Sección de Filosofía
Universidad de La Laguna

Contra los mecanismos de control automatizado:

El uso de la IA en los ámbitos de la vigilancia y el castigo.

Natanael Galván Ortega.
Tutor: Miguel Mandujano Estrada.
Cuarto curso del Grado en Filosofía.
(2022/2023)

Agradecimientos

A Miguel Mandujano Estrada por tutorizar este trabajo. A María Inmaculada Perdomo Reyes por proporcionarme bibliografía relativa al ámbito del uso de las tecnologías de reconocimiento facial. A Antonio Luis Terrones Rodríguez por realizar su seminario *Una ética aplicada a la IA* en la Universidad de La Laguna. A Daniel Innerarity. Ikerbasque por impartir su conferencia *Igualdad algorítmica. Una inteligencia artificial no discriminatoria* durante el IV Congreso Internacional de Derechos Humanos y Globalización organizado por la Universidad de Sevilla y la Universidad de La Laguna. Y en especial a Ana Suárez Suárez, mi compañera de clase, compañera de estudio, y compañera de vida, por acompañarme durante todos los meses que llevó la preparación de este trabajo.

Índice

1. Introducción, el uso de las inteligencias artificiales de reconocimiento facial en el ámbito de la vigilancia

En la actualidad vivimos en una época donde la preocupación por el rápido avance de la sofisticación y el uso de las Inteligencias Artificiales (IA) en los ámbitos de la vida cotidiana ha llegado no solo a los círculos académicos, sino a su vez a las conversaciones de a pie, poblando así las discusiones entre amigos, familiares o conocidos de todo el mundo[1]. De entre todos sus principales usos se nos presentan con especial interés y preocupación los ejercidos por los cuerpos de seguridad del Estado a través de sistemas de reconocimiento facial, así como por los agentes del poder judicial. Precisamente lo que haremos a continuación es articular una crítica filosófica referida a este uso concreto de la inteligencia artificial, mostrando los peligros que implica, poniendo especial énfasis en su herencia y amplificación del sesgo racista, y abordando el debate que su utilización suscita acerca de cuáles deberían ser algunos de los ámbitos de la vida política que deben permanecer aún en el campo de lo analógico.

Para ello, primeramente, nos remontaremos al uso que se hacía de las cámaras por parte de las instituciones policiales en el siglo XIX, y además definiremos el concepto de raza, cuya conceptualización nos resultará especialmente útil a la hora de comprender las aportaciones posteriores. Tras esto, presentaremos algunos de los eventos más destacados en referencia a los vigentes problemas que comprende el uso de las IA y las tecnologías de reconocimiento facial. Después dedicaremos un apartado a la discusión del tema, donde apoyaremos nuestra postura en la adquisición de los conceptos anteriormente tratados, así como en la teoría foucaultiana de la vigilancia, y en las críticas realizadas al solucionismo tecnológico. Finalmente, a modo de conclusión, determinaremos, a su vez en clave foucaultiana, qué ámbitos de la vida política no deben ser relegados a las IA, expandiendo así sobre la tesis de Innerarity acerca de la automatización de la democracia.

Antes de abordar estas cuestiones sería preciso comenzar definiendo brevemente qué es realmente una inteligencia artificial, puesto que la gran mayoría de definiciones que se nos presentan sobre ellas suelen ser de carácter técnico, cercanas a los ámbitos de la programación y la informática, sin embargo, aquí queremos realizar un análisis filosófico de

[1] Algunas de polémicas recientes más notables incluyen la prohibición del algoritmo ChatGPT en Italia (Pacho, 2023), los fallos en los coches autónomos de empresas como Tesla (Helmore, 2022) o la creciente preocupación por la pérdida de empleo que supone la incursión de las IA en el ámbito laboral (Vallance, 2023).

la cuestión, y para ello debemos partir de una definición propiamente filosófica de lo que es una IA. Como suele suceder con la mayoría de los conceptos referidos a realidades complejas, lo cierto es que resulta realmente arduo elaborar una definición completa de lo que es un IA. En su seminario *Una ética aplicada a la IA*, impartido en la Universidad de La Laguna durante los días 24, 25 y 26 de abril de 2023, Antonio Luis Terrones Rodríguez comentó que no sería correcto definir la IA estrictamente como software, pues esto ignora el tremendo consumo que implica su uso. Comentó además que muchas definiciones populares resultan problemáticas, porque ignoran la dimensión emocional de la inteligencia, y nuestro conocimiento del mundo no se articula sólo de manera racional, sino también emocional. Las definiciones de este estilo desembocan un fenómeno denominado "datificación", esto es, la concepción de que todo es susceptible de ser traducido en datos.

Esta serie de cuestiones nos obligan a presentar una definición filosófica de la IA, que sea capaz de comprenderla más allá de sus aspectos técnicos. A este respecto quizás uno de los trabajos más relevantes sea el de la autora Margaret A. Boden quien, ya en 1977, aporta una descripción realmente compleja de la cuestión en su obra *Inteligencia artificial y hombre natural* (1977). En esta obra se trata de definir la IA sin atender a los intereses específicos del personal que las programa, y se hace de la siguiente forma:

> Por "inteligencia artificial", en consecuencia, entiendo el uso de programas de computadora y de técnicas de programación para proyectar luz sobre los principios de la inteligencia en general y de la inteligencia humana en particular. En otras palabras, uso la expresión como término genérico que abarca a toda la investigación sobre máquinas que de un modo u otro es pertinente para el conocimiento y psicología humanos, sin importar el motivo declarado de la programadora particular del caso (Boden, 1984, p. 23).

De esta forma, partiendo de la concepción de "inteligencia" comprendida por la propia Boden como "la capacidad de manipular símbolos creativamente, o de procesar información dados los requisitos de la tarea del caso" (Boden, 1984, p. 38), podemos afirmar que "la inteligencia artificial es el uso de los programas como herramientas en el estudio de los procesos inteligentes, herramientas que ayudan en el descubrimiento de los procesos de pensamiento y de las estructuras epistemológicas que emplean las criaturas inteligentes" (Boden, 1984, pp. 38-39).

2. Antecedentes

Para comprender propiamente cuales son los riesgos que nos presenta el uso de la tecnología de captación de imagen debemos remontarnos a sus usos penitenciarios originarios, los cuales comenzaron a ejercerse a través de la fotografía ya a mediados del siglo XIX. A su vez, con el fin de tratar los sesgos que dicha captación implica, prestando especial atención a la cuestión de la discriminación racial, definiremos el concepto de raza, que tanta discusión a supuesto durante las últimas décadas, a través de la obra del teórico cultural y sociólogo jamaicano Stuart Hall.

2.1. El uso de la fotografía en el siglo XIX

En su obra *El peso de la representación* (1988) John Tagg comienza denotando que, frente a la visión tradicional que entiende la fotografía como un medio de representación fidedigna, esto es, como una prueba de lo realmente sucedido, tal como indica el popular dicho "una imagen dice más que mil palabras", realmente:

> cada fotografía es el resultado de distorsiones específicas, y en todos los sentidos significativas, que hacen que su relación con cualquier realidad anterior sea algo sumamente problemático, y plantean la cuestión del nivel determinante del aparato material y de las prácticas sociales dentro de las cuales tiene lugar la fotografía. (…) La naturaleza indicial de la fotografía -el vínculo causativo entre el referente prefotográfico y el signo- (…) no puede garantizar nada en el ámbito del significado. Lo que establece el vínculo es un proceso técnico, cultural e histórico discriminatorio (...) para organizar la experiencia y el deseo y producir una nueva realidad: la imagen en papel que (…) puede llegar a tener significado de muchas maneras posibles. (…) La fotografía no es una "emancipación" mágica, sino un producto material de un aparato material puesto en acción en contextos específicos, por fuerzas específicas, con unos fines más o menos definidos. (Tagg, 2005, pp. 8-10).

Ahora que hemos visto como la imagen no resulta en ningún caso inocente ni objetiva, pues no puede ser desvinculada del contexto en el que ha sido producida, y por tanto incurre siempre en un carácter aberrante o desfigurante de lo representado, ponemos de manifiesto que para tratar la cuestión fotográfica debemos apelar entonces a un análisis de carácter histórico y no esencial, esto es, el objeto fotográfico remite a una realidad discursiva, que es la que debemos abordar. Así, siguiendo los pasos de Tagg, veremos qué papel específico jugó la fotografía durante la segunda mitad del siglo XIX, en el contexto del surgimiento de nuevas instituciones disciplinarias, como la policía o las prisiones modernas.

Estas instituciones se valieron de estas nuevas técnicas de vigilancia para influir en el cuerpo social, diseñando así una nueva estrategia de gobierno (Tagg, 2005, p. 12).

Todo comienza con el abaratamiento de los costes de producción fotográfica y la producción masiva de cámaras sencillas, lo cual trajo consigo en las décadas de 1880 y 1890 un cambio de gran interés, así, "el eje político de la representación se había invertido por completo, dado que ser reproducido en imagen ya no era un privilegio, sino el lastre de la nueva clase de los vigilados" (Tagg, 2005, pp. 78-79). Se había articulado el retrato como un medio de control ciudadano, a través de su condición de "prueba". Esto aumentó la ya creciente popularidad que había obtenido la actividad de fotografiar a los presos, hasta el punto en que hoy en día resulta indispensable para el desarrollo del sistema judicial, y es de carácter obligatorio en la mayoría de países. En esta serie de imágenes se pone de manifiesto un proceso de suma relevancia pues, como menciona Tagg, no se remiten solamente a exhibir a un delincuente, sino que se vuelven "un retrato del producto del método disciplinario: el cuerpo hecho objeto; dividido y estudiado" (Tagg, 2005, p. 101). La cámara se tornaba la herramienta más útil para representar el cuerpo, porque se consideraba la más fidedigna, que no desfiguraba el rostro como sí lo hacía la caricatura, y no podía incurrir en errores como sí sucedía con el lenguaje pues, como suele decirse, la fotografía habla por sí misma.

