Trabajo Final de Grado en Humanidades. Estudios Interculturales

De la Inteligencia Artificial a la Moral Artificial

Posibilidades y consideraciones éticas

Autor: Iván Rodríguez Tena

Tutor: Daniel Pallarés Domínguez

Data de lectura / Fecha de lectura: Octubre de 2019

Resumen: En este trabajo se tratará de hacer una aproximación a los desafíos a los que se enfrenta la investigación en inteligencia artificial, para ello se hará desde, primero un marco conceptual sobre la inteligencia, centrándose en la teoría de las inteligencias múltiples y las posibilidad de emular éstas artificialmente. Seguidamente, se hará un acercamiento a la realidad de las investigaciones actuales y sus repercusiones, especialmente las respuestas que se generan desde organizaciones y gobiernos, así como de la opinión pública. En la última parte se tratará de como los estudios sobre IA han hecho aparecer un nuevo campo en la investigación, la moralidad artificial, y de como por ello se buscan maneras de que las aplicaciones diseñadas sean puedan ser seguras y beneficiosas para la sociedad. Todo esto implicará una necesidad de que la investigación en este campo vuelva la mirada a las ciencias humanas y sociales para poder responder a los desafíos que se le presenten.

Palabras clave: Cerebro, mente, ética, inteligencia artificial, moral artificial.

Índice

Abstract

Humankind faces great challenges when it comes to applications of artificial intelligence systems in society. This, sometimes, causes fear and, at least, it causes concern to people at every level of society, this is a good reason to approach the theory, design and application of this systems and do it with an ethical point of view. New technologies and their impact in society is a very interesting and important issue due to the speeding in their development and the problems it causes when people are unprepared and this kind of technologies don't consider certain perspectives as gender or socio-economic status.

The objective will be, then make an approach to this field of science taking account of ethics, specifically, and human science broadly. And focusing in the ethic implications that development of autonomous machines has in its application, considering the impact in people and at the end, the kind of treatment people will be given to this machines.

To do so, it will be used actual papers and chapters of books about theories of the concept of intelligence; documentation about two of the most prominent projects in the study of the brain; classic literature about moral, ethics and philosophic concepts concerning this matters. Also in a more technical point of view will be used papers about design and impact of artificial intelligence into people and its ethics considerations.

This work will be structured in three parts. First part focus on the conceptual framework of intelligence, the possibility of emulation and a first glimpse of the issues that can be rise from this field. Next part will made an approach to the two big projects about the study of the brain and their goals and implications. Also, it will focus on some of the real applications of IA systems in real life and the ethical concerns that appears when man and decision making machines collide. Last part is going to dig into the importance IA applications has, and will have as it develops, in human lives. This creates a new field of study moral intelligence and rise the issue to higher levels as it is seen how complex it becomes when IA gain more and more features similar to human beings.

To conclude, it is seen that human sciences has to be side by side with these new technologies to be able to answer complex questions about autonomy, ethical implications and to define how are going to be the new relationships between human and machine.

Introducción

La preocupación de la humanidad frente a que sus creaciones la puedan superar y supongan un peligro para ella misma siempre ha estado en el imaginario humano. La ciencia ficción ha mostrado muchas posibilidades sobre el futuro de la humanidad y uno de estos posibles futuros ha sido el de la creación de una inteligencia artificial que acabara por convertirse en la némesis de la raza humana amenazando con su extinción.

Sin llegar tan lejos en la especulación, multitud de historias han tratado sobre las relaciones entre el ser humano y las máquinas pensantes, ya fueran robots u ordenadores dotados de superinteligencia. Para prevenir estos peligros ya antes de que el propio nombre de Inteligencia Artificial apareciera, Isaac Asimov inventó las leyes de la robótica (y el mismo termino robótica) que protegerían a los humanos de posibles amenazas que provinieran de los seres artificiales creados por los humanos, insertándolas en su propia esencia para que así no pudieran saltárselas y dañar o desobedecer a sus creadores.

Por otro lado esta idea se ha mantenido en el ámbito académico y en la preocupación de organismos e instituciones internacionales, pero en cambio en el ámbito práctico, militar o empresarial, no parece que se estén teniendo en cuenta en absoluto.

Debido a la relevancia que tiene toda la tecnología relacionada con la Inteligencia Artificial, impulsando nuevos avances científicos, prometiendo un mayor progreso de la tecnología humana, se puede pensar que es algo que está más cerca de la ciencia ficción que de la realidad cotidiana, pero esto no es así.

Los servidores de correo electrónico, que se utilizan todos los días por millones de personas, basan el filtrado de correo *spam* en el trabajo de Inteligencias artificiales, ya que ahora mismo es una tarea imposible para los seres humanos, *Google Photos* utiliza una IA para crear animaciones o fotografías panorámicas basándose en tu galería fotográfica y a través del reconocimiento de patrones las une sin que apenas se puedan percibir fallos a primera vista.

En los servicios de seguridad chinos o americanos se utilizan para el reconocimiento facial e incluso para el reconocimiento de actitudes sospechosas de terrorismo, el momento en el que pensar cuando se normalizaría el uso de esta tecnología entre la

sociedad ha pasado, su uso ya está generalizado, unas veces de maneras más sutiles que otras.

Por todo esto, este trabajo tratará de hacer un estudio filosófico, crítico y argumentativo sobre las implicaciones éticas que tiene el desarrollo y la implantación de inteligencias artificiales en la sociedad, ya sean IA débiles, concretas, creadas para trabajos específicos como las que ya existen, o IA fuertes, que estén dotadas de una inteligencia general, capaz de superar, en teoría a los seres humanos. Con el objetivo de poner en valor la necesidad de que en todos los pasos de diseño, desarrollo e implantación de sistemas de IA se tiene que mantener una perspectiva ética y de responsabilidad social, debido, como se ha visto al gran impacto que todo esto produce en la sociedad.

Para ello se partirá de la definición de un marco conceptual que abordará desde diferentes puntos de vista lo que puede considerarse inteligencia y la complicación que puede tener adaptar esta característica humana a un sistema artificial. Asimismo se tratará la forma en la que se puede intentar demostrar si una máquina ha alcanzado este logro o no. Se tratarán los escollos a los que se enfrentan estas investigaciones y los caminos que se están tomando.

En la segunda parte se hará una aproximación a la realidad física del desarrollo de mentes artificiales, revisando los dos grandes proyectos que en la actualidad están en marcha dotados de fondos billonarios, uno en Europa y otro en Estados Unidos, además de esto se hará una aproximación a las respuestas que este tema está despertando a lo largo del mundo, principalmente debido a las implicaciones morales que tienen sus aplicaciones en la sociedad, posiciones que van desde una visión optimista que quiere impulsar cualquier avance científico hasta aquellos que reclaman que se actúe con cautela debido a los peligros que pueden darse al implantar estos sistemas sin un control sobre ellos.

Se abordará también un proyecto que trata de buscar respuestas, aunque sean superficiales, para solucionar los problemas éticos que puedan surgir de dilemas a los que se enfrenten los sistemas autónomos y las conclusiones que se puede sacar de esto. Esta parte terminará mostrando las posturas que se toman desde instituciones

internacionales como son la Unión Europea y la UNESCO, que movidas por la importancia que tienen estas tecnologías para el futuro de la humanidad, no quieren quedarse atrás, ni tampoco quieren que se convierta en una tecnología que se salga de control o que quede reducida a beneficiar a unos pocos, para ello diseñan directrices y financian proyectos relacionados con todo esto.

En el último punto, después de haber tratado y puesto el foco en la importancia que tiene la IA en el desarrollo humano y en cómo puede afectar y cambiar las sociedades donde se implanten sus sistemas, se centrará en el objetivo de valorar como es de importante empezar a hablar de moralidad artificial. Se abordarán diferentes posturas sobre como implantar este nuevo factor en el diseño de sistemas autónomos, y de las diferencias entre los modelos de Inteligencia Artificial y de Moralidad Artificial. Y se pondrá de manifiesto la importancia también de revalorizar las ciencias humanas y sociales dentro de estas investigaciones que en principio pueden parecer puramente tecnológicas, ya que los problemas que como se verán van surgiendo van más allá de las ciencias empíricas enfocadas a la tecnología y debido a la magnitud de los cambios que se pueden dar en la sociedad requieren enfoques mucho más humanistas.

Además de todo esto, este nuevo factor de la moralidad artificial implica que se tiene que empezar a tratar las maquinas, que se vayan a dotar de inteligencia artificial, de otra manera diferente a herramientas, ya que las características que pudieran adquirir o de las que habría que dotarlas para que fueran verdaderamente autónomas harían cambiar aun más la sociedad y modificarían incluso la manera en las personas se tratan las unas a las otras.

1. Marco conceptual para la inteligencia y sus posibilidades

Durante estos últimos años se han dado dos debates sobre la inteligencia artificial: si es posible alcanzar el objetivo de conseguir crearla y si la consecución de esto es algo muy peligroso, tanto como para llevar incluso a la extinción del ser humano. Filósofos, científicos y personalidades de la política y la cultura se han posicionado en los extremos de ambos debates, mientras que la cultura popular se ha llenado de los posibles futuros que aportaba la ciencia ficción en la literatura y el cine.

En los dos primeros capítulos se tratará principalmente el tema de la posibilidad de lograr el diseño y la creación de una verdadera inteligencia artificial y de los diversos proyectos y posturas que se toman en el mundo de la ciencia y de la filosofía.

Para empezar a hablar sobre la inteligencia artificial primero se debería empezar definiendo la misma palabra inteligencia, para así poder encuadrar las investigaciones dirigidas a la creación de inteligencias artificiales que se están llevando a cabo en informática. Algo que en principio puede parecer poco complejo, pero cuanto más se ahonda en el concepto y se intenta definir también como implementarlo en una máquina se vuelve una tarea muy complicada.

También es importante saber cómo se relaciona el concepto de inteligencia con las máquinas y que características deberán tener estas para que se las considere inteligentes, de esta manera se retrocederán unas décadas para ver cómo uno de los genios y pioneros en la computación trató de explicar este paso de la naturaleza a la máquina.

1.1. Alan Turing, pionero en la ciencia informática

Desde el mismo momento en el que Alan Turing publicó su artículo *Maquinaria computacional e inteligencia* en 1950, dónde trataba de ver si las maquinas creadas por el hombre podrían pensar en algún momento, ya dedicó gran parte de el tan solo a reformular la pregunta ¿Pueden las máquinas pensar? Para así poder establecer el criterio de su famoso test. En el cual, realmente, la pregunta era si una máquina era capaz de engañar en una conversación a un ser humano de tal manera que éste fuera incapaz de distinguir a la máquina de otro ser humano.

Turing definía una maquina, susceptible de pensar, como se podría definir cualquier

ordenador a día de hoy, salvando las distancias temporales en potencia computacional. Colocaba esta máquina en la situación de enfrentarse al juego de la imitación, y así reformulaba por primera vez su pregunta de si las máquinas pueden pensar a si una computadora digital puede funcionar bien en el juego de la imitación.

Ya durante estos momentos tan tempranos en los que se planteaban si las máquinas podrían parecerse a los seres humanos, Turing tenía en consideración múltiples opiniones contrarias a que ésto se pudiera conseguir. Y defendía su postura de que se podría dar el caso en unos 50 años, frente a distintos argumentos: como por ejemplo que las máquinas no tenían alma y el pensamiento era una función de ésta. Otro contraargumento, ya antes de que se acuñara la palabra inteligencia artificial, era el miedo a conseguirlo y la esperanza de que no ocurrieran las terribles consecuencias que esto traería. También valoraba las limitaciones de la lógica matemática y que esto haría que dieran resultados erróneos, pero argumentaba que los humanos somos falibles y la máquina trata de imitar al ser humano no de hacer, en principio, una versión infalible de éste.

Algunos contemporáneos a Turing también objetaban que los ordenadores carecerían de consciencia, lo que no les permitiría alcanzar un nivel de inteligencia comparable al ser humano, pero para él la consciencia era un misterio que no hacía falta resolver antes de conseguir una máquina pensante, y que la programación de la computadora trataría de hacerla pasar por un ser humano para ganar en el juego de la imitación y no tenía, realmente, que ser como un ser humano en sus procesos internos. De esta manera lo que importaba era que, en la práctica, pudiera parecer que era inteligente, dando como resultado que un juez humano no pudiera distinguirlos. Este resultado demostraría que la computadora a efectos prácticos es inteligente.

Otras objeciones al éxito en el desarrollo de la inteligencia artificial serían: la incapacidad de creación artística original, la imposibilidad de separar el pensamiento del sistema nervioso, la de conseguir un comportamiento informal o incluso, algo que estaba de moda en ese momento, la incapacidad de desarrollar una percepción extrasensorial (Turing, 1950).

Todos estos argumentos remarcan la relevancia que tienen para la sociedad los

estudios que trabajan en el diseño de máquinas que reproduzcan la inteligencia humana. El debate sobre esto, los miedos y el escepticismo se dan en diferentes sectores de la sociedad donde se valoran aspectos de todo tipo, religiosos, morales, técnicos, etc. De ahí la importancia que tiene este tema y también la necesidad de que la comunidad científica y los gobiernos en general lo tengan en cuenta antes de que se desarrolle y no *a posteriori* cuando surjan los problemas y sea más difícil de solucionarlo todo.

1.2. El propio concepto de inteligencia

En el contexto que nos ocupa es necesario definir el concepto de inteligencia para poder indagar en estas posibilidades o limitaciones para el diseño de una máquina o programa cuyo objetivo sea recrear artificialmente una inteligencia como la humana. Para esto damos por supuesto que la inteligencia humana sería el ideal de este objetivo. Más adelante se podrá apreciar como quizá este ideal para la máquina no se corresponda tanto como se espera con el ser humano y puede que incluso lo sobrepase.

