero en un cubículo diferente. He llegado a Marte. Examino mi nuevo cuerpo y no encuentro ningún cambio físico.

Esta historia nos muestra que quizá la continuidad corporal puede no ser un criterio para juzgar la identidad de una persona a través del tiempo sino sólo una evidencia importante pero falible. Quizá dé a entender que se prefiere un criterio aristotélico, es decir, que la intuición sobre la identidad personal apunta a cierta estructura inteligible -no material- que, en este ejemplo, se correspondería con la información que se ha grabado y teletransportado. O bien, que el criterio de continuidad corporal ofrecido por el sentido común se revela, simplemente, como falso.

Si la identidad personal se halla constituida por la identidad del cuerpo, entonces la identidad personal no se distinguiría en lo esencial de la identidad de los objetos físicos no animados, que vendría dada por su continuidad espacio-temporal. Por ello, el criterio corporal no puede exigir la continuidad material estricta -de manera similar a la nave de Teseo, en todo ser vivo se destruyen células degradándolas hasta sus sillares más básicos y, a partir de ellos, se construyen nuevas células; el cuerpo que el veinteañero ve en la foto no se corresponde con el suyo actual- sino más bien que el cambio material tenga lugar de una determinada manera. Lo que el criterio corporal requiere para que las identidades de la persona P_2 en el instante t_2 y de la persona P_1 en el instante t_1 sean la misma no es que P_1 y P_2 sean materialmente idénticas sino que la materia que constituye a P_2 sea resultado de la que constituye a P_1 tras una serie de sustituciones de manera que sea correcto decir que el cuerpo de P_2 en t_2 es idéntico al cuerpo de P_1 en t_1. Por ello y por los procesos celulares, estaría justificado decir que el cuerpo, luego la identidad, del veinteañero es el mismo que el del niño de la foto.

Pero aún caben más ejemplos que desafían al criterio corporal y que los defensores de los criterios de la continuidad material dicen solventar de manera efectiva. Sidney Shoemaker (1931-) muestra el siguiente experimento mental:

Supongamos que la cirugía ha alcanzado unos niveles de sofisticación nunca antes imaginados. Supongamos que la técnica para operar tumores cerebrales

consiste en extraer el cerebro del cráneo, mantenerlo vivo mientras dura la intervención y volver a colocarlo en su sitio restableciendo todas sus conexiones originales. Cierto día, en una clínica, operan a los señores Brown y Robinson mediante el procedimiento descrito. Pero, por despiste, han reinsertado el cerebro de Brown en el cuerpo de Robinson y el cerebro de Robinson en el cuerpo de Brown. Uno de esos hombres, el que tiene el cerebro de Robinson y el cuerpo de Brown, fallece inmediantamente. El hombre con el cerebro de Brown y el cuerpo de Robinson sobrevive. Llamémoslo Brownson. Al despertar, Brownson recupera la conciencia y se horroriza al verse en el espejo. No reconoce su rostro. Tampoco su timbre de voz. Exige que le llamen Brown, tiene los recuerdos de la vida de Brown y, desde luego, pretende que le lleven a casa de Brown con la familia de Brown, no a casa de Robinson con la famila de Roinson, a la que no reconoce.

Casi todos coincidiríamos en que Brownson es Brown. Algunos defensores de los criterios de continuidad material alegarían que lo que necesario para la identidad personal no es la identidad del cuerpo completo sino la identidad del cerebro -en tanto que es el órgano que parece controlar la memoria, el carácter, la personalidad-. Por tanto, el *criterio cerebral* afirma que P_2 en t_2 será la misma persona que P_1 en t_1 únicamente en caso de que P_2 en t_2 tenga el mismo cerebro que P_1 en t_1.

Ahora bien, existen reflexiones en torno a estudios con pacientes comisutorizados -casos de bisección cerebral que derivan en la desconexión de hemisferios como tratamiento de una grave epilepsia– que permiten pensar que no basta con todo el cerebro sino con una porción suficiente que asegure la identidad personal. Así, el *criterio físico* plantea que la persona P_2 en t_2 es la misma persona que P_1 en t_1 si y solo si suficiente cerebro de P_1 en t_1 sobrevive en P_2 en t_2.

Según estos criterios, el cerebro es responsable del carácter, la personalidad y la memoria pero, ¿los estados mentales son estados cerebrales? ¿El cerebro es responsable de la conciencia? ¿Puede definirse el concepto de *conciencia*?

