Cuestiones éticas sobre la implantación de la inteligencia artificial en la administración pública

Pedro Juan Baquero Pérez
Profesor asociado de la Universidad de La Laguna y jefe de servicio de informática y comunicaciones del Gobierno de Canarias
https://orcid.org/0000-0002-5545-0450

RESUMEN: Este artículo explora las ramificaciones éticas y la responsabilidad moral al implementar la inteligencia artificial en el ámbito público. El texto aborda conceptos clave como inteligencia artificial, ética y responsabilidad, analizando diferentes teorías éticas aplicables, la toma de decisiones morales y la existencia de una ética específica para la administración pública. Se cuestiona si estamos transmitiendo un mensaje adecuado sobre la IA a la sociedad y se examinan las consideraciones morales para su implementación, como la privacidad, la seguridad, la explicabilidad, la justicia, el impacto en los trabajadores y otros efectos sociales. Además, se discute la posibilidad de programar la ética, cómo tratar los peligros y cómo abordar las decisiones morales en la IA. Finalmente, se reflexiona sobre la atribución y distribución de responsabilidades morales en la IA y los retos que enfrentan las administraciones públicas en términos de qué hacer, cómo y cuándo actuar, y quiénes deben estar involucrados en el proceso.

Palabras clave: Inteligencia artificial (IA), administraciones públicas, ética, responsabilidad, privacidad, seguridad, explicabilidad, toma de decisiones

ABSTRACT: This paper explores the ethical ramifications and moral accountability in deploying AI within the public domain. The text addresses key concepts such as artificial intelligence, ethics, and responsibility, analyzing different applicable ethical theories, moral decision-making, and the existence of specific ethics for public administration. It questions whether we are conveying an appropriate message about AI to society and examines the moral considerations for its implementation, such as privacy, security, explainability, fairness, impact on workers, and other social effects. Additionally, the possibility of programming ethics, how to address hazards, and how to tackle moral decisions in AI are discussed. Finally, the paper reflects on the attribution and distribution of moral responsibilities in AI and the challenges public administrations face in terms of what to do, how and when to act, and who should be involved in the process.

Keywords: Artificial intelligence (AI), public administrations, ethics, responsibility, privacy, security, explainability, decision-making

1. INTRODUCCIÓN

En los últimos años se ha producido un desarrollo notable en las tecnologías relacionadas con la inteligencia artificial (IA). El ejemplo paradigmático más reciente de este progreso es ChatGPT, un modelo de lenguaje desarrollado por la empresa *OpenAI* basado en modelos de redes neuronales. ChatGPT es capaz de generar respuestas coherentes y contextuales a preguntas o solicitudes de los usuarios, lo que lo convierte en una herramienta valiosa en una amplia variedad de aplicaciones, desde asistentes virtuales y chatbots hasta generación de contenido y análisis de texto. Este avance en IA se debe a una combinación de factores, incluidos los grandes volúmenes de datos disponibles, la mejora en las arquitecturas de redes neuronales y el aumento en la capacidad de procesamiento. Los modelos de lenguaje, como ChatGPT, se entrenan con vastas cantidades de texto, lo que les permite aprender patrones y relaciones semánticas y sintácticas en el lenguaje humano, permitiéndoles generar respuestas relevantes y precisas en función de las preguntas o solicitudes planteadas. A pesar de estos avances, aún existen desafíos en el campo de la IA, como la comprensión profunda del contexto, la ética y la privacidad, y la capacidad de adaptarse y aprender de forma autónoma en entornos dinámicos y cambiantes. Sin embargo, con el desarrollo continuo y la investigación en inteligencia artificial, está claro que veremos aún más mejoras y más aplicaciones innovadoras en el futuro.

Las administraciones públicas también han adoptado y se han beneficiado de los avances en inteligencia artificial en los últimos años. Al integrar estas tecnologías en sus procesos y servicios, pueden mejorar la eficiencia, reducir costes y proporcionar mejores servicios a los ciudadanos (Zuiderwijk et al., 2021). Son

indudables los beneficios que supone la incorporación de la IA en las organizaciones (Khanzode y Sarode, 2020), no solo por la incorporación de la gran capacidad de acceso, proceso y almacenamiento de una mayor cantidad de datos, sino que podemos delegar en la IA muchas de las tareas que hacen las personas. Muchas de estas tareas están asociadas a la toma de decisiones, las cuales tradicionalmente han estado en manos de personas. De esta forma, una persona puede hacer uso de un servicio público, por ejemplo, para solicitar una ayuda para estudios, sin que en ningún momento esta solicitud llegue a ser examinada por un funcionario. Desde que la IA ni se cansa ni se aburre y trabaja las 24 horas todos los días de la semana, realizando sus tareas en tiempos mucho más reducidos que los que necesita un humano, se puede afirmar que cuando delegamos en las máquinas las tareas que hacemos los humanos tiene grandes ventajas. Por otra parte, la inteligencia artificial puede ser considerada neutral, dado que no se ve afectada por conflictos de interés o situaciones de corrupción que a veces están involucradas las personas. En cualquier caso, este nuevo paradigma de servicio público requiere de reflexión. Así, a pesar de los beneficios potenciales de la IA en las administraciones públicas, también existen desafíos y preocupaciones en aspectos técnicos, regulatorios y éticos (Baquero Pérez, 2023), dentro de estos últimos destacamos la privacidad, la ética, la equidad y la transparencia. Por tanto, es esencial abordar y reflexionar sobre estos temas a medida que las administraciones públicas plantean, adoptan y expanden el uso de la IA para garantizar que se utilice de manera responsable y se maximicen los beneficios para los ciudadanos. Una reflexión en la que tenemos que plantear ciertas cuestiones éticas.

Este artículo trata de contestar a una serie de preguntas sobre la IA desde el punto de vista ético. Inicialmente nos interesa conocer qué entendemos por la IA y por ética, para a continuación entrar en una reflexión moral sobre la implantación de la IA en las administraciones públicas. Son muchas las cuestiones que nos podemos plantear, pero que podemos sintetizar en unas pocas grandes preguntas. La primera pregunta es si realmente una administración pública está enviando un mensaje adecuado a la sociedad cuando indica que está utilizando la IA en sus servicios. La siguiente pregunta viene sobre la reflexión sobre qué peligros y problemas morales nos podremos encontrar cuando utilizamos la IA para que se tomen decisiones por las aplicaciones. Otra pregunta de índole práctica es si realmente podemos programar las aplicaciones tanto para solucionar o mitigar los problemas éticos o para implementar aspectos morales. También, es fundamental abordar la cuestión sobre cómo atribuimos y distribuimos las responsabilidades en esta dinámica donde se delegan en la IA la toma de decisiones. Por último, la última gran cuestión viene sobre cómo las administraciones públicas pueden enfrentarse a los retos éticos de la IA.

2. ¿QUÉ ENTENDEMOS POR LA IA?

La IA no se puede asociar con una tecnología concreta, realmente es más un campo científico o un concepto que una tecnología. Wirtz (2019) parte de distintas

definiciones de distintos autores acerca de la IA para hacer una definición integradora: *"la IA se refiere a la capacidad de un sistema informático de mostrar un comportamiento inteligente similar al humano, caracterizado por ciertas competencias básicas, como la percepción, la comprensión, la acción y el aprendizaje"*[1]. Dentro de esta definición, cuando se habla de un comportamiento inteligente similar al humano nos referimos a la capacidad humana de resolver problemas en un tiempo determinado. En este sentido, la IA intenta replicar el pensamiento y el aprendizaje humano en algoritmos que puedan ser tratados por una máquina. Consideramos un algoritmo como un sistema de reglas que define un proceso compuesto de pasos concretos que resuelven un problema o alcanzan un resultado.

Por otra parte, para poder contestar a la pregunta sobre qué se entiende por la IA tenemos que concretar lo que sería un sistema de IA a través de las partes que lo contienen. De forma general, podemos considerar con el concepto de IA como un sistema con el que nos aproximamos al comportamiento humano. En este sentido, intuimos que esta aproximación puede tener diferentes grados, por ello, es importante contestar a una pregunta sobre si toda la IA se aproxima en la misma medida al comportamiento humano. Por último, el campo científico de la IA hace uso de una gran diversidad de técnicas, por ello, para tener una visión más clara de la IA es necesario conocer y clasificar los distintos tipos de técnicas que se utilizan.

2.1. ¿Qué partes podemos localizar en la IA?

La IA se concreta en un sistema compuesto por distintas partes, donde a este sistema también muchas veces se le denomina con el término de máquina inteligente o, en nuestro contexto, simplemente, máquina. A partir de la anterior definición podríamos plantear cuatro partes que conforman un sistema de IA o una máquina: la percepción, la comprensión, el aprendizaje, el modelo y la acción.

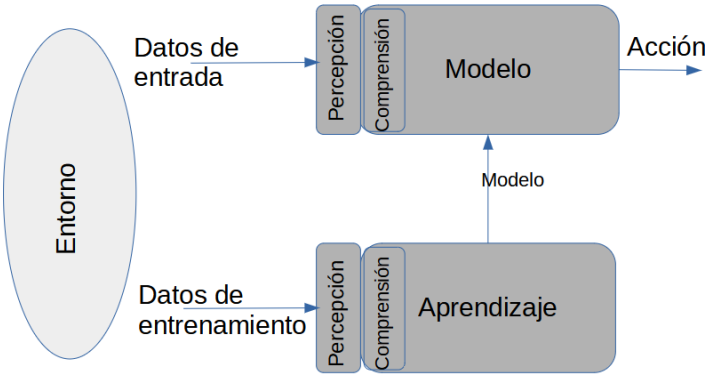

Figura 1: Partes de un sistema de inteligencia artificial

[1] La cita de Wirtz ha sido traducida por el autor del presente artículo, al igual que el resto de las citas cuyo original está en inglés.

La **percepción** es básicamente la parte que permite la obtención de datos del entorno. Como se observa en la figura 1, estos datos pueden ir destinados para ser tratados por el **aprendizaje** o por el **modelo**. Los datos destinados para el aprendizaje son aquellos datos donde, además, se incluyen resultados asociados, los cuales son ya conocidos. Estos serían los datos de entrenamiento; por ejemplo, si los datos son imágenes de animales, cada imagen vendría etiquetada con el tipo de animal de cada imagen, es decir, con el resultado asociado. En cambio, los datos destinados al modelo son aquellos datos que no conocemos el resultado y son los datos de entrada; por ejemplo, si los datos también son imágenes, tendríamos que pronosticar qué tipo de animal contiene la imagen a partir del modelo que se ha generado con el aprendizaje. En general, la fuente de los datos del entorno es heterogénea, tanto refiriéndonos al origen como a la naturaleza de los datos. El origen de los datos en una administración pública puede provenir de los mismos expedientes administrativos, pero también de su entorno asociado, como estadísticas públicas o privadas, encuestas, redes sociales y la monitorización del mismo entorno. Con la naturaleza de los datos entendemos lo que está más relacionado con el formato, el cual puede ser estructurado, como lo que nos proveen las bases de datos, o no estructurado, como imágenes, audios, vídeos, conversaciones y documentos. Es parte de la **percepción** las funciones que permitan captar estos datos, como la recuperación de la información, las interfaces hombre-máquina o máquina-máquina, el reconocimiento del habla y de imágenes. En general, también se intenta convertir los datos no estructurados en estructurados.

La **comprensión** es la capacidad para entender los datos que se han adquirido con el fin de generar mayor conocimiento. En otras palabras, es dar algún significado a los datos. La comprensión no deja de ser un hecho problemático dado que en cierto sentido se requiere que la máquina tenga conciencia de los datos que maneja. Por ello, dentro de la IA hay que acotar este término. Una máquina no solo tiene que adquirir los datos, sino que tiene que saber qué hacer con ellos en función de lo que significan. Esta comprensión puede abarcar diversos aspectos. Por un lado, una máquina debe entender las necesidades del humano que interactúa con la máquina, de forma que en función del tipo del dato haga una cosa u otra. Por otra parte, a los datos se les debe, en alguna forma, etiquetarlos, por ejemplo, si una máquina lee un texto, y se encuentra con la palabra "como", debe saber si se refiere a una conjunción o al verbo "comer".

El **aprendizaje** es la capacidad del sistema de adquirir conocimiento del entorno conocido. Básicamente, este aprendizaje puede ser supervisado o no supervisado en la medida que un humano le indica a la máquina qué es lo que tiene que aprender o es la máquina autónomamente quién aprende. Por ejemplo, pensemos en los sistemas de clasificación de documentos. En el caso de aprendizaje supervisado, una máquina aprenderá a partir de los datos de entrada, los cuales son un conjunto de documentos donde cada uno se ha etiquetado previamente por un humano; entonces la máquina aprende que para los documentos que

tiene la misma etiqueta cumplen con unos patrones específicos, por ejemplo, la máquina detecta que existe un patrón específico para una etiqueta, como podría ser que exista una abundancia de términos específicos de deportes. En el caso de aprendizaje no supervisado los datos de entrada son un conjunto de documentos sin etiquetar; la máquina aprende que existe cierto subconjunto de documentos que tienen un patrón específico, por ejemplo, la máquina detecta que se repiten ciertas palabras con mucha frecuencia en este subconjunto de documentos. Por otra parte, existen otros tipos de aprendizaje, como el denominado por refuerzo, donde la máquina descubre las acciones que dan mejores resultados cuando se prueban. En definitiva, el aprendizaje se basa en procesar los datos del entorno que se tengan con sus resultados conocidos y en la medida que se dispongan de más datos y de mayor calidad, las máquinas aprenderán mejor. Por ello, la IA necesita de una correcta gestión del dato de origen si queremos que aprenda bien.