Las principales características por la que los seres humanos se distinguen de los animales, si se le preguntara a cualquier persona, diría que es la inteligencia y la moralidad. Por ejemplo Francisco J. Ayala (2006) en el capítulo “Las raíces biológicas de la moral”trata de especificar cuáles son las condiciones de posibilidad, necesarias y suficientes, que ha de darse para que un ser pueda tener conductas que califiquemos de morales o éticas. Señala tres condiciones, y lo hace como describiéndolas como capacidades que precisan de una serie de facultades y atributos con determinado nivel de complejidad, que son, asimismo, condiciones de posibilidad de la moralidad. Estas tres capacidades son:

La capacidad de anticipar los propios actos. Que hace uso de las facultades de la imaginación, la memoria y el pensamiento abstracto, para así proyectar la situación actual hacia posibles futuros hipotéticos, de manera que se pueda elegir el moralmente más apropiado.

La segunda capacidad sería la de hacer juicios morales. Que trata de juzgar las razones que se tienen para actuar de determinada forma, esto hace necesario el uso de la metacognición que se asocia a las facultadas mencionadas en la anterior capacidad, así

como a las de la razón y el lenguaje. A esto se le sumarían las emociones, las facultades sociales y la prosocialidad, no solo se basarían estos juicios en lo racional. Estos aspectos de la emoción y lo social servirían para buscar lo mejor más allá de uno mismo, también hacia los demás.

Por último estaría el libre albedrío. Que necesita de dos condiciones. La primera sería la posibilidad real de escoger de entre, al menos, dos opciones, para que pueda existir libertad. Y la segunda es que sea el sujeto, el *yo*, la causa de la elección, pues la libertad es autodeterminación. Pese al debate y a las investigaciones en curso sobre libre albedrío y determinismo, se tiene que asumir la existencia de libertad de acción como condición necesaria para considerar a alguien un agente moral.

Si se habla dela inteligencia o de la planificación, según como se la defina, encontramos muchos animales que pueden ser calificados de muy inteligentes. Si ahondamos en las diferencias se podría decir que los humanos tenemos imaginación para poder planificar acciones que nos proporcionen beneficios en el futuro, pero hay ejemplos de animales, tan inferiores (supuestamente) como los insectos, las hormigas específicamente, que acumulan hojas para cultivar hongos dentro del hormiguero y así tener reservas de comida. Lo que sorprende de esto es que preparan un producto secundario, el hongo, no es una simple acumulación de alimento.

Por otro lado, quizá más sorprendente aun, están los cuervos de Nueva Caledonia, que no solo utilizan herramientas, muchos animales las usan de una manera directa y para una tarea específica, como pueden ser rocas para romper cascaras de huevos o simientes; estos cuervos construyen sus herramientas. Normalmente utilizan estas herramientas para alcanzar comida que por si solos son incapaces. Pero también se han documentado casos de cuervos adaptando un alambre para poder alcanzar la comida e incluso modificando el tamaño y la forma para diferentes situaciones. Lo más importante de este estudio con los cuervos de Nueva Caledonia es el uso de herramientas para interactuar con objetos que les son novedosos, supuestamente por si acaso entrañan algún tipo de peligro (Wimpenny, 2011; Knabe, 2017).

Como se puede ver con los ejemplos sobre inteligencia animal, hay especies que interactúan frente a su entorno de una manera muy parecida a la humana, probablemente

mediante procesos mentales muy distintos a los de los seres humanos, pero logrando objetivos o resultados muy similares a los que conseguirían estos.

Así pues, filósofos y programadores se plantean la idea de que no sería necesario que una máquina pensara realmente como un humano, sino que tan solo valdría con que pareciera que pensara como tal, de manera que, a efectos prácticos, si puede ofrecer unos resultados de adaptación y de resolución de problemas similares a los de los humanos o incluso superiores, no haría falta que sus procesos internos o su diseño tuvieran algo que ver con los de un ser humano. De esta manera se abren muchas posibilidades a la hora de implementar algún tipo de inteligencia artificial, pudiendo evitar que esta inteligencia interiormente sea como la humana y ampliando el margen sobre el que pueden trabajar los programadores o ingenieros.

Se tienen que tener en cuenta varias cosas a partir de este momento, primero que clase de inteligencia va a querer ser emulada artificialmente, y luego si esto es posible de implementar en un soporte artificial.

El concepto inteligencia se ha definido de diversas maneras, una de estas, la más clásica es aquella en la que es una sola característica general, medible mediante test, y que deja a los individuos en una posición dentro de una linea que va desde lo menos inteligente a lo más inteligente según el Cociente Intelectual resultado de esos test. Para Howard Gardner esta forma de medir y definir el concepto de inteligencia deja de lado aspectos muy importantes del pensamiento humano y ha servido principalmente, y aun se utiliza, para asignar presupuestos en educación y para excusar argumentos racistas, clasistas y eugenésicos que reducen la inteligencia a un rasgo hereditario que no se puede potenciar con acciones sociales o educativas, donde esos mismos test tienen sesgos que ya condenan a ciertos sectores de la sociedad a la mediocridad (Gardner, 2010:14-15, 20-24).

Para Gardner esta forma de ver la inteligencia, como un solo rasgo distintivo, era muy incompleto e inició el estudio de la ella desde una aproximación neurológica, relacionando diferentes capacidades potenciales de los humanos con las partes del cerebro que podían gestionarlas. De esta manera dividió el concepto en varias inteligencias, dando pie a la teoría de las inteligencias múltiples.

Esta teoría se basa en que hay individuos que se desenvuelven mucho mejor en unos aspectos de la vida que en otros, de manera que alguien que puede ser aparentemente muy inteligente, a la hora del trato social podría ser un inepto, de manera que destacar en una de las inteligencias no haría que esto pasara en las demás. Alguien que manejaba muy bien la palabra puede sufrir una lesión cerebral y perder esta capacidad sin, por ejemplo, perder a su vez las capacidades matemáticas.

A partir de estas premisas y de multitud de experimentos y fuentes de información Gardner dividió las inteligencias en siete. Definiendo el concepto de inteligencia como «un potencial biopsicológico para procesar información que se puede activar en un marco cultural para resolver problemas o crear productos que tienen valor para una cultura» (Gardner, 2010: 42-43).

Las siete inteligencias que definió Gardner son: la lingüística, la lógico-matemática, la musical, la corporal-cinestésica, la espacial, la interpersonal y la intrapersonal (Gardner, 2010: 51-54). Mas adelante valoraría aumentar estas a 8 con la naturalista.

Este potencial del que habla Gardner hace que se distancie de otras definiciones ya que también entra en valor la educación y el desarrollo de la persona y no lo deja en una simple característica prácticamente invariable y dependiente solo de la genética. Además cuando se habla de dotar de inteligencia a las máquinas el aspecto que más parece que cueste conseguir y por el que los seres humanos, por ahora, son superiores a cualquier logro en este campo, es la capacidad de adaptación a nuevos problemas o entornos.

Dividir la capacidad intelectual de las personas en estos distintos tipos de inteligencias permite facilitar el estudio de su implementación o simulación informática de manera que consigamos emular la inteligencia general mediante ordenadores. De esta manera se divide el problema y se llega rápidamente a la conclusión de que hay capacidades más fáciles de emular que otras.

A día de hoy se puede afirmar que ya hay inteligencias artificiales funcionando: filtros para el *spam* del correo electrónico, o autobuses que se conducen solos como los que hay en Suecia desde 2018, donde la empresa Ericsson también tiene en cuenta las reacciones de la gente e intenta que se normalice como un transporte más (Ericsson,

2019) unas IA que no podrán engañar a los seres humanos haciéndose pasar por estos, pero que sí que realizan tareas que ya resultan imposibles para la inteligencia humana. Se puede decir que ya nos superan totalmente, pero aun así, no llegan a ser tan adaptativas como los humanos y siguen teniendo una serie de limitaciones que aun no se han conseguido resolver.

A continuación, se podrán ver las distintas inteligencias y su relación con los avances que se dan en el campo de la inteligencia artificial.

Si se toma en cuenta la división de Gardner respecto a las inteligencias, el aspecto lógico-matemático puede parecer que está superado, por lo menos en ciertos aspectos de la física o el calculo los seres humanos no podemos avanzar apenas un paso más sin el apoyo de IA y ordenadores cada vez más potentes. Pero aquí nos encontramos con los problemas irresolutos de las matemáticas, como el de la hipótesis del continuo que no puede ser probada ni rebatida. Algo que afecta al aprendizaje de las maquinas, donde cuando se busca definir la capacidad de aprendizaje de una máquina en un problema no muy complejo, inicialmente, de conjuntos se topan con la misma paradoja que la de la hipótesis del continuo[1]. (Ben-David, 2019)

La inteligencia espacial, aquella que «supone la capacidad de reconocer y manipular pautas en espacios grandes […] y en espacios pequeños […].» (Gardner, 2010: 52) es otro tipo de inteligencia a la que se ha dedicado grandes esfuerzos para emular y con mucho éxito, a través de software de conducción y de reconocimiento de patrones en mapas y espacios diversos. En el nivel de los espacios pequeños, aun cuando pueda fallar en la creatividad, hay una gran evolución en algoritmos de cirugía, y de construcción y diseño. Es un campo en el que los sistemas de aprendizaje profundo han funcionado muy bien. Aunque su potencial aun puede llegar más lejos, sobre todo en los aspectos de la inteligencia artificial aplicada a la salud donde se ha avanzado mucho en los modelos predictivos de diagnostico de enfermedades donde:

> las técnicas de inteligencia artificial, en particular las que se engloban dentro del aprendizaje automático (*machine learning*), como es el caso del aprendizaje

1 Esta hipótesis trata de explicar una paradoja que se da en teoría de conjuntos, lo hace desde una pequeña historia donde un barbero no da abasto para afeitar a toda la población (es el único) y el rey entonces prohíbe que se afeite a cualquiera que pueda afeitarse a si mismo. Llega el momento en el que el barbero se ha de afeitar a si mismo, ¿puede o no puede hacerlo?

> profundo (*deep learning*), han llegado al nivel de madurez suficiente para su aplicación en el ámbito de la salud (*Conversation*, 2019).

De todas maneras, a estas aplicaciones les queda camino por recorrer, ya que aun es «el personal médico quien debe interpretar el resultado, ponerlo en contexto y utilizarlo como información adicional para diagnosticar» (Conversation, 2019).

En el caso de la inteligencia corporal-cinestésica «supone la capacidad de emplear partes del propio cuerpo (como la mano o la boca) o su totalidad para resolver problemas o crear productos» (Gardner, 2010: 52), se puede ver como aunque hay cosas como el caminar a dos patas que aun tiene problemas de equilibrio, hay multitud de robots que aprenden y manejan sus partes mecánicas de maneras absolutamente precisas, siendo capaces de realizar tareas que requieren coordinación visual-mecánica de manera casi perfecta.

Cuando se habla de la inteligencia lingüística se puede ver como se vuelve de nuevo al test de Turing, ya hay robots que tratan de aparentar humanidad en las conversaciones, como la famosa Sofia. También hay algoritmos que escriben según el estilo de un escritor del que tienen toda su obra en una base de datos, todos estos tienen sus limitaciones pero avanzan rápidamente en su desarrollo y pronto será muy difícil distinguirlos de una persona. Hay que tener en cuenta que muchas veces se piensa que la maquina fallará o se equivocará al interpretar algún argumento, pero los seres humanos fallamos muchas veces y malinterpretamos a nuestros interlocutores muy a menudo.

Ejemplos de como se trasladan estos errores a las IA que se diseñan en la actualidad se pueden ver en dos casos muy claros:

Uno el de un algoritmo de la empresa *Microsoft* llamada Tay. Esta IA estaba diseñada para mantener conversaciones con jóvenes a través de la red social *Twitter*. En 24 horas paso a hacer comentarios sexistas y xenófobos, insultando a colectivos y deseándoles la muerte mientras alababa a Hitler. Según la compañía se debió a un ataque coordinado para que adquiriera esas características (Mundo, 2016).

El otro es el problema que tenía *Google* para etiquetar individuos en las fotos personales. Cuando lo hacía, la IA etiquetaba a las personas negras como gorilas. La solución que tomaron no estaba exenta de polémica, ya que lo único que hicieron fue

eliminar de las etiquetas las palabras gorila chimpancé y mono. En este caso más allá de lo ofensivo de la situación no parecía dar pie a mayores peligros, pero este hecho demuestra como pueden llegar de sesgados por los prejuicios los datos que alimentan a los algoritmos. También es muy importante resaltar la opacidad del desarrollo de las IA. El problema real llega cuando las aplicaciones pueden llevar a casos de vulneración de los derechos o problemas más dañinos físicamente como que no se reconozcan a ciertos individuos en algoritmos de los coches automatizados (Salas, 2018).

La inteligencia musical para Gardner implica «la capacidad de interpretar, componer y apreciar pautas musicales» (Gardner, 2010: 52). Interpretar o componer ya hay máquinas que lo hacen de maneras insospechadas y de una forma que es indistinguible para los legos en música. El problema vendría con el aspecto de apreciar pautas musicales, si solo lo dejamos en reconocer esas pautas las IA ya hacen cosas más difíciles dentro del ámbito de reconocimiento de patrones, pero si se le da a apreciar el sentido de valorar, ya habría que dotar al algoritmo de algo más complicado como pueden ser los valores o los gustos.

Por último, las inteligencias que Gardner llamaba personales: la interpersonal «la capacidad de una persona para entender las intenciones, las motivaciones y los deseos ajenos, y , en consecuencia su capacidad para trabajar eficazmente con otras personas» (Gardner, 2010: 52); y la intrapersonal que «supone la capacidad de comprenderse a uno mismo, de tener un modelo útil y eficaz de uno mismo –que incluya los propios deseos, miedos y capacidades– y de emplear esta información con eficacia en la regulación de la propia vida» (Gardner, 2010: 53). Estos dos tipos de inteligencia tienen mucho que ver con la autoconciencia, con el sentido del yo y con los sentimientos. Por ahora esto es algo que si que parece que queda muy lejos de alcanzar a comprender de una manera que se pudiera implementar informáticamente. Y pese a que si hay manera de que una IA reconozca unos estados de animo en las personas, lo hace de manera, se podría decir que médica, reconociendo pautas que tiene almacenadas como datos, pero no puede compararlo con nada que pueda sentir ella misma, ya que por ahora no se pueden programar sentimientos en las computadoras.