1.2. La continuidad de la conciencia.

Los siguientes criterios se basan en la conciencia como juicio de identidad. Como veremos, también Roger Penrose [Pen15] basa su filosofía y posición en el problema mente-cuerpo, sumamente ligado al problema de identidad personal, en la inexplicabilidad y la inmaterialidad de la conciencia.

Parece que es de común acuerdo que el conjunto de facultades cognitivas consiste en una sustancia inmaterial, separada e independiente del cuerpo. Por tanto, si una persona es identificada con su mente y no con su cuerpo y su mente es una sustancia no material, entonces la identidad personal a través del tiempo debe estar conectada a la persistencia de dicha mente o conciencia a pesar del cambio físico constante.

Según los desarrollos de John Locke (1632-1704) en el capítulo XXVII de [LN75], la identidad de las criaturas vivientes no depende de la materia de la que están compuestas sino de alguna otra cosa, pues, en ellas, la variación de grandes cantidades de materia no altera la identidad. Es por eso que un potro que crece

hasta ser un caballo es, en todo instante, el mismo animal a pesar del enorme cambio en la materia que lo constituye.

En lo que difiere un caballo de una masa de materia inerte es que la masa de materia no es otra cosa que la cohesión de partes de materia y su manera de estar unidas. Sin embargo, el caballo está constituido por la organización particular de sus partes en un cuerpo coherente y, además, dichas partes participan de una vida común. El caballo continúa siendo el mismo en tanto que continúa participando de la misma vida, aunque dicha vida sea comunicada a nuevas partes de materia unidas de forma vital al propio animal gracias a una organización que se mantiene en el tiempo y que resulta conveniente para el animal. Esta organización, presente en todo instante, hace que cualquier conjunto de materia sea distinguible del resto y es constituyente de una vida individual que existe tanto hacia atrás como hacia adelante en el tiempo ya que es la misma continuidad de las partes del cuerpo vivo la que lo permite.

Algo similar ocurriría con las máquinas. Según la teoría de Locke, la nave de Teseo no sería otra cosa distinta a la organización o ensamblaje de sus partes dispuestas adecuadamente para que, al aplicar la fuerza del viento, pueda cumplirse el objetivo de navegar. Si suponemos que la nave es un cuerpo continuo cuyas partes o piezas se reparan, en una vida común, tendremos algo semejante al cuerpo de un animal. Si bien, en el cuerpo del animal, la organización y el movimiento -que es la esencia de la vida- comienzan al mismo tiempo y provienen del interior. En cambio, el movimiento de la nave de Teseo estaría causado por una fuerza exterior –el viento– y esta puede estar ausente aún cuando la nave está en orden y lista para recibirla.

Por ello, según Locke, quienes sitúen la identidad personal del hombre en otra cosa que no sea, al igual que en los animales, la participación de las partes en una organización vital continua, encontrarán difícil que un embrión pueda transformarse en un anciano.

Además, una persona es un ser pensante, inteligente, provisto de razón y reflexión, que puede considerarse a sí mismo como una misma cosa pensante en diferentes ubicaciones espacio-temporales. Esto sólo es posible porque la persona tiene conciencia, que es inseparable del pensamiento. Estar provisto de conciencia es sinónimo de estar provisto de pensamiento y eso es lo que hace que cada uno sea lo que él llama *sí mismo* y de ese modo se distingue a sí mismo del resto de cosas pensantes. Según Locke, la identidad personal radica en tener conciencia.

Pero, dado que el hecho de tener conciencia se ve continuamente interrumpido -bien por el olvido, bien por el sueño profundo que, aunque permite el pensamiento, no permite los pensamientos acompañados de conciencia-, cabe preguntarse si somos siempe, o no, la misma cosa pensante, es decir, la misma sustancia material o física. Locke afirma que esto no afecta al problema de identidad personal ya que este no radica en saber si es la misma persona la que piensa siempre en la misma sustancia material idéntica.

Es por esto que el veinteañero que mira la foto puede reconocerse a sí mismo en tanto que, a pesar de ser sustancias distintas, posee la misma conciencia. Con independencia de que sus pensamientos sean distintos y de que su actitud ante

la vida difiera radicalmente, dichos pensamientos están acompañados de una conciencia única que se ha mantenido en el espacio y el tiempo.

Por ello, Locke sugiere que diversas sustancias pueden estar unidas en una misma persona –ser pensante- por medio de una misma conciencia de la que participen. La identidad personal sólo depende de tener conciencia con independencia de que se circunscriba a una única sustancia material individual o a un conjunto de sustancias distintas. Posteriormente analizaremos las consecuencias de esta afirmación.