El **modelo** surge por medio del proceso de aprendizaje, el cual permite al sistema registrar lo que ha aprendido. Será el modelo el que procese los datos del entorno para obtener un resultado que lo hemos denominado como acción. La acción tiene mucho que ver con la toma de decisiones. Las decisiones se toman entre un conjunto de posibilidades o acciones. Estas posibilidades pueden conducir a consecuencias, las cuales pueden ser buenas o dar resultados deseables, consecuencias malas o resultados no deseables, o dar resultados neutros. Por ejemplo, en un sistema de conducción automática, el sistema de visión detecta que el semáforo está en rojo con un 10% de probabilidad, en amarillo con un 15% y en verde con un 75%; el sistema tiene dos posibilidades: decidir si frena o continua. En este caso, las consecuencias pueden ser buenas (no hay accidente o no incurre en infracción si el semáforo realmente estuviese en verde) o malas (hay accidente o incurre en infracción si el semáforo estuviese en rojo). En este sentido, la acción no deja de ser el fin último de un sistema de IA, donde muchas veces se requiere la predicción ante nuevas situaciones, lo cual se puede utilizar para la toma de decisiones, el control y la planificación.

2.2. ¿Toda la IA se aproxima de la misma forma al comportamiento humano?

La concepción de que el comportamiento inteligente de una máquina sea similar al humano no deja de ser problemática desde que en la actualidad la IA es incapaz de hacer muchas cosas que son muy sencillas para los humanos. No toda la IA se aproxima al comportamiento humano de la misma manera, Kaplan (2016) divide a la IA en tres tipos: la IA débil o específica, la IA fuerte o general y la superinteligencia artificial.

La IA débil trabaja en tareas y en contextos muy concretos dejando de ser útil en otros contextos y situaciones y, además, es programada por los humanos. Esta IA es la que tenemos actualmente. Por otra parte, la IA fuerte tiene la capacidad de aprender por sí misma, transmitiendo su experiencia a otras tareas

y contextos sin la ayuda de personas. De esta forma, su inteligencia es capaz de trabajar en una amplia gama de contextos. Algunos investigadores, como Fjelland (2020) y Larson (2021), son pesimistas y consideran que no se podrá alcanzar este tipo de inteligencia en su concepción más amplia, estos autores creen que con el conocimiento actual no llegaremos a ver este tipo de inteligencia en las máquinas. Sin embargo, autores como Klinger et al. (2018) afirman que se puede alcanzar este tipo de IA, incluso se han planteado marcos para avanzar en este tipo de inteligencia (Rosa et al, 2016). Por otra parte, autores como Pearl y Mackenzie (2018), han iniciado caminos de investigación para aproximarnos hacia este tipo de inteligencia tratando de vencer las dificultades que tienen las máquinas de gestionar la causa y el efecto. Otro ejemplo, es la misma empresa que ha desarrollado ChatGPT, que tiene entre su misión conseguir una IA general. Por último, con la superinteligencia artificial tendremos máquinas que superen la capacidad humana y lleguen a tener consciencia, donde se denomina singularidad a ese punto o momento a partir del cual se consigue este tipo de inteligencia.

Tal como se ha dicho, actualmente el estado del arte de la IA se encuadra dentro de la IA débil, incluso ChatGPT (Taecharungroj, 2023), que, sin duda, puede considerarse como un gran avance hacia la creación de una inteligencia artificial fuerte, y su desarrollo podría desencadenar importantes transformaciones en la sociedad humana en los próximos años. Aunque ChatGPT es un modelo avanzado basado en arquitecturas GPT (Radford, 2019) y es capaz de comprender y generar texto de manera coherente en una amplia gama de temas, sigue siendo limitado en sus capacidades y no tiene la comprensión general, la autoconsciencia o el razonamiento profundo que define a la inteligencia artificial fuerte o general. ChatGPT está diseñado para tareas específicas, como responder preguntas, mantener conversaciones, ofrecer sugerencias y generar contenido textual. Está entrenado en un conjunto de datos extenso y puede proporcionar respuestas coherentes y útiles en muchos casos, pero no puede adaptarse o aprender de manera autónoma en la misma medida que un ser humano o una inteligencia artificial general lo haría.

2.3. ¿Las técnicas utilizadas por la IA son similares?

En cada una de las partes anteriormente descritas se aplica algún tipo de técnicas de razonamiento, así, se puede aplicar tanto a la percepción (por ejemplo, para reconocer una cara), en la comprensión (por ejemplo, para determinar si una palabra es un nombre, verbo, etc.), en el aprendizaje (por ejemplo, para hacer una clasificación por tipos de documentos) o en las acciones (por ejemplo, para tomar una decisión). Con la IA actual, es decir, con la IA débil, se puede hacer una clasificación de la IA en base al tipo de técnicas que se aplican. A grandes rasgos, las técnicas pueden basarse en reglas, técnicas estadísticas y redes neuronales, a cada uno de conjuntos de técnicas se les asocia con un paradigma, de-

nominados, respectivamente como simbólico, estadístico y conexionista. Estas técnicas han experimentado un proceso de desarrollo, que de forma simplificada se puede decir que empezó con el paradigma simbólico en los años 50 hasta los años 90. En la fase siguiente toma mayor importancia el paradigma estadístico hasta la primera década de este siglo, a partir de entonces empieza un importante desarrollo del paradigma conexionista. Ninguno de estos paradigmas es excluyente, en el sentido de que actualmente se aplican los tres paradigmas y en un mismo sistema pueden estar todos incluidos.

En el paradigma simbólico son las reglas lo que predomina. Las reglas se pueden aplicar de forma distinta. Así, se puede extraer el conocimiento humano de un experto en forma de reglas tomadas a mano. Por ejemplo, uno se puede sentar con un médico y a través de preguntas se puede extraer los pasos y reglas que lleva a cabo para diagnosticar una enfermedad: son estas reglas lo que se implementan en un sistema informático. A este tipo de soluciones se les suele o solía denominar como sistemas expertos. También, en el procesamiento del lenguaje natural las reglas pueden asociarse con las reglas sintácticas, por ejemplo, "un sujeto va seguido de un verbo". Este paradigma tiene como principal ventaja su interpretabilidad, además, las causas y los efectos suelen estar bien definidos. Por contra, su mayor desventaja de este paradigma es que en gran parte de los escenarios el universo de casos es intratable, aplicándose solo a un dominio restringido de casos.

En el paradigma estadístico se aprende de los datos del entorno, en lugar de definir las reglas de forma manual. En otras palabras, podemos hablar de pasar de un sistema racionalista a un sistema empirista. Este paradigma surge del importante desarrollo de las técnicas estadísticas y la cada vez más disponibilidad de datos. A partir de una gran cantidad de datos se pueden llegar a establecer reglas empíricas a través de correlaciones, en lugar de causas y efectos. Por ejemplo, el diagnóstico de una enfermedad, en lugar de sentarse con el experto en medicina, a partir del análisis de base de datos se obtienen reglas, las cuales no tienen que estar sujetas a causa y efecto; o en el caso del procesamiento del lenguaje, a partir del análisis de corpus de textos, se podría establecer reglas asociadas con la probabilidad que después de un sujeto o una palabra exista un verbo o se prediga siguiente palabra. La principal ventaja de estas técnicas es que permite modelar reglas y descubrir información que en muchos casos está oculta a los humanos, con lo que es más versátil que aprender o recoger las reglas manualmente. Su principal desventaja viene de los posibles sesgos si los datos no tienen la suficiente calidad y en cantidad. Otra de las desventajas es que las técnicas estadísticas no tienen por qué recoger las relaciones de causa y efecto, lo que muchas veces dificulta la interpretabilidad.

En el paradigma conexionista el concepto básico en que se basa son las redes neuronales. Podemos considerar las redes de neuronas como modelos matemáticos que intentan emular muy simplificadamente el funcionamiento de las co-

nexiones entre las neuronas en el cerebro, y a la misma neurona se la simplifica con reglas matemáticas simples de forma que se activa o no una neurona. Este paradigma, aunque existía la base teórica desde los años 70, es a partir de 2010 con el gran desarrollo de la capacidad de cómputo de los ordenadores cuando se empiezan a aplicar predominantemente. La principal ventaja de este paradigma es que es más capaz de tratar situaciones complejas, lo que permite más versatilidad a solucionar problemas diversos aplicando las mismas arquitecturas de redes neuronales. Por contra, el entrenamiento de estas redes consume grandes recursos informáticos, necesitando en muchas ocasiones recursos que están solo disponibles por grandes empresas. Además, estas redes neuronales muchas veces son cajas negras en el sentido de que no sabe cómo están tomando las decisiones lo que dificulta en gran medida su interpretabilidad, más aún que en el paradigma estadístico.

3. ¿QUÉ ENTENDEMOS POR ÉTICA?

No existe una definición única de la ética. En este trabajo utilizaremos la definición de Van de Poel y Royakkers (2011, p. 71) como *"la reflexión sistemática sobre la moral"*, donde la moral la consideramos como *"el conjunto de opiniones, decisiones y acciones con las que las personas, individual o colectivamente, expresan lo que consideran bueno o correcto"*. Por tanto, la ética y la moral son dos conceptos distintos, aunque relacionados. Con la ética se procura iniciar un proceso para buscar la moral correcta, con el objetivo de hacer frente a los problemas morales. En verdad, la ética no nos provee con una guía que nos dé respuestas claras, sino que nos permite reflexionar sobre las opciones que podemos tomar. El estudio de la ética lo podemos abordar tanto desde un punto de vista descriptivo como prescriptivo. En el primer caso, principalmente trata de la descripción de la moral existente, incluyendo las actuales costumbres, así como lo que se considera bueno, malo y neutral. En cambio, la ética prescriptiva o normativa juzga la moralidad, mientras la descriptiva no emite juicios sobre esta. Con la ética prescriptiva se tratará sobre cómo nos debemos comportar para ajustar nuestras ideas a los valores y normas, para lo que se harán argumentos que se valdrán de diversas teorías éticas, lo que nos permite discutir críticamente sobre aspectos morales. Cuando aplicamos la ética a problemas concretos hablamos de ética aplicada.

Por tanto, cuando tratamos de aplicar la ética sobre el campo concreto de la IA tenemos que pensar en la ética aplicada. Por ello, hay que conocer qué teorías tienen un mayor consenso en el campo de la ética aplicada y, en base a estas teorías, hay que comprender cómo se toman las decisiones morales. Por otra parte, existen cuestiones que concretan en mayor medida la aplicación de la ética de la IA que tratan sobre la existencia de hechos morales objetivos, su aplicación a la administración pública y cómo tratar la responsabilidad de las decisiones morales.

3.1. ¿Qué teorías de la ética aplicamos?

En la ética aplicada existe un conjunto amplio de teorías éticas sobre cómo actuar, apareciendo contradicciones entre ellas. Existe cierto consenso en que las principales teorías en la ética aplicada son el consecuencialismo, la deontología y la ética de la virtud (Camps, 2022: pp. 394-399). El consecuencialismo se centra en las consecuencias de las acciones, la deontología en las mismas acciones y la ética de la virtud en el actor.

El consecuencialismo tiene su base en el utilitarismo, creado por Jeremy Benthan (1748-1832) y corregido por John Stuart Mill (1806-1873). El utilitarismo se basa en el principio de utilidad que busca la mayor felicidad para el mayor número, en otras palabras, necesita de realizar un balance moral para conocer qué acción tiene las mejores consecuencias. Esta teoría utilitarista no adolece de problemas. Una de las críticas es que puede conducir a la explotación de las mayorías sobre las minorías. Otra crítica es que la acción que conduce a la mejor consecuencia puede ser moralmente mala. Por ello, con el fin de mejorar el utilitarismo inicial se ha incorporado una serie de conceptos que ha conducido al consecuencialismo. Así, al utilitarismo se le ha incorporado el principio de libertad, que nos indica que cada uno es libre de buscar su propio placer, siempre que no impida el placer de los demás. También, se ha incorporado el utilitarismo de las reglas que se centra en mayor medida en la utilidad de las reglas de acción en lugar de en la utilidad de los actos individuales. Por último, una de las críticas de esta teoría viene de la dificultad de medir las consecuencias, como la felicidad o el placer.

La deontología, también denominada ética del deber o de principios, tiene la base en la ética de Immanuel Kant (1724-1804). Básicamente, su teoría se puede resumir en el principio de universalidad y en el principio del fin en sí mismo. El principio de universalidad nos dice que hay que actuar de manera que también en todo momento sirva como ley universal. El principio del fin en sí mismo nos dice que hay que actuar de manera que trates a la humanidad siempre como un fin y no como un medio. Las principales críticas a esta teoría son su rigidez y los conflictos entre las normas. Con el fin de contrarrestar estas críticas se plantea que las normas se consideren no como normas universales sino como punto de partida.

La ética de la virtud se remonta a Aristóteles (384 a.C-322 a.C), se centra en el actor y no en sus acciones o consecuencias. Son las virtudes que tiene el actor las que son la base para que se actúe moralmente. Entre las críticas a esta teoría está que no se nos dice cómo se tiene que actuar y que las virtudes que se poseen las personas consideradas virtuosas no tienen que ser incondicionalmente buenas.

3.2. ¿Cómo tomar decisiones morales?