Así se puede ver como podría llegar a simular una inteligencia interpersonal,

adaptándose al estado de animo de sus interlocutores, pero lo haría asépticamente, sin poder entender los sentimientos que reconoce en los demás.

1.3. ¿Es posible la creación de una inteligencia artificial?

Una vez vistos algunos conceptos de inteligencia humana y la manera clásica de valorar si una máquina se puede considerar que piensa o no, se pasará a analizar algunas de las aplicaciones en el ámbito de la inteligencia artificial así como alguno de los que piensan que no se va a poder llegar a alcanzar esos objetivos tan optimistas que se tienen, empezando a ver también los nuevos problemas que aparecen en este ámbito de la ciencia y que se revisarán más a fondo en la segunda parte.

Se puede ver como hay una diferencia entre el funcionamiento de la inteligencia, ya sea humana, animal o artificial, y los resultados que pudiera tener una simulación de algún tipo de inteligencia independientemente de su funcionamiento interior. Un ejemplo de esto son los programas que logran derrotar a seres humanos a juegos como el ajedrez (Deep Blue de IBM) o el G*o* (AlphaGo de Google), algo que ya esta superado en inteligencia artificial. No habría mucha diferencia aparente entre saber jugar muy bien, de manera humana, a que una maquina tuviera en su base de datos todos los movimientos posibles que se pueden dar en ese juego, de esta manera podría prever todo posible desenlace y ganar siempre. Para un ser humano esto sería imposible mediante ese método.

El problema con estos algoritmos que saben jugar al ajedrez o al G*o* es que sus experiencias no son extrapolables a otros problemas o tareas, tienen programaciones especificas que aprenden dentro de un marco limitado. Por ejemplo el *Deep Blue* de IBM fue capaz de vencer a G. Kasparov a través de la mera fuerza bruta de calcular muchos posibles escenario por delante del jugador. En el caso del *AlphaGo*, si se cambiara una sola regla del juego el maestro de *Go* sería capaz de adaptarse mucho más rápido que el algoritmo (Hecht, 2018).

Estas máquinas preparadas para una tarea especifica superan a los seres humanos de la misma manera que una calculadora los supera en operaciones de calculo matemático, pero les falta aun el nivel de adaptación para que puedan actuar en situaciones que se

salgan de un marco especifico. Aun así, se pueden encontrar inteligencias artificiales en la vida cotidiana más a menudo de lo que la gente suele pensar.

Los filtros de *spam* del correo electrónico, programas que juegan al ajedrez, algoritmos para publicidad dirigida, coches autónomos, todo esto son aplicaciones concretas y específicas para resolver problemas dentro de un marco relativamente limitado, donde realizan una función correcta y donde cada vez son más eficientes.

Hay voces que dicen que nunca se podrá pasar de ese punto, por ejemplo Miguel Benasayag, filósofo y psicoanalista argentino, critica la idea de que una inteligencia artificial llegue a ser como la humana, debido a la diferencia básica entre las dos que es la vida, y que la inteligencia humana no está comprendida dentro del cerebro, sino que es el conjunto de cuerpo, deseos, errores, etc. De esta manera la diferencia no es cuantitativa, por mucho que avance la ciencia en el cartografiado del cerebro o en el desarrollo de maquinas que aprendan, no se podrá alcanzar a superar esa brecha, que es cualitativa, entre el pensamiento humano y otro tipo de pensamiento creado artificialmente (Benasayag, 2018).

John Searle también fue muy crítico con Turing y con el optimismo por todo lo que se podría alcanzar con el desarrollo de la IA. El mismo propuso la división de inteligencia débil refiriéndose aquella que servia para un propósito específico e inteligencia fuerte para aquella que se adaptaba a diversas situaciones, emulando a la inteligencia humana. Su opinión sobre el tema la dejaba bien clara después de haber definido su propia pregunta:

> ‘¿Puede un computador digital, tal como se ha definido, pensar?’ [...] la respuesta es claramente ‘no’. Y es ‘no’ por la razón que hemos puesto de manifiesto reiteradamente, a saber: el programa de computador está definido de manera puramente sintáctica. Pero pensar es algo más que manipular signos carentes de significado, incluye contenidos semánticos significativos. A estos contenidos semánticos es a lo que nos referiremos mediante ‘significado’ (Searle, 1994: 42).

Por otro lado que para ilustrar su crítica propuso otra situación que servia como respuesta al propio Test de Turing en la que quería indicar que no se demostraba que había inteligencia aunque se pasará el Test, proponiendo que:

> La sintaxis sola nunca es suficiente para conseguir una semántica ni, por consiguiente, para generar una genuina vida mental llena de significado, llena de

> todo tipo de contenidos mentales. La auténtica vida mental en su integridad, es decir, una intencionalidad llena de contenidos, con una plena conciencia cualitativa de los mismos, etc., es un producto causal de nuestros cerebros. La sintaxis por sí sola únicamente podrá ofrecer simulacros de vida mental. Simulaciones que jamás conseguirán producir o duplicar una auténtica vida mental. Y que tampoco servirán para explicar nada (Liz Gutierrez, 2009: 120).

El test que proponía Searle y por el que decía que la sintaxis no implicaba semántica, era el Test de la habitación China, que publicó en 1984, este test era una variación del de Turing en la que una persona con un manual, que tenía indicaciones de como contestar a mensajes en ideogramas chinos, se comunicaba con otro, el juez, que tenía que decidir si estaba ante alguien que sabía chino o no; le pasaba mensajes y el otro contestaba según le indicara el manual. Esto podría dar la apariencia de que sabia chino, la sintaxis, pero en realidad, el sujeto, no tenía ninguna idea de lo que estaba contestando o recibiendo. Con esto quería demostrar como la apariencia de inteligencia (en la comunicación) no indicaba que hubiera rasgos de inteligencia verdadera (Searle, 1994: 37-39).

Se puede apreciar que cuando se empieza a profundizar en la idea del diseño y la posibilidad de éxito de una inteligencia artificial, tanto aquellos más optimistas como aquellos más detractores de esta posibilidad proponen sus argumentos, se tendría que valorar que en lugar de hablar solo de inteligencia se tendría que ampliar el concepto a vida artificial inteligente. Incluso cuando se abre un debate sobre si esto es algo muy peligroso para la raza humana o no, también se está hablando de algo que va más allá de la inteligencia; esta, por si sola, parece no ser más que una calculadora de resolución de problemas independiente de sí misma (sin consciencia) pero cuando personalidades como Elon Musk alarman del peligro del desarrollo de una superinteligencia (Palazuelos, 2017), ademas de ser algo que resuelve problemas, la dotan de características humanas y vitales, como deseos, autoconsciencia y un instinto de conservación que le llevaría a defenderse de la humanidad para no ser terminada y ademas también a consumir recursos para mejorarse en plena competición con el ser humano, de ahí el peligro.

Estos aspectos de vida artificial inteligente se verán en más profundidad en el último

capitulo donde se abordan las implicaciones éticas morales y sociales de la implementación de la inteligencia artificial.

En definitiva, se puede ver como realmente las IA están a nuestro alrededor y cada vez se utilizan en más ámbitos de la vida cotidiana de la sociedad. Esto ocurre tanto en formatos más individuales, de uso personal, pero también en la gestión general de las economías, ya sea en el ámbito privado, por empresas para crecer o mantenerse al día en la era de la información, como en el ámbito público en la gestión de la seguridad y la sociedad como ocurre en china con los putos de ciudadanía y los sistemas de reconocimiento facial.

Aun así este tipo de IA no son los que la cultura popular ha visto en la ciencia ficción, ni sobre los que nos alertan o se ilusionan algunas personalidades de la tecnología. Estas otras IA, aquellas IA fuertes, o generales aun parece que estén lejos de llegar. Pero viendo la velocidad de los avances que se dan día a día no es tan descabellado pensar que sea algo que se puede conseguir, ahora mismo hay inteligencias artificiales que hacen cosas inimaginables para la gente de hace unas décadas, los avances en informática aun están lejos de llegar a su límite y aunque Turing hubiera quedado corto en su previsión de que ya tendríamos una computadora que pasara su test, por otro lado se han conseguido muchos avances que el no hubiera creído posibles.

Para poder hacer un acercamiento a estos avances, en el siguiente capitulo se hará un recorrido a la investigación en inteligencia artificial, así como a los grandes proyectos que se están llevando a cabo en el mundo sobre el funcionamiento del cerebro, también se mostrarán los puntos de vista que se tienen respecto al peligro y las consecuencias que pueden traer las investigaciones en este campo sobre la sociedad humana.

2. Acción reacción de las investigaciones actuales en IA

La investigación sobre la Inteligencia Artificial es un campo relativamente nuevo. Ya a mediados del siglo XX se acuño el termino y se empezó a trabajar en ella, incluso se recibieron fondos desde el gobierno de los Estados Unidos para trabajar en proyectos que se basaban en parte en el trabajo de Alan Turing por la atención que suscitó su articulo sobre el juego de la imitación.

Para ilustrar la importancia de este ámbito de la ciencia en este apartado se hará primero un breve recorrido histórico de la investigación en inteligencia artificial, para continuar mostrando los dos grandes proyectos que se están desarrollando en la actualidad sobre el cerebro humano, algo necesario para poder comprender su funcionamiento y así poder emularlo de alguna manera. Esto pondrá la situación de la investigación en IA dentro de un contexto real actual.

Una vez establecido el contexto práctico actual se pasará a analizar, en los siguientes dos apartados, las reacciones que se dan desde organizaciones interesadas en hacer un acercamiento más o menos cauto a estas investigaciones y también a las posturas de algunos países e instituciones trasnacionales.

Para poner la investigación en IA dentro de un contexto histórico Jean-Gabriel Ganascia (2018) divide la historia de la inteligencia artificial en seis etapas:

Una primera etapa, llamada de los profetas sobre finales de los años 50, donde la atención que recibe es muy alta y se especula mucho sobre lo que se podrá conseguir, quizá precipitadas pero no muy alejadas de la realidad como que serían imbatibles en el ajedrez.

A esta le sigue la etapa de los años sombríos, enmarcada en los años sesenta, donde la ilusión de la primera etapa se resiente al haber pocos avances.

La tercera etapa sería la de la IA semántica, donde la investigación derivó hacia la psicología y la memoria, así a como se representaría semánticamente el conocimiento. Lo que llevo al desarrollo de sistemas expertos en los 70 que dieron sus frutos en aplicaciones durante los 80.

La cuarta etapa de neoconexionismo y aprendizaje automático. Desarrollo de algoritmo de aprendizaje que permitían que los ordenadores se autoprogramaran. El

comienzo del *machine learning* y las aplicaciones industriales más avanzadas.

La quinta etapa de la IA a las interfaces hombre-máquina. A finales de los 90 se añade el reconocimiento de emociones y aparece la informática emocional, perfeccionando los programas que pueden entrar en conversaciones con humanos como los *chatbots*.

Finalmente desde 2010 hay un resurgimiento de la IA, donde los avances en la informática y la gestión de datos, como el *big data*, permiten nuevas técnicas como el aprendizaje profundo y las redes neuronales lo que crea un nuevo optimismo en los logros que se pueden alcanzar en este campo. Pero además de ese optimismo no se pueden dejar de lado los problemas que surgen cuando se habla de un uso de datos tan masivo que puede tender a reducir a los seres humanos a una mera mercancía , donde sus datos personales tienen mucho valor pero el propio individuo ignora el uso que se hace de ellos. Todo esto da lugar a, por un lado, problemas personales debidos a la hiperinmediatez o la hiperlocalización, por ejemplo, y por otro a los problemas de fraudes económicos, violaciones de la privacidad e influencias muy cuestionables sobre la política de algunos países. Para evitar cosas así ya se proponen aproximaciones éticas a todos estos avances, por ejemplo para la universidad, buscando que se de transparencia, reciprocidad, dialogo, inclusión, autonomía, inteligibilidad, proporcionalidad y responsabilidad. Y que estos aspectos se traten desde el diseño hasta la aplicación para evitar los problemas antes mencionados (Ganascia, 2018: 7-8; Calvo, 2018).

Una vez visto un recorrido superficial por la historia de la investigación en inteligencia artificial se puede hacer una aproximación al momento actual y a los proyectos que se están llevando a cabo.

Después de haber desentrañado los misterios de la genética con el proyecto del genoma humano otro de los misterios más grandes a los que se ha enfrentado el ser humano es la explicación de cómo es posible pensar y como funciona realmente el cerebro. Este vacío en el conocimiento de la humanidad es uno de los principales obstáculos a superar cuando se habla del desarrollo de la inteligencia artificial, poco a poco se van haciendo nuevos descubrimientos, pero la magnitud a la que se enfrentan

los científicos (cien billones de neuronas y cien trillones de conexiones) convierte esta tarea en algo de enormes dimensiones, debido a esto tanto entidades públicas como privadas se han decidido a financiar esta investigación.

2.1. Los dos grandes proyectos sobre la mente humana

Dentro del ámbito del desarrollo de la inteligencia artificial se están llevando a cabo muchos proyectos de todo tipo a lo largo del mundo. Pero hay dos en los que se debe poner el foco debido a la gran inversión y a la magnitud en la que se desarrollan, se espera que «sobre los próximos 10 años, el *Human Brain Project* financie sobre los 5000 estudiantes de doctorado» (Kendel et al, 2013).