Es importante denotar a su vez que el retrato no se limita a describir a un sujeto, sino que indica cuál es su identidad social (Tagg, 2005, pp. 53-54), de esta forma "los cuerpos –trabajadores, vagabundos, criminales, pacientes, locos, pobres, razas colonizadas- son fotografiados uno a uno: aislados en un espacio estrecho, cerrado; convertidos en rostros enteros y sometidos a una mirada sin respuesta posible" (Tagg, 2005, p. 86). Así, las instituciones de poder comenzaron a fotografiar y clasificar a las personas internadas, las celdas policiales, las cárceles, los manicomios, los centros para niños indigentes, las residencias e incluso los colegios se convirtieron en estudios fotográficos donde se retrataba y clasificaba a las personas, según su edad, su altura, peso, sus antecedentes y sus características fenotípicas y raciales.

Uno de los aspectos claves a tratar es que el uso de las cámaras en el campo de la criminología estaba asociado con la posibilidad de adquirir un carácter predictivo de la práctica, es decir, que el trabajo no se limitaba a analizar los rasgos del criminal, sino que pretendía utilizarlos como precedente para detectar quienes serían potenciales criminales en

el futuro, a causa de su semejanza. Para adentrarnos en este caso resulta de gran utilidad mencionar los estudios de quien es considerado el padre de la antropología criminal, Cesare Lombroso, quien consideraba, siguiendo las teorías de Darwin, que los criminales se encontraban más cerca del mono que los sujetos plenamente desarrollados, y esto se ponía de manifiesto en sus características físicas, que eran las propias de un criminal, entre ellas estarían: "escaso desarrollo del sistema pilífero, escasa capacidad craneal, senos frontales muy desarrollados, sinostosis precoz, mayor espesor de los huesos del cráneo, mandíbulas muy desarrolladas, prognatismo, oblicuidad en las órbitas, piel muy pigmentada, cabellera rizada y espesa, orejas voluminosas, diastema dental" (Montiel Álvarez, 2016, párr. 7). Posteriormente, en Francia, Alphonse Bertillon, comienza a utilizar un sistema antropométrico para identificar delincuentes, este se basaba "la medición de diferentes partes del cuerpo: altura; envergadura (...) y contorno de busto con el sujeto sentado; la longitud y anchura de la cabeza; longitud del dedo medio de la mano izquierda; del pie izquierdo y del antebrazo izquierdo" (Montiel Álvarez, 2006, párr. 21). Tal fue la aceptación del sistema que llegaron a abrirse escuelas de antropometría, que enseñaban a los policías a medir correctamente a los detenidos, sin embargo, apenas unos años más tarde, este método fue sustituido por la toma de huellas dactilares de Francis Galton, que resultó ser mucho más efectivo, hasta el punto en que hoy resulta obligatorio dar las huellas cuando se realiza un documento de identidad o un pasaporte. A pesar de que el método de Bertillon cayó rápidamente en desuso algunos de sus sistematismos siguen hoy vigentes, como es el caso de la famosa doble toma fotográfica del preso, que incluye un retrato frontal y uno de perfil.

Lo que sucede en este caso es que los arrestados son enfrentados al frío juicio de la lente, que es tomado por los oficiales como un capturador de verdad. De ahí se da que la toma de fotos en calidad de "prueba", ya sea el retrato del rostro de un presunto criminal, o los elementos de la escena de un crimen, deban realizarse con sumo cuidado, buscando siempre la mayor "objetividad" posible. La perspectiva, la iluminación o la calidad de la imagen no pueden deformar la "realidad" que está siendo captada. Esto sucede, según Tagg, porque la forma dominante de significación de la sociedad burguesa es la realista, y "el realismo ofrece una fijeza en la que el significante es tratado como si fuera idéntico a un significado preexistente y en la que el papel del lector es meramente el de consumidor" (Tagg, 2005, p. 129). En este caso toda la serie de elementos del proceso que dotan de significado, esto es, el conjunto de elementos lingüísticos y simbólicos presentes detrás de "lo real" o "la verdad", son ignorados por completo, la fotografía como "prueba" se presenta, y es consumida,

enteramente como verdad. Así, parece incluso que se borra el significante y queda sólo el significado, el signo es tratado falsamente como naturaleza, la imagen se identifica con el mundo (Tagg, 2005, p. 129). Los individuos retratados son limitados entonces a sus rasgos, de manera aparentemente objetiva, sus datos son recogidos y archivados, significando una automática pérdida de la privacidad del detenido, que ha sido reducido ahora a una serie de medidas y rasgos que se encuentran en posesión de la policía, lo cual hace de su rastreo una tarea mucho más sencilla.

Las bases de datos que aún utiliza la policía se siguen basando en fotografías de los arrestados, y a su vez esas bases de datos son utilizadas como método de aprendizaje para las inteligencias artificiales. En EEUU han surgido numerosas quejas acerca de las posibles problemáticas que enfrenta este sistema, en su mayoría enfocadas a la desproporcionada cantidad de imágenes de gente perteneciente a minorías raciales que las IA detectan como criminales de manera errónea. Para comprender como se articula esta problemática es preciso definir primero qué es la raza, con la finalidad de clarificar qué es exactamente lo que sucede en esta falsa adjudicación de criminalidad.

2.2 Sobre la raza y el fenotipo

El concepto de raza en términos estrictamente científicos no obtiene ningún tipo de validez cuando se trata de clasificar a los seres humanos, esto es, el conocimiento científico avala el archiconocido tópico que versa: “sólo existe una raza, la raza humana”. Así, decimos que, en lo referido a su definición biológica, la raza es “el conjunto de rasgos característicos comunes y hereditarios que distinguen a un tipo particular dentro de una especie; grupo común definido biológicamente por dichos caracteres hereditarios; noción difícilmente aplicable al hombre debido a las mezclas y cruces” (Russ, 1999, p. 327). Por tanto, no es preciso hablar de distintas razas dentro de nuestra especie, como sí sucede en otras, tales como los perros. De hecho, existe mayor diferencia genética entre las poblaciones del este y el oeste del continente africano que entre los considerados “europeos” o “caucásicos” y los denominados “asiáticos”. Sin embargo, partir de esta definición implica no tener en cuenta la función social que ha ocupado la raza, esto es, como “ficción útil” (Balibar y Wallerstein, 1991), en la historia de la humanidad. Por ello, es preciso atender a una definición del término que sí atienda a su realidad social pues, no podemos considerar que, como las “razas humanas” carecen de fundamentación científica, estas no poseen ningún efecto sobre la vida de los sujetos y, por tanto, es a partir de esa propia experiencia discriminatoria, esto es, a

partir de la función social e histórica de la raza, que esta debe ser definida, como mencionábamos, en términos de ficción útil.

Stuart Hall (2021) trata la raza como un constructo discursivo, un "significante flotante" o un "significante deslizante". En este sentido, al considerar la raza como una categoría discursiva, se muestra a su vez que, como vimos anteriormente, cualquier intento de sostener dicho concepto sobre bases biologicistas resultará siempre de nula eficacia. Y, sin embargo, aunque sea de manera más bien abstracta o poco rigurosa, a partir de los rasgos fenotípicos de una persona (especialmente su pelo, su tono de piel, y los elementos que conforman su cara) todos clasificamos a las personas dentro de "grupos raciales", entonces, ¿cómo es posible que lo que realmente cuente a la hora de definir el concepto no sea una base biológica sino su función social e histórica? Para comprender esta definición debemos entender la raza como una especie de sello, como menciona Hall, una *badge*, que se traduce como insignia o marca distintiva. Si se entiende entonces la raza como un signo, queda claro que esta debe recibir en su estudio un tratamiento no de tipo biológico, sino lingüístico (Hall, 2021, p. 362).

Lo que Hall afirma no es que los rasgos fenotípicos que conforman lo que popularmente denominamos como raza no existan, sino que estos comprenden un funcionamiento lingüístico, en efecto, que la raza funciona como una lengua, de manera significativa. Así, nos dice Hall:

> Los significantes refieren a los sistemas y conceptos de clasificación de una cultura, a sus prácticas de creación de sentido. Y esos elementos adquieren sentido, no por lo que contienen en su esencia, sino por las relaciones cambiantes de diferencia que establecen con otros conceptos e ideas en un campo significante[2]. (Hall, 2021, p. 362).

Puesto que la definición del sentido no refiere en ningún caso a esencialismos, sino que se construye de manera comparativa, se da un constante flujo de sentido, una resignificación perpetua del concepto en el espacio y el tiempo. Por esta razón Hall no se limita a describir la raza como un significante, sino que la describe como un "significante flotante" o deslizante, porque nunca puede adquirir un carácter estático.

[2] Originalmente: "And signifiers refer to the systems and concepts of the classification of a culture, to its practices of *making meaning*. And those things gain their meaning, not because of what they contain in their essence, but in the shifting relations of difference, which they establish with other concepts and ideas in a signifying field". (La traducción es propia).

Se acepta entonces que existen todo tipo de diferencias entre sujetos, pero que estas solo cobran sentido cuando son organizadas a través del lenguaje y pueden así influir en la cultura y la conducta. De esta forma, lo que nos determina racialmente no son nuestros aspectos fenotípicos en sí mismos, sino el sentido lingüístico que estos cobran en nuestra cultura, lo cual nos permite clasificar a los sujetos y actuar consecuentemente respecto a esa distinción. El mundo sólo es inteligible para nosotros una vez ha sido trasladado al ámbito del lenguaje. Así, igual que le atribuimos a la conjunción de fonemas /a/ /r/ /β/ /o/ /l/ un significado dentro de nuestra lengua, una conjunción de rasgos fenotípicos arbitrarios también constituye, junto con la atribución de un significado histórico-social, una unidad de sentido que denominamos identidad racial. Por tanto, determinamos que el significado concreto no es en ningún caso intrínseco a los términos, sino atribuido a los mismos para dotarlos de sentido, que las diferencias entre sujetos sólo adquieren sentido cuando son clasificadas a través del lenguaje en un discurso específico, que es lo que influye en nuestra conducta y nuestra cultura (Hall, 2021, p. 364). Hall afirma que incluso los rasgos externos (pelo, piel, etc) son significantes, y que podemos leer el cuerpo como un texto, que es, de hecho, un texto, que inspeccionamos continuamente como críticos literarios, que "somos lectores de la raza, eso es lo que hacemos, somos lectores de la diferencia social"[3] (Hall, 2021, p. 368).