Las distintas teorías nos pueden dar respuestas diferentes, aunque cada teoría nos puede servir como guía para analizar los aspectos éticos en cada situa-

ción. Tal como afirma Camps (2022: p. 395), no son opuestas la ética de los principios y la ética de las consecuencias, sino complementarias, y un complemento a ambas es la ética de las virtudes, ya que se carecería de la mediación entre la teoría y la práctica que consiguen las virtudes: "*Las virtudes se asientan en el sentimiento y se materializan en hábitos, en costumbres, que se traducen en tendencias a actuar bien o mal*" (Camps, 2022: p. 398).

Por tanto, las distintas teorías nos sirven para tener un marco de análisis y para disponer de argumentos para actuar de una forma u otra. En otras palabras, a pesar de que el uso de estas teorías no tiene que garantizar que se llegue a la decisión correcta, lo que sí nos permite es tomar decisiones que no ignoren las teorías éticas. En definitiva, las decisiones éticas no consisten en que nos guiemos por una sola teoría, sino que se tengan en cuenta las diversas teorías.

3.3. ¿Existen los hechos morales?

A pesar de que tenemos diversas teorías que nos pueden conducir a diferentes respuestas sobre lo que está bien o mal, lo que sí es cierto es que muchos hechos están bien o mal para todas las teorías. Frente al relativismo de los valores que, en su esencia, postula que no hay un ningún orden superior que nos indique lo que está bien o mal, Gabriel (2021, 33) defiende el nuevo realismo moral, donde su primera tesis básica es que "*existen los hechos morales que son independientes de las opiniones personales y colectivas; su existencia es objetiva*". Por otra parte, y tal como nos dice Camps (2022: p. 405), aunque la ética a lo largo de su historia nos ha legado un marco kantiano-utilitarista que debe ser gestionado por personas virtuosas, si no existiese un conjunto de normas universales que nos sirvan de referencia caeríamos "*irremediablemente en la falta de criterios éticos y en el relativismo cultural*".

Además, estos hechos morales se pueden comprender en lo esencial, están destinados a las personas y nos dan una guía sobre lo que es bueno, malo o neutro. También, estos hechos morales que tienen existencia objetiva son correctos siempre. En cualquier caso, Gabriel defiende que no existen reglas que permitan resolver para siempre los problemas morales, ya que nos podemos equivocar. En otras palabras, no siempre es fácil encajar las acciones humanas en buenas (lo que se debería hacer), neutras (lo que está permitido hacer) o malas (lo que hay que evitar hacer). En cualquier caso, en la actualidad tenemos los Derechos Humanos Universales que, tal como nos indica Valcárcel (2002: pp. 49-71), serían nuestra tabla de mínimos.

Por otra parte, existe el progreso moral, en el sentido de que la humanidad va acumulando cada vez más conocimiento moral pero que no resolvemos los problemas éticos de forma definitiva, de esta forma, se va avanzando hacia los hechos morales verdaderos (Gabriel, 2021: p. 277).

3.4. ¿Existe una ética para la Administración Pública?

Tal como nos indica Weber (1946: 117-124), una administración pública debe pensar en las consecuencias de sus actuaciones, es decir, su ética podría considerarse que tiene que ser utilitarista, aunque, al mismo tiempo, afirma que el político debe tener principios. Por tanto, ambas éticas, la consecuencialista y la deontológica, deben actuar al unísono. Por otra parte, los principios de actuación que debe incorporar una administración pública democrática deben ser concretados y como mínimo debería incorporar los Derechos Fundamentales. Así, en la UE existe un marco ético donde están reflejados los valores que los Estados miembros consideran fundamentales para el funcionamiento de los servicios públicos (Radhika, 2012). Estos valores son el Estado de Derecho, es decir, la legalidad, la imparcialidad y objetividad, la transparencia, la rendición de cuentas, la profesionalidad, y el deber de atención, la fiabilidad y la cortesía. Como valores fundamentales, estos deberían estar incorporados en todas las actuaciones de las administraciones públicas. En este sentido, una administración pública, siguiendo la legalidad, debe cumplir las leyes, y esta será su principal guía en sus actuaciones, como es el caso cuando se implanta la IA.

3.5. ¿Cómo tratar la responsabilidad?

Cuando se trata de sistemas que puedan dar lugar a acciones (donde entre las acciones también está incluida la inacción) o efectos indeseados es importante la rendición de cuentas. La responsabilidad se vincula al rol que tiene una persona o entidad sobre una situación concreta. Van de Poel y Royakkers (2011, pp. 10-21) diferencia entre dos tipos de responsabilidad: la activa y la pasiva. La responsabilidad activa es la que se incurre antes de que ocurra algo y está relacionada con el deber de cuidar de las personas o con otras consecuencias. Mientras que la responsabilidad pasiva se trata después de que algo indeseable haya sucedido, estando relacionadas la responsabilidad y la culpabilidad, ya que *"nadie es responsable de algo si no está en su mano haberlo evitado"* (Valcárcel, 2002: p. 261) .

Cuando hablamos de implantar tecnologías, la responsabilidad activa principalmente trata de hacer las cosas bien. Es decir, se basan en ideas o esfuerzos que tienen como objetivo alcanzar un óptimo o un máximo. Estamos hablando de distintos aspectos, como el de disponer entusiasmo tecnológico, en el sentido de asumir retos tecnológicos y desarrollar nuevas posibilidades tecnológicas; de ser eficiente y efectivo; y el de buscar el bienestar humano.

Por otra parte, la responsabilidad pasiva está relacionada con la rendición de cuentas y con la culpabilidad. El rendir cuentas significa justificar las acciones de la persona o de la entidad hacia los demás. Mientras que la culpabilidad trata sobre las consecuencias de sus acciones. Para que alguien sea culpable, a parte de una conducta incorrecta, debe tener una contribución causal, ser previsibles las consecuencias y que se tenga libertad sobre sus acciones.

En general, una persona cuando tiene un rol tiene una relación con los demás que pueden incurrir en responsabilidades. Una persona puede tener varios roles, por ejemplo, su papel dentro de su familia, como gerente, como ingeniero, o como experto, donde cada rol tiene sus propias responsabilidades, pudiendo estar establecidas formal o informalmente. Al existir distintos roles, existen diferentes responsabilidades. Por otra parte, existe una responsabilidad moral que no está limitada al rol que se desempeña en una situación dada, sino que se basa en las obligaciones, normas y deberes que surgen de las consideraciones morales.

Por último, en un contexto de implantación y de desarrollo de tecnologías existen un conjunto de actores que en cierta medida disminuye la responsabilidad de cada actor por separado al reducir la contribución causal de las acciones en conjunto. En el caso de la responsabilidad de los ingenieros se complica aún más por el contexto social del desarrollo tecnológico. Así, aparte de los ingenieros, hay toda una serie de actores que participan en el desarrollo tecnológico: tenemos los directivos que marcan las pautas, los usuarios que formulan los requerimientos de funcionamiento, los usuarios que pueden hacer un mal uso de las implantaciones, las mismas organizaciones que formulan las reglas o normativas, o los grupos de interés que se esfuerzan por sacar ventajas sobre las implantaciones. Esto disminuye la responsabilidad de cada actor por separado, ya que se reduce su contribución causal a la tecnología y la previsibilidad de las consecuencias. Al mismo tiempo, introduce responsabilidades adicionales, ya que deben tener en cuenta a otras partes interesadas y sus intereses en el desarrollo de nuevas tecnologías.

4. ¿ESTAMOS ENVIANDO UN MENSAJE ADECUADO DE LA IA A LA SOCIEDAD?

En la actualidad se hace un uso intensivo de la palabra "inteligencia" dentro del ámbito tecnológico. No solo se utiliza en la misma denominación de la inteligencia artificial, sino también se usa para hablar de dispositivos inteligentes, teléfonos inteligentes, etc. Es indudable que su uso se ha extendido y mientras en el sector privado no deja de ser una estrategia de marketing, en el ámbito público se tiene que reflexionar si se está dando un mensaje adecuado a la sociedad, en el sentido de si estamos asociando adecuadamente inteligencia a las aplicaciones. En este sentido, es necesario reflexionar sobre el significado de qué se entiende por inteligencia, si realmente la IA es inteligente y, más concretamente, si la IA entiende lo que está haciendo.

4.1. ¿Qué es la inteligencia?

Una concepción de la inteligencia trata de su capacidad de pensar. Esta concepción de la inteligencia nos lleva a excluir a las máquinas o, en otras palabras, a

considerar directamente que la IA no es inteligente. Por otra parte, Gabriel (2019: p. 100) nos dice que "*en el caso de la IA no se trata de pensamiento sino de un modelo de pensamiento*", es decir, que la IA intenta asemejarse a lo que modela, a la inteligencia humana. En este sentido, existen otras definiciones de la inteligencia donde podremos encuadrar la inteligencia dentro de la IA. Rosa et al. (2016) hace su definición de inteligencia, aunando un conjunto de perspectivas, como "*una herramienta de resolución de problemas que busca soluciones en entornos dinámicos, complejos e inciertos*". En esta definición se permite un amplio abanico de escenarios que resuelven problemas, buscando soluciones de forma eficaz y eficiente, en otras palabras, que se resuelvan problemas en un tiempo determinado y utilizando los menos recursos posibles. La problemática de esta definición es sobre qué entendemos por resolución de problema. Pongamos un ejemplo. Si queremos encontrar el camino óptimo entre dos puntos, podríamos tener varias opciones: analizar visualmente todos los caminos, aplicar un algoritmo manualmente, programar y ejecutar el algoritmo o usar un programa informático existente. La mejor solución dependerá de factores como la complejidad del problema, la representación de los caminos, la disponibilidad de algoritmos y habilidades del agente que resuelve el problema. Realmente, la realidad no funciona así. Si quiero llegar de un punto a otro es por algo. El problema original de llegar de un punto a otro puede venir de múltiples causas, por ejemplo: se quiere llegar al aeropuerto desde donde estoy, se necesita comprar un medicamento, se quiere visitar la Catedral de X, etc. Así, hallar el camino óptimo es la consecuencia de formular el problema de que estoy aquí y quiero llegar allí de la mejor forma posible. Incluso el concepto de óptimo no tiene que ser el camino más corto, por ejemplo, podría ser el que tiene mejores vistas. Pongamos que el problema sea que se quiere visitar la Catedral: este problema puede venir de la formulación de otro problema original: "Como mañana no tengo nada previsto: ¿Qué puedo hacer para pasar el día?"

En definitiva, la misma resolución de problemas que busca soluciones en entornos dinámicos, complejos e inciertos al mismo tiempo realmente parte de una formulación de problemas. El agente que resuelve el problema debe tener una serie de habilidades y, en función de estas, la forma de resolver los problemas y dar soluciones serán diferentes. Las habilidades que posee el agente son de todo tipo, ya sean de conocimientos, heurísticas y disponer de trucos, que le permiten buscar la mejor formulación y solución del problema. Rosa et al. (2016) argumentan que entre las habilidades más útiles está la misma capacidad para adquirir habilidades, incluyendo la reutilización de las destrezas que dispone y la automejora. En definitiva, la capacidad de aprendizaje dentro de la inteligencia es una de las habilidades más importantes.

4.2. ¿Es inteligente la IA?

Esta pregunta se puede reformular como: ¿un programa puede resolver problemas de forma que busque soluciones en entornos dinámicos, complejos e

inciertos? Pongámonos en el ejemplo anterior. Está claro que el problema inicial de buscar el camino óptimo por parte de un programa, como el que nos provee Google Maps, puede ser resuelto si le damos el punto inicial y el punto destino, además, no solo nos da el camino más corto, sino también otras opciones. Este programa ha sido programado por un humano a través de la implementación de un algoritmo que permite resolver problemas de localizar los mejores trayectos en prácticamente todo el mundo. Por tanto, se puede decir que un programa puede resolver problemas en un tiempo limitado, mejor incluso que un humano.

Por otra parte, en el caso del ejemplo, el programa ha sido diseñado e implementado por un humano, que primero ha formulado el problema de forma general (¿cómo hallar el camino óptimo entre dos puntos en un grafo?), ha buscado una solución general para este problema (ha seleccionado un algoritmo entre los distintos que existen) y lo ha programado. En este sentido, la inteligencia del programa es limitada, y solo se restringe a resolver un problema muy acotado, sin tan siquiera formular el problema.

En verdad, en la actualidad la IA resuelve problemas concretos y, tal como afirma Gershman et al. (2015), tiene una racionalidad limitada. Los problemas los tienen las personas y son estas quienes formulan el problema, siendo la IA una herramienta tecnológica de la que disponen los humanos para, aprovechando las capacidades de cómputo y almacenamiento, resolver los problemas de manera más eficiente. Por tanto, para que un programa podamos calificarlo como inteligente y que se aproxime a la inteligencia humana, deberíamos pedirle, además, que sea capaz de formular problemas, y que sea capaz de crear algoritmos por sí mismo que puedan resolver el problema. Estamos hablando de al menos disponer de una IA fuerte, que en la actualidad los sistemas actuales no la poseen.

En resumen, utilizar el término inteligencia dentro de la IA débil se puede considerar problemático, ya que podemos transmitir un mensaje a la sociedad de que un programa dispone de capacidades inteligentes y humanas. Por ello, se considera que una administración pública debería transmitir mensajes más reales y hablar de términos más próximos a la realidad, por ejemplo, utilizando términos como algoritmos, en lugar utilizar términos que transmitan un mensaje erróneo, y reservar el término inteligente cuando realmente se disponga de una IA fuerte.

4.3. ¿Entienden las máquinas lo que hacen?