Se puede ver cómo existe un gran interés que permitirá que se avance en muchos aspectos en la investigación de muchos de los aspectos teóricos y prácticos de la inteligencia artificial. El interés por el estudio de estos proyectos radica en como buscan desentrañar los misterios del funcionamiento del cerebro humano, cada uno desde un enfoque algo diferente. Y esto es muy importante para el desarrollo de una inteligencia artificial que básicamente tratará de emular la mente humana.

Si tomamos la idea de que «todos los fenómenos mentales, ya sean conscientes o inconscientes, visuales o auditivos, dolores, cosquilleos, picazones, pensamientos, toda nuestra vida mental, están efectivamente causados por procesos que acaecen en el cerebro» (Searle, 1994: 22) se puede ver como de importantes son los estudios del cerebro para la investigación en inteligencia artificial.

Se empezará con una aproximación al proyecto europeo, el *Human Brain Proyect*, sus principales características y los diferentes ámbitos en los que se mueve. Luego se verá el proyecto americano, *The BRAIN Iniciative*, que aunque similar difiere en la aproximación que hacen al cerebro humano y tiene un enfoque más directo a la medicina.

Uno de estos grandes proyectos es el *Human Brain Project (HBP)*, uno de los mayores abanderados de la Unión Europea, llamados *Future and Emerging Technologies FET*, el otro gran abanderado es el estudio del grafeno. El *HBP* tiene una proyección de 10 años, que comenzó en 2013 con un billón de euros de fondos y cuyo

objetivo es el estudio del cerebro humano para poder mapearlo y replicarlo informáticamente.

Aprovechando estas infraestructuras el proyecto se divide en seis plataformas: neuroinformática, acceso a los datos compartidos del cerebro; simulación cerebral, replicar la arquitectura cerebral y su actividad en computadoras; analítica y computación de alto rendimiento, proveer las capacidades analíticas y de computación requeridas; informática médica, acceso a datos de pacientes y la identificación de la firma de las enfermedades; computación neuromórfica, desarrollo de una computación inspirada en el cerebro; y por último neurorobótica, para el uso de los robots en las pruebas de simulaciones cerebrales (*Human Brain Project*, n.d.).

Este proyecto, como se ha visto, se basa en la simulación de un cerebro humano, lo lleva a cabo reproduciendo las neuronas y las sinapsis electrónicamente mediante procesadores. La ventaja que se consigue con esto es la velocidad de proceso, que hasta ese momento en simulaciones informáticas se daba unas mil veces más lenta que los procesos biológicos. Otro de los objetivos más interesantes es la comprensión de la cognición humana y cómo se relacionan dentro del cerebro los conceptos a través de la relación del aprendizaje, la memoria, la atención y los comportamientos orientados a los objetivos (*Human Brain Project*, n.d.).

Pero a parte de los objetivos meramente técnicos, hay un factor muy importante que han tenido en cuenta, y este es la parte ética y social del proyecto que está financiado en gran medida por fondos públicos. Ellos mismos ya lo declaran sus intenciones al marcar como objetivos la diseminación de los datos y el conocimiento adquiridos, así como asegurarse de que todo su trabajo se lleva a cabo con responsabilidad y con la finalidad de beneficiar a toda la sociedad.

En su propia página web tienen una sección llamada *live papers* que permite acceder a artículos científicos sobre los trabajos que realizan, además de esto se pueden encontrar multitud de documentos e información detallada de los procesos que llevan a cabo, programas de educación y colaboración así como exposiciones y museos que muestran los logros que consiguen, un ejemplo de estos es el *Mind The Brain*, una exposición lanzada en el museo de Jerusalén este mismo año 2019 (*Human Brain*

Project, n.d.).

El *HBP* hace una aproximación directa hacia la inteligencia artificial, que se basa en que si pueden representar y reproducir enteramente el cerebro humano neurona a neurona y sinapsis a sinapsis, esta reproducción se comportará como un cerebro pensante. Solo poniendo el foco en la cantidad de dinero invertido y de los recursos que se están dedicando a este proyecto se puede ver la importancia que le dan los gobiernos y la sociedad a la mente humana y por extensión al desarrollo de la inteligencia artificial.

Este proyecto además es una gran oportunidad para fundamentar investigaciones relacionadas con el cerebro y sus enfermedades mediante simulaciones y reconstrucciones, esto permitirá ir salvando las distancias que separan los avances en el conocimiento teórico del cerebro de una ciencia práctica que aporte beneficios directos a ala sociedad. Dando pie a un salto entre la neurociencia y la medicina clínica, así como grandes avances en nuevas tecnologías informáticas. Todo esto gracias, sobre todo a la búsqueda de una visión unificada de todos los datos y teorías que se dan en estos ámbitos. Para ello aportan datos y estudios ochenta centros de neurociencia de toda Europa (Kandel et al, 2013: 661).

Otro de los grandes proyectos relacionados con el estudio del cerebro es *The BRAIN (Brain Research through Advancing Innovative Neurotechnologies®) Initiative*, este proyecto, de fondos americanos, impulsado en 2013 por el gobierno estadounidense, donde éste ha inyectado «inicialmente 100 millones de dólares para gastar en el curso del año 2014 […] con unas expectativas de gasto de 3 billones de dólares a 10 años» (Kandel et al, 2013: 663).

Este proyecto tiene como objetivo desarrollar nuevas y más completas herramientas de información para comprender el cerebro humano y sus funciones, tanto en un estado saludable como en uno enfermo. Tiene un enfoque más médico que el del proyecto europeo, por algo el mayor impulsor de este es el Instituto Nacional de Salud (*NIH*, en sus siglas inglesas) estadounidense. Pero al igual que aquel, todo lo que se deriva de la linea central tiene muchas aplicaciones en muy diversos campos de la neurobiología y de la computación aplicada a la inteligencia artificial (*The BRAIN Initiative*, n.d.).

La idea principal de este proyecto es salvar las barreras que existen entre investigadores de distintos ámbitos y las instituciones, creando problemas en el flujo de la información, entorpeciendo los avances. De esta manera se trata de buscar el enfoque necesariamente interdisciplinar que permita afrontar el complicado problema de estudiar el cerebro (Kandel et al, 2013: 660).

Una de las mayores diferencias entre este proyecto y el europeo es que este está enfocado directamente al funcionamiento propio del cerebro, al análisis de las conexiones neurales y a buscar las bases biológicas del funcionamiento de los procesos cerebrales. Todo ello para hacerse con nuevas herramientas que les permita crear un mapa detallado y dinámico que monitorice un cerebro en funcionamiento mediante su actividad bioeléctrica (Bargmann et al, 2014: 6-10).

Este proyecto no muestra tan fácilmente la información y el acceso a datos como lo hace el europeo, pero también define y tiene abierto al público parte de sus resultados y herramientas. Así como dedica un grupo entero de trabajo, el *Neuroethics Working Group*, a los aspectos legales, sociales y éticos de sus investigaciones.

Estos proyectos pese a tener una gran financiación y a permitir la cooperación de muchas organizaciones y centros de estudio, hacen frente a un desafío mucho mayor que el que se dio en el proyecto del Genoma Humano, ya que este último en comparación era un simple proyecto de ingeniería, donde los objetivos estaban perfectamente definidos al igual que sus principios. En cambio estos dos proyectos que están en marcha representa desafíos mucho mayores (Kandel et al, 2013: 661).

Por otra parte los beneficios también serán muy importantes, incluso sin llegar a tener un éxito completo, todo lo que se saque de investigaciones paralelas reportará muchos nuevos conocimientos. El hecho de que a través del Human Brain Project se pueda hacer posible realizar experimentos sobre los procesadores que representarán el cerebro humano (o algunas de sus partes), unos experimentos que serían imposibles sobre el tejido biológico, permitirá grandes avances en el estudio del cerebro (Kandel et al, 2013: 662)

Se puede apreciar que hay un interés y una preocupación en las repercusiones éticas que puede tener la investigación sobre el cerebro o la inteligencia artificial. Cualquier

salto en el progreso científico y tecnológico tiene importantes repercusiones, pero cuando se habla de inteligencia artificial entran en juego aspectos sobre la toma de decisiones y esto afecta de manera muy directa al concepto de libertad y de libre albedrío de los seres humanos, además de a los conceptos de definición de persona y a la cuestión de donde recae la responsabilidad de los actos que se lleven a cabo cuando las decisiones no las toma un ser humano.

Sobre la misma autonomía del ser humano en la toma de decisiones desde una perspectiva cerebral se dan debates filosóficos sobre si la libertad es una ilusión y son procesos más automáticos que luego se intentan razonar[2], etc. estos grandes proyectos también pueden aportar algo de luz, siempre desde la perspectiva cerebral, sobre el tema.

Por todo esto, paralelos a estos grandes proyectos que sirven para ejemplificar la importancia en el desarrollo tecnológico humano que tienen este tipo de investigaciones, aparecen estudios y grupos de personalidades que advierten sobre las consecuencias de obviar los aspectos éticos cuando se aborda el tema de la inteligencia artificial.

2.2. Preocupaciones éticas y reacciones a ellas

Una vez vista la realidad de los grandes proyectos que tratan el cerebro humano y las posibles vertientes aplicables a la tecnología informática, se puede apreciar que dentro de ambos ya existe una preocupación por las implicaciones éticas en este ámbito de la ciencia, se pasará a observar como esta preocupación también se extiende a todo tipo de aplicaciones de la inteligencia artificial en la vida real. No solo en las posibles aplicaciones futuras sino también en las que ya se están intentando poner en marcha.

A continuación, se mostrarán, primero algunas de las instituciones que han aparecido para alentar o alertar sobre este tipo de avances tecnológicos y también algún estudio que se está dando a nivel mundial sobre la moral y la ética alrededor de máquinas que toman decisiones construidas por el ser humano.

Cuando se diseñan máquinas que van a tomar decisiones sin la supervisión de un ser

2 Se hace referencia a los famosos experimentos de Hans Kornhuber y Lüder Deecke donde medían las diferencias entre tiempos cuando se toman decisiones y se mueve en estos casos los dedos (Deecke, Scheid y Kornhuber, 1968).

humano, surgen cuestiones que preocupan, tanto a los que las producen como a los que las van a consumir, y por extensión a todos aquellos que se verán afectados por ellas. En el caso de los vehículos autónomos, debido a lo avanzado y a la inmediatez con la que parece que se quieren poner en funcionamiento en las carreteras, la preocupación afecta prácticamente a toda la población.

El público en general no está dispuesto a aceptar que una maquina le lleve por la carretera sin que ninguna persona tome decisiones durante el trayecto, un ejemplo de la aversión a esto se puede ver entre el público de Estados Unidos, un país donde estos proyectos están más avanzados.

> En abril de 2018, una encuesta de la empresa Gallup encontró que solo el 9% de los adultos estadounidenses querría usar un coche sin conductor en cuanto las regulaciones del gobierno lo certificaran como seguro. Otro 34% decía que ellos esperarían un tiempo desde la implantación antes de subirse a bordo. Mientras que un 52% decía que nunca usarían un vehículo sin conductor (Hetch, 2018: 142).

Por un lado se ve una gran desconfianza en el hecho de perder la responsabilidad sobre las decisiones y las consecuencias al ir en uno de esos coches, pero cada vez que se esta dejando a otra persona conducir, es algo tan asumido y natural. En ese caso no aparecen todos estos reparos o prejuicios y se confía en esa persona que va a tener la vida de los ocupantes del vehículo en sus manos. La gran mayoría de accidentes en la carretera son debidos a errores humanos, pero aun así la desconfianza en que la máquina decida es generalizada y el debate inevitable.

Pero no solo es una cuestión de confianza, también hay una sensación de perdida de identidad. La oposición a este tipo de vehículos se organiza también en asociaciones para preservar el derecho humano a conducir. Esto se debe a que si se implantara para que la seguridad en la carretera fuera mayor, esta debería estar reservada a los vehículos autónomos, algo que es muy posible que fuera instaurándose hasta desaparecer el conductor humano. De aquí que hayan aparecido asociaciones como *Human Driving Association* o la campaña *Save Driving*[3]. Que atraen a miles de amantes de la conducción (Hetch, 2018: 143).

Esta preocupación y la necesidad del público a conocer cómo deciden estos coches

3 Juego de palabras que confunde “conducción segura” con “salvar a la conducción”.

autónomos, así como de los ingenieros a aplicar los algoritmos para que sean seguros, eficientes y generen confianza entre la gente, ha dado pie a un proyecto mundial impulsado desde el Instituto Planck para el desarrollo humano. El proyecto se llama *Moral Machine*, dio comienzo en el año 2015 para ver cómo se debería responder a los dilemas que pueden surgir durante el funcionamiento de los coches autónomos cuando la IA que los conduce tuviera que tomar decisiones de vida o muerte que afecten al conductor y a otros viandantes.

El estudio es plenamente interactivo y a través de la web del proyecto cualquiera, en casi cualquier parte del mundo, puede participar mediante un juego de elecciones sobre la decisión que debería hacer el vehículo autónomo cuando se enfrenta a dilemas en los que alguien podría acabar muerto (Scalable Cooperation Group, n.d.).

El objetivo que tienen es formar una imagen de la opinión de muchos seres humanos a lo largo del mundo y así dar pie al debate sobre escenarios que para resolverse dan lugar a consecuencias morales.

Esta forma de tratar este asunto desde un punto de vista de dilema moral se ve muy ejemplificado en el dilema del tranvía, su forma moderna la definió Philippa Foot en su articulo *The Problem of Abortion and the Doctrine of the Double Effect* (1967). Con la expresión de doble efecto se refiera a los dos efectos que puede producir una acción, el efecto que se busca y el efecto que se puede prever pero no era buscado. En este articulo entre muchos dilemas en los que debía elegir una solución con más o menos victimas mortales e incluso se dan elecciones cualitativas donde se sopesa la culpabilidad o inocencia de las victimas, aparece el famoso dilema donde un conductor que lleva un tranvía descontrolado debe elegir entre una vía donde hay un trabajador o una vía donde hay cinco trabajadores. El punto sobre el que Foot quiere incidir es en como se diferencia que se prevea una muerte como causa de la dirección que se toma debido a una elección y otra muy distinta que se busque la muerte de alguien como parte del plan en curso.

