Escuela
Politécnica
Superior

Ética en la Inteligencia Artificial

Grado en Ingeniería Informática

Trabajo Fin de Grado

Autor:
Pablo Giner Hidalgo (alumno)
Tutor:
Francisco Antonio Pujol López (tutor)

Universitat d'Alacant
Universidad de Alicante

Diciembre 2024

Ética en la Inteligencia Artificial

Autor
Pablo Giner Hidalgo (alumno)

Tutor
Francisco Antonio Pujol López (tutor)
Tecnología Informática y Computación

Grado en Ingeniería Informática

Escuela Politécnica Superior

Universitat d'Alacant
Universidad de Alicante

ALICANTE, Diciembre 2024

Preámbulo

Cuando estaba en el bachillerato no sabía muy bien qué estudiar en la universidad, hasta que un día vi la película ”Blade Runner” y algo se encendió en mí. La escena del test de Turing me generó la duda: ¿habrá algún día en que no podamos diferenciar un robot de una persona? Realmente, de eso trata la película: hasta qué punto un androide puede ser más humano que una persona. Fue esa duda, ese miedo, el que me hizo elegir esta carrera. Quería entender si algo así podía suceder y, de ser así, cómo podría suceder.

Conforme fui avanzando en la carrera y aprendiendo conceptos, problemas y soluciones, me fui interesando cada vez más por la inteligencia artificial. El concepto de crear un ”ser inteligente”, una máquina capaz de pensar como un ser humano, me recordó a esos replicantes. Sin embargo, al profundizar en el tema, sentí que faltaba algo, lo que solemos llamar ”alma”. Esa alma, para mí, es la ética.

Investigando más sobre la ética en la inteligencia artificial, descubrí que ya existían estudios al respecto, pero la gran mayoría de la información que encontré (artículos, libros, etc.) estaba escrita por psicólogos o filósofos, y eran muy pocos los ingenieros que trataban el tema. Entonces pensé: ¿cómo es posible que nosotros, los creadores de estos robots e inteligencias, estemos hablando tan poco sobre el tema? ¿No deberíamos también dar nuestra opinión y visión?

Este estudio se basa en eso, en ofrecer, desde el punto de vista de un futuro ingeniero, una opinión sobre la ética en la inteligencia artificial. No solo respetando los estudios psicológicos y filosóficos del asunto, sino abrazándolos y entendiendo que son cruciales para el tema. Para alcanzar este objetivo, todos los distintos departamentos debemos colaborar: filosofía, psicología, ingeniería, jurisprudencia, etc.

Agradecimientos

Este ha sido un camino duro, y no habría sido posible sin el apoyo de personas a las que quiero agradecer lo que han hecho por mí durante el mismo.

En primer lugar, agradecer a la universidad por todos estos años, por los momentos más complicados y por los más divertidos de mi vida. Gracias a todos estos se ha conformado la persona que soy hoy en día, cosa de la que estoy muy orgulloso. Agradecer, por supuesto, a Francisco Antonio por su ayuda y apoyo desde el primer momento, por haber confiado en mí y en mi propuesta, y por haber sido tan comprensivo durante este trabajo. Desde el día que tuve la idea de este proyecto pensé en él como tutor, pues además de su experiencia en el tema, me transmitió en su clase algo que ningún profesor me había transmitido aún: pasión por su trabajo. Ganas de transmitir sus conocimientos a sus alumnos. Es algo que, por lo que he experimentado, es de una rareza extraordinaria, y estoy seguro de que la educación ganaría mucho si todos los profesores tuviesen ese entusiasmo.

En segundo lugar, agradecer a la gente que he ido conociendo durante todos estos años. Muchos de ellos son y serán amigos durante el resto de mis días. Otros muchos se perdieron por el camino, pero dejaron una huella imborrable en mí. El día que vine a estudiar aquí, era un chico bastante tímido y con miedo a la soledad que podía encontrar al mudarme aquí, solo. Pero encontré todo lo contrario: un montón de personas muy distintas a mí, a lo que conocía hasta ahora, que me ayudó a abrir mi mente de una manera inmensa. Personas sin las que este camino hubiese sido muchísimo más duro, que me han apoyado en mis momentos más bajos y han creado recuerdos que tengo grabados de manera permanente. Gracias a todos los que habéis hecho de esta mi casa.

En tercer lugar, agradecer a mis amigos de siempre su apoyo incondicional. Toda la gente con la que me he criado, que me han forjado como persona y me han impulsado a creer que puedo conseguir lo que me proponga. A todos los que han escuchado mis lamentos, mis alegrías, mis dudas y mis ocurrencias. Sabéis que estoy y siempre he estado muy agradecido con vosotros, y que lo que hemos vivido no es nada comparado con lo que nos queda por vivir. Sabéis que siempre estaré ahí, igual que siempre habéis estado vosotros.

Por último, y de forma más importante, agradecer a mi familia todo lo que han hecho por mí. A todos los que me habéis apoyado, habéis creído en mí. A todos los que me han llamado en algún momento para preguntarme qué tal o para felicitarme por algún costoso aprobado. Gracias por vuestro ánimo incondicional, que me ha llegado siempre, aún estando lejos. Me gustaría mencionar a algunas personas en concreto que han sido mis mayores pilares estos años. Para empezar, a mi amigo Dani, quien trascendió ese umbral hace tiempo para convertirse en mi hermano. No sé dónde ni cómo estaría si no nos hubiésemos conocido, pero seguro que no hubiese llegado tan lejos. Gracias por todo. También quiero agradecer a mi abuelo, que siempre ha sido un ejemplo de esfuerzo para mí y una fuente de motivación con

su inagotable confianza. Siempre has creído en mí, tanto que has conseguido que hasta yo me lo crea. Te quiero. Y por supuesto, a mis padres y a mi hermano, que son la fuerza que me impulsa cada día. Siempre habéis luchado por mí, me habéis levantado cuando me caía y me habéis sostenido cuando estaba en pie. Siempre me habéis escuchado, comprendido y apoyado, sin importar momento o circunstancia. La persona que soy hoy día es gracias a vosotros, y vuestro orgullo es lo más importante que he recibido en toda mi vida.

Es a ellos a quienes dedico este trabajo.

Podemos ver sólo un corto trecho del futuro,
pero podemos ver lo suficiente para darnos cuenta
de que hay mucho que hacer.

Alan Turing.

Índice general

1 Introducción

La tecnología está avanzando a pasos agigantados en la actualidad, y sin duda estamos viviendo una revolución, quizá la más grande de nuestra historia. Desde la aparición de ChatGPT en nuestras vidas, la **Inteligencia Artificial (IA)** se ha vuelto algo conocido y al alcance de toda la sociedad. Algunos le tienen pavor, otros la consideran una especie de 'heraldo del fin'. Sin duda es un cambio, como cualquier revolución, que ha llegado para cambiar nuestras vidas, y los cambios comportan una parte negativa (obvia y muy discutida, en este caso) y una parte positiva (que como ingenieros defendemos y promulgamos). Sin embargo, hay una **delgada línea** que las separa, una línea sobre la que debemos trabajar y debemos respetar para que todos los beneficios que nos propone no sean completamente opacados por sus perjuicios. Esta línea es la **ética**.