Que una máquina disponga de cierto entendimiento es otro aspecto problemático desde que existen sistemas que han sido entrenados para comportarse como humanos, como los asistentes virtuales con los que se puede establecer una conversación, siendo paradigmático el caso de ChatGPT. Incluso recientemente ha aparecido la noticia de que un ingeniero de Google ha afirmado que

el sistema LaMDA dispone de conciencia. Google desmintió dicha afirmación e incluso sancionó a dicho ingeniero. LaMDA es un modelo de lenguaje natural para aplicaciones de diálogo que fue diseñado en 2017 por Google en base a una arquitectura específica de redes neuronales (Thoppilan, 2022). Resumidamente, al igual que ChatGPT, este sistema se basa en entrenar una arquitectura específica de redes neuronales usando una gran base de datos de diálogos y de contenidos. El resultado es que si alguien pregunta algo a este sistema, su respuesta se aproxima a lo que tiene grabado, es decir, en base a diálogos reales y existentes previamente.

Así, si nos encontramos un programa con el que podríamos establecer una conversación completa, como ChatGPT, incluso pudiendo superar la prueba de Turing (French, 2000), la cual analiza la capacidad de un programa para disponer de un comportamiento inteligente similar a una persona, nos deberíamos de preguntar: ¿realmente este programa entiende lo que le preguntan y lo que contesta? Esta pregunta se aproxima a si realmente la máquina tiene consciencia. Searle (1984) contesta a esta pregunta afirmando que una máquina puede realizar acciones sin realmente entender lo que está haciendo. Searle recurre a un experimento mental para demostrar esto. Simplificadamente, este experimento se basa en una habitación donde está Juan que no entiende el chino, pero tiene una serie de diccionarios y manuales que le indican qué hacer, es decir, hace algo como esto: "si aparecen tales palabras, entonces, escribe estas palabras". De tal forma, que se puede traducir al chino pareciendo al hablante de chino que está hablando con alguien que entiende el chino. Searle afirma que ni Juan, ni los manuales, ni la habitación en conjunto entiende el chino.

En definitiva, alguien o un programa puede ejecutar mecánicamente un algoritmo para el chino sin realmente entender el chino: lo que hace el programa es que simula que lo entiende. En otras palabras, los programas no entienden ni saben lo que hacen: no tienen consciencia.

5. ¿CÓMO SABEMOS CUÁNDO ES MORAL IMPLANTAR LA IA?

Es indiscutible que la IA puede tener muchos beneficios y que podemos utilizarla para mejorar los servicios públicos. Sin embargo, aunque las intenciones sean buenas, pueden aparecer problemas morales que son consecuencias no deseadas de su aplicación. No todas las aplicaciones que usan la IA se enfrentan con los mismos problemas éticos. Pongamos el caso de un servicio jurídico que está organizado en tres secciones: uno para resolver los temas de personal, otro para los temas de medio ambiente y el último para el resto de los temas. Existe una aplicación que clasifica automáticamente los documentos que le llegan en cada uno de los tres temas utilizando técnicas de inteligencia artificial. En realidad, en este caso no se tiene que considerar problemas morales. Sin considerar que puedan existir datos personales en los documentos, el mayor problema que se puede encontrar es que se produzcan errores de clasificación, por ejemplo,

un 5% de los documentos se etiquetan erróneamente. En verdad, este problema no tiene repercusiones morales, en el sentido de que si un documento está mal etiquetado, la sección correspondiente lo devolverá, y se enviará a la sección correcta.

Ahora pongamos el caso de un servicio que se dedica a tramitar inspecciones tributarias, donde un clasificador inteligente recibe un expediente de una persona, crea un perfil y decide si se inicia o no una inspección o no. En este caso nos podemos encontrar que existen una serie de sesgos que hacen que ciertos colectivos tengan más probabilidades que se le inicie una inspección, por ejemplo, por residir en una zona donde se maneja más dinero negro. También, nos encontramos con problemas de privacidad, por ejemplo, se cruzan datos de otras bases de datos que facilitan la creación de un perfil. Por último, nos encontramos con el problema de la explicabilidad, por ejemplo, una persona puede preguntar a la administración cuáles son los criterios que se han utilizado para determinar a quién se le realiza una inspección.

En definitiva, al igual que las aplicaciones se categorizan de forma diferente en función de los datos personales que se tratan, las aplicaciones que usan técnicas de IA se pueden categorizar de forma distinta en función de los problemas morales con los que se pueden enfrentar, existiendo aplicaciones donde no se plantean problemas morales. Así, los problemas morales pueden ser de distinta índole y grado. En este sentido, así como se tiene que hacer una reflexión diferente para cada problema moral, la respuesta normativa no va a ser igual sobre lo que puede afectar a la privacidad, a la seguridad, la rendición de cuentas y transparencia, a la justicia de las decisiones, a cómo afecta a los trabajadores y otros efectos que pueden tener en la sociedad.

5.1. ¿Cómo puede afectar a la privacidad?

Las administraciones públicas disponen de una gran cantidad de datos y muchos de carácter personal. Por ello, existe una constante preocupación por la protección de datos personales. En general, cuando usamos las tecnologías existe una preocupación ética desde el momento que se recopilan los datos personales y, posteriormente, con todo el procesamiento que se hace de ellos, que incluye su posible acceso y compartición. En este sentido, la protección de datos personales intenta respetar la privacidad de las personas, dando el derecho a las personas que sepan que se están recopilando datos suyos, qué se hace con sus datos, y que se opongan a su recopilación y procesamiento.

Con la recopilación de grandes cantidades de datos, por ejemplo, a través del desarrollo de Internet de la Cosas, y con la entrada de aplicaciones que usan la IA aparecen nuevos desafíos en la protección de la información personal. Así, cuando se aplican técnicas de aprendizaje automático, previamente se parte de la recopilación de datos cuyo origen puede estar en un contexto y que luego se

utilizan en otro contexto diferente. En general, para garantizar la privacidad en estas aplicaciones se procede a anonimizar los datos, es decir, se eliminan o modifican los datos que identifican las personas. En cualquier caso, sus datos, aunque eliminando datos como el DNI, el nombre y apellidos, se mantienen y se pueden utilizar para otros fines y otros contextos. En estos casos, las administraciones están dando valor a los datos, pudiendo las personas no ser conscientes de que sus datos están siendo usados para otros fines. En este sentido, las personas tienen el derecho de ser informadas y que éstas den su consentimiento, siendo esto una de las obligaciones de la administración.

Sin embargo, entre los mayores peligros que tiene la IA es su gran capacidad de crear perfiles o inferir si un individuo determinado está o no presente en un conjunto de datos, que a la vez permite que se puedan reconstruir atributos sensibles haciendo uso de correlaciones. El fin de la creación de perfiles podría parecer ser en principio loable. Pensemos en el caso que a través del cruce datos se pueda determinar que ciertas personas son susceptibles de recibir algún tipo de ayuda. Estas personas pueden estar agradecidas que se les informen directa y proactivamente que tienen derecho a estas ayudas, pero, el gestor se plantea la problemática de en qué medida puede cruzar datos cuyo origen están en otros contextos sin que las personas sean conscientes que una IA está procesando sus datos.

Por otra parte, es más problemática esta creación de perfiles cuando hacen predicciones de comportamiento, por ejemplo, utilizando la IA para la vigilancia. Pongamos el caso de que se pueden predecir ciertos perfiles que son propensos al fraude o a delinquir. Incluso los perfiles pueden ser utilizados para fines políticos de manipulación. En estos casos, podemos convertir a la administración en un Estado que "*todo lo ve*". Tal como nos argumenta Coeckelbergh (2020a, p. 100) el peligro es que la IA nos conduzca a "*nuevas formas de manipulación, vigilancia y totalitarismo, no necesariamente en forma de política autoritaria, sino de una manera más oculta y altamente eficaz*".

Por último, la privacidad o el derecho a la intimidad es un derecho fundamental universal, esto es, es aplicable a todas las personas y, por tanto, su aplicación tiene su sustento en la ética de los principios, es decir en la deontología. Por otra parte, en una sociedad democrática se requiere la contribución de cada persona para conseguir el bien común, lo que puede llevar al sacrificio individual en beneficio del interés colectivo. Por ello, en algunos casos hay que plantear un equilibrio entre enfoques utilitaristas y deontológicos. Esto es, conocer cuándo el derecho a la intimidad es más importante que el interés colectivo. Por ejemplo, Correia et al. (2021), en el campo de la salud, trata el caso de la difusión de resultados sanitarios que por razón de interés público, aún en conflicto con el derecho a la intimidad, cuando se trata de enfermedades transmisibles, como el COVID-19. En cualquier caso, parece que en muchos casos es más razonable la selección de la deontología, ya que, tal como afirma Prabhumoye et al. (2020), "*proporciona normas éticas claras y se ajusta a la idea jurídica del Estado de Dere-*

cho en el sentido de que estas normas éticas obligan a todas las personas por igual, en lugar de cambiar las normas para conseguir un determinado resultado".

5.2. ¿Son los sistemas IA seguros?

La seguridad es una preocupación creciente sobre todo en los sistemas informáticos. A parte de que la IA se puede utilizar para realizar acciones de pirateo y hackeo más sofisticadas y peligrosas, lo que llevará a disponer de sistemas de protección también más sofisticados, la seguridad de las aplicaciones que usen la IA debe tener en cuenta los nuevos peligros. En la medida que las aplicaciones sean en mayor medida una caja negra que no conozcamos muy bien la explicabilidad de sus decisiones, son más susceptibles que sus sistemas sean atacados internamente sin que seamos conscientes. En este sentido, los ataques pueden ser tan sutiles como tan solo modificar ligeramente algunos datos específicos de entrenamiento. Como se ha visto, el entrenamiento se basa en disponer de datos del contexto. Así, una simple manipulación de estos datos puede resultar que el programa nos dé resultados diferentes. Por ejemplo, si en una aplicación para detectar fraude se manipulan los datos de entrada para que no se tengan en cuenta ciertos perfiles, estos perfiles pueden quedar fuera de la lucha contra el fraude.

En este caso de seguridad, cada una de las principales teorías éticas sugiere diferentes cursos de acción. Desde la perspectiva del consecuencialismo, las acciones que maximicen el bienestar y minimicen el daño son las más éticas, y en este caso, la implementación de sistemas de protección más sofisticados podría ser vista como una acción ética, ya que puede reducir el riesgo de ataques y minimizar el daño potencial. La deontología sugiere que hay ciertos principios éticos fundamentales que deben ser seguidos, como la protección de la privacidad y la libertad de las personas, y la manipulación de los datos de entrenamiento para obtener resultados fraudulentos es moralmente incorrecta. Por lo tanto, las acciones deben enfocarse en mitigar estos peligros y proteger estos principios éticos. Por último, la ética de la virtud sugiere que las personas involucradas en el desarrollo y uso de estos sistemas deben desarrollar ciertas virtudes, como la integridad y la responsabilidad. Los desarrolladores y usuarios de sistemas de IA deben ser responsables de su uso y asegurarse de que no se utilicen para dañar a otros. En resumen, cada teoría ética tiene su propia perspectiva y sugiere diferentes acciones para abordar los peligros y preocupaciones de seguridad en relación con los sistemas informáticos y la IA.

5.3. ¿Con la IA se pueden explicar las decisiones?

No siempre está claro cómo la AI toma sus decisiones. Cuando utilizamos un sistema basado en reglas, la toma de decisiones está explícitamente diseñada,

con lo que una vez que se ha tomado una decisión se puede seguir el proceso que nos ha conducido a la decisión final. También, cuando se utiliza una técnica estadística basada en un árbol de decisión se puede explicar la decisión tomada. El principal problema surge cuando se aplican técnicas que utilizan redes neuronales y técnicas estadísticas avanzadas. En estos casos, o no es tan evidente o a veces es prácticamente imposible explicar los resultados, al menos por una persona no experta.

No hay que confundir explicar una decisión concreta que explicar cómo funciona el programa. Cuando se diseña y desarrolla una aplicación, aparte de conocer el código, se conoce cómo funciona, es decir, se conocen los algoritmos y por qué se ha elegido, y, también, se han seleccionado los datos para el entrenamiento y cómo mostrar los resultados. Con el aprendizaje en la IA se obtienen unos patrones en los datos, estos de alguna manera se seleccionan para que sirvan de base para la decisión, pero el desarrollador o usuario de la aplicación no conoce los patrones que la IA ha seleccionado. Por tanto, no es capaz de explicar los resultados. Por ejemplo, al final de entrenar una red neuronal los pesos entre las neuronas se quedan fijos, cuando aplicamos nuevos datos de entrada, la decisión obtenida depende de estos pesos. De forma general, la relación de estos pesos aporta poca información para explicar la decisión, no existiendo relaciones causales que puedan dar algún tipo de información sobre el razonamiento que ha seguido la aplicación. En definitiva, tenemos una caja negra que nos da una decisión, pero no sabemos explicarla (London, 2019).

Los usuarios de estas "cajas negras" pueden no saber lo que están haciendo cuando utilizan la IA para decidir y actuar por ellos. Por tanto, tal como afirma Coeckelbergh (2020b) esta falta de transparencia y explicabilidad es moralmente problemática porque crea ignorancia por parte de las personas que utilizan la IA. Así, la falta de explicabilidad en muchos casos de la IA puede conducir a tener menos confianza, tanto para el responsable de la aplicación como para los destinatarios de las decisiones. De esta forma, un funcionario responsable no sabría explicar a un ciudadano el motivo por el que se le ha excluido de una ayuda y el ciudadano se sentiría indefenso ante la administración. Los ciudadanos tienen el derecho a la explicación, pero la IA tiene entre uno de sus problemas que no siempre es fácil dar explicaciones. Por tanto, uno de los peligros que tiene la IA es que puede limitar la transparencia y la rendición de cuentas.