Ya comentamos la relevancia de las huellas dactilares, luego de las manos, como criterio de evidencia corporal. Locke objetaría que los miembros del cuerpo son partes de la persona en sí misma. Si, por ejemplo, se le corta una mano y por ello se separa a la persona de la conciencia que tenía acerca de lo que la mano experimentaba, entonces la mano ha dejado de ser parte del sí mismo de la persona. Luego la sustancia material en la que consistió el sí mismo -la persona- en un determinado momento puede modificarse y que la persona siga siendo la misma.

Es más, para Locke, el sí mismo es esa cosa consciente y pensante independientemente de que la sustancia que lo constituya -sea esta o no material-. Es, además, sensible al placer y al dolor, capaz de experimentar felicidad y desgracia. Esta cosa consciente se refiere a sí hasta donde se extienden los límites de su conciencia. Siguiendo con el simil de las manos, supongamos que, dada una persona –ser consciente y pensante-, se le corta el dedo meñique. La persona era consciente de su dedo meñique antes de que se le arrebatase. Pudiera ocurrir que la conciencia de la persona en sí misma acompañara al dedo y abandonase al resto del cuerpo. Entonces, sería evidente que ese dedo sería la misma persona y el sí mismo ya nada tendría en común con el resto del cuerpo.

Puesto que son la felicidad y la desgracia aquello por lo que cada uno se preocupa de sí mismo, es esta identidad personal -conciencia– el pilar sobre el que se fundamentan el derecho y la justicia. Se debe a que cada persona busca el bien para sí sin importar lo que le pueda ocurrir a cada una de sus partes. Puesto que la identidad personal se basa en la conciencia y no en la sustancia, al sujeto se le puede atribuir responsabilidad moral y se podría justificar el castigo y la culpa. Del mismo modo, el sujeto consciente es el que recibiría el premio y la recompensa. Si la conciencia se va con el dedo meñique, el mismo sí mismo sería aquel que antes se preocupaba por todo el cuerpo –pues el meñique era parte de la totalidad del cuerpo– y tendría que reconocer como suyas las acciones perpetradas por el cuerpo en su totalidad.

Si el resto del cuerpo cobrase repentina conciencia tras la extirpación del meñique, y dicha conciencia fuera ajena al conocimiento del meñique, entonces el sí mismo que se fue con el meñique no se ocuparía del resto del cuerpo como parte suya, no reconocería como propias ninguna de las acciones del cuerpo y no podrían serle imputadas –ni obtener beneficio por ellas-.

Por ello y dado que la identidad personal sólo consiste en el tener conciencia, castigar a una persona por lo pensó dormida y de lo cual no tiene conciencia despierta no sería más justo que castigar a un hombre por los actos de su hermano gemelo sólo porque su apariencia exterior se asemejara tanto que fuesen indistinguibles corporalmente.

Sin embargo, podría objetarse la siguiente situación. Supongamos que un hombre pierde totalmente la memoria sobre ciertas partes de su vida. ¿No es ese hombre la misma persona que aquella que realizó las acciones y tuvo conciencia de ellas en cierto momento? Locke apostilla, aunque con matices, que si es posible que un hombre tenga varias conciencias incomunicadas en instantes distintos, entonces un hombre podría ser diferentes personas en momentos distintos. Esto encajaría con esas expresiones en las que se dice que alguien no "está en sí mismo" –por ejemplo, en un arrebato de ira– o que alguien, "por fin, se ha encontrado a sí mismo" -por ejemplo, cuando abraza cierta espiritualidad o religión-. Estas frases indican, para quienes las emplean, que el sí mismo habría sufrido un cambio y que lo que constituye ese sí mismo de esa persona ya no está en ese hombre.

Pero, si un hombre comete un delito ebrio, ¿por qué se le castiga cuando está sobrio y dice no ser consciente de haber cometido el delito, precisamente, por estar ebrio? Se debe a que las leyes humanas castigarían a ambos –al hombre ebrio y al sobrio– ya que no pueden distinguir con certeza qué es lo real y qué es lo simulado, de modo que la ignorancia del ebrio no es un atenuante. Porque, aún siendo cierto que el castigo -y la recompensa- va unido a la identidad personal y esta al ser consciente, los jueces condenan al borracho porque el hecho se ha probado en su contra y la falta de conciencia no puede ser probada por parte del borracho. Veremos las implicaciones que esto puede tener en el mundo tecnológico. Locke apuesta porque, en el futuro, cuando no haya secretos, nadie será responsable de algo que desconocía totalmente sino que recibirá su sentencia en función de si, en el momento de la acción, tuviera o no conciencia de la misma.