Es importante que nos quedemos con la concepción de que la raza se articula como un sello, una marca asociada a una serie de valores, un signo, o significante, como menciona Hall, que conlleva a su vez un significado. Lo que esto nos permite comprender es cómo ese signo, que es histórica y culturalmente asociado a un significado al ser captado por un elemento no humano, pero programado por un humano, asociará a su vez dicho significante con el significado histórico-social que conforma el sesgo epistémico de su programador. Sería un error además considerar que la influencia de lo extra científico no tiene cabida en este ámbito pues, como explica Hall, durante mucho tiempo fue la religión la que puso en duda la igualdad entre los europeos o "hijos de Dios" y los habitantes del llamado Nuevo Mundo, y fue sólo a posteriori que esta serie de sesgos ya presentes en la sociedad se trataron de justificar científicamente, como mencionábamos a través de la frenología. Así, si la ciencia determinaba la inferioridad de ciertos grupos raciales, como decíamos en realidad

[3] Originalmente: "we are readers of race, that's what we are doing, we are readers of social difference". (La traducción es propia).

arbitrariamente delimitados, entonces la superioridad europea sería ya indiscutible. Si en este caso seguimos considerando la labor científica de manera inocua, impermeable respecto a los aspectos histórico-sociales que guían la investigación, corremos el riesgo de incurrir en el mismo error.

3. Estado actual

En la actualidad la incursión de las IA en los ámbitos de la vida cotidiana ha suscitado grandes controversias, marcadas por las posibles consecuencias que podría tener el uso de algoritmos como ChatGPT en los ámbitos de la educación o la política, sin embargo, la amplitud de esta problemática nos obliga a reducir nuestro estudio a cuestiones más específicas, que en este caso referirán exclusivamente a los entornos de la vigilancia y el castigo. A continuación, presentaremos dos casos controversiales, que resultan especialmente esclarecedores a la hora de comprender cómo el uso de la IA en estos campos ha generado grandes polémicas. Por un lado, abordaremos el vituperado uso del algoritmo *COMPAS* en los juzgados estadounidenses, y por otro, trataremos el debate que suscitó la prohibición del uso de las tecnologías de reconocimiento facial en la ciudad de San Francisco.

3.1. El algoritmo *COMPAS*

Uno de los usos más comunes de la IA es el de producir sugerencias comerciales basadas en los historiales de búsqueda de los usuarios, así, las actividades realizadas por estos son comparadas con las de otros usuarios con historiales similares y, en base a estos datos, reciben sugerencias de compra. Más allá de las implicaciones que esto supone en términos de privacidad esta serie de cuestiones no parecen implicar un gran desafío ético para el presente, sin embargo, la capacidad predictiva de las IA implica complicaciones mucho mayores cuando trata de utilizarse en ámbitos más complejos. Tal es el caso de los algoritmos empleados por universidades o empresas para decidir quiénes merecen ser aceptados en sus instituciones, que a su vez realizan una labor predictiva basada en el historial de personas con perfiles similares a los del sujeto juzgado (Farid, 2018). De entre todos estos casos nos resultan especialmente sugerentes los algoritmos utilizados por los poderes ejecutivo y judicial del Estado, a la hora de tratar de predecir el futuro comportamiento de los arrestados.

Actualmente, en varios estados de EEUU, se utiliza una IA para tomar decisiones acerca de la fianza en el momento del arresto, así, la información de una persona que ha sido

arrestada por acometer un supuesto crimen es introducida en el algoritmo, y ese algoritmo genera una evaluación de riesgo en función de dichos datos, el cual refleja la probabilidad que tiene ese individuo de volver a cometer un crimen en el futuro, y lo refleja en una escala del uno al diez. En caso de que el algoritmo indique que se trata de un sujeto de alto riesgo el juez puede utilizar esta información para denegarle la fianza, manteniéndolo en prisión hasta el momento del juicio. Igualmente, si la IA indica que el sujeto arrestado es de bajo riesgo este puede ser liberado hasta que suceda el juicio (Farid, 2018). A este respecto, el pasado 2016, los periodistas de investigación de ProPublica (Angwin et al., 2016) publicaron un informe sobre el algoritmo COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), que, de entre todos los que cumplen la función antes mencionada, es el que se ha ido extendiendo rápidamente por juzgados de toda Norteamérica, hasta el punto en que versiones similares están siendo implementadas a día de hoy en lugares[4]. Lo que ProPublica encontró es que este algoritmo perjudicaba especialmente a los acusados negros, los cuales recibían hasta el doble de probabilidades de ser falsamente diagnosticados como futuros reincidentes. Así, la falsa predicción de una reincidencia se daba en un 44’9% en sujetos negros y en un 23’5% en blancos; de igual manera la IA se equivocó al predecir que algunas personas no reincidirían, cuando sí lo hicieron, lo cual se dio en un 28’1% de las personas negras y un muy superior 47’7% de las personas blancas (Angwin et al., 2016). Esto nos muestra que los acusados negros tienen una probabilidad mucho mayor de ser erróneamente juzgados por este algoritmo, lo cual implica que son vetados injustamente de su derecho a la fianza más a menudo, siendo obligados a continuar encarcelados.

Es preciso mencionar que el funcionamiento de este algoritmo es un secreto corporativo, que no conocen ni siquiera los propios juzgados. Sin embargo, a través de un proceso de ingeniería inversa basado en puede deducirse su funcionamiento (Dressel y Farid, 2018). Sabemos entonces que los dos factores fundamentales para realizar el cálculo que indica la capacidad de reincidencia del/la acusado/a son la edad y el número de crímenes previamente cometidos (Dressel y Farid, 2018). Bien, podemos afirmar con total seguridad que la edad es un factor no racializado, sin embargo, el número de crímenes previos es un factor severamente influenciado por el seguimiento racista de la policía estadounidense, esto implica que, de manera clara, el funcionamiento del algoritmo se ve severamente afectado por el conjunto de datos que recibe, el cual ha sido previamente recabado por humanos, en

[4] Actualmente en Cataluña ha comenzado a utilizarse el sistema RisCanvi, con la finalidad de decidir qué presos merecen la libertad condicional.

este caso agentes de policía, con el sesgo que ello implica. Por tanto, el número de crímenes acaba siendo inadvertidamente un indicador racial (Dressel y Farid, 2018). Tenemos entonces que el funcionamiento del algoritmo sería aproximadamente de esta forma: si una persona joven posee un largo historial criminal se asume que posee un alto riesgo de reincidencia, consecuentemente, una persona mayor que haya cometido un número muy reducido de crímenes sería de bajo riesgo. Este sistema resulta sorprendentemente simple, lo cual ha sido considerado a su vez problemático, puesto que la empresa que lo suministra lo promociona como un sistema de gran complejidad, pretendiendo hacer ver que sus cálculos resultan más sofisticados de lo que realmente son. Esto hace que las personas, en este caso los jueces, tomen sus decisiones en mayor estima de la que lo harían si supieran cuan simple es realmente su mecanismo, lo cual aumenta su capacidad de influencia (Farid, 2018). Inclusive, "en cierto caso, los jueces depositaron más confianza en el algoritmo COMPAS que en los acuerdos a los que llegaron la defensa y el fiscal" (Coeckelbergh, 2021, p. 14).

Tenemos entonces que el algoritmo es visto como un mecanismo poseedor de la supuesta capacidad de determinar la probabilidad de reincidencia de un acusado, de igual manera, como vimos anteriormente, las cámaras se utilizaban para decretar quiénes serían potenciales criminales en el futuro. Lo que esto nos indica es que la tecnología se ha utilizado desde sus inicios por parte de los poderes ejecutivo y judicial del Estado no sólo como prueba de delito, como puede ser la existencia de una cámara de seguridad al uso, un radar de velocidad o un alcoholímetro, sino que además se le ha pretendido dar un carácter predictivo. De acuerdo con esta consideración no debería resultar controversial o alarmante que la tecnología se utilizara hoy en día con un carácter predictivo, sin embargo, encontramos el caso del algoritmo COMPAS una diferencia clave, la máquina ya no es un instrumento auxiliar en lo respectivo a la predicción, esto es, que, a diferencia de la cámara, lo que caracteriza al algoritmo no es su calidad de "prueba" sino su capacidad de decisión autónoma. La vigente polémica no viene dada entonces por el uso de las tecnologías con fines predictivos, sino porque estas ya no se utilizan con la finalidad de recabar datos posteriormente utilizados en el juicio, sino que emiten juicios por sí mismas.

3.2. San Francisco prohíbe la tecnología de reconocimiento facial

El pasado 15 de mayo del 2019 el ayuntamiento de San Francisco aprobó, para sorpresa de muchos (pues esta ciudad se ha posicionado durante las últimas décadas como una de las grandes capitales de la innovación tecnológica), la primera legislación que prohíbe

el uso de las herramientas de reconocimiento facial a través de cámaras. Esta decisión se tomó por mayoría (de 8 a 1) en un pleno del ayuntamiento, donde la junta alegó que la privacidad y la protección de minorías debía primar sobre el libre uso de esta tecnología, especialmente si es utilizada por las fuerzas del orden con el fin de reconocer delincuentes (Mendoza Gonzales, 2019, párr. 1). A dicha reunión acudieron representantes pertenecientes a organizaciones en favor de los derechos civiles, quienes alegaron que el uso de esta tecnología puede provocar que la policía realice arrestos erróneos, que afectan especialmente a las minorías éticas. Se apoyan en un estudio realizado por el Centro de Georgetown para la Privacidad y la Tecnología (Mendoza Gonzales, 2019), que asegura que el reconocimiento facial encuentra mayor porcentaje de fallo con rostros de piel oscura.