Este tipo de dilemas éticos no permiten una aproximación exhaustiva y formal al estudio de una moralidad general de la humanidad, ni responderán a si hay una solución universal a esos mismos dilemas. La realidad es mucho más compleja que una respuesta

dicotómica a un dilema moral.

Aun así se han llevado a cabo muchos estudios como por ejemplo los de Marc D. Hauser donde se ve la diferencia de reacción frente a dilemas personales o impersonales, donde se aprecia como se tiende a cuidar de los individuos que sentimos más cercanos que de aquellos que están más lejanos, pero si se simplificara así tal cual no sería posible criticar el nepotismo y la corrupción. La ética no puede quedarse ahí, la ética tiene que servir para recordar que podamos cuidar de los cercanos pero sabiendo que tenemos la capacidad de llegar a los que están más lejos (Cortina, 2013: 68-72).

Estos dilemas y su estudio también han servido para recalcar la dificultad que tendría la aplicación de una lógica formal a la toma de decisiones de un coche autónomo debido al contexto cultural.

Esta web permite el acceso a los resultados divididos por países, donde se pueden comparar dos de ellos mientras se ve la media mundial. Este estudio recoge mas de 40 millones de muestras y se encuentra en 233 países. Pese a lo simple y limitado que pueda parecer, los resultados revelan las grandes diferencias que se dan entre los países, mostrando como cambian ciertos valores entre unos y otros y lo que en un sitio se podría aceptar que estuviera implementado en el vehículo autónomo, en otros sería inaceptable.

Además de este tipo de proyectos que buscan reflexionar sobre las implicaciones éticas que se dan en la investigación sobre máquinas autónomas e inteligencias artificiales, también aparecen organizaciones que buscan promover y dar buena imagen a los progresos tecnológicos más punteros y también otras que piden precaución o por lo menos actuar con cautela y valorando todas las consecuencias que se pueden dar.

En el primer caso se encontraría la *Singularity University* que se definen a si mismos como «una comunidad de aprendizaje global e innovación que utiliza tecnologías exponenciales para enfrentarse a los mayores desafíos globales y construir un mejor futuro para todos.» (Singularity University, n.d.)

Mientras que estos parten de una visión más optimista sobre los beneficios que los grandes saltos tecnológicos pueden proporcionar a la humanidad, se pueden encontrar a otros que, como Nicholas Agar, piensan que las soluciones al control de las IA serán

dependientes del desarrollo de las propias IA, ya que no podemos comprender aun lo que implica una IA verdadera, de manera que no se debe crear una excesiva preocupación por lo que conlleva el desarrollo de una superinteligencia, es curioso en cambio como Agar alerta de los peligros apocalípticos de los humanos mejorados (Agar, 2016).

Pero por otra parte también han aparecido voces que alertan de lo peligroso que puede ser la falta de control en el desarrollo de este tipo de tecnologías. En la imaginería popular y en las ideas que la cultura del ocio ha creado en la mente de la gente a través del cine y la literatura, se ha ido formando un relato de como la creación de inteligencias artificiales puede llevar a la extinción del ser humano[4].

A raíz de este temor una de las organizaciones que más relevancia a adquirido en la opinión pública es el instituto *Future of Life (FLI* también por sus siglas*)*, que nace según sus propias palabras con la misión de «catalizar y apoyar investigaciones e iniciativas para la salvaguarda de la vida y el desarrollo de visiones optimistas del futuro, incluyendo vías positivas para que la humanidad dirija su propio curso considerando las nuevas tecnologías y desafíos» (Future of Life Institute, n.d.).

Aunque a través de muchos de los medios de comunicación se les ha tratado de alarmistas, ya que suelen tender a asociar las declaraciones de sus miembros más famosos como Elon Musk o Stephen Hawkings con la idea de que se generará una superinteligencia que, compitiendo con los humanos por los recursos, los aniquile.

Esta idea de que la IA se aumente a sí misma y supere a los humanos la ha desarrollado bastante el filósofo Nicholas Bostrom, advirtiendo que una superinteligencia (artificial) fallara y por una mala programación en sus objetivos destruyera a la humanidad o la sojuzgara (Bostrom, 2011) . Él advierte de que de las dos maneras de alcanzar una superinteligencia, que deje tan atrás a lo que ahora se considera inteligencia, se puede dar de dos formas, a través de la IA y a través de la mejora del ser humano por la biotecnología. En estos temas tiene la visión totalmente contraria a la que se ha visto antes de Nicholas Agar con el que debate a través de artículos y respuestas a

4 Películas como *Terminator* o series como *Battlestar Galactica*, sin contar las innumerables novelas sobre el tema, donde normalmente por el mal uso, o abuso, por parte de los humanos, la IA se convierte en el enemigo de la humanidad.

estos.

Pero Bostrom opina que el camino de la IA para llegar a la superinteligencia es muy peligroso, ya que esta tiene enormes ventajas sobre la inteligencia biológica, principalmente lo que no tiene son las limitaciones a las que llegaría la otra, en cuestión de edición, duplicabilidad y expansión de memoria o proceso de datos. Sobre todo alerta del peligro de una explosión de inteligencia que haga que se salga del control humano sin tiempo para una respuesta (Brundage, 2015).

Esta visión apocalíptica que los muchos medios asocian al Instituto *Future of Life* no se corresponde tanto con el trabajo que tratan de hacer. Sus preocupaciones surgieron del hecho de como se ha empezado a aplicar toda esta tecnología para la industria armamentística y la guerra con los peligros que esto conlleva.

Basándose en un informe de *PAX*[5] llama la atención del publico sobre como se está financiando de manera millonaria la investigación de aplicaciones de IA para asuntos militares en siete grandes países como son: Estados Unidos, China, Rusia, Reino Unido, Francia, Israel y Corea del Sur.

El mayor problema sobre el que alerta el instituto es el del desarrollo de sistemas autónomos de armamento, como el proyecto *ATLAS* (*Advanced Targeting and Lethality Automated System*) americano. Pero quizá lo más importante de este informe es la posición que toman estos países frente a la preocupación de organizaciones independientes o de instituciones como la ONU, como se verá en el apartado siguiente.

Partiendo de esta preocupación principal el *FLI* fue ampliando el espectro de posibilidades y de consecuencias que tiene el desarrollo de esta tecnología. Por ello organizaron una conferencia[6] de la que surgieron los 23 *Asilomar AI Principles.* A parte de la preocupación por que las IA puedan salirse del control humano, una de las principales cuestiones que consideran relevantes es la transparencia y la importancia que se le tiene que dar a este tema para que no se recorte en medidas de seguridad y de control, que son las primeras cosas que se dejan de lado cuando países o compañías

5 Esta es una organización independiente donde se encuentran numerosas organizaciones sociales e iglesias que trabaja en pos de la dignidad humana y la solidaridad con activistas por la paz y víctimas de la violencia de la guerra.

6 En esta sección de la web de Future of Life Institute: https://futureoflife.org/bai-2017/ se pude encontrar la información adicional y específica de ella y videos de los conferenciantes.

compiten por ser los primeros en conseguir resultados, para el *FLI* esto es uno de los grandes peligros. Aun así el pilar en el que se sustentan estos principios (se pueden consultar en el anexo) es el de que esta tecnología debe beneficiar a toda la humanidad no solo a los que primero la consigan y que se tiene que llegar a acuerdos que hagan cumplir esto.

Estos principios no son inamovibles, el mismo instituto alienta al debate y a la discusión para que se puedan limar fallos o llamar la atención sobre los problemas que pueden derivar de estos, como los que señala Patrick Lin respecto al principio del beneficio compartido y de los problemas que esto puede implicar por que puede tener las mismas contrapartidas que el consecuencialismo o que el utilitarismo y servir como excusa para hacer daño a unos pocos por ampliar el beneficio de muchos (Conn, 2018).

2.3. Posturas de instituciones nacionales y trasnacionales

Se acaba de ver en el apartado anterior como ciertas organizaciones privadas se posicionan respecto a las consecuencias que pueden aparecer con el desarrollo de sistemas de inteligencia artificial. Ahora se abordarán algunas de las posiciones que adoptan ciertos países y después también las directrices y consejos que proponen instituciones trasnacionales como la UNESCO o la Unión Europea.

Ante esto hay variedad de posiciones entre los países más poderosos del mundo, ya que una de las reclamaciones de las organizaciones que se han visto en el apartado anterior y que abogan por la paz es la prohibición de estos sistemas autónomos de armamento. Estados Unidos alega que es pronto y que no va a entrar en debates. China sí desea negociar y llegar a una conclusión, pero mientras tanto no hace más que avanzar en su desarrollo. Como Israel que no quiere que unas normas rígidas detengan los beneficios de su desarrollo. Rusia dice que no hay que olvidarse de los beneficios que pueden aportar estos sistemas y que se fijará en las normas legales internacionales de la actualidad.

Por otro lado, el Reino Unido sí que se posiciona en que la última palabra para cualquier sistema armamentístico tiene que estar en manos de un humano. Francia tiene una posición similar donde el uso de fuerza letal debe estar supervisado por humanos.

Mientras Corea del Sur tampoco quería que las discusiones sobre esto hicieran que se resintiera el desarrollo en el área civil y se preocupaban de las implicaciones en el área militar, actualmente han ampliado su preocupación por estas limitaciones también a su uso en defensa (Gronlund, 2019).

Se puede apreciar a través de este informe que hay una carrera internacional en el desarrollo de la IA aplicada a la guerra y que ningún país quiere quedarse atrás en la investigación debido a las grandes ventajas que les puede aportar en futuros conflictos y como urge que se lleguen a acuerdos internacionales, ya que la situación es tremendamente similar a la que se daba el siglo pasado en el desarrollo de armamento nuclear. Una diferencia que preocupa a los miembros de este instituto y que debería preocupar a toda la comunidad internacional es el hecho del mismo concepto que se está debatiendo, su autonomía del factor humano más allá de su creación.

Como se verá en la siguiente sección la responsabilidad es un factor clave. Aún hay quien defiende el argumento de que solo se seguían ordenes cuando se lleva a cabo una acción poco ética o directamente contraria los derechos humanos, lo que lleva a intuir que no se debería dejar que la vida humana estuviera en manos de una máquina que decide sin contar con el factor humano, sobre todo, como pasa con los asuntos militares, cuando el desarrollo y la implementación son totalmente opacas. La vida humana no puede depender de situaciones en las que la responsabilidad se ha diluido tanto que sería inviable saber con qué criterios se ha llevado una operación que pudiera ser considerada un crimen de guerra o contra la humanidad, y pudiera escudarse siempre en problemas técnicos.

Aun así, y pese a la divergencia de posiciones, en estos países, que tienden a buscar aumentar su ventaja sobre los demás en las investigaciones de este ámbito algunas instituciones como la UNESCO y la Unión Europea también publican directrices e indicaciones para los países que trabajan en ellas, para buscar puntos en común y objetivos con unas miras en beneficios comunes.

Este mismo año 2019, La Comisión Europea a publicado una comunicación al Parlamento Europeo sobre la IA, donde habla de los beneficios que puede traer su desarrollo pero también de los retos a los que se enfrentará la sociedad debido a su uso.

El interés que se muestra en el comunicado indica también la falta de información y coordinación dentro de la UE sobre este tema, ya que insta a los países miembros a que se de un plan coordinado e indica que se tienen que destinar grandes inversiones de dinero para que «prepararse para las transformaciones socioeconómicas y garantizar el establecimiento de un marco ético y jurídico apropiado» (Comisión Europea, 2019: 2).

En este documento se definen un grupo de directrices y de requisitos que se deben cumplir en el desarrollo de aplicaciones de IA fiables elaboradas por un grupo de expertos en el tema (Comisión Europea, 2018). El requisito principal en el que se basan es el de generar confianza fundamentándose en los valores de la Unión Europea, «respeto de la dignidad humana, la libertad, la democracia, la igualdad, el Estado de Derecho y el respeto de los derechos humanos» (Comisión Europea, 2019: 3).

Organizados de manera distinta, estos principios son muy similares a los que el *FLI* define en los de *Asiomar*, una diferencia es que los de UE valoran más la legalidad vigente y se fijan en que se cumplan cosas más concretas como la protección de datos. Por otra parte coinciden en gran medida en casos como la transparencia, de vital importancia, y la necesidad de que siempre exista la posibilidad de auditar los procesos. Así como valorar en toda medida los posibles impactos negativos de la implantación de la tecnología.

Parece ser que la preocupación de la UE se centra más en aspectos técnicos y en evitar impactos negativos, valorando que la dimensión ética se tenga en cuenta en todos los niveles. Pero pese a hablar de que «la Comisión seguirá esforzándose por llevar el enfoque de la Unión a la escena mundial y establecer un consenso sobre una IA centrada en el ser humano» (Comisión Europea, 2019: 9). No le da importancia al hecho de que se tenga que lograr un beneficio lo más amplio posible y que abarque a toda la sociedad como hace el *FLI* o la UNESCO como se verá a continuación.

La UNESCO como institución también a puesto su mirada en la IA, dedicando un numero entero de su revista el Correo de la UNESCO a este tema o impulsando la primera conferencia mundial para promover un enfoque humanista a la Inteligencia Artificial. Esta institución se ha centrado más en el aspecto de sus aplicaciones para la educación, dentro de la agenda 2030 para la educación global han publicado el

documento *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development* (UNESCO, 2019) en él se valoran muchos de los aspectos que se deben tener en cuenta a la hora de desarrollar IA dedicadas a la educación y cómo esta tecnología debe llegar al máximo número de personas posibles, sin dejar de lado a aquellos con menos oportunidades y medios.