1.1 La ética

Por ponernos en contexto, quiero hablar primero de la ética. ¿Qué es la ética y por qué es tan importante? Bien, según la RAE, la ética es el conjunto de normas **morales** que rigen la conducta de la persona en cualquier ámbito de la vida. El término ética fue empleado por primera vez por Aristóteles para nombrar a uno de los campos de estudio de sus predecesores (Platón y Sócrates). Esta rama de la filosofía estudia lo correcto y lo incorrecto, lo bueno y lo malo, la **moral**. ¿Y en qué afecta esto a la IA? Toda aplicación, robot, máquina, etc. que creamos como ingenieros cumple un propósito por el cual fue diseñado. Pero ahora que estamos creando cada vez herramientas con mayores y más amplios propósitos, debemos empezar a fijarnos en cuáles son las consecuencias **morales** de nuestras creaciones, pues como estamos comprobando, tienen el poder de cambiar la sociedad por completo.

1.2 La revolución

Muchos expertos, como Emilio Gayo, presidente de Telefónica España, aseguran que estamos viviendo la Cuarta Revolución Industrial. Estas son sus palabras:

"La Cuarta Revolución Industrial difiere de las anteriores en que, por primera vez, se centra en derribar las barreras del conocimiento a través de la aplicación del Internet of Things (IoT), para sensorizar y conectar, capturando así datos de cualquier proceso; y tecnologías de IA, para extraer información y conocimiento de los datos, mejorando así la capacidad de tomar decisiones y la inteligencia humana". Emilio Gayo (2019)

Como mencioné anteriormente, ChatGPT ha sido el 'despertar' para la sociedad en general en cuanto al mundo de la IA. Esta ha hecho generar una gran curiosidad, y con ella una gran ola de dudas, a todo ciudadano. Dudas que, normalmente como conocedores de la realidad y funcionamiento de este chatbot (lo que se nos permite saber), y de la IA, conocemos. ¿Cómo puede ser que lo sepa todo? ¿Cómo puede contestar con tal rapidez a mis preguntas? Sin

embargo, hay otras preguntas que tampoco nosotros podemos contestar. ¿Es esto beneficioso o perjudicial para las nuevas generaciones? ¿Acaso es el inicio de la sustitución del ser humano robots? Como creadores de estas inteligencias, debemos ser conscientes de estos miedos e inquietudes y afrontarlos. Por ello considero que la ética en nuestro trabajo es tan importante, porque es lo que moldeará el futuro, lo que le pondrá límites, lo que ayudará a que todos juntos, ciudadano de a pie o experto, sigamos el mismo camino. Sin la ética, el miedo y la duda es segura.

2 Marco Teórico

2.1 Ética en la Inteligencia Artificial (IA)

La ética en la IA es un campo de estudio que examina las implicaciones morales y éticas del desarrollo y uso de tecnologías de IA. Con el rápido avance de esta y su creciente integración en tantos aspectos de nuestra vida cotidiana, es fundamental abordar las preocupaciones éticas que surgen. Durante el tiempo que he estado realizando este proyecto, he descubierto una serie de problemas presentes hoy día, que tocan distintos hábitos de nuestra vida, pero que todos deben resolverse para alcanzar un punto final que nos contente a todos. Para estos problemas presento 6 conceptos, todos ligados entre si, como comprobaremos, a los que he llamado 'principios éticos', basándome en los 'Principios sobre IA' adoptados por los líderes del G-20 en la cumbre de Osaka en 2019 Agustinoy (2024). Los expongo a continuación:

2.1.1 Responsabilidad

Después de asistir a varios congresos sobre la IA (propuestos por profesionales ajenos a la informática, pero con la colaboración de expertos en la materia), y escuchando a la sociedad que nos envuelve, me he dado cuenta de que el ámbito más temido hoy en día en cuánto a la IA es la responsabilidad.

Al comprar un producto, nosotros, como usuarios de este, queremos unas garantías. Para que una aplicación que use IA sea aceptada, debe superar una tasa de acierto previamente marcada. Por ejemplo, en el 'Large Scale Visual Recognition Challenge 2017 (ILSVRC2017)', el ganador tuvo un Average Precision (AP) de 0.731392. Agustinoy (2024) Aunque el AP es algo más complejo que un porcentaje, por simplificarlo diremos que tiene un 73,14% de tasa de acierto. Sin embargo, ninguno podemos imaginar venderse una aplicación de reconocimiento de voz de un 73,14% de tasa de acierto. De hecho, la consultora de IA y Procesamiento natural del lenguaje (NLP) SpeechWare, con contrato Marco para la implantación del reconocimiento de voz con la Unión Europea en Bruselas y Luxemburgo, promete un porcentaje de acierto del 99% con su aplicación DictaLaw, su sistema de reconocimiento de voz jurídico. DigaLaw (2010)

La cuestión es, ¿qué pasa cuando falla? Porqué sí, es difícil que lo haga, pero lo hace, aunque sea solo en un 1% de los casos. Si lo miramos sin preocupación puede ser insignificante, pero detengámonos un segundo a pensar en lo que significa un error. Un error en la detección de un cáncer en una aplicación médica, o un error en una decisión de un coche autónomo que provoque una muerte ¿De quién es la culpa? ¿Quién debe pagar las consecuencias?

Para responder a esta pregunta existe la Ley. Justo este año 2024 se aprobó, el día 14 de marzo, la Ley de IA de la Unión Europea (UE), pionera en realizar una ley de este tipo. No entraré en detalle en ella, pero dice lo siguiente sobre la responsabilidad:

"La Ley de IA prevé el derecho a presentar una denuncia ante una autoridad nacional. Sobre esta base, las autoridades nacionales pueden poner en marcha actividades de vigilancia

del mercado, siguiendo los procedimientos de los reglamentos en la materia.

Además, la Directiva sobre responsabilidad en materia de inteligencia artificial propuesta tiene por objeto facilitar a las personas que pidan una indemnización por los daños causados por sistemas de inteligencia artificial de alto riesgo medios eficaces para identificar a las personas potencialmente responsables y obtener pruebas pertinentes para la reclamación por daños y perjuicios. A tal fin, la Directiva propuesta prevé la divulgación de pruebas sobre sistemas concretos de inteligencia artificial de alto riesgo que se sospeche hayan causado daños.

Además, la Directiva revisada sobre responsabilidad por los daños causados por productos defectuosos garantizará que haya indemnizaciones para las personas que sufran muerte, lesiones corporales o daños materiales a causa de un producto defectuoso en la Unión, y aclara que los sistemas de inteligencia artificial y los productos que integren sistemas de inteligencia artificial también estarán cubiertos por las normas vigentes." Imagenet (2017)

Tuve la suerte de asistir a unas jornadas sobre derecho de IA realizada en esta misma universidad los días 4 y 5 de diciembre de 2023. En la ponencia del Prof. Aurelio López-Tarruella Martínez, Doctor en Derecho desde 2004, sobre el marco jurídico en materia de Inteligencia Artificial, nacional y europeo, habló de este reglamento en marcha. Explicó que la IA ya estaba regulada de forma *ex post*, es decir, una vez sucedido el evento, pero aún no de forma *ex ante*, anterior al evento. Por poner un ejemplo de esto: si tu, en un coche autónomo, tienes un accidente que provoca una muerte, eres ciertamente responsable del hecho pues así lo dicta la legislación vigente, dado que eres el conductor del vehículo. Con esta ley de IA, se 'identificaran las personas potencialmente responsables', por lo que no tendremos porque ser responsables (o, por lo menos, los únicos responsables). Y esto es importante.