En definitiva, Locke, y las teorías neo-lockianas, apuestan porque la identidad personal no se basa en el cuerpo o la sustancia sino en el hecho de tener conciencia. El cuerpo puede cambiar y la conciencia permanecer igual.

Capítulo 2

La identidad personal como problema tecnológico.

En los inicios de los años cincuenta, Grey Walter (1910-1977) mostró al mundo el robot que había creado. Su tortuga, como se le apodó, se deslizaba por el suelo hasta que sus baterías estaban bajas, momento en el que buscaba el enchufe más cercano y se conectaba para recargarlas. Ya con la batería repuesta, la tortuga se desconectaba y volvía a caminar.

Uno de los objetivos de la Inteligencia Artificial (IA) es proporcionar vías hacia el entendimiento de cualidades mentales como la felicidad, el dolor o el hambre. Recordemos que, para Locke, estos eran algunos de los fines que perseguía o evitaba cada uno.

Siguiendo con el ejemplo, la tortuga de Grey parece mostrar un comportamiento similar al del ser humano. Podríamos decir que cuando la tortuga de Grey tiene hambre –se está quedando sin batería– busca el suministro de alimento más cercano al igual que un ser humano va a la nevera. Algo en el interior de la tortuga es sensible al estado de la carga de su batería.

Añadamos atributos y capacidades a la tortuga de Grey para que, de manera simplificada, simule el comportamiento humano. Supongamos que posee una escala de sentimientos que va desde el dolor extremo (-100) hasta el placer absoluto ($+100$). Supongamos que posee algún medio para registrar su puntuación y que sus acciones están ajustadas para maximizarla. También asumimos que es capaz de llevar a cabo otras acciones, aún sin ser muy complejas, distintas a la de buscar un enchufe para así modificar su puntuación. Quizá pueda ponerse al sol para calentarse o buscar el contacto con otras tortugas.

¿Estaríamos en lo cierto al asegurar que la tortuga siente placer cuando su puntuación es alta y que siente dolor cuando es baja? ¿La tortuga podría, simplemente, sentir?

2.1. La Inteligencia Artificial fuerte.

La teoría de la Inteligencia Artificial (IA) fuerte afirma que dispositivos como la tortuga de Grey son inteligentes, poseen una mente y, además, se les puede atribuir cierto tipo de cualidades mentales. Según los teóricos de la IA fuerte, la actividad mental se correspondería con una secuencia bien definida de operaciones -algortimos-.

Pero, el cerebro humano es, a priori, bastante más complejo que los dispositivos hasta ahora diseñados. Esto no es impedimento para la IA fuerte, que sostiene que la diferencia entre el funcionamiento del algoritmo del cerebro humano –incluyendo sus manifestaciones conscientes– y el de cualquier otro dispositivo más simple radica en la complicación, en el orden de estructura o en propiedades autoreferentes. Consideran que toda cualidad mental -como la conciencia-, es una característica del algoritmo que ejecuta el cerebro -entendiendo el cerebro como máquina de ejecución-.

Los defensores de la IA fuerte alegan que aquellos algoritmos cuyo desempeño sea análogo al del cerebro humano podrán experimentar autónomamente sentimientos y tener conciencia propia. Lo importante es el algoritmo en sí y no sus realizaciones físicas particulares. El algoritmo desarrollaría sus cualidades mentales independientemente de si es ejecutado en una máquina de ruedas y poleas, en un dispositivo electrónico o en un cerebro humano.

2.2. Conciencia y creatividad.

El fotógrafo David Slater dejó una cámara en la reserva natural de Tangkoko, Indonesa, para comprobar si los macacos allí residentes podían hacer fotografías. El resultado fue extraordinario. Los selfies de se viralizaron y Slater quiso registrar las fotografías como propias. Sin embargo, en 2014, los tribunales estadounidenses le negaron la autoría de las fotografías alegando que un objeto –las fotografías– que no ha sido creado por un humano no puede ser sujeto de copyritgh. Más aún, la organización animalista PETA exigió que los macacos fuesen los receptores de los beneficios derivados de los derechos de autor. Si bien, los tribunales desestimaron esta alegación objetando que los animales no pueden disfrutar del beneficio económico que reportarían los derechos de autor. ¿Qué tiene que decir la ley ante estas cuestiones sobre la autoría de una obra? ¿Por qué la IA fuerte y la respuesta de Locke al problema de identidad personal están tan ligadas?