En el caso de la falta de explicabilidad y transparencia de la IA en la toma de decisiones, cada teoría ética sugiere diferentes cursos de acción. Según el consecuencialismo, donde las acciones que maximizan el bienestar y minimizan el daño son las más éticas, si la falta de explicabilidad y transparencia de la IA reduce la confianza de los ciudadanos en la administración, esto podría tener consecuencias negativas para la sociedad en su conjunto. La deontología, por otro lado, sugiere que ciertos principios éticos, como la transparencia y la rendición de cuentas, deben ser seguidos al desarrollar y utilizar sistemas de IA en las administraciones públicas. En este sentido, las administraciones públicas

que utilizan sistemas de IA tienen una obligación ética de garantizar que estos sistemas sean transparentes y puedan explicar sus decisiones. La ética de la virtud, por último, sugiere que las personas involucradas en el desarrollo y uso de estos sistemas deben desarrollar ciertas virtudes, como la responsabilidad y la honestidad. Es importante que las personas implicadas en la implantación de estos sistemas sean responsables de su uso, asegurándose de que estos sistemas no se utilicen para dañar a otros y que se mantengan transparentes y explicables.

5.4. ¿Pueden tomar decisiones no justas?

Una administración pública tiene que preocuparse de ser justa. Rawls (1999/1971: pp. 30-36) entiende que existe la intuición de que en una sociedad un valor fundamental y prioritario es la justicia. Rawls considera a la justicia como equidad y que se tiene que corregir los sesgos de las distintas contingencias hacia la igualdad. La consideración de la justicia como equidad supone que se aleja del utilitarismo para evitar que las mayorías restrinjan o sacrifiquen la libertad de las personas. Por ello, la justicia en estos términos se acerca a las teorías deontológicas, debiendo ser la eliminación de los sesgos un principio de actuación de las administraciones públicas. En este sentido, cuando se utiliza la IA hay que tener en cuenta que no solo las personas tienen sesgos, sino que los programas también pueden tenerlos. Los sesgos de los programas pueden venir de los algoritmos que se utilizan, pero principalmente de los datos. Los sesgos pueden venir del mismo entorno, por la misma calidad de los datos, de no disponer de suficientes datos, de la exclusión de determinados colectivos o de recurrir a datos históricos.

El mismo entorno puede tener de forma implícita sesgos. Por ejemplo, aspectos como el sexo pueden venir sesgados del mismo lenguaje o costumbres, así Bolukbas et al. (2016) reportó que cuando se utilizan corpus de los textos disponibles en internet aparecen sesgos de estereotipos de sexo, obteniendo resultados como "*hombre es a un programador informático, como mujer es ama de casa*".

Una fuente de la mala calidad de la fuente de los datos puede provenir de las mismas personas que los recogen al transmitir sus sesgos humanos a los datos. Las personas pueden tener prejuicios, que pueden no ser conscientes de tener estos sesgos, al ser estos aprendidos, como son los sesgos de confirmación (Peters, 2020), donde las personas tienden a interpretar la información en función de sus creencias.

La insuficiencia de datos puede conducir también a sesgos con independencia de que los datos sean de buena calidad. Un conocido problema en estadística es el sobreajuste (Schaffer, 1993). Este sobreajuste puede venir de que no se consideran lo suficiente ciertos colectivos. Así, si tenemos una población pequeña, con distintos colectivos, si un colectivo está poco representado, nos pueden dar predicciones erróneas. Por ejemplo, si se quiere predecir el cáncer en una

población, donde se recoge 1000 muestras de personas, y en estas muestras solo existen dos personas de un colectivo en particular y estas no han desarrollado cáncer, una predicción sobre este colectivo es que no tendrá cáncer.

La exclusión de determinados colectivos en cierta forma está relacionada con el problema anterior. Pensemos en el caso que se entrena un sistema de IA con datos con el fin de predecir o localizar colectivos que necesiten de algún tipo de ayuda pública, por ejemplo, para dar ayudas al estudio a estudiantes. Por ejemplo, si los datos para el entrenamiento se recogen de una zona donde no existen personas en riesgo de pobreza y luego se utiliza esta aplicación para localizar estudiantes en otras zonas, las características de las personas perteneciente a estos colectivos desfavorecidos no se tendrán en cuenta.

Por último, recurrir a datos históricos también puede dar lugar a sesgos. Fuchs (2018) pone como ejemplo el sistema COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), el cual es un sistema de ayuda a la decisión para los jueces en EE.UU., que ayuda a los jueces para determinar si un acusado mientras espera un juicio queda libre o debe permanecer en la cárcel. Este sistema fue entrenado con datos históricos, incluyendo entre estos los datos de sistemas policiales, que con frecuencia muestra un sesgo hacia la selección de barrios de bajos ingresos y lugares con mayor concentración de minorías. De esta forma, las decisiones están sesgadas hacia determinados colectivos.

5.5. ¿Cómo afecta a los trabajadores?

Entre uno de los principales beneficios de la IA es que permite avanzar en la automatización de la administración pública. Recientemente, se ha visto los beneficios de la automatización en la pandemia reciente. Por ejemplo, Doğuç (2021) nos indica las potencialidades de los sistemas RPA (*Robot Process Automation*) para dar respuestas rápidas automatizando procesos en la administración. En los sistemas RPA se trasladan las tareas repetitivas que hacen las personas a lo que se denomina un robot, que es en realidad un programa instalado en el ordenador. Cuando existen acciones que tienen la forma de toma de decisiones, entonces se puede plantear la incorporación de un algoritmo, basado en técnicas de IA, que tome la decisión. Así, las administraciones pueden mejorar la eficacia y, al mismo tiempo, aliviar a las personas de las tareas repetitivas.

Coeckelbergh (2020a, pp. 136-137) nos indica que esta automatización junto con la IA nos plantea problemas no sólo sobre el futuro y el significado del trabajo, sino también sobre el futuro y el significado de la vida humana para la sociedad en general, que podemos trasladar a las mismas administraciones públicas. A nivel de la sociedad existe la preocupación que la IA pueda suponer una importante destrucción de empleo, pero también está la cuestión sobre qué tipo de empleos serán los más afectados, y qué personas serán las más perjudicadas y las más beneficiadas. En cualquier caso, existen diferentes valoraciones, tanto

positivas como negativas, de las consecuencias de la incorporación de la IA en la sociedad (Muñoz Vela, 2022: pp. 35-48).

Una administración pública cuando, a consecuencia de la automatización, reduce la carga de sus empleados, puede destinar estos empleados a otro tipo de trabajo o reubicar los trabajadores en otros servicios. En la medida que los empleados públicos tienen una mayor estabilidad en el empleo, las consecuencias sobre el empleo, al menos en el corto plazo, no tiene que verse en gran medida afectado. En este sentido, la gestión de la velocidad de la incorporación de la IA es un aspecto que tiene que ser tratado.

Otro aspecto es lo relacionado con el significado del trabajo. Tal como argumenta Coeckelbergh (2020a, pp. 141-142), no siempre existe una tarea que deba evitarse desde el punto de vista de un trabajador. El trabajo en sí mismo puede tener valor para el trabajador que le da "*un propósito y un significado, y que tiene varios beneficios como las conexiones sociales con otros, la pertenencia a algo más grande, la salud y las oportunidades de ejercer la responsabilidad*". En estos casos, algunas tareas se tienen que plantear que sean reservadas para las personas, no siendo necesario que la IA abarque todas las tareas sino aquellas que sean menos significativas. También, en lugar de delegar las tareas a la IA, se puede elegir colaborar con ella. Por ejemplo, la IA se puede utilizar para desarrollar imágenes creativas (Anantrasirichai y Bull, 2021), en este sentido se le puede dejar sola o que colabore con personas.

En cuanto a la automatización y la IA en la administración pública, cada teoría ética sugiere diferentes cursos de acción. Según el consecuencialismo, las acciones que maximizan la eficacia y la productividad son las más éticas. Por ejemplo, si la automatización y la IA mejoran la eficiencia en la administración pública, esto puede tener beneficios positivos para la sociedad, como una menor burocracia. La deontología, por otro lado, sugiere que ciertos principios éticos, como la justicia y la equidad, deben ser seguidos al implementar la automatización y la IA en la administración pública, y que se deben considerar los impactos sobre los empleados y las personas más vulnerables de la sociedad. La ética de la virtud, por último, sugiere que las personas involucradas en el desarrollo y uso de la IA deben desarrollar ciertas virtudes, como la responsabilidad y la consideración por el bienestar de los empleados y las personas más vulnerables de la sociedad. Los desarrolladores y usuarios de sistemas de IA deben ser responsables de su uso, asegurándose de que no dañen a otros y que se mantengan justos y equitativos. Además, es importante preservar ciertas tareas para las personas para que los trabajadores encuentren significado y propósito en su trabajo.

5.6. ¿Qué otros efectos pueden tener en la sociedad?

Cuando la administración pública incorpora la IA no solo le tiene que afectar internamente, sino que puede tener repercusiones en la sociedad. Se puede

esperar que tenga efectos positivos si se gestionan bien los aspectos que se han tratado anteriormente, es decir, los que afectan a los sesgos, privacidad y transparencia. En cualquier caso, puede afectar negativamente en materias como el empleo. En el caso de la incorporación de la automatización en aquellas tareas se han externalizado, podría tener repercusiones en el empleo en la sociedad. Pensemos en los servicios de atención de llamadas masivas, donde muchos de estos servicios están externalizados, por ejemplo, la incorporación de asistentes virtuales. Los asistentes virtuales pueden plantear problemas éticos en cuanto a la calidad de la atención y el trato interpersonal que reciben los ciudadanos, como menciona Bies (2001) en relación con la justicia interaccional. Desde la perspectiva del consecuencialismo, se podría argumentar que la implementación de sistemas de atención automatizados puede ser ética si mejora la eficacia y la eficiencia de la administración pública. En cuanto a la deontología, se podría argumentar que los principios éticos, como la justicia y el respeto a los derechos de los ciudadanos, deben ser considerados en la implementación de estos sistemas, teniendo en cuenta los valores fundamentales como el respeto y la dignidad. Finalmente, desde la perspectiva de la ética de la virtud, se podría argumentar que las personas involucradas en el desarrollo y uso de estos sistemas deben desarrollar virtudes como la empatía y la responsabilidad social para garantizar una atención respetuosa e inclusiva. Por otra parte, el uso de asistentes puede afectar el empleo que genera externamente el sector público. Por tanto, hay que añadir otras cuestiones morales anteriormente tratadas, que afecten a la sociedad, como puede ser el empleo, incluso con el medio ambiente.

6. ¿PODEMOS PROGRAMAR LA ÉTICA?

Se ha visto que cuando se implantan sistemas de IA nos podemos encontrar con una serie de peligros, principalmente sobre privacidad, sesgos y rendición de cuentas. Además, en estos mismos sistemas están incorporadas decisiones y acciones que hay que programar, donde muchas de estas decisiones podrían tener consecuencias morales. Por tanto, cuando hablamos de programar la ética tenemos que pensar, por una parte, en cómo tratar los peligros de la IA y, por otra parte, en implementar algoritmos que tomen decisiones morales.

6.1. ¿Cómo tratar los peligros?

Vimos que entre los peligros de implantar la IA teníamos una disminución de la privacidad, la aparición de sesgos y una reducción de la transparencia. Tal como argumenta Júdez y Gracia (2001) dentro del campo de la salud, los problemas éticos consisten en conflictos de valor, siendo la parte más compleja la deliberación sobre qué curso de acción es mejor tomar desde un punto de vista moral. Hay que tener en cuenta que elegir entre dos valores supone una pérdida irreparable de un valor sobre el otro. Por ejemplo, no aplicar la IA para evitar

los riesgos de privacidad, podría suponer renunciar al principio de eficiencia. Por ello, la deliberación sobre el mejor curso de acción a tomar debe ir en el sentido de poder salvaguardar en la medida de lo posible todos los valores en conflicto. En este sentido, una estrategia es la mitigación de los peligros. Con el fin de reducir estos efectos se pueden aplicar medidas técnicas, las cuales se basan en desarrollar programas o algoritmos específicos dentro de las mismas aplicaciones de IA.

En relación con la privacidad, la medida más conocida es la aplicación de lo que Dwork et al. (2006) denominó privacidad diferencial, que trata de caracterizar la cantidad de información de los datos de una persona determinada que es relevada a través de algún tipo de cálculo, en otras palabras, la privacidad diferencial no es más que garantizar que un atacante no pueda inferir si una persona en particular estaba o no presente dentro de un conjunto de datos o que se pueda reconstruir sus atributos sensibles. La aplicación de la privacidad diferencial se aplica cada vez más en aquellos servicios que tratan con gran cantidad de datos sobre los que se puede inferir datos personales, como por ejemplo en aquellas aplicaciones de Internet de la Cosas (Jiang et al, 2021). Uno de los mecanismos que se utilizan es la incorporación de ruido o redundancia con el fin de conseguir más privacidad (Tran et al., 2021), aunque una de sus consecuencias es que esto puede afectar en mayor medida a algunos colectivos sobre otros, en otras palabras, aumenta el sesgo. En cualquier caso, aunque en ciertos casos el problema de la privacidad no está del todo resuelto, se puede considerar que existe un importante conocimiento sobre esta problemática y, en determinados casos, cómo tratarla (Dwork y Roth, 2014).