Nosotros, como individuos y usuarios de cualquier aplicación, máquina, robot, etc. con IA, tenemos una responsabilidad por el mero hecho de usarla, obviamente siendo conscientes del riesgo que comporta. Usando el ejemplo de antes, pongamos que el porcentaje de fallo de la IA de nuestro coche es menor al 1%, haciendo que debamos ser conscientes que sigue existiendo la posibilidad de fallar. Esto ya nos hace responsables, por una parte. Si además el vehículo cuenta con la posibilidad de conducción manual, con más razón aún lo somos.

Sin embargo, gracias a esta ley esta responsabilidad no recae solo en nosotros, pues estamos de acuerdo en que sería algo completamente injusto. Gracias a este reglamento llegarán una serie de gestión de riesgos, datos de calidad, documentación técnica, trazabilidad, instrucciones de uso, vigilancia humana, precisión, solidez, etc. que se resumirán en una seguridad por parte del usuario, de la sociedad e incluso de las propias empresas en cuánto a la responsabilidad jurídica del asunto.

Como ejemplo de uso, tenemos el caso de TESLA. Rivera (2024) Esta compañía ha salido indemne de cantidad de denuncias sobre sus coches autónomos, para algunos aprovechándose del 'vacío legal' que suponían aun estos automóviles. La compañía fue acusada bajo demanda de negligencia y ocultación intencional debido al fallecimiento del marido de la denunciante un par de años antes. Aunque en el artículo, de principios de año, critica una legislación tardía, esperemos que la propuesta por la UE cambie el panorama de la IA en los juzgados.

2.1.2 Transparencia

Ya hemos mencionado este principio, pues al fin y al cabo todos están ligados entre ellos. La transparencia en el contexto de la IA se refiere a la capacidad de explicar y comprender cómo

y por qué un sistema de IA toma decisiones, procesa datos o realiza acciones específicas. Esto incluye la claridad en los algoritmos utilizados, los datos empleados, los procesos de toma de decisiones y la lógica subyacente al sistema. La transparencia permite a los usuarios, desarrolladores y reguladores acceder a información comprensible sobre el funcionamiento interno de la IA, lo que es esencial para el principio de responsabilidad.

Para que una IA sea aceptada por un usuario, no solo hace falta un principio claro de responsabilidad, si no que, como cualquier otro producto, necesita la confianza del consumidor, en este caso del usuario. Sin transparencia, es imposible conseguir esa relación de confianza entre usuario y desarrollador, que es aún más importante teniendo en cuenta en que aspectos se está utilizando la IA: salud, justicia, empleo, seguridad, etc. ¿Como va a aceptar una persona que necesite un tratamiento un medicamento recetado por una IA cuándo no confía en su criterio?

De hecho, la transparencia podría englobarse dentro de la responsabilidad si la subdividiésemos en responsabilidad jurídica y ética. La jurídica sería la que hemos hablado en el apartado de responsabilidad. En cuanto a la responsabilidad ética, los desarrolladores de IA tienen una responsabilidad significativa en la creación de sistemas que no solo sean eficientes y efectivos, sino también éticamente sólidos. La responsabilidad ética implica diseñar y programar IA que respete la dignidad humana, los derechos fundamentales y que minimice los posibles daños, alineándose por ejemplo con el principio hipocrático tradicional en la medicina.

Que una IA sea explicativa es un buen y necesario comienzo. Esto significa que los desarrolladores deben ser capaces de proporcionar razones o justificaciones comprensibles y claras sobre cómo llega su creación a sus decisiones o conclusiones. Sin esta regla, la IA no tendría cabida en aplicaciones críticas como la salud, el derecho, las finanzas y otras áreas donde sus decisiones pueden tener impactos significativos. Estamos obligados como desarrolladores a garantizar seguridad y confianza a nuestros usuarios, y a mantener una responsabilidad sobre nuestras creaciones, pues sin ello es imposible que avancemos de una forma correcta. Esto se liga con otro principio, el de la justicia, del que hablaremos más adelante.

Obviamente, esta transparencia tiene unos claros desafíos. El primero es la clara complejidad técnica. Muchos sistemas de IA, especialmente aquellos basados en redes neuronales profundas, son intrínsecamente complejos y funcionan como "cajas negras", donde incluso los desarrolladores tienen dificultades para explicar cómo se llegan a ciertas decisiones. Esta complejidad técnica hace que la transparencia sea difícil de lograr. De hecho, este es uno de los mayores problemas a los que se enfrenta la IA, aunque, si lo pensamos, puede llegar a ser algo obvio. Aunque profundizaré en ello más adelante, una forma de pensar en la IA es la de un niño que poco a poco va aprendiendo. En este caso, cuando a un niño le enseñas algo (o crees que lo haces) y luego ves al niño ejecutando ese conocimiento (pongamos de ejemplo mirar a los lados antes de cruzar la calle) realmente no puedes asegurar que ha llegado a ese gesto debido a tus enseñanzas, pues podría estar haciéndolo porque su mejor amigo lo hace, o porque ha visto hacerlo al protagonista de su película favorita. Por eso creo que no se aleja tanto del problema que presenta la IA en este aspecto.

Otro desafío es la propiedad intelectual y los secretos comerciales que las empresas pueden tener con sus productos. Las empresas que desarrollan sistemas de IA a menudo se enfrentan a un dilema entre mantener la transparencia y proteger su propiedad intelectual. La divulgación completa de los algoritmos y modelos utilizados puede comprometer la ventaja competitiva

de una empresa. A esto hay que sumarle que la IA necesita datos, y estos datos pueden entrar en conflicto con la privacidad, que al mismo tiempo entra en conflicto con la transparencia, siendo un claro problema para las empresas que desarrollen IA.

Está claro que este tema es difícil de tratar. Desarrollar modelos de IA que sean explicables e interpretables es una forma clave de lograr transparencia. Los modelos interpretables son aquellos cuyas decisiones pueden ser fácilmente comprendidas por humanos. Por ejemplo, en lugar de utilizar una red neuronal profunda, que es difícil de interpretar, se podría optar por modelos más simples como árboles de decisión o reglas lógicas en ciertos contextos. Sin embargo, esto significaría dar pasos atrás en cuanto a la evolución, desarrollo y uso de la IA, teniendo que valorar que aspecto pesa más en la balanza.

En cuanto a casos que sirvan para representar este tema, los veremos en los siguientes principios, pues la transparencia y los mismos estarán, como veremos, ligados de forma completa.