Los algortimos, digamos, clásicos se basan en seguir ciertas reglas codificadas que dadas por el programador. Si bien, parece que ahora los algoritmos pueden ir más allá de lo que sabemos cómo ordenarles y aprender una tarea concreta a su manera. Pueden incluso dar lugar a nuevas creaciones inesperadas.

Fue el caso de AlphaGo. La idea original de Demis Hassabis (1976-) era diseñar un meta-programa que pudiese escribir a su vez un programa que jugase al Go. Dicho meta-programa sería creado según aprende jugando al Go. Su construcción se basaría en la experiencia adquirida durante las partidas. Tiempo después,

el equipo de DeepMind, ente ellos Hassabis, crearon un algortimo capaz de jugar a Go que aprendiese a través de partidas y se adaptase a las mismas.

En octubre de 2015, probaron el algoritmo contra el campeón europeo Fan Hui. AlphaGo ganó por cinco juegos a cero. Cuando la derrota de Fan Hui llegó a la prensa asiática, despreciaron la victoria del algoritmo alegando el bajo nivel de los jugadores europeos.

Fan Hui continuó jugando contra AlphaGo y, por tanto, enseñándole a jugar. El algoritmo se colocó en torno a la posición 300 en el ranking mundial de jugadores de Go. Tras ver las virtudes de su algoritmo, invitaron al multicampeón Lee Sedol a una partida de 5 juegos que se llevarían a cabo entre los días 9 y 15 de marzo de 2016 en una localización secreta y que sería retransmitida por internet. El ganador recibiría un millón de dólares.

En el primer juego, bastante clásico, el algoritmo ganó al surcoreano. La sorpresa llegó cuando, en el segundo juego, mientras Sedol fumaba un cigarrillo en la azotea del hotel en el que se celebraba la competición, el algoritmo ordenó dar 5 pasos hacia adelante. Un movimiento tremendamente arriesgado. AlphaGo había roto la ortodoxia habitual realizando un movimiento inexplicable para cualquier ser humano. Por supuesto, AlpaGo ganó el segundo juego. Hizo lo propio en el tercer y quinto juego. Lee Sedol sólo ganó el cuarto juego, y empleando estrategias sumamente rompedoras, que, desde entonces, serían aprendidas por AlphaGo y, por tanto, quedarían prácticamente inutilizadas para volver a atacar al algoritmo.

Siguiendo el razonamiento de la IA fuerte, AlphaGo parece replicar el cerebro humano y, por tanto, tener conciencia autónoma. De hecho, es capaz de ejecutar movimientos calificados como suicidas por los humanos, luego parece tener inteligencia y criterio propio. AlphaGo es capaz de realizar movimientos que los seres humanos no comprendemos, no tenemos la capacidad o la intución que parece tener el algoritmo. Puesto que, siguiendo a Locke, la conciencia otorga identidad personal y sólo los sujetos con identidad personal pueden ser objeto de recompensa, entonces el algoritmo es quien ha ganado la partida y es quien debe ser recompensado con el premio. Pero, ¿es realmente una conclusión acertada? ¿Sería más adecuado otorgarle el premio a los programadores del algoritmo?

Aparte de considerar castigos y recompensas, cabe preguntarse cuál sería el trato adecuado para con un algoritmo que posee conciencia autónoma.

Los retratos de Rembrandt (1606-1669) no tienen parangón. La expresividad que conseguía es halagada por otros pintores como Van Gohgh (1853-1890): *Rembrandt goes so deep into the mysterious that he says things for which there are no words in any language* [Van03].

Un grupo de data scientists de Microsoft y Delft University of Technology se plantearon si, usando los cuadros disponibles de Rembrandt, podrían entrenar un algoritmo capaz de nuevas obras. El equipo estudió 346 obras del pintor para explorar las proporciones de los rostros que en ellos aparecían. También pretendían aprender el trato de la luz de Rembrandt, habitualmente centrada en un área determinada del cuadro, como si la luz surgiese de un foco. Pero el algoritmo no estaba diseñado para crear obras que replicasen a Rembrandt sino para que las crease como si pudiese ver el mundo a través de los ojos del propio pintor. Tras 18 meses de trabajo, el 5 de abril de 2016, mostraron al

mundo su intento de resucitar a Rembrandt. Es innegable que la obra creada por el algoritmo -y ejecutada, pintada por una máquina- capta el estilo del pintor holandés. Casi todo el mundo afirmaría que se trata de una obra de Rembrandt. Expertos cono Ernst van de Weterigns (1938-2021) rechazaron la idea del nuevo Rembrandt creado por un algoritmo pero quedaron sorprendidos por el resultado. Weterings criticó ciertas inconsistencias y sutilezas en la obra: las pinceladas pertenecían el estilo del pintor en el año 1652 mientras que el cuadro en su conjunto podía enmarcarse en el primer peiodo de Ámsterdam (1932-1936). Matices inapreciables por el público general.