El acta fue redactada por Aaron Peskin, el concejal promotor del proyecto, quien a su vez mencionó su intención de ejercer mayor presión a nivel estatal ampliando estas medidas a Oakland o Massachussets. Debido a ello, Peskin afirmó lo siguiente: “esta no es una política antitecnológica. Se trata de poder exigir responsabilidades en torno a la tecnología de vigilancia, de garantizar que se hace un uso seguro de ella, (…) se puede tener seguridad sin convertirse en un estado policial” (EFE, 2019, párr. 3). Peskin citó el estudio llevado a cabo en julio de 2018 por la Unión Estadounidense de Libertades Civiles (ACLU, en inglés), la cual halló que el reconocimiento facial había identificado incorrectamente a 28 congresistas estadounidenses (en su mayoría pertenecientes a minorías étnicas) como criminales al comparar sus fotografías con imágenes de un archivo policial. Matt Cagle, un abogado perteneciente a la ACLU, afirma que el uso del reconocimiento facial es un atentado contra la privacidad, inadmisible en una democracia sana. Pone a China como ejemplo de un Estado que hace uso de estas tecnologías para llevar a cabo el seguimiento de minorías (especialmente Uighurs, una minoría mayoritariamente musulmana) y la excesiva invasión de la privacidad ciudadana. Basándose en estos mismos principios la ACLU le pide a Amazon que deje de vender esta tecnología a las fuerzas del orden tras la prueba, propiciada por el estudio Gender Shades, de que sus sistemas llamados Rekognition reportaron reconocer con mayor precisión los géneros de los rostros de piel clara que de piel oscura (Conger, Fausset y Kovaleski, 2019, párr. 10).

A pesar del apoyo recibido la decisión tomada por el ayuntamiento de San Francisco no ha estado exenta de crítica. Jonathan Turley (experto en ley constitucional por la Universidad de George Washington) alega que esta tecnología mejoraría en gran medida la

seguridad en grandes eventos, como conciertos o aeropuertos. A su vez, el presidente de la Unión de Policía de San Francisco, Tony Montoya, dijo que la prohibición de esta herramienta frenaría la investigación policial, y que, a pesar de no ser perfecta, se encuentra todavía en desarrollo, y ha sido efectiva a la hora de encauzar investigaciones policiales. En la misma línea Ed Davis, comisionado de la policía de Boston, dijo que la prohibición había sido demasiado prematura, y referenció además como, en 2013, esta tecnología permitió a la policía de Boston atrapara a los dos terroristas que perpetraron el atentado del maratón de Boston. Dijo que, en un principio, nadie quería seguir el modelo chino, pero la utilidad de la tecnología resultó innegable (Conger et al., 2019, párr. 18-37).

Para adentrarnos con mayor profundidad en la cuestión debemos comprender cómo funciona la tecnología de reconocimiento facial, para identificar así cómo se producen los errores hasta ahora mencionados. Por norma general, los procesos de reconocimiento facial siguen los siguientes cuatro pasos: primero, tu cara es captada por una cámara, en una fotografía, un video, o en tiempo real. Segundo, el software mide tus rasgos faciales, llamados "*landmarks*" (puntos de referencia) o "*nodal points*" (puntos nodales), que marcan, por ejemplo, la distancia entre los ojos, la anchura de la boca, la distancia entre la frente y la barbilla, etc. (Cada sistema utiliza puntos nodales distintos). Tercero, la información adquirida se convierte en una fórmula matemática que representa tu «firma facial», la cual es única, y funciona como una especie de huella dactilar de la cara, una "*faceprint*" en lugar de una "*fingerprint*". Y cuarto, esa firma es comparada con un banco de imágenes en busca de coincidencias, lo cual puede suceder en apenas segundos. (Insider Tech, 2019). Para que este proceso funcione la IA debe haber sido entrenada previamente haciendo uso de un banco de imágenes, lo cual le permite aprender a reconocer caras. Si la base de imágenes no es lo suficientemente variada entonces la IA puede tener problemas a la hora de reconocer rostros que difieran excesivamente del banco de imágenes inicial (por ejemplo, por su tono de piel).

Esta serie de cuestiones han preocupado a su vez a diversos autores, entre ellos se encuentra el experto en ética e IA Mark Coeckelbergh, quien ha mostrado interés por la posible capacidad de las tecnologías de reconocimiento facial de reconocer inclusive estado de ánimo, a través de las expresiones de los sujetos captados. A este respecto nos dice lo siguiente:

> Con cámaras en la calle y en otros espacios públicos, podemos ser identificados y «leídos», incluso por lo que respecta a nuestro estado de ánimo. Mediante el análisis de nuestros datos,

> se puede predecir nuestra salud mental y corporal (sin que lo sepamos). Los empresarios pueden usar la tecnología para monitorizar nuestro rendimiento (Coeckelbergh, 2021, p. 14).

4. Discusión y posicionamiento

Resulta imposible realizar una crítica a la automatización de la vigilancia sin hacer mención de la archiconocida obra *1984* de George Orwell (1949), que se ha posicionado como el buque insignia del género distópico. En ella, como ya es sabido, se nos plantea la idea de una sociedad sometida a una permanente vigilancia, ejercida por la figura del *Big Brother*, cuya función es la de ejercer temor sobre la población, para que eviten disidir de los dictámenes establecidos por el partido único. En este caso se muestra claramente como la vigilancia está dictaminada por una serie de ideales previamente seleccionados, esto es, que la observación de la población se realiza con el fin de asegurarse de que la ideología dominante es acometida a toda costa. Esta serie de consideraciones encuentran su formulación filosófica última en la obra de Foucault, quien, en algunos de sus escritos más célebres, invierte el esquema ideológico de la relación entre el sujeto y las instituciones que componen su contexto sociopolítico, así, en un sentido foucaultiano, decimos que no son las instituciones las que reflejan los ideales sociales, sino que, por el contrario, son los sujetos los que adoptan las ideologías propias de las instituciones. En cualquier caso, lo que ambos autores presentan es la concepción de una vigilancia ideologizada, cuya función es la de asegurarse de que todo lo no normativo es perseguido, y en última instancia sometido o eliminado.

En este caso trataremos de mostrar cómo, en primera instancia, los sistemas de vigilancia y de castigo automatizados se corresponden necesariamente con una serie de valores preestablecidos, que el algoritmo hereda ya sea consciente o inconscientemente de su programador; y segundo, que un sistema de vigilancia y castigos más eficiente no implica necesariamente un mejor funcionamiento atendiendo a sus cuestiones éticas, estipulando así que los principios de eficiencia y moralidad no son equiparables.

4.1. Inteligencia artificialmente humana

Resulta sencillo considerar que una máquina programada por un ser humano pueda ser susceptible de cometer actos perversos, pues estaría a merced de los intereses de su programador, sin embargo, la situación que se nos presenta actualmente resulta un tanto distinta, pues los algoritmos poseen sistemas de autoaprendizaje, lo cual nos inclina a

considerar que, puesto que estos aprenden de manera aparentemente autónoma, no requieren de ninguna fuerza externa, esto es humana, susceptible de fallar o de introducir sus propios intereses en el sistema. La concepción que subyace de esta idea es la de que las máquinas resultan esencialmente, no solo más eficientes, sino a su vez, más justas y objetivas que los humanos, lo cual implicaría que la posibilidad de sustituir la acción humana por la algorítmica supondría una notable mejoría en términos éticos, o inclusive el inicio de un nuevo estadio moral para la humanidad. Sin embargo, los sistemas de autoaprendizaje de las inteligencias artificiales requieren de una base de datos que ya ha sido sesgada de antemano, por tanto, aunque las máquinas sean capaces de aprender por sí mismas, todavía no pueden aprender por sí mismas de sí mismas; precisan necesariamente de un componente externo que les proporcione los datos sobre los cuales se articula el proceso de aprendizaje, y es precisamente esa información externa la que nunca puede ser objetiva, pues está ligada a los conceptos históricos, sociales, culturales, políticos o éticos de las personas. A este respecto el ya citado autor John Tagg menciona lo siguiente acerca de la fotografía:

> La fotografía como tal carece de identidad. Su posición como tecnología varía con las relaciones de poder que la impregnan. Su naturaleza como práctica depende de las instituciones y los agentes que la definen y la ponen en funcionamiento. (…) Al igual que el Estado, la cámara nunca es neutral. Las representaciones que produce están sumamente codificadas y el poder que ejerce nunca es el poder propio (Tagg, 2005, p. 85).

Ocurre lo mismo con las IA, es por ello que el estudio sobre las mismas, como mencionó Antonio Luis Terrones Rodríguez en su seminario *Una ética aplicada a la IA*, no debe ser de carácter técnico, sino político. Tenemos entonces que "Los sesgos de la inteligencia artificial no existen un vacío ni son producto de los algoritmos. Existen en nuestra propia cultura –incluyendo el lenguaje- y luego se oscurecen y propagan a través del uso de sistemas inteligentes"[5] (Sartori, Theodorou, 2022). Por tanto, los algoritmos no superan las limitaciones morales humanas, sino que las heredan y las automatizan, lo cual las hace más difíciles de detectar y, en última instancia, de combatir, amplificando así su influencia. Es preciso entonces cultivar nuevas formas de relacionarnos con las IA, guiadas por un principio fundamentalmente ético, que "serene un ímpetu técnico que ha venido asimilándose como un solucionismo tecnológico" (Terrones Rodríguez, 2023, p. 9).

[5] Originalmente: "Biases in AI do not exist in a vacuum nor are the product of algorithms. They exist in our own culture—including language—and are then obscured and propagated through the use of intelligent systems". (La traducción es propia).

De entre todos los sesos discriminatorios presentes en nuestros días uno de los más presentes es el sesgo racista que, como vimos anteriormente, sigue afectando a las poblaciones racializadas a través del uso de IA tales como el algoritmo *COMPAS* o las tecnologías de reconocimiento facial. A continuación, ahondaremos con mayor profundidad en como las IA heredan y amplifican las características propias de la discriminación racial.