Como posibles acercamientos a una solución para alcanzar el objetivo de la transparencia, veo cuatro posibles opciones, todas ellas compatibles. La primera es una documentación detallada y exhaustiva sobre cómo se diseñó, entrenó y probó un sistema de IA, incluyendo información sobre los datos utilizados, los procesos de desarrollo, las pruebas de sesgo y las evaluaciones de impacto ético.

La siguiente sería el uso de herramientas de visualización que permitan a los usuarios visualizar cómo funciona un modelo de IA, como gráficos que muestren la importancia de diferentes variables en la toma de decisiones, que pueden mejorar la comprensión y la transparencia.

Como tercera opción, realizar auditorías y evaluaciones independientes, que garantizarían confianza, evaluando tanto los algoritmos como los datos utilizados, y siendo sus resultados compartidos con el público o con los reguladores.

Finalmente, el desarrollo por parte de las empresas que implementen o desarrollen IA de unas claras políticas de divulgación y comunicación sobre cómo y qué información se transmitirá al público, y cómo se comunicarán las decisiones tomadas por los sistemas de IA a los usuarios.

2.1.3 Privacidad

El siguiente punto es la privacidad. Sé que puede resultar confuso, ¿cómo puede ser un principio la transparencia y el siguiente la privacidad? Como he explicado anteriormente, la transparencia tiene un problema en cuánto a la privacidad de los datos de sus usuarios, y es a esta privacidad a lo que me refiero. Aunque los sistemas, deben también respetar la privacidad de los datos de los usuarios, asegurando que la información personal se maneje de manera segura y ética, sin ser explotados de manera indebida o accesibles a partes no autorizadas. . En un mundo donde los datos son el "nuevo petróleo", proteger la privacidad es crucial para evitar abusos y violaciones de los derechos humanos.

La privacidad es crucial por varias razones. El primero es obvio: es un derecho fundamental, reconocido en múltiples instrumentos internacionales, como la Declaración Universal de Derechos Humanos. Los individuos deben tener el control sobre su información personal para proteger su dignidad y autonomía. Esto hace que existan leyes y regulaciones en muchas jurisdicciones que protegen la privacidad de los datos, como el Reglamento General de Protección de Datos (GDPR) en la UE Europea (2016) o la Ley de Privacidad del Consumidor de California (CCPA). Bonta (2024) Los sistemas de IA deben cumplir con estas regulaciones para operar legalmente. Esto en cuanto al margen jurídico (vinculado a la responsabilidad).

Alejándonos del mismo, encontramos puntos importantes como la ya mencionada confianza del usuario. Esta confianza en las tecnologías de IA depende en gran medida de la capacidad de estas para proteger la privacidad de los usuarios. Si los individuos creen que sus datos personales no están seguros, es menos probable que confíen en los sistemas de IA o los utilicen. Por si no fuese suficientemente importante que un usuario confié en el producto del que hace uso, o del mismo productor, podemos sumarle a la parte más ética de este principio la prevención de abusos y la mitigación de riesgos. Está claro que, sin salvaguardias adecuadas, la información personal puede ser utilizada de manera indebida para fines como la vigilancia masiva, la manipulación, la discriminación o la explotación comercial, y la recopilación y el procesamiento de grandes cantidades de datos personales por parte de sistemas de IA pueden exponer a los individuos a riesgos como la pérdida de privacidad, el robo de identidad, el fraude y otras formas de abuso. Es por ello que al desarrollar o emplear cualquier uso de IA, debemos tener un firme compromiso con la privacidad del usuario.

Estos mismos hechos son parte, a su vez, del desafío del principio en si. Está claro que los sistemas de IA necesita una recopilación masiva de datos, lo que aumenta el riesgo de violaciones de la privacidad. Aunque los datos pueden ser anonimizados para proteger la privacidad, existe el riesgo de reidentificación, donde los datos aparentemente anónimos pueden ser vinculados de nuevo a individuos específicos mediante técnicas avanzadas de análisis de datos.

Hoy en día existen ciertas respuestas retóricas muy utilizadas por los firmes defensores de estos sistemas: desde el "¿Qué más da que usen tus datos? ¿Tienes algo que esconder?" hasta el "¡Tus datos no le interesan a nadie, no eres Bill Gates!". El caso es que, aunque las respuestas son debatibles, los sistemas de IA pueden acceder y procesar datos extremadamente sensibles, como información de salud, biometría o comportamiento en línea. El uso inadecuado de estos datos puede tener graves consecuencias. Además, estos datos personales pueden ser compartidos entre diferentes sistemas y aplicaciones de IA, a menudo sin el conocimiento o el consentimiento del usuario, lo que incrementa el riesgo de exposición y abuso de información. Mas adelante abordaremos este aspecto.

Además, es necesario recalcar que hay unas obvias desigualdades en la Protección de Datos: las personas en diferentes regiones del mundo no tienen el mismo nivel de protección de datos. En algunas jurisdicciones, las leyes de privacidad pueden ser débiles o inexistentes, lo que expone a los individuos a mayores riesgos. Hemos visto las leyes en la UE, por ejemplo, pero países como Afganistán, Yemen o Siria están completamente desprotegidos ante esto, aunque otros como Corea del Norte lo están en beneficio de su propio gobierno.

Combinando transparencia y privacidad, tenemos el caso de Cambridge Analytica. AGENCIAS (2018) En resumen, en 2018 se descubrió que esta consultora de marketing y publicidad para fines políticos y corporativos obtuvo los datos de 50 millones de usuarios de Facebook para idear "la herramienta" para influir en las elecciones presidenciales de Estados Unidos de 2016, lo que resultó en la victoria de Trump. Incluso algunos medios aseguran que el proceso podría haber moldeado también el 'Brexit'. Según la noticia, 'A través de un inofensivo test de personalidad diseñado por el profesor de Psicología de la Universidad de Cambridge, Aleksandr Kogan para la empresa Global Science Research, Cambridge Analytica recopiló información sobre la personalidad de los usuarios de la red social, sus gustos, sus inclinaciones políticas y sus ideales. Su aplicación fue descargada por alrededor de 2.700 personas, y también tuvo acceso a los datos de los amigos de sus usuarios.'.

Obviamente nuestra privacidad (y prácticamente libertad) es innegociable, ¿pero qué estrategias podemos seguir para protegerla? Empezando por las más simples, tenemos la minimización de datos. Esta táctica sugiere que solo se deben recopilar y procesar los datos estrictamente necesarios para el propósito específico de la IA. Al reducir la cantidad de datos recopilados, se disminuyen los riesgos asociados con la privacidad. También, aunque ya hemos dicho que la anonimación de los datos no es infalible, puede ayudar a proteger la privacidad al eliminar o enmascarar identificadores personales de los datos utilizados. A continuación, planes como el consentimiento informado, donde los usuarios son informados de manera clara y comprensible sobre qué datos se recopilan, cómo se utilizan y con quién se comparten, son esenciales para garantizar que los individuos tengan control sobre su información personal. Mantener un enfoque de privacidad por diseño (llamado Privacy by Design) de Protección de Datos (2023), el cual implica incorporar consideraciones de privacidad desde las primeras etapas del desarrollo de un sistema de IA, en lugar de tratarlas como un aspecto secundario o posterior, así como el permanente cifrado de los datos, que asegura que los datos no puedan ser leídos o utilizados por partes no autorizadas en caso de una brecha de seguridad, son métodos indispensables para una privacidad óptima. Como en todos los casos, el cumplimiento de las normas y regulaciones ya establecidad (y la creación de estas en los países y lugares donde no las haya, o estas sean débiles) y la realización de auditorías y evaluaciones de impacto en la privacidad, deben ser una parte integral del ciclo de vida del desarrollo y uso de la IA.