Otros como Jonathan Jones (1976-) son más vehementes: *What a horrible, tasteless, insensitive and soulless travesty of all that is creative in human nature. What a vile product of our strange time when the best brains dedicate themselves to the stupidest "challenges" when technology is used for things it should never be used for and everybody feels obliged to applaud the heartless results because we so revere everything digital.*

Jones sostiene que el proyecto denigró y mancilló el espíritu creativo de Rembrandt. Calificó la nueva obra como un atentado contra el arte perpetrado por unos locos. Su crítica completa puede consultarse en [Jon16].

Según Locke, distintas sustancias pueden ser parte de una misma conciencia, luego, siguiendo el punto de vista de la IA fuerte y asumiendo que el algoritmo, puesto que es capaz de nuevas creaciones con un estilo muy marcado, posee conciencia, cabe preguntarse si esta es la de Rembrandt. O quizá sea una nueva conciencia autónoma. En cualquier caso, la IA fuerte afirmaría que tiene conciencia y, por tanto, puede sentir. Así que es lícito cuestionar si críticas desdeñosas como la de Jones son éticas. ¿Es moral tratar de ese modo al algoritmo? ¿Podría el algoritmo aprender con el feedback de los críticos y mejorar hasta silenciarlos?

2.3. Conciencia y metaverso.

Uno de los primeros mundos paralelos que intentaban imitar la vida real fue Second Life. En mayo de 2007, el ARD, un canal de televisión alemán, descubrió que un grupo de avatares frecuentaban un club virtual en el que mantenían sexo con menores cibernéticos. Esta investigación hizo que Peter Vogt, fiscal del Departamento de Prevención de Pornografía infantil, declarase *Vamos a descubrir quién está detrás de todo esto* [Cóz07].

Estas declaraciones, siguiendo las teorías expuestas, presentan múltiples problemas. El principal es que se trata de delitos cometidos por avatares. ¿Tienen estos conciencia? Puesto que han sido creadas por personas –seres con conciencia– y son manejadas por las mismas ¿poseen la misma conciencia que los creadores o una autónoma? Si existe el delito, ¿quién es el responsable?

Lo más obvio sería considerar que ni los pedófilos ni las víctimas existen realmente, algo que la IA fuerte cuestionaría. Si los avatares no existen realmente y no se puede cometer un crimen si los hechos no ocurren realmente, entonces el caso está zanjado. Si alguien se siente atacado no tiene más que apagar el ordenador.

Pero los problemas van aún más allá. Según las normas de Second Life, no podían entrar menores. Es decir, los usuarios detrás de los avatares que supuestamente fueron abusados eran mayores de edad, luego libres y capaces de, a priori, decidir si sus avatares querían tener sexo online. Por tanto, se plantea un problema en la autenticación de las identidades.

Si dichos avatares tuviesen conciencia, el problema de autenticación sería aún más complejo. Al igual que con Rembrandt, ¿la conciencia de los avatares sería autónoma o sería la de sus creadores?. Ya Locke, en su ejemplo del hombre ebrio y sobrio, apunta a la imposibilidad de verificar si alguien tiene o no conciencia en un determinado momento y anhela *un mundo en el que se hagan patentes los secretos de todos los corazones* [LN75]. Si esa verificación es complicada, no podemos imaginar las múltiples respuestas a la pregunta planteada al inicio del párrafo. Aunque quizá la ley sea más tajante.