4.2. Herencia y amplificación del sesgo racista

La cuestión acerca de la no inocencia del hacer científico se encuentra ya presente en la obra de unos de los autores antes mencionados, Stuart Hall. Este comenta que la ciencia es habitualmente utilizada para justificar la separación entre seres humanos, lo cual consigue con gran efectividad gracias a su carácter oscurantista. Así, refiere al conocimiento científico como una especie de caja negra que, por ejemplo, a la hora de justificar las diferencias raciales entre seres humanos, alude a la existencia del código genético, un elemento que es comprensible en un laboratorio pero que no puede ser percibido a simple vista (Hall, 2021, p. 366). En su caso, Hall se refería a la justificación científica que durante un largo tiempo sostuvo la concepción de la existencia de "razas humanas" diferenciadas, pero la situación que él nos presenta puede ser extrapolada al ámbito de las IA, en el cual hablamos a su vez en términos de codificación, esto es, nos referimos aquí a los códigos presentes en los algoritmos que rigen el comportamiento de las IA, los cuales no son perceptibles a simple vista, obligándonos a creer que la información que generan es siempre verídica, así, el carácter oscurantista del hacer científico lo hace en cierta forma incuestionable. Si la injusticia se basa en un principio invisible esta se torna incuestionable, por ello, cuando el mecanismo permanece oculto su "naturalización" resulta más sencilla.

Así, es preciso apuntar que, a pesar de la gran popularidad que han adquirido las IA durante los últimos años, la gran mayoría de la población ignora totalmente cómo funcionan los algoritmos que las regulan. Esta ignorancia se debe principalmente a dos factores: una decisión consciente de diseño, dada por razones sociales o económicas, o una imposibilidad técnica, ya que algunos de los enfoques de aprendizaje algorítmico de la actualidad escapan incluso a la comprensión de sus propios programadores (Ananny & Crawford, 2018). De esta manera se forma la naturaleza de "caja negra" de las inteligencias artificiales, la cual desemboca, por parte de la población, en la atribución de muy poca o demasiada confianza en dichos sistemas (Lee & See, 2004). A su vez, muchas compañías se aprovechan de este

desconocimiento para hacer uso de vacíos legales[6]. Como vimos en el caso del algoritmo COMPAS, la empresa mantiene en secreto el funcionamiento del sistema, para que su simpleza no afecte a su credibilidad.

A este respecto Hall menciona que "las funciones *culturales* de la ciencia, en los lenguajes y el discurso del racismo, han sido las de proveer precisamente esa garantía y certeza de absoluta diferencia que ningún otro sistema de conocimiento hasta ahora ha logrado proveer"[7] (Hall, 2021, p. 366). Lo aquí se vindica no es que la práctica científica posea una maldad intrínseca, sino que esta se encuentra siempre a merced del discurso cultural hegemónico, que la utiliza habitualmente con el fin de justificar la diferencia, de normativizarla generando una sensación de esencialismo. Por ejemplo, actualmente se utilizan los datos de índice de criminalidad y de arrestos policiales como forma de probar que las poblaciones negras, indígenas y latinas en EEUU se comportan de manera problemática. Si a esto le sumamos que no es un humano quien realiza los arrestos sino una máquina, podrían eliminarse entonces las acusaciones de racismo que reciben habitualmente los agentes de policía porque, ¿cómo puede ser racista una máquina?, si es un sistema algorítmico, esto es objetivo, quien decide los arrestos, ¿no será acaso que realmente los grupos racializados tienden, de manera esencial o "genética", a la criminalidad? La respuesta es clara, lo que realmente sucede es que la máquina es heredera del sesgo racista de su programador.

Esta cuestión se observa claramente en el funcionamiento de las tecnologías de reconocimiento facial, las cuales clasifican racialmente a los sujetos de acuerdo con las características atribuidas por los humanos, basadas en los aspectos más burdos de la apariencia física. Decíamos anteriormente que esta serie de rasgos visibles son leídos e interpretados por el sujeto, que son significantes a los que se les atribuyen ciertos significados, según Hall, es la gran obviedad de la visibilidad de la raza lo que nos muestra

---

[6] Tal es el caso de los sistemas discriminatorios de tarjeta de crédito de Apple. Para ampliar información véase: Nasiripour, S., Natarajan, S. (2019). Apple co-founder says Goldman's apple card algorithm discriminates. Bloomberg. Retrieved from https://www.bloomberg.com/news/articles/2019-11-10/apple-co-founder-says-goldman-s-apple-card-algo-discriminates. Otro caso notable refiere al algoritmo "sexista" diseñado por Amazon para reclutar trabajadores. Para ampliar información véase: Dastin, J. (2018). Amazon scrapped a secret AI recruitment tool that showed bias against women. Reuters 10 October 2018.

[7] Originalmente: "the cultural functions of science, in the lenguages and discourse of racism, have been to provide precisely that guarantee and certainty of absolute difference which no other systems of knowledge up until that point have been able to provide". La traducción es propia

que esta significa algo (Hall, 2021, p. 370), y ese significado es trasladado del lenguaje humano al digital. Esto sucede porque las tecnologías de reconocimiento facial requieren de una base de datos compuesta por imágenes que han sido previamente seleccionadas, las cuales contienen necesariamente un sesgo, ya que se debe decidir qué imágenes son aportadas y cuáles no, de acuerdo a un criterio específico. Normalmente, cuando se pretende entrenar a un algoritmo de reconocimiento facial, se le suele "alimentar" con dos tipos distintos de imágenes, las que representan rostros y las que no, así, progresivamente la IA comienza a ser capaz de distinguir entre ambas, y de esta forma aprende a detectar rasgos faciales y a distinguirlos de todos los demás elementos percibidos por la lente de la cámara (Kantayya, 2020). El problema surge cuando la base de datos solo posee rostros de personas blancas, en cuyo caso aprende a diferenciarlas con facilidad, desarrollando a su vez una incapacidad para distinguir de manera clara los rostros que presentan rasgos distintos a estas, lo cual tiende a generar confusiones entre personas, llegando a relacionar a sujetos inocentes con historiales criminales que no le perteneces (Kantayya, 2020).

Sin embargo, cuando la IA no se basa en tecnologías de reconocimiento facial, como sucede con el algoritmo COMPAS, ¿cómo puede llegar a relacionar un significante con un significado concreto? Al no poseer una imagen de la persona acusada, carecería a su vez de significante. Bien, como ya hemos vimos, lo que sucede es que, partiendo del significado el propio algoritmo puede intuir el significante, por ejemplo, utilizando un dato como puede ser el número de crímenes previamente cometidos, de manera indirecta, se está tomando la raza como categoría de relevancia a la hora de determinar, por ejemplo, la capacidad de reincidencia del/la acusado/a, puesto que las poblaciones negras o indígenas poseen, de media, historiales delictivos mayores, lo importante es considerar que este factor se debe a sus condiciones de vida y al *racial targeting* que sufren, factores que la IA no tiene en cuenta. Esto es, si uno entrena a la IA para detectar potenciales criminales partiendo de los principios que se utilizan habitualmente con este mismo fin, lo que sucede es que la máquina reproduce necesariamente los mismos sesgos que recibe, incrementando su peligrosidad debido a la automatización. A esta función se la suma la denominada "policía predictiva", que pretende que lo algoritmos sean capaces de "predecir dónde es probable que ocurran los delitos (por ejemplo, en qué área de una ciudad) y quién puede cometerlos, pero el resultado puede ser que un grupo específico socioeconómico o racial esté señalado desproporcionadamente por la vigilancia policial" (Coeckelbergh, 2021, p. 14). Esta práctica ya ha comenzado a realizarse tanto en Estados Unidos como en Europa.

Para ilustrar esta cuestión acerca de cómo puede llegar a ser racista la tecnología podemos hacer uso de la obra *La ballena y el reactor* (1986), escrita por Langdon Winner, en la cual se menciona lo siguiente:

> Los objetos denominados "tecnologías" constituyen maneras de construir orden en nuestro mundo. Muchos artefactos y sistemas técnicos que son importantes en la vida cotidiana contienen posibilidades para ordenar la actividad humana de maneras muy diversas. En forma consciente o inconsciente, deliberada o involuntariamente, las sociedades eligen estructuras tecnológicas que influyen en la forma de trabajar de la gente, en su forma de comunicarse, de viajar, de consumir, etcétera, durante mucho tiempo (…) En ese sentido las innovaciones tecnológicas son similares a los decretos legislativos o las fundaciones políticas que establecen un patrón para el orden público que perdurará por muchas generaciones. (…) Los temas que dividen o unen a las personas en la sociedad se resuelven no sólo en las instituciones y las prácticas de la política propiamente dicha, sino también, en forma no tan obvia, en arreglos tangibles de acero y hormigón, cables y semiconductores, tuercas y tornillos (Winner, 1987, p. 45).

Winner, a modo de ejemplo, realizó un conocido estudio sobre la construcción racista de los puentes de Long Island, en Nueva York, proyecto que corrió a cargo del arquitecto Robert Moses. Esta serie de más de doscientos puentes poseía una altura de apenas tres metros, lo cual hacía imposible que el transporte público, compuesto por guaguas de más de cuatro metros, accediera a Long Island. El objetivo último de esta decisión de diseño era el de alejar a la población negra, en su inmensa mayoría usuaria del transporte público, de las zonas verdes y costeras de Nueva York. Así, el mar pasaba a ser una zona reservada para los blancos de clase alta, que eran quienes poseían suficiente solvencia económica como para mantener un vehículo particular. De esta forma podemos decir que un puente, aparato inerte aparentemente alejado de las decisiones de la vida política, un conjunto de hormigón al que le es imposible discriminar a nadie, se trata realmente una construcción racista, diseñada con intereses discriminatorios. De igual manera una cámara o un algoritmo esconden intereses humanos tras su diseño, su uso y su posicionamiento en el mundo. Así:

> Las máquinas inteligentes replican, duplican, injusticias ya existentes porque, para empezar, se basan en conjuntos de datos sesgados. Automatizan y amplifican las injusticias ya

> existentes (…) la comunidad técnica de la IA requiere transparencia y explicabilidad, capacidad de réplica y responsabilidad (Sartori y Theodorou, 2022, párr. 1)[8].