2.1.4 Justicia

No hablaré de la justicia como derecho o razón, pues eso corresponde a los juristas, ni como principio moral, que correspondería a los filósofos. Hablaré de la justicia como igualdad. La justicia implica que los sistemas de IA deben ser diseñados y operados de manera que no discriminen injustamente a ningún grupo o individuo, sin sesgos o discriminaciones indebidas. Esto implica que las decisiones automatizadas por IA deben ser imparciales, justas y no deben favorecer a ciertos grupos sobre otros injustamente. Los sistemas de IA deben ser diseñados y utilizados de tal manera que reconozcan y respeten las diferencias individuales y sociales, y que no perpetúen o exacerben desigualdades existentes.

Dado el potencial de sesgos en los algoritmos, garantizar la justicia es esencial para evitar resultados que perpetúen o agraven desigualdades existentes. La justicia es un principio crítico, debido a las siguientes razones. Para empezar, es fundamental para proteger los derechos humanos. Los sistemas de IA que discriminan o tratan a las personas de manera injusta violan derechos fundamentales, como el derecho a la igualdad, a la no discriminación y a la dignidad. Además, la IA tiene el potencial de impactar significativamente en diversas áreas de la vida, desde la justicia penal hasta la atención médica, pasando por el empleo y la educación. Si no se abordan adecuadamente, los sesgos en la IA pueden perpetuar o amplificar desigualdades sociales y económicas, como veremos a continuación.

La justicia en la IA es clave para construir y mantener la confianza pública en estas tecnologías, pues si las personas creen que los sistemas de IA son injustos o discriminatorios, es probable que desconfíen de ellos, lo que puede obstaculizar su adopción y utilidad. Las empresas y organizaciones que desarrollan y utilizan IA tienen la responsabilidad ética de asegurar que sus sistemas no causen daño ni traten a los individuos de manera injusta. Esto es especialmente importante en contextos donde las decisiones automatizadas pueden tener

un impacto significativo en la vida de las personas. Estos dos últimos puntos han sido tratados en el principio de transparencia y tratan de lo que puede también llamarse 'responsabilidad ética', siendo flagrante la simbiosis entre todos los principios.

Asegurar la justicia en la IA es un desafío complejo debido a varios factores. El primero es el sesgo en los datos. Los sistemas de IA se entrenan utilizando grandes volúmenes de datos. Si estos datos están sesgados, ya sea porque reflejan desigualdades sociales o porque contienen errores sistemáticos, el sistema de IA puede reproducir y amplificar estos sesgos. Esto puede resultar en decisiones injustas o discriminatorias, justamente lo que se pretende evitar. Como vimos con la transparencia, la opacidad de los algoritmos es un problema. Muchos sistemas de IA, especialmente aquellos basados en técnicas como el aprendizaje profundo, son opacos o funcionan como "cajas negras". Esto significa que es difícil, incluso para los desarrolladores, entender cómo se toman las decisiones. Esta falta de transparencia complica la identificación y corrección de injusticias. La falta de diversidad en los equipos que desarrollan sistemas de IA puede también contribuir a la creación de algoritmos sesgados. Si los desarrolladores no consideran diversas perspectivas y contextos, los sistemas de IA pueden no reflejar las necesidades y realidades de todas las personas, lo que lleva a decisiones injustas.

Puede parecer más o menos sencillo arreglar estos tres puntos, pero es el último el que es, para cualquier individuo, imposible de solucionar: las desigualdades estructurales. Son estas desigualdades estructurales en la sociedad, como las diferencias de acceso a la educación, la atención médica o el empleo, las que pueden ser reflejadas y exacerbadas por los sistemas de IA. Sin un enfoque consciente en la justicia, la IA puede perpetuar estas desigualdades en lugar de mitigarlas, aunque depende de una sociedad entera poder cambiarlas perpetuamente.

Son bien conocido casos donde la IA ha demostrado tener unos claros sesgos. Dos ejemplos muy conocidos son el algoritmo COMPAS y la herramienta de reclutamiento de Amazon.

Sin entrar en detalle, el primero de los mencionados (Correctional Offender Management Profiling for Alternative Sanctions) es un sistema de evaluación de riesgos utilizado en el sistema judicial de Estados Unidos para determinar la probabilidad de reincidencia de los acusados. Este mostró un sesgo racial significativo, y varios estudios[1] demostraron que el algoritmo tendía a sobreestimar el riesgo de reincidencia en personas negras y subestimar el riesgo en personas blancas. Esta discrepancia reflejaba un sesgo racial en la toma de decisiones, lo cual generó fuertes críticas y cuestionamientos sobre el uso de IA en el sistema judicial. Sin embargo, este algoritmo sigue siendo utilizado en varios estados a pesar de las evidencias.

El segundo fue una herramienta de reclutamiento desarrollada por Amazon basada en IA que fue entrenada para revisar currículums y seleccionar a los mejores candidatos. Sin embargo, el sistema mostró un sesgo contra las mujeres porque fue entrenado usando datos históricos que reflejaban una fuerza laboral predominantemente masculina, lo que llevó al sistema a favorecer candidatos masculinos y a penalizar aquellos que mencionaban palabras relacionadas con el género femenino, como "mujeres". La misma compañía descubrió estos sesgos, pero esta vez sí decidió descontinuar el proyecto al no ser capaces de eliminar completamente estos sesgos.[2]

Creo que no es necesario decir donde se ha sido ético, y donde no. Está claro que en el proceso de alcanzar un producto éticamente aceptable comportará problemas y fallos, como es el caso de Amazon, pero corresponde a un desarrollo ético el aceptar el error y

[1] Estudio realizado por ProPublica sobre el algoritmo COMPAS
[2] Noticia de VICE sobre la herramienta de Amazon

seguir trabajando en él, o descartarlo y buscar otras vías. Un estudio del MIT aseguró que, poniendo de ejemplo el caso del COMPAS, es imposible un algoritmo judicial justo.[3] Y probablemente, tirando del hilo, nos demos cuenta de que es muy difícil garantizar justicia (esta vez sí, el término filosófico), ética o cualquier concepto abstracto en cualquier algoritmo que desarrollemos. Sin embargo, como creadores de estos algoritmos, debemos trabajar por llegar a ello y conseguir el resultado más cercano a estos conceptos.

Para empezar, el primer paso para promover la justicia es nuestro segundo principio: aumentando la transparencia y la explicabilidad de los sistemas de IA, los usuarios y afectados por las decisiones de IA podrán entender cómo y por qué se tomaron estas decisiones, lo que facilita la identificación de injusticias y la rendición de cuentas, ergo, la responsabilidad. Por supuesto, a esto le seguiría el cumplimiento y creación de normas y leyes que aboguen por la defensa de esta justicia, cosa que nunca hay que dar por sentada.