De esta manera es necesario hacer un acercamiento humanista para el uso de la IA en la educación, controlada por personas y centrada en las personas. Hacer que la implementación de todo esto sea más equitativa, inclusiva, abierta y personalizada, para que nadie quede fuera su alcance. Evitando que al excluir a algún colectivo de todos estos avances se tienda a una homogeneización social mayor y la exclusión de dichos colectivos aumente.

Hacen hincapié en que la brecha digital no debe hacer mella en la implantación de estas medidas, teniendo en cuenta la edad, el genero, cualquier tipo de discapacidad, el estatus social o económico y que se preparen las condiciones para que los países que no tienen los medios para implantarlo puedan hacerlo (UNESCO, 2019).

Viendo todas estas posturas se puede apreciar como hay un gran interés en el desarrollo de todo tipo de investigaciones sobre IA y en sus aplicaciones. Además se da a una escala global, organizaciones e instituciones reclaman que se empiece a llegar a consensos internacionales sobre los limites y regulaciones que se deben dar en estás lineas de investigación debido a su importancia y a su impacto en la todas las sociedades. Asumiendo estas la importancia que tienen todos estos avances y que radica en cómo la implantación y el desarrollo de este campo va cambiando la sociedad en todos sus aspectos y en como continuará haciéndolo cuanto más se desarrolle.

Ya se ha hablado de las posibilidades y repercusiones de una Inteligencia Artificial general o sencillamente de IA débiles aplicadas a cualquier aspecto de la sociedad, pero cuanto más avanza la investigación y la autonomía de los sistemas aumenta, aparece un factor muy relevante a tener en cuenta. No es la misma Inteligencia Artificial, ya sea como especulación futura (la IA fuerte) o como algo cotidiano como son las IA que nos rodean, el asunto que cada vez tiene mas importancia es la posibilidad de crear una Moralidad Artificial, si esto se puede dar y si es comparable a la creación de una IA, así

como la manera en la que condicionará la vida en las sociedades humanas, la aceptación o el rechazo debido a la confianza que genere, y también el hecho de como la misma moralidad humana se pondrá a prueba al intentar aplicarla a las máquinas.

3. La Moralidad Artificial, ética, valores y derechos

Se ha hecho un recorrido a lo largo de este trabajo al contexto que rodea la investigación en el campo de la inteligencia artificial. Se ha partido de valorar si es posible la creación de una inteligencia artificial, partiendo de conceptos diferentes de inteligencia como el clásico o como el de las inteligencias múltiples de Gardner. A continuación se ha revisado el contexto práctico en el que se mueve la investigación, a través de los grandes proyectos en marcha, para pasar a ver diferentes posturas que se generan en torno a todo esto, se ve pues como la sola posibilidad de que se llegue a crear una inteligencia artificial genera grandes preocupaciones a todos los niveles de la sociedad, tanto públicos como privados, pero resaltando la excepción que se da en la rama militar de los gobiernos.

Se puede observar como esta preocupación ya aparecía en la literatura de ciencia ficción desde antes de que se acuñara el término Inteligencia Artificial, Isaac Asimov en sus novelas de robots establecía las tres leyes de la robótica:

> –Bueno, veamos, empecemos por las tres reglas fundamentales de la robótica; las tres reglas enterradas en lo más hondo del cerebro positrónico de cualquier robot. –En la oscuridad, fue marcando cada punto con sus dedos enguantados –. Tenemos: uno, un robot no debe dañar a un ser humano ni, por inacción, permitir que un ser humano sufra daño.
> –¡Correcto!
> –Dos –continuó Powell–, un robot debe obedecer las órdenes que le sean dadas por un ser humano, salvo cuando dichas órdenes contravengan la Primera Ley.
> –¡Correcto!
> –Y tres, un robot debe proteger su propia existencia, siempre y cuando dicha protección no contravenga ni la Primera ni la Segunda Ley (Asimov, 2019: 192).

Con esta escena Asimov plasmaba en 1942 la preocupación que podría tener la sociedad frente a robots que pudieran tomar decisiones por si mismos sin esperar una confirmación humana. Eran unas leyes sencillas que a parte de velar por la seguridad de las personas, se preocupaba de que los robots que se fabricaran gozaran de la confianza del público, algo sobre lo que el autor ya puso el foco de atención ya que era casi tan importante la seguridad como que la sociedad aceptara la existencia de robots en su seno.

Estas leyes limitaban la actuación de esos constructos autónomos de manera que

fueran seguros para su interacción con los seres humanos, una preocupación que llega hasta el día de hoy. Es un aspecto del estudio de la inteligencia artificial sobre el que llaman la atención, sobre todo aquellos, como el *FLI*, que buscan que se tenga una actitud más cauta y crítica en la aplicación de estas tecnologías.

Durante este parte en el primer apartado se recorrerá el paso de la búsqueda de una seguridad para que las aplicaciones de la inteligencia artificial no sean peligrosas, en un principio, para los usuarios, y, en general para la sociedad; como esta preocupación hace que solucionarlo sea más complicado de lo que parece e incluso de pie a un nuevo campo académico. Esto dará lugar a un debate que vuelve a la idea de si una IA ha de ser inteligente en esencia o basta con que lo parezca, esta vez el debata se centrará en la moralidad, donde algunos optarán por la opción práctica de que basta con que la máquina actúe de una manera moral de cara a una percepción exterior o a la opción por la que abogan otros donde esta vez es necesario que la máquina sea verdaderamente un agente moral.

En el segundo apartado se verá como la situación se complica aun más al ir añadiendo características a estas creaciones humanas para así convertirlas en agentes morales. Al añadirles estas nuevas características la visión de éstas IA y el trato que se les debería dar en la sociedad cambia de una manera en la que puede que ya no deban ser tratadas como meras herramientas si se lleva hasta el final el objetivo de crear máquinas verdaderamente autónomas. Con todo esto, se aproximará al objetivo de mostrar la necesidad de que todas las investigaciones en este ámbito de la ciencia tengan que tener en cuenta a las ciencias humanas y sociales para poder responder a los desafíos que la aplicación de la inteligencia artificial en la sociedad supone.

3.1. De la seguridad a la moralidad artificial

Cuanto más se profundiza en el estudio del impacto de las nuevas tecnologías en la sociedad se tiende a valorar los aspectos éticos de esa tecnología, desde su origen hasta su aplicación, de manera que se va ampliando la preocupación por la simple seguridad hasta llegar a una preocupación por cualquier impacto ético que se produzca sobre la sociedad, en general o por lo menos, sobre todos aquellos que se vean afectados por esa

tecnología.

Si se parte de las tres simples leyes de la robótica de Asimov se puede ver enseguida como están muy limitadas y pueden llevar a situaciones paradójicas o en las que el robot se quedaría colgado. Asimov ya asumía estas limitaciones en los cuentos sobre robots que escribía, mostrando las paradojas y problemas que surgían al aplicar este tipo de reglas a una máquina. Incluso hizo que uno de esos robots redefiniera su propia esencia creando una nueva ley para poder enfrentarse a situaciones más complejas y que así pudiera tomar una decisión.

La Ley Cero de la robótica, «Hay una ley que es superior a la primera ley. "Un robot no puede lastimar a la humanidad o, por falta de acción, permitir que la humanidad sufra daños." La considero ahora la ley Zeroth de la robótica» (Asimov, 1992: 322).

Aun así se puede ver como, si se valora esta Ley como se pudiera valorar el utilitarismo de John Stuart Mill, al interpretarlo como la mayor felicidad (o beneficio) para el mayor numero de personas (Mill, 2002), tendría el mismo problema de poder justificar atrocidades por un bien mayor. De hecho en la misma novela ponen en tela de juicio esta ley por el mismo motivo. Así se puede ver como no es sencilla la tarea de dotar a una inteligencia artificial con unas medidas de seguridad aceptables y claras que se admitan en la sociedad actual y que también, acorde a los estándares actuales, entren dentro de unos límites éticos o morales admisibles.

Existe de por si un sentido de la responsabilidad sobre las creaciones humanas, o, por lo menos, debería existir. De este sentido de la responsabilidad surge una necesidad de dotar de una moralidad, yendo más allá de unos simples mecanismos de seguridad, a las máquinas autónomas. Esto se debe a que si se trata de que las máquinas actúen sin supervisión, para poder interactuar con los humanos en la sociedad se les ha de dotar de algún mecanismo que se lo permita hacer de una manera correcta para con el resto. Por ello en la actualidad el foco del problema ya no se centra en la inteligencia artificial de por sí, algo que pasa a formar parte más de los aspectos técnicos sino en esos mecanismos de moralidad.

Sumado a los grandes avances que se están dando en la actualidad y a la cantidad de aplicaciones que ya funcionan mediante sistemas autónomos junto a los que están por

venir, está haciendo:

> emerger un nuevo campo académico conocido como "moralidad artificial", "ética artificial" o "ética de las máquinas", una disciplina que entrecruza la inteligencia artificial, la robótica, la ética y la filosofía e intenta implementar los juicios morales humanos con la acción moral en sistemas artificiales (Kukita, 2015: 27).

Con la aparición de este nuevo campo, el debate que se daba en un principio sobre la inteligencia: si bastaba con simularla para que los resultados sean parecidos a los que daría una inteligencia genuina o, si es necesario que intrínsecamente tenga que ser también como la inteligencia humana, se amplia a los aspectos morales que se deriven de estas IA. De manera que ahora la cuestión es si bastará con que un sistema autónomo parezca moral para serlo o hace falta algo más. En torno a este tema se han ido posicionando filósofos y científicos de todas clases, de los que se mostrarán algunas de sus posturas más relevantes a continuación.

Siguiendo la linea que marcaba Asimov, científicos computacionales como Mathias Scheutz o Selmer Bringsjord proponen crear modelos lógicos para resolver cuestiones éticas, asociando la ética a la sola razón. Junto a Susan Leigh Anderson asumen que la ética puede ser computable y que su trabajo es conseguir un programa que pueda resolver dilemas éticos. Anderson desarrollo una IA ética llamada MedEthEx una máquina con autoaprendizaje que podía enfrentarse a situaciones desconocidas (Kukita, 2015: 29-30).

Se puede apreciar como este tipo de modelos puede caer en grandes contradicciones, por seguir un proceso de pura lógica, como podría ocurrir con las leyes de la robótica, unas meras reglas que se tengan que cumplir para la toma de decisiones. Pese a que estos nuevos modelos sean más completos y complejos, el mayor problema es el que se ve con todo dilema ético, y es que ciertas respuestas pueden no ser correctas en según qué culturas como demuestran las grandes diferencias en los resultados de *The Moral Machine* que se han visto anteriormente.

Estos autores que intentan convertir la moral en algo programable mediante razonamientos lógicos o cálculo «intentan justificar esta reducción de la moralidad indicando que su objetivo es hacer maquinas con un uso práctico, no agentes artificiales moralmente completos.»(Kukita, 2015: 31) Pero hay muchos factores que indican que

esto no es tan sencillo de conseguir: como por ejemplo pueden ser las emociones o la confianza, que muchas veces va de la mano de esas emociones. Como se puede comprobar en encuestas o análisis de mercado donde el público suele preferir un conductor humano frente a una máquina pese a que todas las pruebas que el factor humano es la mayor causa de accidentes en la carretera, por ejemplo.

Estos autores dividen la aproximación a la moral artificial de dos formas, una aproximación de arriba a abajo y otra de abajo a arriba. La primera es aquella que se basa en reglas y que sigue las teorías y principios para los procedimientos de decisión que parten de ideales religiosos, códigos morales o sistemas filosóficos como La Regla de Oro, el utilitarismo la deontología kantiana, etc. Sería lago parecido, aunque más elaborado, a las tres leyes de la robótica de Asimov. Esta manera de aplicar la moral artificial la consideran poco robusta y muy dada a que se creen conflictos entre las reglas. El problema común que acarrea este tipo de aplicaciones de la moral es la gran dificultad en programar prácticamente como reunir y comparar toda la información necesaria para que se aplique a tiempo real (Wallach, Smit y Allen, 2005: 152).

Para Wendell Allan y Colin Allen no importa si la máquina es moral en esencia o no, lo único relevante es si los resultados que produce son morales o no. Para ello diseñan un test similar al de Turing, donde lo que el juez valorará es si las respuestas que da la maquina a ciertas situaciones o desafíos éticos se pueden considerar morales o no (Kukita, 2015: 30).

Esto podría pecar del mismo error de perspectiva que S. L. Anderson ya que dependiendo de la cultura del juez se podría dar una valoración u otra. Por otro lado existe la dificultad que se daría juzgando a un ser humano real, algo que tampoco daría unos resultados aceptados universalmente. Hay que tener en cuenta también que más que intentar recrear una moral humana de manera artificial, donde se tendría que asumir lo falibles que son los humanos, se está buscando una maquina que se comportara como el ideal de humano. Solo esta idea ya parece inalcanzable cuando es difícil asumir unos valores de humanidad universales y perfectos que pudieran ser aceptados por todos.

La gestión de la moralidad de las IA como una serie sistemas artificiales basados en reglas se ha rebatido como muy insuficiente y quedan muy limitados a tareas de

«inferencia lógica y calculo, pero tienen dificultades con tareas como reconocimiento de imagen o navegación en situaciones del mundo real» (Kukita, 2015: 28).

Por otra parte, la aproximación de abajo-arriba sería aquella que, en lugar de usar reglas o imponer un sistema moral específico, lo que hace es, en un ambiente especifico, premiar el comportamiento que se quiere desarrollar, mediante una evolución de los mecanismos de elección con resultados de éxito y error. Esto sería como la emulación de la educación moral de un niño dentro de un contexto social. Este sistema también tiene sus problemas, entre ellos la lentitud del desarrollo. Mas allá de los desafíos que supone, esta aproximación da el salto de la inteligencia artificial a la vida artificial, llevando a cabo experimentos de evolución de algoritmos artificiales como los que se llevan a cabo en la plataforma *Alife* (Wallach, Smit y Allen, 2005: 153).