Realmente los mayores retos que afronta la IA en el presente no están relacionados con sus limitaciones técnicas, sino con sus sesgos discriminatorios. Para ser capaces de enfrentarnos a esta cuestión debemos partir de la base antes mencionada, una comprensión y estudio de la IA en términos políticos y no solo en términos técnicos, que nos permita acercarnos a la cuestión desde una perspectiva más completa, que muestre su verdadera complejidad. Para finalizar este apartado es preciso traer a coalición unas últimas palabras de Winner, que resultan especialmente clarificadoras: "Las dolorosas sutilezas de las medidas y el diseño ocultan fallas vergonzosas en el juicio humano. Nos hemos vuelto cuidadosos con los números, insensibles a todo lo demás. Nuestro rigor metodológico se está convirtiendo en un *rigor mortis* espiritual" (Winner, 1987, p. 199).

4.3. La vigilancia según Foucault

Para continuar en nuestro estudio sería conveniente realizar una breve mención a algunos conceptos fundamentales acuñados por Michel Foucault en lo referido a la vigilancia, ya que este es uno de los autores que la ha tratado de manera más célebre e incisiva. Si pretendemos abordar la cuestión de la vigilancia desde una perspectiva foucaultiana es preciso comentar el denominado "modelo del panóptico", un proyecto formulado por el filósofo británico Jeremy Bentham a finales del siglo XVIII, el cual pretendía reformar el sistema penitenciario presente europeo de su época. En una entrevista que Foucault concedió a Jean-Pierre Barou y Michelle Perrot (Foucault, 1979) recogida en la obra *Los ojos del poder* (1977), define el panóptico de la siguiente forma:

> en la periferia un edificio circular; en el centro una torre; ésta aparece atravesada por amplias ventanas que se abren sobre la cara interior del círculo. El edificio periférico está dividido en celdas, cada una de las cuales ocupa todo el espesor del edificio. Estas celdas tienen dos ventanas: una abierta hacia el interior que se corresponde con las ventanas de la torre; y otra hacia el exterior que deja pasar la luz de un lado al otro de la celda. Basta pues situar un vigilante en la torre central y encerrar en cada celda un loco, un enfermo, un condenado, un obrero o un alumno. Mediante el efecto de contraluz se pueden captar desde la torre las siluetas prisioneras en las celdas de la periferia proyectadas y recortadas en la luz. En suma, se

[8] Originalmente: "Intelligent machines replicate, duplicate, existing inequalities since they rely on biased dataset to start with. As magnifying glasses, they automate and amplify existing inequalities (…) the AI technical community is calling for transparency and explainability, accountability and contestability". La traducción es propia.

> invierte el principio de la mazmorra. La plena luz y la mirada de un vigilante captan mejor que la sombra que en último término cumplía una función protectora (Foucault, 1979, p. 10).

En su obra *Vigilar y castigar* (1975) Foucault hace referencia a este modelo del panóptico, mencionado que, aunque a priori no lo parezca, el verdadero efecto del sistema no se encuentra en la vigilancia efectiva, sino en la interiorizada. El desconocimiento del sujeto sobre su propia condición de vigilado es lo que, en última instancia, modifica su comportamiento. Este comienza a vigilarse a sí mismo ante la posibilidad de poder estar siendo visto, el desconocimiento de su estado, el no saber si está siendo observado, por temor a ser castigado, le lleva siempre por el camino de la prudencia ante la posibilidad de ser descubierto. A este respecto Foucault comenta que el mayor efecto del Panóptico es "inducir en el detenido un estado consciente y permanente de visibilidad que garantiza el funcionamiento automático del poder. Hacer que la vigilancia sea permanente en sus efectos, incluso si es discontinua en su acción." (Foucault, 2003, p. 185). Lo que sucede en este caso es que dicho sistema "automatiza y desindividualiza el poder" (Foucault, 1979, p. 10).

Foucault argumenta que esta estructura arquitectónica sirvió de modelo para diseñar numerosas construcciones hospitalarias de finales de siglo XVIII, y carcelarias de principio del XIX (Foucault, 2003, p. 186), sin embargo, actualmente muchos autores apuntan a que esta no ha sido precisamente la forma en la que se ha logrado la ultravigilancia, sino de manera más simple, a través de las nuevas tecnologías. Así, este nuevo panóptico digital no se basaría en aislar a los presos, sino en mantenerlos interconectados, sustituyendo el *Big Brother* por el *big data*[9] (Han, 2016). A su vez, el uso de las cámaras ha permitido que este sistema sea enteramente posible, hoy en día incluso las tiendas o los parkings se han convertido en verdaderos panópticos. Lo interesante resulta ver cómo este carácter del panóptico se traslada inherentemente al uso de las cámaras y de la IA, que captan imágenes de manera ininterrumpida cuando actúan a modo de sistemas de vigilancia, obligando así a los sujetos a actuar con cautela ante la presencia de las mismas. Lo curioso, como menciona Foucault, es que la cámara ni siquiera requiere estar encendida para realizar su función, pues su mera presencia ya modifica el comportamiento de los presentes. Además, como habíamos visto anteriormente a través de Stuart Hall (Hall, 2021), el carácter oscurantista de la ciencia, que se traduce en el ámbito de las IA en el desconocimiento del funcionamiento de los

[9] Una de las formas más comunes en las que se ejerce la vigilancia a través de las nuevas tecnologías se basa en el requerimiento de ser vigilado para acceder a ciertas comodidades o servicios, tal como sucede con los teléfonos móviles o las *cookies* obligatorias de las páginas web.

algoritmos, hace de sus productos unos métodos de control especialmente efectivos pues, ante la imposibilidad de saber cómo actúa la máquina, como podría ser un sensor, o de saber cuál es su rango de alcance, como podría suceder con una cámara, la reacción natural será siempre la de modificar el comportamiento, pues el desconocimiento contribuye en todo caso a la impresión de ser vigilado. De esta forma, menciona Foucault, "las relaciones de poder figuran entre las cosas mejor ocultas del cuerpo social, lo cual puede explicar por qué se encuentran entre las menos estudiadas" (Tagg, 2005, p. 90). El poder "se ejerce haciéndose invisible; en cambio, impone a aquellos a quienes somete un principio de visibilidad obligatorio" (Foucault, 2003, p. 174). Así se articula una clara línea que une poder y conocimiento, pues "el ejercicio del poder crea continuamente conocimiento, y a la inversa, el conocimiento induce constantemente efectos de poder" (Tagg, 2005, p. 116).

Esta concepción foucaultiana de que "la producción de nuevos conocimientos desencadenaba nuevos efectos de poder, de igual modo que las nuevas formas del ejercicio del poder producían nuevos conocimientos del cuerpo social en trance de ser transformado" (Tagg, 2005, pp. 12-13), encuentra relación directa con la reciente implementación de las IA en los ámbitos de la vigilancia y el castigo social. Subyace de aquí una idea esencial, la de que el poder no se articula solo a modo de prohibición, sino que se da en su mayoría como producción, como creador de realidad. (Tagg, 2005, p. 115). "Foucault, al estudiar el "nacimiento" de la cárcel, busca los posibles efectos positivos de los sistemas punitivos, en lugar de únicamente sus efectos represivos" (Tagg, 2005, p. 115). La idea central aquí no es entonces que la vigilancia de por sí sea en todos los casos innecesaria o negativa, sino que los vigentes mecanismos a través de los cuales se ejerce, y los posibles mecanismos que se nos presentan en un futuro cercano a este respecto, desembocan en consecuencias indeseables. La cuestión esencial a la que nos enfrenta Foucault no es entonces la de la maldad o la represión de forma autosuficiente, no se trata de hacer una crítica a los sistemas de vigilancia, comprendiendo estos como un mero instrumento que permite a los poderosos someter a los débiles. Se trata de algo más complejo, de "investigar la historia política de la producción de la 'humanidad' como un objeto de conocimiento en un discurso que se concibe como 'ciencia'" (Tagg, 2005, p. 116). Esto es, de cómo los avances de la humanidad han traído consigo toda una serie de consecuencias, de cómo toda nueva fuente de conocimiento puede traducirse como emancipación o como subyugo, y de cómo incluso se dan ambas vertientes simultáneamente, afectando a distintas clases sociales.

Toda esta serie de cuestiones cristalizan en una concepción realmente curiosa, la de la "inversión del eje político de la representación". En la antigüedad ser representado era un privilegio, pero actualmente esto ha cambiado, tras el surgimiento de las cámaras, como menciona Tagg, es la privacidad lo que se ha vuelto un privilegio, y la vigilancia una carga de las masas. Esta es la "inversión del eje político de la representación" (Tagg, 2005, p. 85), que se ve ahora amplificada gracias al uso de las tecnologías de reconocimiento facial. La representación, ahora más que nunca, no actúa como una celebración del sujeto retratado, sino como un sometimiento del mismo. Así, nos dice Tagg acerca de la fotografía, una exhaustiva vigilancia permanente es necesaria para desarrollar las nuevas formas de poder que han surgido a partir del siglo XVIII, las requeridas para la explotación productiva de los cuerpos acumulados en masa (Tagg, 2005). La creciente demanda productiva requiere a su vez de una creciente vigilancia, y esta se debe ejercer a través de un sistema capaz de estar presente en medio de la población trabajadora, como puede ser la policía, la cual se implementó con la excusa de que había surgido una nueva amenaza criminal, fabricada por el propio sistema penitenciario y el periódico sensacionalista, la policía nace entonces como un medio para preservar la integridad de la propiedad privada y la riqueza en forma de medios de producción, ya no en manos de sus propietarios sino de la clase trabajadora (Tagg, 2005, pp. 95-98). De igual manera, la inmensa mayoría de sistemas de vigilancia presentes hoy en día están situados en lugares donde se acumula riqueza, tales como inmuebles particulares, bancos o tiendas de todo tipo (joyas, ropa, coches, etc.), lo cual nos incita a preguntarnos no solo qué es lo que buscamos proteger realmente, sino a su vez qué interés incitan al desarrollo de las nuevas tecnologías de vigilancia.