Contestando a los desafíos que hemos tratado, por supuesto incluir una amplia diversidad de perspectivas en los equipos de desarrollo de IA puede ayudar a asegurar que los sistemas reflejen mejor las necesidades y realidades de diferentes grupos de personas. Esto puede incluir la contratación de desarrolladores de diversos orígenes raciales, de género y socioeconómicos, así como la participación de grupos afectados en el proceso de diseño.

Finalmente, aunque puedo pecar de insistente, reitero mi apuesta por un diseño que tenga la equidad en mente desde el principio y por la realización de auditorías de sesgos. Esto implica considerar cómo los algoritmos y los datos pueden afectar a diferentes grupos y tomar medidas para minimizar las desigualdades, incluyendo el ajuste de algoritmos para corregir sesgos conocidos o la recopilación de datos adicionales para mejorar la representatividad. Las auditorías servirían para identificar y corregir sesgos que puedan llevar a decisiones injustas, pudiendo incluir pruebas con datos diversos y representativos para asegurar que el sistema funciona de manera justa para todos los grupos. Estos dos aspectos, como podéis ver, se repiten para todos los principios, pues me parecen verdaderamente cruciales e infrecuentes.

2.1.5 Beneficios

Este tema puede parecer complejo, pues obviamente, cualquier empresa que use o desarrolle IA espera un beneficio, normalmente económico. Sin embargo, no pretendo oponerme a esto, sino más bien que no sea la único que se busque obtener. Bajo mi punto de vista, la IA debe ser desarrollada y utilizada para promover el bienestar humano, maximizando los beneficios mientras se minimizan los daños. Este principio se enfoca en el uso positivo de la IA, asegurando que sus aplicaciones contribuyan al bien común.

La beneficencia implica que el desarrollo, implementación y uso de la IA deben orientarse hacia la generación de resultados positivos para los individuos y la sociedad en su conjunto. Esto abarca no solo los beneficios directos, como mejorar la eficiencia o aumentar la precisión en diversas tareas, sino también la consideración de los impactos a largo plazo y el bienestar general de todos los afectados. La importancia de este punto se ve reflejada en distintos aspectos.

El primero de ellos es la mejora de la calidad de vida. La IA tiene un enorme potencial para mejorar la vida de las personas, desde avances en la atención médica hasta la optimización de procesos en la educación, el transporte, y otros sectores. Asegurar que estos beneficios

[3] Estudio realizado por el MIT sobre el algoritmo COMPAS

se maximicen y se distribuyan de manera justa es esencial. Por lo tanto, el segundo debe ser la reducción del daño, pues además de promover beneficios, la beneficencia en la IA también implica minimizar cualquier daño potencial que los sistemas de IA puedan causar. Esto incluye evitar decisiones erróneas, prevenir el uso malintencionado de la tecnología, y mitigar los efectos negativos no intencionados. Le siguen aspectos como el fomento de la justicia social, donde la IA puede ser una herramienta poderosa para reducir desigualdades y mejorar el acceso a recursos, siempre que se diseñe y utilice con un enfoque centrado en la beneficencia. Esto implica asegurarse de que los beneficios de la IA no estén limitados a unos pocos, sino que sean accesibles para todos. Finalmente, tenemos la construcción de confianza. Promover la beneficencia en la IA es clave para construir y mantener la confianza del público. Si las personas creen que los sistemas de IA están diseñados y utilizados con el bienestar en mente, es más probable que acepten y adopten estas tecnologías.

Por supuesto, este principio no está exento de desafíos y riesgos que superar. El primero es la evaluación de impactos, es decir, determinar cuáles son los beneficios y riesgos de un sistema de IA. Esto puede ser complejo, dado que algunos impactos pueden ser difíciles de predecir o cuantificar, especialmente cuando se trata de efectos a largo plazo o indirectos. A continuación, el equilibrio entre beneficio y daño puede ser difícil de encontrar. A veces, las decisiones sobre el uso de la IA implican sopesar beneficios y daños potenciales, especialmente en situaciones donde los beneficios para algunos pueden implicar costos o riesgos para otros. Algo importante a tener en cuenta son los sesgos en la definición de bienestar: los desarrolladores y diseñadores de IA pueden tener diferentes ideas sobre lo que constituye el bienestar, basadas en sus propios valores, culturas y experiencias. Esto puede llevar a decisiones que, aunque bien intencionadas, no reflejan adecuadamente las necesidades o deseos de todas las personas afectadas. Para acabar, tenemos el problema de equidad. La distribución de los beneficios de la IA puede no ser equitativa. Algunos grupos pueden beneficiarse desproporcionadamente, mientras que otros, especialmente aquellos con menos acceso a tecnología o recursos, pueden quedar atrás o incluso ser perjudicados.

Como caso de uso, alejándonos de lo negativo como en los principios anteriores, y teniendo en cuenta lo esperanzador que resulta este principio, he recopilado tres casos que hablan de la IA de manera positiva o, como mínimo, alentadora. El primero es de IMF Blog, donde se trata un tema tan común como es el del remplazo del hombre ante la IA[4]. En este se presentan datos, asegurando que la IA cambiará la economía mundial, afectando a un 40% de trabajos, pero manteniendo la fe en que, si es tratada de la manera correcta, esta revolución beneficiaría mucho a la sociedad. El siguiente es un artículo el World Economic Forum titulado 'Cómo la IA puede marcar el comienzo de una economía que dé prioridad a las personas'[5]. Por si el titulo no es lo suficientemente prometedor, el artículo expone, enumerando ciertas necesidades, lo siguiente:

"La inteligencia artificial ya está mejorando nuestra economía y bienestar de múltiples maneras, incluyendo una mayor productividad laboral, diagnósticos de salud más precisos, nuevas formas de descubrimiento de fármacos, y una mejor toma de decisiones, por nombrar solo algunas. Pero esto es solo el comienzo.

Estoy convencido de que las empresas y los emprendedores podrán aprovechar la tecnología de la IA de nuevas maneras que desbloquearán un potencial humano no explotado y

[4]AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity.
[5]How AI can usher in an economy that puts people first

beneficios a una escala sin precedentes. Solo necesitamos abordar la innovación con las motivaciones correctas y mecanismos de responsabilidad adecuados para asegurar que esa promesa se convierta en realidad."

Podemos comprobar que esta rotunda afirmación se alinea perfectamente con nuestros 'principios éticos'. Finalmente tenemos una noticia de Euronews donde se habla sobre la convención anual del World Economic Forum en Davos[6]. Este año, el tema central fue la IA y el, citando literalmente, "rol positivo que podría desempeñar si se encontrara la regulación adecuada".

"Para Kathy Bloomgarden, directora ejecutiva de Ruder Finn, el sentimiento es mucho más optimista de lo que hubiera pensado antes de venir a Davos.

"La gente realmente está viendo la inteligencia artificial y la nueva tecnología e innovación como una forma de resolver muchos de nuestros problemas", dijo ella.