A parte de estas dos aproximaciones, la evolución lógica para evitar sus carencias son los sistemas híbridos de ambas de manera que las investigaciones se aproximen más al objetivo de conseguir una expansión de las elecciones y un comportamiento flexible (Wallach, Smit y Allen, 2005: 154).

Parece que hace falta algo más que ese conjunto de sistemas de reglas o de evolución educativa, ya que entran en juego una serie de factores que hacen que sea necesario dotar a las máquinas autónomas de otras características para poder cumplir con la complejidad que implica el factor moral artificial, como se verá a continuación.

3.2. De la confianza y responsabilidad a una nueva visión de las IA

En este apartado se tratara de valorar la complejidad del diseño de una IA moral, así como de la complejidad de las relaciones de las IA con los seres humanos dentro de la sociedad, haciendo hincapié en la necesidad de ínterdisciplinariedad cuando se trata de investigaciones en el campo de la realidad virtual y de como se tienen que revalorizar las ciencias humanas y sociales dentro de estas investigaciones, ya que pueden aportar mucho y ayudar a responder a los desafíos a los que se enfrentan en ese campo.

Como ya se veía aparecer hace décadas, un factor importante en la integración de IA en la sociedad es el de la confianza. Este aspecto es tan importante como para evitar que se acepte nada parecido a una máquina autónoma entre las personas si no esta dotada de

algún mecanismo que permita que las personas puedan confiar en dicha máquina, como se ve que sucede con los coche autónomos.

> Ya en su tiempo, pensando precisamente en la aplicación de la ética, Kant va mucho más allá y nos ofrece una definición positiva del principio de publicidad donde se busca el consentimiento o acuerdo sobre los resultados alcanzados. Hoy en día podemos interpretar está necesidad de la publicidad desde el seguimiento y control de todos los grupos de interés, esto es, desde su inclusión en un diseño institucional que responda de los esfuerzos realizados por la empresa para acercarse a la participación y al libre acuerdo (Garcia-Marza, 2014: 252).

El principio de publicidad como aspecto para ganar confianza ya se ha mencionado en apartados anteriores de mano de instituciones como la Comisión Europea cuando recalcan la importancia de la transparencia en los procesos de diseño y aplicación de IA. Algo que, como también se ha visto anteriormente en la primera y segunda parte, han dejado de lado muchas empresas que se dedican a implantar aplicaciones de IA en el mercado y que ha llevado a escándalos por las consecuencias que han tenido sobre algunos sectores de la sociedad.

Cuando se habla de ser moral o solo parecerlo, el aspecto de la confianza adquiere aun más relevancia y decanta la balanza sobre la necesidad de ser moral. Es más aceptable para las personas admitir a alguien que parezca inteligente, lo sea o no, que a alguien que parezca que tiene actitudes éticas pero que se sospeche que sus intenciones no lo son. Si se añade a esto el factor de la responsabilidad aun se agrava más la diferencia entre seres humanos y máquinas.

El mismo Asimov en sus relatos ya tenía en cuenta este aspecto, sus tres leyes de la robótica eran intrínsecas a la ingeniería del soporte físico de la IA de sus robots, no se podían modificar y sin ellas no podían funcionar, esto hacía que las personas confiaran en cualquier clase de robot ya que sabían que no estarían en marcha si esas leyes no funcionaran.

Cuando estos aspectos éticos no se tienen en cuenta suceden cosas como con el algoritmo de contratación de *Amazon,* que discriminaba a las mujeres solo por el hecho de serlo, de manera que perdió la confianza de los ejecutivos de la compañía debido a la mala publicidad que les estaba generando al perder la confianza del público. Esto se debía a que tomaba como fuente de aprendizaje la manera de contratación que había

llevado *Amazon* durante los anteriores diez años (Rubio, 2018).

Esto fortalece la idea de lo difícil que es implementar algo así cuando las personas que lo hacen están trabajando con sesgos culturales, de género, etc. y valores que raramente se acercan al ideal que se les está pidiendo a las máquinas. También refuerza la idea de que la ética debe estar presente en todos los pasos y procesos, no solo en el producto final, ya sea un algoritmo, una IA o un robot.

Se puede retroceder al capítulo anterior y recordar que la cuestión principal que recalca la Unión Europea respecto a la IA es la de la confianza y que los ciudadanos puedas depositar la suya en estas tecnologías.

A la hora de implementar la moralidad en máquinas autónomas se tienen que tener en cuenta esos factores anteriormente mencionados, teniendo en cuenta también que las decisiones éticas no son solo aquellas que surgen de dilemas éticos sino que están muy relacionadas en las relaciones sociales y en como las personas se tratan unas a otras, algo que es dinámico, muy cambiante según el contexto espacio-temporal (Kukita, 2016: 32).

Para que se pueda dar una confianza en este producto final tiene que estar clara la cuestión de en quien o en que se deposita la responsabilidad de las acciones de esa máquina autónoma. Este aspecto, el de la responsabilidad es también muy importante ya que como sucede con cualquier organización, se pude hablar de la *responsabilidad social corporativa* (o RSC) ya que al final la implantación de muchas de estas aplicaciones sobre las que se ha hablado a lo largo del texto ha recaído en empresas privadas que, por lo que parece, no han tenido demasiado en cuenta los aspectos éticos de sus acciones. De esta manera se pude ver como «la perspectiva de la ética empresarial nos permitirá apreciar el valor de la RSC como un recurso moral, así como diferenciar entre una gestión estratégica y una gestión ética dela RSC» (García-Marzá, 2014: 240). sin profundizar demasiado en estos conceptos lo que si se puede intentar apreciar es como es muy necesaria esa aproximación ética a la responsabilidad y no solo su uso como un factor más de marketing, repitiéndolo de nuevo, se ha visto las consecuencias que conlleva ignorar esta perspectiva.

Así pues, se puede apreciar como la responsabilidad es un factor decisivo a la hora

de que alguien deposite su confianza en las decisiones de otro. Cuando se espera que un agente tome una decisión moral también se espera de este que asuma las responsabilidades que correspondan por las consecuencias que traerá esa acción (Kukita, 2015: 31).

Estos aspectos de la responsabilidad se están asumiendo a un nivel global de todo el proceso y el diseño de aplicaciones de inteligencia artificial, pero como se ha ido viendo los avances en este campo tienen perspectivas más específicas, como puede ser la creación de máquinas autónomas, que tomen decisiones por si mismas y no estén supervisadas por seres humanos. Se ha hablado ya del debate de la moralidad de este tipo de máquinas. Pero si hablamos de autonomía se podría entrar en un aspecto filosófico muy profundo ya que para que un agente pudiera estar dotado de autonomía, como los seres humanos[7], se le tendría que dotar de unas características que van más allá de inteligencia propiamente dicha, como también se vio en la primera parte se puede hablar de las condiciones para que un agente sea moral o ético. Estas eran tres capacidades: la de anticipar los propios actos, la de hacer juicios morales y el libre albedrío (Ayala, 2006). Respecto a estas tres capacidades se puede ver como la primera no sería difícil de alcanzar por una IA, sobre la segunda se ve como hay un gran debate respecto a como hacerlo, pero la tercera tiene implicaciones que van un poco más allá de lo que se ha venido argumentando hasta ahora.

La complejidad de que una máquina autónoma pueda ser moral o no aumenta en gran medida cuando se habla de responsabilidad. Para que una máquina pudiera tener responsabilidad sobre sus acciones debería poder entender las consecuencias de las acciones que realiza y entender no es lo mismo que conocer, ya que entender implicaría alguna clase de autoconocimiento y de reflexión, que ya se ha visto que, por ejemplo, en el intento de emular ese tipo de inteligencia (con la clasificación de Gardner, que se vio en la primera parte) es de las más complicadas de conseguir.

Aplicar el concepto de entendimiento en una máquina implica que esa máquina debería contar con unos valores dentro de sus procesos de toma de decisiones. Tendría que valorar que significa quitar una vida, en el caso de las armas autónomas letales por

7 Algo que como se ha visto en apartados anteriores también puede dar pie a debate, como se vio con los experimentos sobre potencial de preparación (Deecke, Scheid y Kornhuber, 1969)

ejemplo, y para esto tendría que saber darle valor a la existencia, cosa que debería implicar algún rasgo de autoconciencia. Y este rasgo es algo que ni la ciencia ni la filosofía en la actualidad alcanzan a comprender o por lo menos a ponerse de acuerdo en sus características esenciales y mucho menos en implementarlas en un sistema artificial.

A parte de la propia creación de IA y la complejidad de sus diseño, otro aspecto a tener en cuenta es la relación de estas con los seres humanos.

Como se puede ver es muy difícil crear un manual de comportamiento entre las personas que se pueda programar, ya que las máquinas tenderán a hacer una interpretación literal de ello. Solo hay que fijarse en como se malinterpreta una conversación mediante mensajería de texto frente a una en persona, el tono de voz, la gestualidad que en un momento con una persona es correcta, puede ser en otro momento con la misma persona algo totalmente ofensivo.

Estas situaciones en las que las personas cometen errores continuamente, son muy complejas y dependientes también de aspectos culturales. Pero también crean un problema añadido, un robot con la tecnología que se está desarrollando podría tener, si no lo hace ya, un sistema de sensores que pudiera detectar cuando alguien miente, o cuál es el estado emocional de una persona, con una facilidad mucho mayor de como lo haría otra persona.

Si se asume que el objetivo al construir una máquina es hacerlo lo mejor posible y que cumpla sus funciones de la manera más perfecta. Se acabaría tendiendo a construir un robot con todas esas características, como las de leer a los seres humanos de la manera más perfecta posible. Pero las personas no aprecian la compañía de alguien que siempre sabe si mientes o dices la verdad y cuál es tu estado de animo de una manera tan certera, rompiendo la barrera de la propia intimidad. Este problema es muy importante cuando ya se están empezando a fabricar y comercializar robots de compañía como los que hace la empresa SoftBank y su robot personal *Pepper* (Softbank Robotics, n.p.). Esta la otra faceta de la IA relacionada con la ética trae consigo mayores desafíos e implica una mayor amplitud de miras y una mayor necesidad de incluir la visión humanística en la ingeniera y el diseño de maquinas así.

Cuando el objetivo de la aplicación de una IA son las relaciones sociales no solo hay

que tener en cuenta la moralidad que pueda estar implementada en el robot, sino que también hay que valorar profundamente el impacto que produce en los seres humanos con los que interacciona.

Ahora mismo no hay robots con sentimientos, pero las personas si que desarrollan sentimientos hacia algunos artefactos, teniendo una tendencia a darle cualidades antropomorfitas a las cosas. Se dan casos que llaman mucho la atención, como intentar evitar que un robot creado para detonar minas siguiera haciendo su trabajo porque a los soldados que lo manejaban les parecía algo inhumano. O hacer jugar a unas personas con un dinosaurio robot y tiempo después pedirles que lo destruyeran, respondiendo la mayoría que no lo harían (Kusko, 2019).

Como ocurre en ocasiones, cuando el ser humano trata de plasmar sus características en sus propias creaciones, lo que hace es definirse mejor a sí mismo y alcanzar una mayor comprensión de la condición humana. Además de que como se ha visto anteriormente, el desarrollo de toda esta tecnología ya está cambiando la sociedad, junto con el hecho de que se espera más un ideal que una copia exacta de los seres humanos al intentar crear máquinas inteligentes autónomas y tratar de dotarlas de procesos que tengan en cuenta la ética, la responsabilidad, etc. el ser humano deberá plantearse como va a tratar a esos constructos.

Si se asume la idea de Immanuel Kant de que para que un individuo sea un agente moral debe estar dotado de libertad, de libre albedrío (Kant, 2005: 20), la sociedad, y especialmente aquellos que trabajan en proyectos relacionados con el campo de la IA, deberían tener en cuenta como van a tratar a sus sujetos de investigación. Si cada vez implementan más características humanas a construcciones artificiales se ha de plantear el momento en el que se tenga que redefinir el concepto de persona. Ya que si una IA estuviera dotada de Inteligencia, Moralidad y Libre Albedrío, se tendría que empezar a pensar cómo se le está tratando para poner entonces el foco de nuevo en la moralidad humana.

> Ahora bien, en el caso de que fuera posible construir sistemas con una inteligencia general, se plantearía un tipo de cuestiones éticas muy diferentes de las anteriores. Si fueran seres autónomos, tendríamos que aceptar que son personas y que, en consecuencia, es preciso reconocerles dignidad y exigirles responsabilidad (Cortina,

2019)

Volviendo a la ciencia ficción, que tantas veces se ha adelantado a cuestiones del futuro humano, en un capitulo de Star Trek la Nueva Generación, el Comandante Data, un ser artificial, se enfrenta a un juicio en el que se tiene que definir si es una persona o una propiedad. Se analiza en este momento como define a la raza humana el trato que le darán a una criatura, que aunque haya sido creada por ellos tiene inteligencia, entendimiento y consciencia de si misma. Ese trato, en definitiva, juzgara realmente a los humanos, más que al ser artificial, y si estos son una raza de esclavistas o no (Scheerer, 1989).

De esta manera surgen dos cuestiones que abren dos caminos diferentes: por un lado el desafío de como implementar una moralidad en las IA, algo bastante complejo y que, como se ha visto, genera mucho debate; y por otro lado el impacto que va causando en las relaciones y en la moralidad humana la integración de creaciones artificiales humanas dentro de la sociedad y como esta va cambiando y adaptándose a esto. Así se puede ver la importancia que recae tanto en como entidades artificiales tratarán a los seres humanos, si se las dota de una moralidad artificial; y de como los seres humanos tratarán a esas entidades artificiales si estas acaban estando dotadas de conciencia y libre albedrío.