Con respecto a los sesgos, para conseguir una equidad algorítmica, se ha trabajado en el diagnóstico y la mitigación del sesgo. Para ello, es necesario una definición del sesgo que se pueda medir en base a su cuantificación, lo cual en sí mismo es problemático ya que depende de lo que se considere equidad. Whittaker et al. (2018) hacen un repaso de algunos de los enfoques que tratan de entender y definir la problemática relacionada con la equidad y el sesgo algorítmico. Un enfoque se basa en fijarse en los daños que producen y las oportunidades sobre los distintos colectivos, sin entrar en su reparación. Un segundo enfoque intenta el diagnóstico y la mitigación del riesgo a partir de los datos de entrenamiento o que se utilizan como entrada, ya detectando sesgos como los relacionado con el género, la raza y situación socioeconómica. Un tercer enfoque se basa en lo que se denominan anticlasificación, que trata de omitir los atributos protegidos, por ejemplo, no solo se eliminaría la raza sino otros atributos que pueden inferir la raza. Otro enfoque se basa en la paridad de la clasificación en el sentido de buscar el mismo rendimiento predictivo para las distintas clasificaciones. Por último, están los que se basan en utilizar estrategias de calibración, los cuales se fijan más en los resultados de las decisiones o predicciones en lugar de los datos, de forma que los resultados no dependan de los atributos protegidos.

Tal como Whittaker et al. (2018) exponen, estos enfoques tienen limitaciones, ya que las soluciones algorítmicas requieren de una noción matemática de equidad que es difícil de conseguir. Por ejemplo, tenemos que la estrategia de la paridad de la clasificación cuando se igualan colectivos entre ellos da lugar a la disminución de la precisión para ciertos colectivos. También, en el enfoque de la anticlasificación cuando se eliminan atributos protegidos, esto puede dar lugar a aumentar los sesgos al no considerar las particularidades de los colectivos afectados. Por otra parte, los distintos enfoques pueden excluirse entre ellos, ya que cada enfoque hace más hincapié en lo que se considera más justo. Además, está la misma problemática de definir cuantitativamente el concepto de equidad y la imposibilidad de conseguir que se cumplan al mismo tiempo las diferentes definiciones de equidad (Chouldechova, 2017).

Por último, con el fin de conseguir mayor transparencia en la IA y que las personas tengan un derecho a la explicación, existe un esfuerzo investigador dentro de la IA en disponer de lo que se denomina la IA explicable. Por otra parte, tal cómo se plantea Coeckelbergh (2020a, p. 121), existe la problemática de saber qué puede considerarse una buena explicación y para quién. En cualquier caso, surgen dudas si realmente es siempre posible disponer de una IA que sea transparente (Vilone y Longo, 2020). Hay que tener claro que la transparencia no se refiere a que se pueda analizar el programa, porque, cómo se ha dicho, en muchos sistemas, conociendo el programa no es suficiente para explicar sus decisiones o predicciones.

En función de la técnica de IA que se utilice será más o menos difícil conseguir una buena explicación. Mientras que con los sistemas basados en reglas o árboles de decisión es más fácil de conseguir explicar cada paso en un proceso de decisión, incluido explicar las decisiones que afectan a colectivos concretos, no es así en muchos de los sistemas de aprendizaje automático y profundo (Guidotti et al., 2018). La problemática viene de lo que denomina Coeckelbergh (2020a, pp. 119-120) *"abrir la caja negra"* que, además, puede servir a la ética, ya que permite mejorar los sistemas de IA y aprender de ellos. Así, en campos como la medicina existe una alta preocupación en saber cómo se predicen los resultados de los sistemas basados en redes neuronales. Por ejemplo, un médico no solo se conforma con que un sistema de IA pueda hacer buenas predicciones de padecer glaucoma, sino quiere saber en qué se fija la red neural para hacer sus predicciones. En este campo existen grandes avances, donde fijándose en las capas intermedias de las redes neuronales se puede ver qué atributos son amplificados, por ejemplo, nos podrían indicar qué atributos serían los responsables de padecer glaucoma en el futuro. Por tanto, en ciertos modelos de sistemas de IA complejos podría alcanzarse la explicabilidad, existiendo un importante conocimiento sobre cómo tratarlo (Molnar, 2022). En cualquier caso, en muchos modelos todavía no es fácil conseguir niveles de explicabilidad satisfactorios, que no son suficientes para alcanzar los niveles mínimos de transparencia que exigen las personas a una administración pública.

En algunos casos, se presenta una problemática donde se puede lograr cierto grado de explicabilidad con ciertos modelos, aunque sea limitado, mientras que otros modelos proporcionan predicciones mucho más precisas pero sin posibilidad de explicación en ese momento. En estas situaciones, surge un dilema entre sacrificar el rendimiento, que es un principio de eficacia, en aras de la transparencia, es decir, la ética utilitarista y la ética deontológica entran en conflicto. Otra problemática viene de que la misma transparencia puede ir en contra de los fines que persigue la utilización de la IA. Por ejemplo, un sistema que detecte el fraude tributario, al explicar cómo el sistema lo detecta nos está diciendo cómo se fija en determinados atributos, en este sentido si los colectivos que defraudan conocen estos atributos pueden procurar eliminarlos o enmascararlos.

6.2. ¿Cómo programamos las decisiones morales?

La programación de las decisiones morales supone la traslación de un modelo de toma de decisiones a un algoritmo que tenga en cuenta consideraciones morales. En este sentido, la primera pregunta que nos podemos hacer es qué tipo de modelos tomamos como referencia. Así, podemos tomar como referencia cómo las personas toman sus decisiones o un modelo matemático que busque la solución óptima.

El problema de tomar como modelo de referencia a las personas viene de la racionalidad limitada de la toma de decisiones de las personas. Aparte que existen autores, como Gabriel (2017), que rechazan que la mente humana pueda reducirse a una cuestión cibernética de las decisiones y afirma que "*no existen algoritmos morales, reglas o sistemas de reglas tales que permitan resolver para siempre los problemas morales*" (Gabriel, 2021: p. 37), otros autores analizan esta posibilidad, aunque con dificultades. En este sentido, Simon (1957) propone su modelo de racionalidad limitada a partir de su afirmación que las personas no son completamente racionales en muchas situaciones, en la medida que actúan mediante impulsos emocionales que no son del todo racionales y que aplicar la racionalidad por parte de las personas tiene sus limitaciones al no disponer de toda la información disponible, el de disponer de un tiempo limitado y la misma limitación que tiene nuestra mente en asimilar y generar conocimiento. Gigerenzer (2010) sostiene que el comportamiento moral se fundamenta en heurísticos sociales pragmáticos, en lugar de reglas morales o principios de maximización. Estas heurísticas no son inherentemente buenas o malas, sino que su valor depende del contexto en el que se aplican. De esta manera, una persona podría saltarse las reglas o normas establecidas si considera que se trata de una mejor elección moral. Por tanto, tomar como modelo de referencia a las personas puede humanizar la toma de decisiones, aunque se encuentra con la gran limitación de crear algoritmos que lo modelen.

Por ello, el diseño de algoritmos éticos se basa principalmente en modelos matemáticos o lógicos de racionalidad. Este tipo de algoritmos pueden tomar

decisiones morales como, por ejemplo, decidir si se inicia una inspección tributaria. El modelado de estos algoritmos se puede basar en la aplicación de reglas lógicas como, por ejemplo, "si ocurre *algo*, entonces hago *esto*". La aplicación de reglas lógicas permite la incorporación de las concepciones kantianas (Ulgen, 2017), incluso reforzadas con modelos estadísticos. Los modelos matemáticos de racionalidad que busquen la solución óptima pueden basarse en las concepciones del consecuencialismo como, por ejemplo, utilizando la función de utilidad (Aliman y Kester, 2019). Por otra parte, la utilización de las concepciones de la ética de la virtud puede ser problemática desde que supone la incorporación de la moralidad dentro de agentes artificiales (Gamez et al., 2020).

Una problemática significativa es cuando se enfrenta el programador con posibles situaciones en las que aparecen dilemas morales (Yu, 2018). Aunque existen académicos que afirman que los dilemas morales no existen (Gabriel, 2021: p. 127) (porque realmente se trataría de situaciones trágicas, donde optar por una opción se pierde un valor insustituible, por ejemplo, cuando se tiene que elegir entre dos vidas), lo que es verdad es que estas situaciones ocurren y se tienen que abordar. Júdez y Gracia (2001) afirman que en estos casos la forma más adecuada es a través de la deliberación y aplicando como virtud la prudencia aristotélica para el razonamiento práctico. El problema es cómo programar estas decisiones o deliberaciones, por ejemplo, en la conducción autónoma se pueden dar este tipo de situaciones cuando se tiene que elegir entre dos situaciones en que ambas pueden implicar la pérdida de vidas humanas.

Con independencia del problema de modelar decisiones racionales matemáticamente y de los problemas de los sesgos y la dificultad de su interpretabilidad, estos algoritmos cometen errores (Martin, 2019). Estos errores se pueden clasificar como errores de clasificación o errores de proceso. Dentro de los errores de clasificación tenemos los denominados errores de tipo I, o falsos positivos, por ejemplo, cuando se etiqueta a alguien como mujer cuando es realmente un hombre; y los de tipo II, o los falsos negativos, por ejemplo, cuando se etiqueta a alguien como no mujer cuando realmente es mujer. En general, cuando se entrenan o se elige un algoritmo se busca el mejor resultado o de los errores tipo I o los de tipo II o se llega a un compromiso entre ambos errores. Cuando se aplica la clasificación sobre colectivos, estos errores pueden conducir a sesgos. Por otra parte, los errores de proceso pueden suceder cuando se aplican factores que no deberían ser determinantes para la toma de decisiones, por ejemplo, cuando se utilizan redes neuronales, en el entrenamiento se podrían aprender factores que realmente no deberían considerarse porque son intrascendentes, como el nombre de los padres, o por no ser éticamente elegibles, como la raza.

En definitiva, los algoritmos éticos suponen la modelización de aspectos éticos en los algoritmos, enfrentándose a la problemática que supone la dificultad de su matematización. Además, estos algoritmos son susceptibles de errores, no siendo ético ignorarlos o fomentarlos, por tanto, se requiere una gestión de errores y de su responsabilidad.

7. ¿CÓMO SE ATRIBUYE Y DISTRIBUYE LA RESPONSABILIDAD MORAL EN LA IA?

Cuando hablamos de responsabilidades sobre un sistema tecnológico, como es el caso de un sistema de IA, aparecen los problemas que Coeckelbergh (2020a, pp. 113-114) denomina de muchas cosas y de muchas manos.

Por un lado, el problema de muchas cosas viene de la complejidad de las partes que conforman un sistema de IA. Como vimos más arriba la IA tiene distintas partes, así, por ejemplo, algo puede salir mal porque los datos para el entrenamiento se han recopilado mal, porque se ha entrenado mal el sistema, porque los datos del entrenamiento no son los adecuados, porque los algoritmos no han funcionado correctamente, porque los resultados no se han comprendido bien o porque todo el sistema en sí mismo no está bien integrado.

Por otro lado, el problema de las muchas manos viene de la dificultad de rastrear a todas las personas involucradas en la historia de un sistema de IA. Cada una de las partes de un sistema de IA puede tener historias diferentes. Por ejemplo, un algoritmo puede ser el proceso evolutivo donde han participado diferentes personas, donde puede darse el caso que dicho algoritmo se diseñó para un fin concreto y termine aplicándose a otro contexto diferente. Además, en el diseño y desarrollo de los sistemas de IA no solo intervienen los ingenieros, existen distintas personas con responsabilidades en cada parte del sistema, donde los usuarios forman parte también de este proceso de diseño. Estos usuarios son realmente los responsables de marcar las pautas de diseño de una IA. Por ejemplo, son los que dicen de dónde partir para los datos de entrenamiento o qué porcentajes de errores son los adecuados. También, los gerentes, en la medida que deciden si implantar un sistema de IA sobre una actividad que puede ser crítica, son los responsables de posibles consecuencias malas. Por último, el sistema de IA en sí mismo tiene más capacidad de acción y puede asumir muchas de las tareas y decisiones de los humanos, así, si este se comporta mal moralmente, ¿cómo atribuimos entonces la responsabilidad moral?

En definitiva, para comprender lo que puede salir mal o ha salido mal y para evitar que la responsabilidad, tanto activa como pasiva, se diluyan, por una parte, hay que tener un conocimiento de las distintas partes y determinar quién es el responsable de cada una de las partes y, por otra parte, tener un conocimiento cuál ha sido el rol de los gerentes, usuarios e ingenieros en las acciones necesarias en la decisión de la implantación, del diseño, del desarrollo y del uso del sistema de IA. También, hay que considerar el mismo sistema de IA cuando se comporta autónomamente si realmente puede una máquina ser responsable y por qué es diferente el problema de la responsabilidad en la IA. Por último, para concretar el problema de la responsabilidad debemos entender qué tipo de responsabilidad existen y cómo abordar estas lagunas de responsabilidad.

7.1. ¿Pueden ser responsables las máquinas?

Harris y Anthis (2021) realizaron un estudio entre los debates existentes de académicos sobre si las entidades con IA merecen algún tipo de consideración moral, donde parece que existe un acuerdo generalizado de que algunas entidades artificiales podrían merecer consideración moral en el futuro cuando se disponga de una IA fuerte, aunque no de forma clara en la actualidad. En cualquier caso, se trata de un debate abierto, donde se proponen que se amplíen marcos éticos convencionales consecuencialistas, deontológicos y de ética de la virtud, a enfoques de "ética de la información" y "social-relacional" para abordar mejor este tema.

Aunque se han plantteado estas entidades con una IA fuerte puedan atribuirse alguna consiseración como persona jurídica o como sujeto de derecho en el futuro (Tamayo Haya, 2020: pp 175-228) en la actualidad lo que parece claro es que, aunque la IA puede tomar decisiones y realizar acciones que tiene algún tipo de consecuencia ética, realmente la máquina no dispone de consciencia, de emociones y de la capacidad de formar intenciones, por tanto, no es consciente de lo que hace y tampoco es capaz de pensar moralmente.