4.4. Experimentos mentales

No cabe duda de que un sistema de vigilancia propiciado por una IA traería consigo ciertas ventajas, de entre las cuales podría destacar su efectividad, la cual se traduciría en una notable bajada de los índices de criminalidad. Esto nos enfrenta a una cuestión, a saber, cuando medimos el carácter ético de un sistema de vigilancia, ¿es acaso la efectividad el factor decisivo, o existe un beneficio en la posibilidad de error? Esto a priori parece contradictorio, cómo puede un sistema de vigilancia ser mejor siendo menos efectivo, esto es, permitiendo más "injusticias". Bien, para abordar esta cuestión haremos uso de dos experimentos mentales, que nos pondrán en la tesitura de enfrentarnos a situaciones en las cuales, paradójicamente, resulta más justo un sistema que permite la "injusticia".

En primera instancia vamos a mencionar algunas cuestiones que nos presenta el autor Evgeny Morozov en su obra *La locura del solucionismo tecnológico* (2013). Morozov nos sitúa en una tesitura realmente curiosa, que nos sirve para ilustrar cual es uno de los grandes problemas de la policía predictiva, basada en la "prevención situacional del delito" (SCP, por sus siglas en inglés). Esta es la concepción de que "las oportunidades de delinquir son la causa del delito y la consiguiente creencia de que es necesario diseñar los entornos de manera tal que el delito se vuelva imposible[10]" (Morozov, 2015, p. 217). Más allá de las posibles consecuencias negativas que podría tener una sociedad dominada por las SCP en lo referido al tipo de ciudadanía que cultivaría, esto es, desconfiada e incapaz de realizar juicios morales propios, pues estarían en cualquier caso "obligados" a hacer el bien (Morozov, 2015, p. 222), nos centraremos a continuación en la necesidad de la protesta, a través de un curioso ejemplo.

Morozov imagina una versión de los EEUU idéntica a la conocemos, pero que se encuentra muchos años más avanzada a nivel tecnológico, en cuyo caso en la década de 1950 poseerían la tecnología de, por ejemplo, 2030. En este caso podríamos imaginar que construyen un sistema que se monta sobre una guagua, cuya función sería la de escanear los rostros de los pasajeros situados en las paradas, para compararlos con una base de datos de previos alborotadores. Gracias a este sistema todas las personas que posean un historial de comportamiento conflictivo en el transporte público, ya sea alborotando o no pagando el billete, serán detectadas antes siquiera de subir al transporte y, cuando deseen entrar, las puertas no se abrirán, y no podrán acceder al vehículo. No cabe duda de que este sistema disminuiría en gran medida el número de conflictos sucedidos anualmente en el transporte público. Pero, ahora esta cuestión nos enfrente a una situación extrema, en la década de los años cincuenta muchas guaguas estadounidenses mantenían sistemas de segregación racial, que separaban a las personas blancas del resto de pasajeros. En este contexto la activista Rosa Parks fue arrestada en Alabama el año 1955, por romper esas precisas normas, convirtiéndose en un gran símbolo del activismo por los derechos civiles. Bien, aunque habitualmente Rosa Parks es recordada como una mujer que, bien por equivocación o por cansancio, decidió no ceder su sitio a una persona blanca aquel día, de manera totalmente inesperada, lo cierto es que aquello se trataba de un acto de protesta premeditado, y que Parks era una reconocida

[10] Los sistemas de SCP suelen basarse en cinco principios: "aumentar el riesgo, aumentar el esfuerzo, disminuir la recompensa, disminuir las provocaciones y disminuir las excusas-, y uno o más de esos mecanismos suelen traducirse en intervenciones del entorno material para prevenir el delito (Morozov, 2015, pág. 218). Por ejemplo, hacer que las vallas sean más altas y colocar alambre de espino en la punta, esto aumenta en gran medida el esfuerzo que debe emplear quien pretende traspasarla, quien posee además un alto riesgo de caer desde una gran altura o de ser herido por el alambre.

activista[11] que había sido arrestada con anterioridad. En aquel momento la protesta fue realizada con éxito, y es recordada hoy en día como un gran avance en la historia del país, y del mundo, sin embargo, situándonos en este contexto turbotecnológico, la IA habría detectado a Rosa Parks como una criminal reincidente, y consecuentemente no le habría permitido subir, negándole así el espacio de protesta. De esta forma, una ley inmoral no podría haber sido nunca cuestionada (Morozov, 2015, pp. 231-232).

Imaginemos ahora un caso menos extremo, en el que un sistema de SCP con base algorítmica es utilizado en los tranvías de Tenerife, y que este sistema, de igual manera, impide que las personas que no han validado el bono puedan utilizar el servicio, creando además perfiles de reincidencia registrados en una base de datos. Pongámonos ahora en una situación en la que una señora mayor necesita recargar el bono mensual para poder hacer uso de dicho servicio. Habitualmente este puede recargarse en todas las paradas o a través de los teléfonos móviles, sin embargo, este proceso, que requiere el uso de pantallas, suele resultar muy difuso para la gente mayor que, en su lugar, suele acudir presencialmente a la estación, donde el personal de información les recarga el bono personalmente. Bien, suponiendo que el bono de la señora ha caducado, y que esta no sabe recargarlo en la terminal, proceso de hecho muy habitual, esta debe realizar, como mínimo, un viaje "ilegal" en tranvía, en el que no se puede validar el bono, para poder llegar hasta la estación. Para evitar que los usuarios viajen gratuitamente en el tranvía la empresa que lo regenta utiliza un sistema de vigilancia humano, basado en la implementación de revisores, quienes al entrar en el tranvía revisan que los bonos han sido validados y multan personalmente a quienes no lo han hecho. Lo cierto es que la efectividad de este sistema se ve muy limitada pues, muchas personas son capaces de burlarlo, ya que los revisores no son capaces de estar atentos a todos viajeros que no validan al entrar. Ciertamente, si este sistema fuera sustituido por uno que impidiera entrar a quienes no validaran su bono, el número de viajes "ilegales" se vería reducido en gran medida, mejorando así la efectividad del sistema anterior. Sin embargo, en lo que respecta a la situación antes mencionada de la señora mayor, este nuevo y mejorado sistema implicaría la imposibilidad de recargar su bono. Además, si en última instancia, el impago del billete quedara registrado en una base de datos, la señora mayor pasaría a ser considerada una pasajera conflictiva, en cuyo caso perdería el derecho a utilizar el transporte público. Por el contrario, un revisor humano, podría dialogar personalmente con la señora, llegando así a un

[11] Por aquel entonces Rosa Parks era secretaria del NAACP (National Association for the Advancement of Colored People).

acuerdo con ella, que podría incluir inclusive el ayudarla personalmente a realizar el pago en alguna de las terminales; obviamente también cabe la posibilidad de que la señora reciba una multa, pero la capacidad del revisor para interpretar situaciones complejas hace que esta no sea la única posibilidad.

Queda claro pues que el sistema automatizado resulta más eficiente en lo que refiere a la prevención de la realización de crímenes, sin embargo, paradójicamente, esto no lo vuelve un sistema más justo. Lo que esto nos indica es que, en lo referido a la vigilancia y el castigo, es preferible un sistema que comprenda el "error humano", ya sea en términos de protesta o en respuesta a las especificidades de cada usuario.

5. Conclusiones y vías abiertas

El problema al que nos enfrentamos es entonces el de la sobrestimación del intelecto sintético, delegamos decisiones en los algoritmos porque pensamos que estos las toman, no sólo con mayor velocidad, sino a su vez con mayor objetividad y neutralidad que nosotros. Sin embargo, como hemos visto, los sesgos resultan un problema significativo en el desarrollo e implementación de las IA, porque las bases de datos que estas requieren no pueden ser nunca objetivas. Esto implica que, de manera necesaria, las IA son una extensión de nuestros sistemas morales, y llevarán siempre consigo los problemas presentes en nuestras sociedades, ya sean de carácter moral, político o económico, acarreando además consigo la posibilidad de amplificar y automatizar sus aspectos discriminatorios. (Sartori y Theodorou, 2022). Toda esta serie de asuntos nos acercan necesariamente a una cuestión, que resulta de acuciante relevancia para nuestros días, y que puede ser resumida en una pregunta, a saber, ¿qué aspectos de la vida política pueden ser realmente relegados a las IA? Bien, esta cuestión ha sido abordada con profundidad por Daniel Innerarity quien, en alguno de sus trabajos más recientes (Innerarity, 2020), se ha preguntado por el alcance que debe poseer la inteligencia artificial en el ámbito de la democracia.

Innerarity considera que el creciente proceso de automatización que está experimentando la política resulta deseable en numerosos aspectos, lo que se plantea es cuáles deben ser los límites normativos que marquen el avance de este proceso (Innerarity, 2020, p. 87). A este respecto nos dice que la pretensión de automatizar la vida política no resulta novedosa y que, tanto en sus inicios como ahora, esta siempre ha tratado de alcanzar

un mayor grado de objetividad al vigente. Esta es precisamente la función de la burocracia, cuyos procedimientos andan en busca de una organización política de la sociedad, así, "como la burocracia para el Estado moderno, la inteligencia artificial parece llamada a ser la lógica de legitimación de las organizaciones y los Gobiernos" (Innerarity, 2020, p. 90). Tenemos entonces que, en la actualidad, "los procedimientos de cálculo y algoritmización prometen neutralizar los prejuicios subjetivos mediante procedimientos exactos de decisión" (Innerarity, 2020, p. 89). Ante esta posibilidad han surgido posiciones tanto tecnofílicas como tecnófobas, ambas defensoras de un mismo principio, que "la lógica de la tecnología puede sustituir a la de la política" (Innerarity, 2020, p. 94), lo que habitualmente varía entre ambas posturas es si consideran esta sustitución, de la política por la tecnología, de manera positiva o negativa. Parece ser que durante los últimos años la segunda postura ha acabado por imponerse a la primera, tanto popular como académicamente. Se ha acrecentado la preocupación por el surgimiento de un autoritarismo tecnológico, una postura que hasta ahora parecía relegada exclusivamente a los ámbitos de la ciencia ficción, presentes en obras literarias y cinematográficas. Este miedo ante la posibilidad de vivir bajo un régimen tecnocrático, en el cuál las decisiones son tomadas por IA de manera autónoma, nos acerca a una cuestión de vital importancia, a saber, qué aspectos de la vida política no pueden ser delegados a las máquinas.