Todo esto además abre la puerta de nuevo y hace entrar en valor al debate que se da en esferas políticas sobre si las humanidades son más o menos necesarias en la educación. Se ha podido ver a lo largo de todo el trabajo como están interconectadas la filosofía, la ética, la sociología, la antropología con la búsqueda de respuestas que se intenta dar desde las ciencias más técnicas como la informática a los desafíos que presenta la investigación y desarrollo de la inteligencia artificial y de como su implantación requiere de profundizar en el estudio de la moral y la ética para que al final sea algo funcional y que no cause muchos más problemas y catástrofes que beneficios para la sociedad

Si el ser humano no se conoce a si mismo va a ser complicado que pueda dotar a sus creaciones de algún tipo de sensibilidad hacia el mismo ser humano y la sociedad que van a compartir.

4. Conclusión

Como se ha podido ver hasta ahora, existe un gran debate en torno a la Inteligencia Artificial. Primero se cuestiona la posibilidad de alcanzar el logro de una inteligencia general artificial, parecida a la de los humanos, y a continuación su conveniencia por los peligros que podría conllevar.

Se ha visto en la primera parte de este trabajo el tema de la posibilidad y de si la ciencia y la tecnología humana serán capaces de alcanzar ese logro. Para ello se ha hecho una aproximación al primer texto que analizaba la posibilidad de conseguir crear maquinas que pensaran, para de ahí pasar al mismo concepto de inteligencia de como a su alrededor se genera debate, asumiendo para este trabajo la tesis de Gardner de las inteligencias múltiples. Se finalizaba esta parte valorando diferentes puntos de vista sobre si sería posible alcanzar el logro de una inteligencia general artificial, y viendo también como en algunos aspectos se han superado unas expectativas pero en otros parece que no se puedan alcanzar.

Una vea abordado el marco conceptual, se ha pasado a abordar el marco contextual, sobre la actualidad de las investigaciones sobre el cerebro humano, a través de dos grandes proyectos que tratan ese tema. También se ha tratado en esa parte las diferentes aplicaciones y problemáticas que causa la implantación de la inteligencia artificial en la sociedad y de como se esta dando una respuesta a esto. Organizaciones, gobiernos e instituciones trasnacionales muestran sus posturas y recomendaciones.

Se podría llegar a aceptar que no habría demasiada diferencia entre parecer ser siempre inteligente y serlo a efectos prácticos. Pero quedarse solo con esa idea como se ha visto en el capitulo segundo no basta. Los seres humanos no solo están dotados de una inteligencia que les permite conocer y adaptarse a entornos desconocidos, así como a la creación artística, sino que también poseen un sentido de la moral que les permite convivir unos con otros. Así la cuestión cambia de dirección, de abordar la inteligencia a enfrentarse al desafío de conseguir una moral artificial. Esto es lo que se ha empezado a valorar en la última parte.

También se ha visto como este nuevo desafío hace aparecer nuevas cuestiones sobre las características que se deberían implementar en estas creaciones artificiales que en

definitiva tratan de emular la vida humana, ya sea de manera ideal o con sus limitaciones. Algo que ya se tiene en cuenta en los grandes proyectos que se dan en este campo, así como en instituciones públicas o privadas que reclaman que se ponga el foco de atención en los aspectos éticos de todas estas investigaciones.

Todo ello lleva a concluir que el desarrollo de esta tecnología va a suponer todo un reto en las relaciones entre los humanos y también entre los humanos y las máquinas, algo que tendrá que ser redefinido. Y para que esto pueda tener éxito, se vuelve a poner de relevancia la importancia de las humanidades en la sociedad actual, algo que está siendo denostado por algunos sectores. Pero que somo se ha visto a lo largo del trabajo, estas ciencias humanas están totalmente entrelazadas con el desarrollo de las investigaciones en inteligencia y moral artificial.

Ya que sin esta parte del conocimiento no se podrá lograr que las creaciones humanas aporten beneficios reales a la sociedad.

5. Bibliografía

AGAR, NICHOLAS (2016): Don't Worry about Superintelligence en *Journal of Evolution and Technology*, Vol. 26 Issue 1, Feb. 2016, 73-82.

ASIMOV, ISAAC (2019): Circulo vicioso en *El robot completo*, Madrid, Alamut. 183-198.

ASIMOV, ISAAC (1992): *Robots e imperio*, Barcelona, Plaza & Janés.

AYALA, FRANCISCO J. (2006): *La evolución de un evolucionista*. Valencia, Universidad e València.

BARGMANN, CORNELIA ET AL (2014): *BRAIN 2025. a scientific vision*, The BRAIN Initiative, Washington.

BENASAYAG, MIGUEL; MEYRAN, REGIS (2018): ¡El cerebro no es el que piensa! En *El correo de la UNESCO*, Jul-Sep 2018, 15-17.

BRUNDAGE, MIKE (2015): Taking superintelligence seriusly Superintelligence: Paths, dangers, strategies by Nick Bostrom en *Futures*, Núm 72, 32-35.

BOSTROM, NICK Y ELIZIER YUDKOWSKY (2014): The Ethics of Artificial Intelligence en *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, 316-334.

CALVO, PATRICI (2018): Ética de las cosas (EoT). Hacia una digitalización socialmente responsable y moralmente válida del ámbito universitario. En Andrés, Alicia y Sanahuja, Rosana (Eds.). *Un diseño universitario para la responsabilidad social*, Castellón, Universitat Jaume I, publicacions.

COMISIÓN EUROPEA (2019): *Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las regiones. Generar confianza en la inteligencia artificial centrada en el ser humano*. 8 de abril de 2019. Recuperado de https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=COM:2019:168:FIN. [Consultado el 31 de julio de 2019].

CONISIÓN EUROPEA (2019): High-Level Expert Group o Artificial Intelligence. Recuperado de https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence. [Consultado el 31 de julio de 2019].

CONN, ARIEL (2018): AI Should Provide a Shared Benefit for as Many People as Possible en *Future of Life Institute*, 10 de enero de 2018. Recuperado de https://futureoflife.org/2018/01/10/shared-benefit-principle/. [Consultado el 26 de julio de 2019].

CONVERSATION, THE (2019): De las migrañas al cáncer: inteligencia artificial para diagnosticar enfermedades en *Huffpost.es* 11 de septiembre de 2019. Recuperado de https://bit.ly/2lUtDrk. [Consultado el 17 de septiembre de 2019].

CORTINA, ADELA (6 de junio de 2019): Ética de la inteligencia artificial desde Europa en *elpais.com*. Recuperado de https://elpais.com/elpais/2019/06/05/opinion/1559729489_306891.html. [Consultado 12 de septiembre de 2019]

CORTINA, ADELA (2013): *¿Para qué sirve realmente…? La ética*. Barcelona, Paidós.

DEECKE, LÜDER; SCHEID, PETER Y KORNHUBER, HANS S. (1969): Distribution of Readiness Potential, Pre-modon Positivity and Motor Potential of the Human Cerebral Cortex Preceding Voluntary Finger Movements en *Exp. Brain Res*. Núm. 7, 158-168.

ERICSSON (n.d): Would you take a self-driving bus?. Recuperado de https://www.ericsson.com/en/internet-of-things/trending/driverless-buses-in-stockholm-sweden. [Consultado el 4 de septiembre de 2019].

FOOT, PHILIPPA (1967): The Problem of Abortion and the Doctrine of the Double Effect en Oxford Review, núm. 5, 5-15.

FUTURE OF LIFE INSTITUTE (n.d.) Future of Life Institute website. Recuperado de https://futureoflife.org/team/. [Consultado el 20 de septiembre de 2019].

GANASCIA, JEAN-GABRIEL (2018): Inteligencia artificial: entre el mito y la realidad en *El correo de la UNESCO*, Jul-Sep 2018, 7-9.

GARCÍA-MARZÁ, DOMINGO (2014): La RSC en perspectiva ética en Mediterráneo Económico, núm. 26, 239-254.

GARDNER, HOWARD (2010): *La inteligencia reformulada. Las inteligencias múltiples en el siglo XXI*, Barcelona, Paidós Ibérica.

GRONLUND, KRISTEN (2019): State of AI: Artificial Inteligence, the Military and Increasingly Autonomous Weapons en Future of Life Institute, 9 de mayo de 2019. Recuperado de https://futureoflife.org/2019/05/09/state-of-ai/. [Consultado el 8 de

agosto de 2019].

HECHT, JEFF (2018): Automation. Meeting people´s expectations en Nature, vol 563, Nov 2019, 141-143.

HUMAN BRAIN PROJECT (n.d.). Human Brain Project website. Recuperado de https://www.humanbrainproject.eu/en/. [Consultado el 20 de agosto de 2019].

KANDEL, ERIK R. ET AL (2013): Neuroscience thinks big (and collaboratively) en *Nature*, vol. 14, Sept 2013. 659-664

KANT, IMMANUEL (2005): *Crítica de la razón pura*, Barcelona, Taurus.

KNAEBE, BRENNA ET AL (2017): New Caledonian crows show behavioural flexibility when manufacturing their tools en *Behaviour* jan 2017, Vol 154, Iss 1, 65-91.

KUKITA, MINAO (2015): Differences between artificial intelligence and artificial morality en *Applied Ethics: Security, Sustainability and Human Flourishing*, pp. 27 37.

KUSKO, FEDERICO (2019): Cómo querer a un robot en *SINC*, 6 de agosto de 2019. Recuperado de https://shorturl.at/vI069. [Consultado el 3 de agosto de 2019].

LIZ GUTIERREZ, ANTONIO MANUEL (2009): Simulando a Searle en *Praxis Filosofía* N.º 28, Enero-Junio 2009, 117-141.

MILL, JOHN S. (2002): *El utilitarismo*, Madrid, Alianza.

PALAZUELOS, FÉLIX (2017): Elon Musk:"La inteligencia artificial amenaza la existencia de nuestra civilización" en *El Pais*, 18 de julio de 2017. Recuperado de https://elpais.com/tecnologia/2017/07/17/actualidad/1500289809_008679.html. [Consultado el 5 de septiembre de 2019].

RUBIO, ISABEL (2018): Amazon prescinde de una inteligencia artificial de reclutamiento por discriminar a las mujeres, *El País*, 12 de octubre de 2018. Recuperado de https://elpais.com/tecnologia/2018/10/11/actualidad/1539278884_487716.html. [Consultado el 22 de julio de 2019].

SALAS, JAVIER (2018): Google arregla su algoritmo 'racista' borrando a los gorilas elpais.com Recuperado de https://bit.ly/2mxMxAP. [Consultado el 11 de septiembre de 2019]

SCALABLE COOPERATION GROUP (n.d.): Moral Machine website. Recuperado de http://moralmachine.mit.edu/hl/es. [Consultado el 25 de agosto de 2019].

SCHEERER, ROBERT, SNODGRASS, MELINDA M. (1989): The Measure of a Man en *Star Trek: The Next Generation*, Epi. 9, Temp. 2.

SEARLE, JOHN (1994): *Mentes, cerebros y ciencia*, Madrid, Cátedra.

SINGULARITY UNIVERSITY (n.d.): Singularity University website. Recuperado de https://su.org/. [Consultado el 22 de agosto de 2019].

SOFTBANK ROBOTICS (n.d.): Softbank Robotics website. Recuperado de https://www.softbankrobotics.com/us/. [Consultado el 26 de agosto de 2019].

THE BRAIN INITIATIVE (n.d.). NIH. The BRAIN Initiative website. Recuperado de https://braininitiative.nih.gov. [Consultado el 20 de agosto de 2019].

TURING, ALAN M. (1950): Computing machinery and intelligence en *Mind* 49, 433-460.

Una inteligencia artificial se vuelve racista antisemita y homófoba en menos de un día en Twitter (28 de marzo de 2016) en *Elmundo.es*. Recuperado de https://www.elmundo.es/tecnologia/2016/03/28/56f95c2146163fdd268b45d2.html. [Consultado el 12 de septiembre de 2019].

UNESCO (2019): *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*, UNESCO. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000366994. [Consultado el 1 de julio de 2019]

WALLACH, WENDELL, SMIT, IVA Y ALLEN, COLIN (2005): Artificial Morality: Top Down, Bottom-uo, and Hybrid Approaches en Ethics and Information Technology, 7, Sept, 149-155.

WIMPENNY, JOANNA H. WEIR, ALEXANDER A.S. KACELNIK ALEX (2011): New Caledonian crows use tools for non-foraging activities en *Animal Cognition* May 2011, Vol 14, 459-464.

6. Anexos

6.1. Anexo 1. Los principios para la IA de Asilomar

Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.

Research Issues

1) **Research Goal**: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.

2) **Research Funding**: Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:

- How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
- How can we grow our prosperity through automation while maintaining people's resources and purpose?
- How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
- What set of values should AI be aligned with, and what legal and ethical status should it have?

3) **Science-Policy Link**: There should be constructive and healthy exchange between AI researchers and policy-makers.

4) **Research Culture**: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.

5) **Race Avoidance**: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and Values

6) **Safety**: AI systems should be safe and secure throughout their operational lifetime,

and verifiably so where applicable and feasible.

7) **Failure Transparency**: If an AI system causes harm, it should be possible to ascertain why.

8) **Judicial Transparency**: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

9) **Responsibility**: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

10) **Value Alignment**: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

11) **Human Values**: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

12) **Personal Privacy**: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

13) **Liberty and Privacy**: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

14) **Shared Benefit**: AI technologies should benefit and empower as many people as possible.

15) **Shared Prosperity**: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

16) **Human Control**: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

17) **Non-subversion**: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

18) **AI Arms Race**: An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

19) **Capability Caution**: There being no consensus, we should avoid strong

assumptions regarding upper limits on future AI capabilities.

20) **Importance**: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.

21) **Risks**: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

22) **Recursive Self-Improvement**: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.

23) **Common Good**: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.