"Ya sea en problemas de tecnología climática, ya sea como un motor de crecimiento, para la prosperidad económica, para unir a las personas. Realmente creo que es más optimista basado en la innovación tecnológica que podríamos esperar ver"."

¿Qué estrategias podemos seguir para cumplir este principio? Bien, como podemos imaginar, para ello necesitamos antes cumplir el principio de transparencia y de responsabilidad. El primero porque ayuda a los usuarios a entender cómo funcionan y a confiar en que las decisiones se toman en su mejor interés. La explicabilidad también permite la identificación de posibles problemas antes de que causen daño. El segundo porque los gobiernos y organismos internacionales pueden jugar un papel crucial al establecer regulaciones que promuevan la beneficencia en la IA. Esto incluye la creación de estándares éticos que guíen el desarrollo y la implementación de la tecnología, así como la supervisión para asegurar el cumplimiento.

Como estrategias propias, destacaría la evaluación de impacto ético, realizando evaluaciones durante todo el ciclo de vida de un sistema de IA para ayudar a identificar y mitigar posibles daños. Esto incluye considerar no solo los impactos inmediatos, sino también los efectos a largo plazo y en diferentes contextos. También es importante la capacitación y educación, pues es vital educar y capacitar a los desarrolladores de IA en ética y beneficencia, asegurando que comprendan las implicaciones de sus decisiones y estén preparados para diseñar sistemas que promuevan el bienestar. Asimismo, los usuarios también deben ser educados sobre cómo interactuar con la IA de manera que maximice los beneficios y minimice los riesgos.

Finalmente, como en los anteriores puntos, tener claras a la hora del diseño de la IA las necesidades y el bienestar del usuario final es esencial. Esto implica involucrar a los usuarios en el proceso de diseño, escuchar sus preocupaciones y asegurarse de que la IA se alinee con sus intereses y valores.

2.1.6 Autonomía

Finalmente nos encontramos el ultimo de nuestros principios. Este está mas orientado a las personas, a la sociedad. Hemos empezado este apartado hablando del miedo, del respeto que esta le tiene a este nuevo mundo, a esta nueva revolución. Un miedo basado en la sustitución, en la coerción sobre la libertad misma, en una evolución tan rápida que no da tiempo a adaptarse. Pensando en ello, llegué a este principio, algo más teórico, que creo que puede ayudar a desarraigar ese miedo: el de autonomía. Me explico:

[6] A positive role for AI? Takeaways from the 2024 World Economic Forum in Davos

Siendo realistas con el punto en el que estamos, las personas deben tener el control sobre las decisiones que les afectan, lo que incluye la posibilidad de optar por no ser afectados por sistemas de IA. Respetar la autonomía es fundamental para proteger la dignidad y los derechos individuales en la era de la IA.

La autonomía se entiende comúnmente como el derecho de los individuos a gobernarse a sí mismos, a tomar decisiones personales que reflejen sus valores, creencias y deseos. En la ética de la IA, la autonomía se extiende a cómo estas tecnologías pueden influir o, en algunos casos, erosionar la capacidad de las personas para tomar decisiones autónomas. Por ejemplo, en el ámbito de la salud, los sistemas de IA utilizados para diagnosticar o recomendar tratamientos deben diseñarse de manera que apoyen, en lugar de reemplazar, la toma de decisiones del paciente y del profesional médico. Esto es esencial para garantizar que los pacientes mantengan el control sobre sus propias decisiones de salud.

Uno de los principales desafíos que plantea la IA a la autonomía es la capacidad de estas tecnologías para influir en las decisiones humanas de maneras que no siempre son transparentes o conscientes. Algoritmos de recomendación, publicidad dirigida y perfiles personalizados pueden afectar sutilmente las elecciones de las personas, limitando su capacidad para tomar decisiones realmente libres y autónomas. Además, el uso de IA en la automatización del trabajo puede llevar a una reducción de la autonomía en el lugar de trabajo. Los trabajadores pueden sentirse cada vez más controlados por sistemas automatizados que dictan su ritmo de trabajo, supervisan su desempeño y limitan su capacidad para tomar decisiones autónomas en su labor diaria.

El respeto por la autonomía está profundamente vinculado con los derechos humanos. La Declaración Universal de los Derechos Humanos, por ejemplo, subraya la importancia de la libertad de pensamiento, conciencia y religión, todas ellas manifestaciones de la autonomía individual. En este sentido, la IA debe diseñarse y utilizarse de manera que respete y promueva estos derechos, permitiendo que los individuos mantengan el control sobre sus propias vidas y decisiones.

A medida que la IA continúa avanzando, el principio de autonomía se enfrentará a nuevos desafíos y oportunidades. El desarrollo de IA con capacidades cada vez más avanzadas plantea la posibilidad de que estas tecnologías puedan tomar decisiones más complejas en nombre de los humanos. Esto podría aliviar la carga de la toma de decisiones en situaciones complejas, pero también podría conducir a una dependencia excesiva de la tecnología y una disminución de la autonomía humana. Es esencial que la investigación y el desarrollo de la IA continúen centrados en cómo mantener y fortalecer la autonomía humana, asegurando que estas tecnologías sean herramientas que empoderen a las personas en lugar de reemplazarlas.

Para ver porque es importante esta autonomía, fijémonos en Google DeepMind, concretamente en un artículo publicado por ellos mismos sobre 'la ética de los asistentes de IA avanzads'[7]. DeepMind es una compañía inglesa de investigación y desarrollo de IA. La empresa ha creado varios desarrollos, como una red neuronal que aprende cómo jugar a los videojuegos de una manera similar a la de los seres humanos, así como una máquina de Turing Neural, o una red neuronal que puede ser capaz de acceder a una memoria externa como una máquina convencional de Turing, lo que resulta en una computadora que imita la memoria a corto plazo del cerebro humano. En el artículo se nos habla de un futuro ultra conectado mediante estos asistentes avanzados, creando "una nueva fase de interacción hu-

[7] The ethics of advanced AI assistants

mana con IA". Hablan sobre la influencia de estas IA y sus ventajas, y de los límites que hay que marcar por sus inconvenientes. Pero es un apartado en concreto que quiero remarcar:

"Capaces de comunicarse fluidamente en lenguaje natural, los textos escritos y las voces de los asistentes avanzados de IA pueden volverse difíciles de distinguir de los de los humanos.

Este desarrollo plantea un conjunto complejo de preguntas sobre confianza, privacidad, antropomorfismo y las relaciones humanas apropiadas con la IA: ¿Cómo podemos asegurarnos de que los usuarios puedan identificar de manera confiable a los asistentes de IA y mantener el control sobre sus interacciones con ellos? ¿Qué se puede hacer para garantizar que los usuarios no sean indebidamente influenciados o engañados con el tiempo?

Se deben implementar salvaguardas, como las relacionadas con la privacidad, para abordar estos riesgos. Es importante que las relaciones de las personas con los asistentes de IA preserven la autonomía del usuario, apoyen su capacidad para desarrollarse y no dependan de manera emocional o material."

Finalizada la explicación de los principios y con estas palabras en la cabeza, pasemos al siguiente punto.