Bien, en esta línea, Innerarity considera que podría suponer un gran riesgo para la democracia el considerar que los sistemas algorítmicos son capaces de superar la capacidad humana para la gestión de recursos, pues esto podría implicar el surgimiento de una "nueva especie de populismo tecnológico (que) podría extenderse bajo la promesa de una mayor eficiencia" (Innerarity, 2020, p. 95). Lo que sucede en este caso es que, a pesar de la gran utilidad que la IA puede aportar a la democracia, a través, por ejemplo, de la recabación masiva de datos, esta no puede encargarse enteramente de ella porque:

> la democracia no es la traducción inmediata y agregada de lo que decidimos individualmente; el carácter dinámico y transformador de la vida democrática incluye un elemento de cambio, descubrimiento y emergencia para lo que no sirve un sistema pensado para hacernos descubrir únicamente lo que ya sabemos (Innerarity, 2020, p. 97).

Lo que Innerarity concluye es que, en el ámbito de la democracia, la IA no debe ser quien tome las decisiones políticas, porque los algoritmos realizan predicciones siempre en base a experiencias pasadas; esto es, mayoritariamente, cuando un algoritmo realiza una

predicción considera que el futuro será similar al pasado. Esto hace que nuestro comportamiento vigente sea plasmado por la máquina como un ejemplo a futuro, lo cual puede generar, como mencionaba Morozov, la imposibilidad de corregir el vigente estado del comportamiento social, el cual puede ser, y de hecho es, sexista, homófobo, tránsfobo, xenofóbico, clasista y, como hemos tratado en profundidad, sórdidamente racista. Sin embargo, como menciona Innerarity, "la política no aspira solo a reflejar lo que hay, sino a cambiar ciertas cosas de algún modo intencional" (Innerarity, 2020, p. 98). Sucede entonces que la vida política comprende aspectos que resultan impredecibles, y los algoritmos sólo son predictivos en tanto que, como mencionábamos anteriormente, asumen una relación de continuidad entre nuestro pasado y nuestro futuro, sin atender a aspectos tales como lo deseos o las aspiraciones de los sujetos. En última instancia la legitimidad de la democracia no procede de su efectividad con respecto al resto de sistemas, sino de su poder de decisión, se haga buen o mal uso del mismo.

Ampliando sobre esta tesis lo que aquí argumentamos es que tanto la acción de vigilar como la de punir comparten en cierta forma relación con la de votar, requieren de una capacidad para interpretar y decidir en entornos complejos y volátiles, que deben ser abordados de la manera más "personalizada" posible; este es precisamente el carácter de los atenuantes judiciales, el de adecuar la ley al sujeto, el cual merece ser tratado justamente, esto es, no según un proceso de igualdad sino de equidad, que comprenda su situación y necesidades particulares. La promesa de una mayor eficiencia, como menciona Innerarity, no puede convencernos para defender un régimen democrático automatizado, y de igual manera tampoco lo puede de un sistema automatizado de vigilancia o de castigo. Seguramente estos resultarían más eficientes que unos llevados a cabo por humanos, pero, en lo referido a situaciones que requieren de tal interpretabilidad, entran en juego cuestiones que requieren de una visión no algorítmica, la cuestión de la permisibilidad resulta crucial a la hora de diseñar sistemas de control, porque esta termina por ser más ética que el principio de eficiencia, tanto en lo referido a la posibilidad de réplica, como a la comprensión de situaciones de excepción. A priori podría parecer que esta tesis resulta excesivamente catastrofista o tecnófoba, sin embargo, lo que aquí defendemos no es la necesidad de abandonar los nuevos avances producidos en los ámbitos de la IA, sino en tratarlos con especial cautela, evitando en cualquier caso implementar su uso automáticamente, esto es, sin haber sopesado previamente sus posibles efectos. Evidentemente, el rápido avance que están experimentando las IA en la actualidad traerán consigo toda una renovada serie de cuestiones morales que será requerido

tratar y reconsiderar a la par que sucede este proceso, lo importante es no permitir que la implementación preceda a la reflexión, pues esta debe servir de guía necesaria para la acción.

6. Bibliografía citada.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. https://doi.org/10.1177/1461444816676645.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016, 23 de mayo). Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. *ProData*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Balibar, É y Wallerstein, I. (1991). *Raza, Nación y Clase*. Iepala.

Boden, M. A. (1984). *Inteligencia artificial y hombre natural*. (Traducción de Armero Sanjosé, J. C). (Trabajo original publicado en 1977). Tecnos.

Coeckelbergh, M. (2021). *Ética de una inteligencia artificial*. (Traducción de Álvarez Canga, L). (Trabajo original publicado en 2020). Cátedra. https://es.scribd.com/document/512363355/Mark-Coeckelbergh-Etica-de-La-Inteligencia-Artificial-Ediciones-Catedra-2021#.

Conger, K., Fausset, R., Kovaleski, S. F. (2019). San Francisco Bans Facial Recognition Technology. *The New York Times*. https://www.beaude.net/traces/downloads/San%20Francisco%20Bans%20Facial%20Recognition%20Technology%20-%20The%20New%20York%20Times.pdf.

Dressel, J y Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1). https://pubmed.ncbi.nlm.nih.gov/29376122/.

EFE. (2019, 15 de mayo). San Francisco prohíbe el uso del reconocimiento facial para identificar a criminales. *La Vanguardia*. https://www.lavanguardia.com/internacional/20190515/462256381193/san-francisco-reconocimiento-facial-prohibe.html.

Foucault, M. (1979). El ojo del poder. En Pierre Belfond (Eds.), *Jeremias Bentham el panótico. Michel foucault el ojo del poder*. (pp. 9-26). Las Ediciones de la Piqueta. https://iedimagen.files.wordpress.com/2012/02/bentham-jeremy-el-panoptico-1791.pdf.

Foucault, M. (2003). *Vigilar y castigar*. (Traducción de Garzón del Camino, A.). (Trabajo original publicado en 1975). Siglo veintiuno. https://www.ivanillich.org.mx/Foucault-Castigar.pdf.

Hall, S. (2021). Race, the Floating Signifier: What More Is There to Say about "Race"? En P. Gilory y R. W. Gilmore (Eds.), *Selected writings on race and difference* (pp. 359-374). Duke University Press.

Han, B. (2016). *En el enjambre*. Herder.

Helmore, E. (2022, 9 de agosto). Tesla's self-driving technology fails to detect children in the road, group claims. *The Guardian*. https://www.theguardian.com/technology/2022/aug/09/tesla-self-driving-technology-safety-children.

Innerarity, D. (2020). El impacto de la inteligencia artificial en la democracia. *Revista de las Cortes Generales* /109/, 87-103. https://doi.org/10.33426/rcg/2020/109/1526.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Mendoza González, S. (2019, 15 de mayo). San Francisco, primera ciudad en prohibir la tecnología de reconocimiento facial en EE UU. *El País*. https://elpais.com/tecnologia/2019/05/15/actualidad/1557904606_766075.htm.

Montiel Álvarez, T. (2016, 23 de enero). *La Fotografía Policial Del Siglo XIX. El sistema Bertillon*. Fotografía Histórica.com. https://fotografiahistorica.com/2016/05/18/486/.

Morozov, E. (2015). *La locura del solucionismo tecnológico*. (Traducción de Viviana Piñeiro, N). (Trabajo originalmente publicado en 2013). Katz.

Pacho, L. (2023, 31 de marzo). Italia bloquea el uso de ChatGPT por incumplir la normativa de protección de datos. *El País*. https://elpais.com/tecnologia/2023-03-31/italia-bloquea-el-uso-de-chatgpt-por-incumplir-la-normativa-de-proteccion-de-datos.html.

Sartori, L. y Theodorou, A. (2022). A sociothecnical perspective for the futurre of AI: narratives, inequalities, and human control. *Ethics and Information Technology*, 24(4). https://doi.org/10.1007/s10676-022-09624-3.

Tagg, J. (2005). *El peso de la representación*. (Traducción de Fernández Lera, A). (Trabajo original publicado en 1988). Gustavo Gili.

Terrones Rodríguez, A. L. (2023). Ética para la inteligencia artificial sostenible. *ARBOR Ciencia, Pensamiento y Cultura*, 198(806), 1-12. file:///C:/Users/usuario/AppData/Local/Temp/Temp1_El%20encuentro%20necesario%20entre%20%C3%A9tica%20e%20IA.zip/El%20encuentro%20necesario%20entre%20e%E2%95%A0%C3%BCtica%20e%20IA/3_Sensibilidad_moral_e_IA_sostenible/E%E2%95%A0%C3%BCtica%20para%20la%20inteligencia%20artificial%20sostenible.pdf.

Terrones Rodríguez, A. L. (2023, 24 de abril). *Una ética aplicada a la IA.* Camino de Extensión Universitaria. ULL. San Cristóbal de la Laguna, España.

Vallance, C. (2023, 28 de marzo). AI could replace equivalent of 300 million jobs – report. *BBC News*. https://www.bbc.com/news/technology-65102150.

Winner, L. (1987). *La ballena y el reactor*. (Traducción de Casals, E. B). (Trabajo original publicado en 1986). Gedisa.

7. Enlaces y referencias audiovisuales.

Insider Tech. (2019, abril, 07). *What's Going On With Facial Recognition? | Untangled*. [Video]. YouTube. https://www.youtube.com/watch?v=BqQT4sIOYA0&t=1s.

TEDx Talks. (2018, octubre, 02). *The danger of predictive algorithms in criminal justice | Hany Farid | TEDxAmoskeagMillyard*. [Video]. https://www.youtube.com/watch?v=p-82YeUPQh0.

Kantayya, S. (Directora). (2020). *Sesgo Codificado* [Documental]. 7th Empire Media, Ford Foundation - Just Films, Chicken & Egg Pictures.