En este sentido, Coeckelbergh (2020a, p. 111) afirma que una máquina no puede ser considerada moralmente responsable. Entonces, al igual que en nuestro sistema legal, los niños o los animales no pueden ser responsables legales, sino que la responsabilidad se traslada a sus tutores, parece que la responsabilidad legal debería ser trasladada a los responsables de las máquinas, ya sean humanos o una organización. Esta perspectiva está respaldada por el informe "Informe sobre responsabilidad derivada de la inteligencia artificial y otras tecnologías digitales emergentes" publicado el *Expert Group on Liability and New Technologies* (2019). Es importante recalcar que dicho grupo rechazó otorgar personalidad jurídica a los robots, como se evidencia en este informe. Los expertos consideraron que no era necesario ni conveniente otorgarles una personalidad jurídica propia, ya que sus acciones tendrían inevitablemente consecuencias para una persona física o jurídica preexistente.

7.2. ¿Por qué es diferente el problema de responsabilidad en la IA?

Por regla general, la responsabilidad moral o legal de las consecuencias del mal funcionamiento de una máquina son atribuidas a las personas o a una organización. Esto es así, básicamente, desde que tradicionalmente una persona tiene el control de la máquina y, además, es capaz de responder y dar explicaciones. Cuando utilizamos aplicaciones que usan técnicas de la IA la situación cambia porque el responsable de la aplicación puede no ser capaz de predecir las decisiones de una máquina, ya que una máquina podría tomar decisiones por sí misma. Por ejemplo, una vez que ChatGPT ha sido puesto a disposición del público, sus diseñadores y desarrolladores ya no pueden predecir las respuestas que ChatGPT muestra a las preguntas de sus usuarios.

Además, en muchos casos, las acciones de la aplicación son difíciles de explicar. Por tanto, cada vez más existirá un conjunto de acciones de las máquinas que en la que la forma tradicional de atribuir responsabilidades no es compatible con el concepto de justicia y moral que ha venido imperando en nuestra sociedad, desde que las personas pueden no tener el control mínimo que se requiere para que una persona pueda ser responsable de las acciones de las máquinas. Cuando esto ocurre, Matthias (2004) lo denomina brechas de responsabilidad.

7.3. ¿Qué tipos de responsabilidades se ven amenazadas con la IA?

Santoni de Sio y Mecacci (2021) identificaron cuatro tipos diferentes de brechas de responsabilidad: la brecha de culpabilidad, la brecha de rendición de cuentas moral, la brecha de rendición de cuentas pública y la brecha de responsabilidad activa. Los tres primeros tipos se encuadran dentro de la responsabilidad pasiva.

El concepto de brecha de culpabilidad es importante para abordar el problema de la atribución de responsabilidad cuando algo sale mal y se incurren en consecuencias moralmente malas. Es importante dado que existe el riesgo de que aparezcan lagunas de responsabilidad y nadie pueda ser considerado culpable de los resultados no deseados de las acciones realizadas por los sistemas de IA y las partes perjudicadas no sean compensadas o no reciban justicia.

Por otra parte, distinguen dos formas de la brecha de la rendición de cuentas. En el caso de la "brecha de la rendición de cuentas pública", la amenaza es que los ciudadanos no puedan obtener una explicación de las decisiones tomadas por los organismos públicos. En cambio, la "brecha de la rendición de cuentas moral" es más amplia, y considera el problema que tienen los humanos al verse reducida su capacidad de explicar el comportamiento de los algoritmos opacos, lo que también incluye las dificultades que se encuentran las personas implicadas en el proceso de desarrollo tecnológico, como los ingenieros, en conocer y comprender el objetivo y el significado de este proceso.

En lo que conciernte a la "brecha de responsabilidad activa", esta aborda el riesgo asociado a la posibilidad de que quienes diseñan, emplean e interactúan con la inteligencia artificial no posean suficiente conciencia, habilidades y motivación para reconocer y cumplir con sus obligaciones morales en relación al comportamiento de los sistemas que crean, supervisan o utilizan. Estas personas tienen la responsabilidad de que la implantación de la IA no suponga solo beneficios para la sociedad, sino que no afecten negativamente a los derechos e intereses de otras personas.

Por último, es diferente el caso de cuando el resultado de las decisiones proviene de los datos de entrenamiento. El tratamiento de estos datos puede "*tener consecuencias negativas o provocar prejuicios a personas o patrimonios*" (Váz-

quez de Castro, 2020: 270-273). En estos casos, será necesario determinar si ha existido una mala praxis en la elección o en el tratamiento de los datos.

7.4. ¿Cómo abordar las brechas de responsabilidad?

Las brechas de seguridad se han abordado desde distintos enfoques, a los que Santoni de Sio y Mecacci (2021) los han denominado fatalistas, deflacionistas y solucionistas. Los fatalistas, los presentan como un problema nuevo e insoluble, y se centran en una comprensión limitada de la culpabilidad de las máquinas. Los deflacionistas lo consideran un falso problema, subestiman la novedad de la IA y su implicación en las atribuciones de culpabilidad, los riesgos de las lagunas en la responsabilidad moral y pública de los diseñadores de sistemas, y la responsabilidad activa. Y los solucionistas afirman que se pueden resolver a partir de nuevas herramientas técnicas o jurídicas, por ejemplo, incorporando la IA explicable y otras mejoras tecnológicas y revisando los regímenes de responsabilidad jurídica, pero sin entrar a describir cómo deberían cambiar las prácticas morales y sociales y las normas jurídicas. De acuerdo con los mencionados autores, estos enfoques no logran abordar adecuadamente la complejidad de los problemas relacionados con las brechas de responsabilidad. Por ello, esbozan lo que sería un enfoque más integrado y global que aborde las brechas de responsabilidad para abordar las lagunas de responsabilidad con la IA con mayor amplitud, para lo que sugieren en basarse en el enfoque denominado "control humano significativo" (CHS) (Santoni de Sio y Van den Hoven, 2018) que en su esencia trata de que los humanos tengan un control significativo sobre los sistemas de IA.

CHS parte del marco teórico de Fischer y Ravizza (1998) que está basado en la teoría filosófica de la responsabilidad y el control moral y tiene una visión orientada al diseño sensible al valor. Su objetivo es minimizar las posibles brechas de responsabilidad, para lo que actúa tanto sobre el diseño organizativo y legal como sobre el diseño técnico. Para lograr esto se debe cumplir con dos condiciones que denominan seguimiento y rastreo que describen cómo deben ser las características que debe reunir un sistema hombre-máquina y la relación de control y para mantener la responsabilidad humana sobre el sistema.

El seguimiento requiere que la combinación de elementos técnicos, humanos y organizativos esté diseñada de manera que sea posible identificar al menos un agente humano a lo largo de la cadena de diseño, desarrollo y uso. Esta persona debe tener un conocimiento de las capacidades y limitaciones del sistema de IA, además de una conciencia moral de su papel para dar una respuesta por el comportamiento de este sistema. Por otra parte, el rastreo requiere que exista una alineación de las capacidades de las personas, cada una en su rol, con el sistema de IA. De esta forma, la IA estará bajo control humano desde que exista una conexión causal fiable entre los comportamientos de las personas y las máquinas.

8. ¿CON QUÉ RETOS SE ENFRENTAN LAS ADMINISTRACIONES PÚBLICAS?

Si bien es cierto que la implantación de la IA en las administraciones públicas no está libre de problemas éticos y de peligros no es motivo para que como en cualquier proceso de transformación y modernización la administración pública no puede quedar al margen de aprovechar sus beneficios y oportunidades. Por ello, no hay motivo para que las administraciones no encaren estos problemas éticos. En este sentido, la Unión Europea (UE) ha tomado especial interés en la IA, y en ser un referente mundial, sobre todo en disponer de una política para la IA con un enfoque ético, que esté centrada en las personas y que sea segura, ética, segura y de acuerdo con los valores fundamentales (Ulnicane, 2022). En 2018 se creó un nuevo *Grupo de Expertos de Alto Nivel en Inteligencia Artificial* (siendo las siglas, en inglés, HLEG-AI) con el fin de proponer un conjunto de directrices y principios sobre IA. Este grupo publicó el documento "*Directrices éticas para una IA fiable*" (HLEG-AI, 2019). La UE enmarca su objetivo con el término de IA fiable, que a la vez se refiere a tres componentes: la IA debe ser lícita, en el sentido de que debe cumplir con el marco legal; debe ser ética, con el fin de garantizar los valores y principios éticos; y debe ser robusta, tanto a nivel social como técnica, sin producir daños.

Básicamente, su propuesta se basa en los principios éticos de hacer el bien y no producir daños, de ser justo, salvaguardar la capacidad de la acción humana y de actuar con transparencia. Desde la perspectiva del consecuencialismo, se puede argumentar que el principio de hacer el bien y no producir daños es fundamental para maximizar el bienestar de la sociedad en general. Por lo tanto, las administraciones públicas deberían implementar sistemas de IA que cumplan con este principio ético. Desde la perspectiva de la deontología, el principio de justicia y la protección de los colectivos más vulnerables también son fundamentales y que, además, no atenten contra la privacidad, por ejemplo, evitando la elaborando perfiles. En este sentido, se podría argumentar que las administraciones públicas tienen una obligación ética de garantizar que los sistemas de IA sean justos y no discriminativos. Además, el principio de transparencia se basa en la explicabilidad, lo que también es un valor deontológico, de forma que la IA pueda ser auditable e interpretable por las personas en sus distintos niveles de comprensión, tanto a nivel técnico como en los procesos que intervengan las personas. Por último, desde la perspectiva de la ética de la virtud, se podría argumentar que los responsables de la aplicación de la IA deben desarrollar virtudes como la responsabilidad y la consideración por el bienestar de los ciudadanos y colectivos vulnerables. En este sentido, se tienen en cuenta los principios éticos en todas las etapas del desarrollo y uso de la IA. En base a estos principios, las administraciones públicas se tienen que plantear las preguntas sobre qué y cómo se deben aplicar, cuándo y por quién.

8.1. ¿Qué y cómo se debe hacer?

En cuanto a lo qué se debe hacer, se ha visto que todavía existen muchas cuestiones pendientes. Lo que sí está claro es que ya existen directrices y principios generales, los cuales deben ser la guía de las acciones a realizar. Tenemos distintas problemáticas, muchas no tienen una solución clara, tales cómo sobre mitigar los problemas de privacidad, de los sesgos, de transparencia y de rendición de cuentas. Sin embargo, no está tan claro qué debe hacerse, qué curso de acción preciso debe tomarse. Por ejemplo, no está tan claro cómo tratar la transparencia o la parcialidad, los prejuicios existentes en la sociedad y las opiniones divergentes sobre la justicia y la equidad. En cualquier caso, un aspecto importante es disponer de una definición y clasificación clara de la IA, e incluso si se debiera cambiar este término por otros que realmente no transmitan un mensaje confuso a la sociedad, en el sentido de que se utilicen términos que parezcan entes autónomos con una inteligencia próxima a la humana. Así, primero es separar aquellos sistemas que realmente conduzcan a problemas éticos y, sobre estos, clasificarlos en función de sus repercusiones éticas. Como respuesta a esto, el Parlamento Europeo y el Consejo Europeo han solicitado una propuesta de Reglamento de Inteligencia Artificial (Comisión Europea, 2021). En esta propuesta se perfecciona la definición de los sistemas de IA, establece prohibiciones concretas y formula clasificaciones claras, incluyendo criterios específicos para los sistemas de IA de alto riesgo.

Sobre cómo hacerlo, los instrumentos principales son las medidas regulatorias y disponer de códigos de conducta y buenas prácticas, así como, de educación. En cuanto a las medidas regulatorias, estas son amplias, pero estas deben basarse en las directrices y principios que estén establecidos. Las leyes, decretos y reglamentos son los instrumentos principales de las administraciones públicas. En cualquier caso, las primeras preguntas que deben hacerse son si la normativa actual es suficiente, si hay que reforzar o si hay que desarrollar una nueva normativa, aunque lo más seguro que será una combinación de estas medidas. Muchos aspectos pueden estar bajo paraguas de normas actuales, como es lo relacionado con la protección de datos y la seguridad de la información, pero que, posiblemente, requerirá de interpretaciones de esta normativa, no solo en lo concerniente a las medidas para garantizar la estos dos aspectos, si que exista una clara atribución de responsabiliades (Vázquez de Castro, 2020: pp. 263-269). En cualquier caso, aspectos cómo la transparencia de los algoritmos requerirá desarrollos normativos que creen marcos para la implantación de sistemas de IA en la misma medida que ya existen marcos o guías específicos en otros aspectos como, por ejemplo, para el cumplimiento de los Esquemas Nacionales de Seguridad o de Interoperabilidad. En cuanto a las medidas para disponer de códigos de conducta y buenas prácticas, queda mucho por desarrollar dentro de las administraciones públicas. Estas medidas cubrirán aquellos aspectos donde la normativa no sea suficiente y se profundice en lo relacionado con la responsabilidad activa. Sobre esto, en otros ámbitos ya existen experiencias, como las de IEEE (Instituto de Ingenieros Eléctricos y Electrónicos), que ha desarrollado un

documento de lo que tiene que ser un diseño alineado con la ética (IEEE, 2017), que puede servir como referente. Por último, en lo relacionado con la educación, está claro que para el conjunto del colectivo de empleados públicos, no solo en el ámbito tecnológico, se deben incluir en los planes de formación medidas para que conozcan, entiendan, apliquen y usen la IA.