Para abordar el problema de autenticación, rechazaremos la postura de la IA fuerte y asumiremos que los avatares no poseen conciencia. Nos preguntamos cuál sería modo de identificar a los abusados, pues estos están fingiendo una identidad falsa. A día de hoy, los métodos de autenticación mas empleados son los siguientes:

- **Autenticación QR.** Utiliza la cámara del dispositivo para escanear un código que permite el acceso. Sólo los usuarios con un código adecuado podrían entrar a un supuetso metaverso.
- **Doble Factor.** Sistema que requiere dos medios de identificación diferentes para garantizar el acceso.
- **OTP SMS.** El código de un sólo uso es un factor de autenticación para transacciones confidenciales. Cada vez que alguien intentase entrar en el metaverso debería generar uno.
- **Biometría.** Basada en las características biológicas únicas de cada uno.
- **Certificado digital.** Emitido por una Autoridad de Certificación, acredita la identidad de un usuario.

Siguiendo los puntos de vista sobre el problema de identidad personal, los métodos biométricos quedarían inmediatamente descartados -el cuerpo es un criterio de evidencia falible-. Para el resto, habría que comprobar que aquel que está haciendo uso de ellos posee conciencia y hace uso de la misma en el momento de autenticarse. Ya hemos visto la imposibilidad de esto.

El metaveso propuesto por Mark Zuckerberg promete un mundo virtual totalemetne inmersivo al que los usuarios se conectarán mediante una serie de dispositivos y tendrán la sensación de estar realmente dentro de dicho mundo. Podrán interactuar con los elementos allí presentes. Será como una teletransportación, recordemos al viajero a Marte, a una realidad alternativa y se podrá experimentar allí las mismas, quien sabe si nuevas, sensaciones que en el mundo real. ¿Quién las experimentaría realmente: el usuario o su avatar? En ese mundo alternativo, ¿qué leyes imperarían? ¿Se necesitaría una policía en ese mundo?, ¿valdría la del mundo real?

2.4. La habitación china de Searle.

El filosofo estadounidense John Searle (1932-) se opone radicalmente al punto de vista de la IA fuerte argumentando que el atributo mental de la comprensión y la conciencia están plenamente ausentes del algoritmo -o máquina que lo ejecute-.

La cuestión es si un acierto d ella máquina es indicio de comprensión por parte de la máquina o del propio algoritmo Por ello, propone el experimento mental de la habitación china.

Supongamos que el ser humano ha sido capaz de construir una máquina capaz de entender el idioma chino. Esta recibe como entrada una frase escrita en idioma chino, la procesa y produce una respuesta convincente capaz de hacer que un hablante de chino dé la respuesta por válida y, por tanto, la máquina supera el Test de Turing. Ahora, supongamos que Searle, que no sabe chino, se mente en el interior de dicha máquina equipado con manuales y diccionarios que le indican las reglas ortográficas y gramaticales del idioma chino. Searle está completamente aislada del exterior salvo por una rendija por la que pueden entrar y salir, mediante hojas de papel, textos en chino. Supongamos que, fuera de la máquina, se encuentra el hablante chino que creyó que la máquina comprendía su idioma. El hablante introduce por la rendija una serie de preguntas en chino y todas son respondidas de manera adecuada.

Searle deja claro que no entiende una palabra de chino, de modo que no entiende nada de lo que le introducen como input y tampoco entiende nada del output que ofrece. Sigue sin entender nada al salir de la máquina. Se ha limitado a ejecutar cierto algoritmo en la búsqueda de sus respuestas.

Parece evidente que la ejecución adecuada de un algoritmo no implica que haya habido comprensión o conciencia alguna.

Pero esta no es la única dificultad que plantea la IA fuerte. Según los teóricos, lo único importante es el algoritmo. No importa qué ejecuta el algoritmo. Da igual si es ejecutado por un cerebro, una computadora, una comunidad de monjes o un sistema de ruedas y poleas. La idea reside en que es simplemente la estructura lógica del algoritmo lo significativo del estado mental que se supone que representa. La encarnación física de dicho algoritmo no influye.

Si bien, como apunta Searle, esto parece ser una forma de dualismo. Ya en el s. XVII, René Descartes (1596-1650) argumentaba que había dos tipos de sustancia: sustancia mental y sustancia o materia ordinaria. Se supone que la sustancia mental no está compuesta de materia, algo que también recoge Locke. Estos filósofos argumentan que la sustancia mental –conciencia– existe con independencia de la sustancia material.

La sustancia mental de la IA es la estructura lógica del algoritmo, la encarnación de este es irrelevante. El algoritmo tiene un tipo de existencia no material ajena a cualquier realización particular del mismo en términos físicos. Por tanto, los defensores de la IA parecen creer que los algoritmos forman parte de la sustancia de sus propios pensamientos, sentimientos, entendimiento y conciencia. Resulta paradójico que la IA fuerte desemboque de forma extrema en el dualismo: precisamente la opinión con la que menos deberían estar de acuerdo sus defensores.

Conclusiones.

Al igual que Roger Penrose y en contra del criterio de la IA fuerte, me resisto a creer que la conciencia sea un proceso meramente algorítmico. Es cierto que hay visiones que minimizan la actividad inconsciente en la resolución de problemas y en los procesos creativos, pero también hay quienes sospechan que son muchos los procesos inconscientes que operan sin directrices conscientes pero que, periódicamente, necesitan supervisión [Mad].

Es célebre la historia de cómo Henri Poincaré (1854-1912) descubrió las funciones fuchsianas mientras subía a un autobús. En palabras del mismo, *invención es discernimiento y elección*, pero dónde y cómo se hace dicha elección es una cuestión enigmática. J.E. Littlewood (1885-1977) afirma que *la incubación es el trabajo del subconsciente durante el tiempo de espera, que puede durar varios años. La iluminación, que puede ocurrir en una fracción de segundo, es la manifestación de la idea creativa en la ciencia. (...) La iluminación implica alguna relación misteriosa entre el subconsciente y el consciente, de otro modo tal manifestación no podría darse. ¿Qué es lo que enciende la bombilla en el momento oportuno?*

Suponiendo que los algoritmos no pueden alcanzar una conciencia autónoma -al menos por ahora- y, por tanto, tampoco el habitual trabajo inconsciente del que hablan muchos autores y que necesita ser dirigido periódicamente de manera consciente, deberían ser sus creadores los que fuesen beneficiarios de las recompensas que derivasen de sus algoritmos y también los responsables de los perjuicios que pudieran surgir a partir de ellos. Estarían en la obligación ética y moral de dar solución a los problemas que generasen sus algoritmos.

Aunque se trate de ficción, es ese mismo imperativo moral el que hace que Victor Frankenstein tenga que ir al confín del mundo para acabar con la criatura que ha creado [She06]. Por si fuera poco, el ser diabólico al que da vida Frankenstein tiene conciencia propia, luego podría objetarse si el acto de asesinar a la creación es o no objeto de castigo. Dudas que se despejarían asumiendo que el algoritmo no puede tener conciencia propia y, por tanto, su destrucción no generaría ningún conflicto -al menos en cuanto a la adecuación o no de matar a un ente consciente-.

En definitiva, pienso que no entender cómo funciona un algoritmo no permite que podamos atribuirle una conciencia. Nuestra falta de comprensión y fascinación ante outputs imprevisibles no justifica que el algoritmo tenga intuición, sentimientos o conciencia. Simplemente hace patente el carácter limitado de nuestro conocimiento.

Bibliografía

[Cóz07] Álvaro de Cózar. «Delitos reales de seres virtuales». En: *EL PAÍS* (2007) (vid. pág. 12).

[Du 19] Marcus Du Sautoy. *The creativity code*. Harvard University Press, 2019.

[Jon16] Jonathan Jones. «The digital Rembrandt: a new way to mock art, made by fools». En: *The Guardian* (2016) (vid. pág. 12).

[LN75] John Locke y Peter H Nidditch. «John Locke: An essay concerning human understanding». En: (1975) (vid. págs. 5, 13).

[Lóp+22] Álvaro López García y col. *Apuntes de Seguridad, privacidad y aspectos legales*. 2022.

[Mad] Universidad Complutense de Madrid. *¿Actividad subconsciente? Algunos testimonios*. URL: http://www.mat.ucm.es/cosasmdg/cdsmdg/04vida/gocesesteticos/seminarioestrategiasppm/parapensarmejor/subconscientetestimonios.htm (vid. pág. 15).

[Pen15] Roger Penrose. *La nueva mente del emperador*. DEBOLS! LLO, 2015 (vid. pág. 5).

[Poi18] Henri Poincaré. *La invención matemática: cómo se inventa: el trabajo del inconsciente*. KRK Ediciones, 2018.

[Rod03] Mariano Rodríguez González. *El problema de la identidad personal. Más que fragmentos*. 2003.

[She06] Mary Wollstonecraft Shelley. *Frankenstein o el Prometeo moderno*. Ediciones Colihue SRL, 2006 (vid. pág. 15).

[Tec06] El País Tecnología. «Policías y leyes propias para Second Life». En: *EL PAÍS* (2006).

[Van03] Vincent Van Gogh. *The letters of Vincent van Gogh*. Penguin UK, 2003 (vid. pág. 11).