8.2. ¿Cuándo?

Con respecto a la dimensión temporal, hay que plantearse diversas cuestiones. En primer lugar, ¿cuándo empezar a aplicar las políticas específicas para la IA? Parece obvio que cuanto antes mejor, ya que una vez que la IA esté incorporada en las administraciones públicas, podría ser demasiado tarde o ser costoso adaptar los sistemas establecidos a las políticas establecidas. En segundo lugar, ¿hasta cuándo sería el horizonte temporal de las políticas? En este caso, es difícil prever el grado de evolución de la IA, sobre todo si se llegará a disponer de una IA fuerte en el corto plazo. Lo que sí parece aconsejable es centrarse en la IA débil y que las políticas que se apliquen deberían estar dotadas de la suficiente flexibilidad para que se vayan adaptando a la evolución tecnológica de la IA.

8.3. ¿Por quién?

La última cuestión es quién debería actuar. Aquí la mayor problemática proviene de la existencia de diversos actores y su ámbito de actuación. Una cuestión es si se quiere que exista un planteamiento universal. Esto podría ser aconsejable y, de hecho, la OCDE y la UNESCO han desarrollado orientaciones éticas sobre la IA. También parece que, tal como argumenta Coeckelbergh (2020a, pp. 156-157), aunque existen diferencias culturales entre países y regiones, las políticas sobre la ética de la IA "*son notablemente similares*", con independencia que algunos países tienen sutiles diferencias, como el caso de China que hace mayor hincapié en la estabilidad social y el bien colectivo. Por ello, parece que los actores internacionales pueden tomar un papel importante en la definición en una tabla de mínimos, siendo cada país o región las que pueden concretar las políticas, así como desarrollar la normativa pertinente sobre esta materia.

Otro de los actores principales está al lado de la parte tecnológica, es decir, los que diseñan y desarrollan los sistemas de IA. En este lado, los centros encargados de definir las políticas y normas de carácter técnico de cada administración parecen que deberían ser los responsables de la definición de códigos de actuación o de conducta, así como de buenas prácticas. En cualquier caso, en España el planteamiento pasa por la creación de una agencia específica, en el mismo sentido que existen agencias en otros ámbitos, como en la protección de datos. En este sentido, dentro de la disposición adicional centésima trigésima de la Ley 22/2021, de 28 de diciembre, de Presupuestos Generales del Estado para

el año 2022 se autoriza al Gobierno a impulsar una Ley para la creación de la Agencia Española de Supervisión de Inteligencia Artificial en España, dentro de sus principales objetivos está la minimización de riesgos de la IA.

En definitiva, aunque las políticas y las directrices internacionales tienen una alta convergencia, las administraciones públicas, en la medida que se incorpore la IA, se enfrentan a un importante reto en la interpretación de la actual normativa, su modificación y la elaboración de nueva normativa, así como, los códigos de conducta y de buenas prácticas. Todas estas acciones podrán ser asuntos que en el corto plazo puedan llegar a ser controvertidos, debido a muchas de las cuestiones sin resolver que giran en torno a la ética de la IA.

9. CONSIDERACIONES FINALES

Ciertas cuestiones que se han planteado en este trabajo han sido contestadas, pero también otras cuestiones han quedado abiertas. Se ha pretendido dejar claro muchos de los problemas éticos con que se enfrentan la implantación de la IA, pero lo que no ha llegado a contestarse completamente es sobre cómo muchos de estos problemas se pueden resolver definitivamente. Cuanto más compleja es la problemática, los planteamientos para solucionarlos son más inciertos, desde que en la actualidad no existen soluciones tanto algorítmicas ni normativas que los resuelvan. En cualquier caso, aunque se está lejos de resolver muchos problemas, sí existe un conocimiento de qué problemas se pueden o no se pueden resolver con el conocimiento actual.

Muchas de las cuestiones sin contestar aparecen cuando se intenta solucionar un problema surgen otros problemas, en otras palabras, nos encontramos con compromisos. Así, al mitigar los problemas de privacidad pueden aflorar problemas de sesgos, o cuando abordamos los problemas de sesgos nos encontramos con problemas de efectividad que, por otra parte, pueden abundar en otros problemas de sesgos. También, cuando se pretende mejorar la explicabilidad o interpretabilidad de la IA pueden venir acompañada con una disminución de la eficacia. Por tanto, la implantación de la IA cuando nos encontramos con problemas éticos, la solución más que técnica tendrá que venir acompañada de deliberaciones, donde estén implicados los distintos interesados, los cuales tendrán que balancear diferentes principios, como el de privacidad, justicia, transparencia y eficacia.

Por tanto, para abordar los problemas éticos relacionados con la implantación de la IA, es necesario encontrar un equilibrio entre los diferentes principios éticos. Desde la perspectiva del consecuencialismo, se podría argumentar que se deben buscar acciones que maximicen el bienestar y minimicen el daño, incluso si esto implica encontrar un compromiso entre diferentes principios éticos. Por ejemplo, en el caso de los problemas de privacidad y sesgos, se podrían tomar acciones que minimicen ambos, aunque no se logre erradicar completamente ninguno de los dos. Desde la deontología, se podría argumentar que hay ciertos

principios éticos que deben ser seguidos, como la protección de la privacidad y la justicia, y que no se deben sacrificar en favor de la eficacia o la eficiencia. Por último, desde la ética de la virtud, se podría argumentar que las personas involucradas en el desarrollo y uso de la IA deben desarrollar ciertas virtudes, como la responsabilidad y la consideración por el bienestar de los afectados, y que se deben tomar acciones que aseguren que la implantación de la IA sea justa y equitativa a largo plazo. En definitiva, la implantación de la IA exigirá deliberaciones para establecer el mejor curso de acción a seguir, considerando los diferentes principios éticos involucrados.

En cualquier caso, aunque en este artículo se han destacado muchos de los problemas éticos y riesgos asociados con la IA, también es cierto que no todas las implantaciones tienen que enfrentarse a esta problemática. En estos casos, se podría argumentar desde la perspectiva del consecuencialismo que se debería permitir el uso de la IA, siempre y cuando se hayan abordado adecuadamente los aspectos técnicos necesarios. Sin embargo, es importante tener en cuenta que cuando se plantea la utilización de la IA, se deben considerar cuidadosamente los posibles cursos de acción a tomar y preguntarse en qué medida su implantación podría comprometer los aspectos éticos tratados en este artículo, como la privacidad, la transparencia y la justicia. Desde la deontología, se podría argumentar que la protección de estos principios éticos debe ser una consideración fundamental en la toma de decisiones relacionadas con la IA. En resumen, aunque no todas las implantaciones de IA presentan problemas éticos, es importante abordar estos aspectos con prudencia para asegurar que la IA se utilice de manera ética y responsable.

10. REFERENCIAS

Aliman, N. M., & Kester, L. (2019). Requisite variety in ethical utility functions for AI value alignment. *arXiv preprint arXiv:1907.00430*.

Anantrasirichai, N., & Bull, D. (2021). Artificial intelligence in the creative industries: a review. *Artificial Intelligence Review*, 1-68.

Baquero Pérez, P. J. (2023). Retos de la implantación de la inteligencia artificial en las Administraciones Públicas. En M.ª C. Campos Acuña & I. Expósito Suárez (Eds.), *La transformación de las Administraciones Públicas de Canarias* (pp. 95-105). La Ley.

Bies, R. J. (2001). Interactional (in) justice: The sacred and the profane. *Advances in organizational justice*, *89118*.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Camps, V. (2022). *Breve historia de la ética*. RBA Libros.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Coeckelbergh, M. (2020a). *AI ethics*. Mit Press.

Coeckelbergh, M. (2020b). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*(4), 2051-2068.

Comisión Europea. (2021). *Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión* (COM/2021/206 final). https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A52021PC0206

Correia, M., Rego, G., & Nunes, R. (2021). The right to be forgotten and COVID-19: Privacy versus public interest. *Acta bioethica*, *27*(1), 59-67.

Doğuç, Özge. Robotic process automation (RPA) applications in COVID-19. En *Management Strategies to Survive in a Competitive Environment*. Springer, Cham, 2021. p. 233-247.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer, Berlin, Heidelberg.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends®* in *Theoretical Computer Science*, *9*(3–4), 211-407.

Expert Group on Liability and New Technologies (2019). *Informe sobre responsabilidad derivada de la inteligencia artificial y otras tecnologías digitales emergentes* (Report from the Expert Group on Liability and New Technologies). Comisión Europea]

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control : A theory of moral responsibility*. Cambridge University Press

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, *7*(1), 1-9.

French, R. M. (2000). The Turing Test: the first 50 years. *Trends in cognitive sciences*, *4*(3), 115-122.

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, *2*(1), 1.

Gabriel, M. (2017). *I am Not a Brain: Philosophy of Mind for the 21st Century*. John Wiley & Sons.

Gabriel, M. (2019). *El sentido del pensamiento*. Madrid: Pasado y presente.

Gabriel, M. (2021). Ética para tiempos oscuros. Valores universales para el siglo XXI. Barcelona: Pasado & Presente.

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & SOCIETY*, *35*(4), 795-809.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273-278.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in cognitive science*, *2*(3), 528-554.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42.

Harris, J., & Anthis, J. R. (2021). The moral consideration of artificial entities: a literature review. *Science and Engineering Ethics*, *27*(4), 1-95.

HLEG-AI (Grupo de Expertos de Alto Nivel sobre Ia IA), *"Directrices éticas para una IA fiable"*, Unión Europea, Bruselas,8 de abril 2019.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.2017. *"Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems,"* Version 2. IEEE.

Jiang, B., Li, J., Yue, G., & Song, H. (2021). Differential privacy for industrial internet of things: Opportunities, applications, and challenges. *IEEE Internet of Things Journal*, *8*(13), 10430-10451.

Júdez, J., & Gracia, D. (2001). La deliberación moral: el método de la ética clínica. *Medicina clínica*, *117*(1), 18-23.

Khanzode, K. C. A., & Sarode, R. D. (2020). Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A Literature Review. *International Journal of Library & Information Science (IJLIS)*, *9*(1), 3.

Klinger, J., Mateos-Garcia, J., & Stathoulopoulos, K. (2018). Deep learning, deep change? Mapping the development of the Artificial Intelligence General Purpose Technology. *arXiv preprint arXiv:1808.06355*.

Larson, Erik J. *The Myth of Artificial Intelligence*. Harvard University Press, 2021.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, *49*(1), 15-21.

Martin, K. E. (2019). Designing ethical algorithms. *MIS Quarterly Executive June*.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175-183.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.).https://christophm.github.io/interpretable-ml-book

Muñoz Vela, J. M. (2022). *Retos, riesgos, responsabilidad y regulación de la inteligencia artificial: Un enfoque de seguridad física, lógica, moral y jurídica*. Editorial Aranzadi.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Peters, U. (2020). What is the function of confirmation bias?. *Erkenntnis*, 1-26.

Prabhumoye, S., Boldt, B., Salakhutdinov, R., & Black, A. W. (2020). Case study: Deontological ethics in NLP. *arXiv preprint arXiv:2010.04658*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radhika, D. (2012). Ethics in public administration. *Journal of Public Administration and Policy Research*, *4*(2), 23-31.

Rawls, J. (1999). *A Theory of Justice*. Cambridge: Harvard University Press. (año de publicación del libro original; 1971).

Rosa, M., Feyereisl, J., & Collective, T. G. (2016). A framework for searching for general artificial intelligence. *arXiv preprint arXiv:1611.006*

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 15.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 1-28.*

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine learning, 10(2), 153-178.*

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences, 3(3), 417-424.*

Simon, H. A. (1957). *Models of man; social and rational.* New York: Wiley

Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing, 7(1), 35.*

Tamayo Haya, S. (2020). Los robots como entes inteligentes. En J. I. Solar Cayón (Ed.), *Dimensiones éticas y jurídicas de la inteligencia artificial en el marco del Estado de derecho*. Editorial Universidad de Alcalá.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). LaMBA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239.*

ran, C., Fioretto, F., Van Hentenryck, P., & Yao, Z. (2021). Decision Making with Differential Privacy under a Fairness Lens. *In IJCAI (pp. 560-566).*

Ulgen, O. (2017). Kantian ethics in the age of artificial intelligence and robotics. *QIL, 43, 59-83.*

Ulnicane, I. (2022). Artificial Intelligence in the European Union: Policy, ethics and regulation. In *The Routledge Handbook of European Integrations*. Taylor & Francis.

Valcárcel, A. (2002). Ética para un mundo global: una apuesta por el humanismo frente al fanatismo. Temas de hoy.

Van de Poel, I. R., & Royakkers, L. M. (2011). *Ethics, technology, and engineering: An introduction. Wiley-Blackwell.*

Vázquez de Castro, E. (2020). Aproximación a la responsabilidad derivada de los riesgos de la inteligencia artificial en Europa. En J. I. Solar Cayón (Ed.), *Dimensiones éticas y jurídicas de la inteligencia artificial en el marco del Estado de derecho* (pp. 1-367). Editorial Universidad de Alcalá.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093.*

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., y Schwartz, O. (2018). *AI now report 2018 (pp. 1-62). New York: AI Now Institute at New York University.*

Weber, M. (1946), *From Max Weber: Essays in Sociology, traducción, compilación e introducción de H. H. Gerth y C. Wright Mills, Nueva York, Oxford University Press.*

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, *42*(7), 596-615.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953*.

Zuiderwijk, A., Chen, Y. C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3), 101577.