2.2 Enfoques Filosóficos

La ética aplicada a la inteligencia artificial no surge en un vacío; está profundamente arraigada en corrientes filosóficas que han debatido la naturaleza del bien, la moralidad y la justicia durante siglos. En este apartado, examinaremos la visión de un filósofo al que tuve la suerte de escuchar y veremos cómo tres enfoques filosóficos principales —el utilitarismo, la deontología y la ética de la virtud— se aplican a la IA, y cómo cada uno puede ofrecer perspectivas valiosas (y a veces contradictorias) sobre los dilemas éticos que enfrenta esta tecnología.

2.2.1 Introducción

El 10 de mayo de 2024 se realizó la XIX Jornada de Comités de Bioética de la Comunidad Valenciana en el Hospital General Universitario de Castellón, y tuve la fortuna de acudir de asistente. El tema de esta jornada era 'Inteligencia Artificial: Ética, derecho y salud'. Fue aquí donde tuve el placer de escuchar al Prof. Vicente Bellvert, Catedrático de Filosofía del Derecho y Filosofía Política y Director del Departamento de Filosofía del Derecho y de la Política de la Universitat de València. A partir de su enfoque como filósofo me gustaría empezar a tratar la ética vista desde el punto de uno.

Según Vicente, la IA parte de una premisa errónea, pues esta debe ser neutral y la valoración ética depende de su uso. Esta falacia, según él, se debe a que la tecnología esta moralmente cargada en diseño. Sus argumentos son los siguientes, basados en los niveles de normatividad de una IA: el primer nivel es la ética del usuario, a quién muchas veces se le impone como usarla. El segundo nivel es la regulación jurídica de la actividad, limitada completamente por las big-tech. A continuación, tenemos la normatividad social y la tendencia del ser humano a ceder ante los efectos negativos a cambio de los 'positivos'. Finalmente nos encontramos con el diseño de la tecnología, que siempre busca algún objetivo concreto.

Para el profesor, el relato y la finalidad tienen simplemente un barniz ético. Recalcando su afirmación de que la IA es una gran falacia, pregunta: ¿Acaso no depende el avance del

propio ser humano? ¿Debemos optar por que no se puede poner puertas al campo? o ¿por qué sin normas no hay humanidad? Además, el enfoque filosófico es muy importante debido a los efectos colaterales de la IA. ¿Son una contingencia inevitable que debe ser regulada, con prevención y compensación? ¿O son signos de un futuro inhumanista, desigual y alienado, debiendo ser la IA revisada integralmente en su diseño?

2.2.2 Modelos, gobiernos y principios

Dejando estas preguntas en el aire, a las que volveremos más tarde, Bellvert habla de los distintos modelos de IA que encuentra. Estos son el oriental, el anglosajón y el europeo. El oriental presenta un desarrollo competitivo orientado completamente al control social. El anglosajón, una regulación mínima a cambio de una máxima libertad, fomentando la iniciativa privada. El europeo, sin embargo, apuesta por una regulación garante de derechos humanos sin limitar el desarrollo.

Esto nos devuelve a los niveles de normatividad que he mencionado antes. Incluso con el mismo objetivo, el desarrollo de IA, hay diferentes modelos que buscan diferentes objetivos por diferentes medios. Si son mejores o peores no lo trataremos, al menos de momento, pero nos daremos cuenta de que es recurrente encontrarnos con este problema: diferentes opiniones, caminos y fines.

En cuanto a tipos de gobiernos de IA, el profesor distingue cinco, siendo el primero los principios. Estos son lineas directivas, no juzgan, simplemente dirigen. El segundo son los conceptos, que presentan una inclusión y una exclusión. Los terceros, las leyes y las políticas, sí permiten los juicios. El cuarto son los mecanismos de autorregulación, que son decisiones propias, de individuos o colectivos. Finalmente, el quinto, son las certificaciones, que dependen al completo de un veredicto externo.

Como hemos visto en el principio de responsabilidad, hemos pasado de las directivas y conceptos al inicio de una ley (esperemos que sea el principio de muchas). Aunque esto es algo positivo como sociedad, debemos tener en cuenta que también significa tomar un camino, el que elijan los legisladores como correcto, y apartarnos del resto.

A continuación, Vicente enumeraba los principios de la Organización para la Cooperación y Desarrollo Económicos (OCDE) en cuanto a la IA. Como dije al principio de este marco teórico, son estos en los que me he basado para la propuesta de los principios éticos anteriores, así que no profundizaré en ellos. Sus críticas iban dirigidas a tres puntos concretos. En el primero, donde la IA debe estar al servicio de las personas y del planeta (nuestra beneficencia), duda de su posibilidad. De hecho, lo considera ambiguo y paradójico. En cuanto a la transparencia y la responsabilidad, considera que es demasiado fácil de decir, pero demasiado difícil de cumplir.

Bueno, hemos tratado ya estos temas y sin duda que podemos estar de acuerdo con su última afirmación. Sin duda, cualquier principio ético es fácil de decir, pero su realización es algo realmente complejo. Es por esto justamente que debemos empezar ya a tratar y trabajar estos principios. También son comprensibles sus dudas sobre la beneficencia, alineadas con los desafíos expuestos sobre el principio, pues, como pretendo recalcar en este punto, cada individuo, colectivo, empresa o país tiene una opinión, que puede oponerse a la de al lado. Esto, sin duda, puede resultar imposible o como mínimo paradójico.

2.2.3 Consideraciones colaterales

Nombrando a Günther Anders, filósofo germano-austríaco autor de la obra 'La obsolescencia del hombre', y haciendo mención a la misma, el profesor habla de unas consideraciones colaterales de la IA que considero muy interesantes.

La primera es el 'utopismo invertido'. Siendo el utopismo la tendencia a planes, proyectos, doctrinas o sistemas ideales que parecen de muy difícil realización, Vicente considera que en el aspecto de la IA le hemos dado la vuelta a la situación: ahora tenemos los medios para llegar a lo que antes parecía inalcanzable, pero no tenemos un fin real. En mi opinión, aunque no tenemos los medios para llegar donde queremos llegar (hablando de una IA completamente ética), sí es cierto que el punto en el que estamos parecía imposible hace años, y que aunque en este ensayo tenemos un claro fin, que es el máximo acercamiento posible a una IA ética, realmente en cuanto a la IA en general no existe un fin concreto, consensuado ni general. Este punto me parece realmente cautivador, pues sugiere la posibilidad de que, intentando abarcar tanto a la vez, nos hayamos alejado del camino correcto.

El segundo es la 'vergüenza prometeica'. Esto se refiere al sentimiento de inferioridad o vergüenza que los seres humanos experimentan al comparar sus capacidades con las de las máquinas o tecnologías que han creado. En términos más concretos, la vergüenza prometeica implica que el ser humano se siente cada vez más inadecuado frente a las máquinas y tecnologías avanzadas, que superan en muchos aspectos las habilidades humanas, ya sea en términos de precisión, velocidad, resistencia o capacidad de procesamiento. El ser humano, en su capacidad creadora, ha dado origen a máquinas tan poderosas que, irónicamente, lo hacen sentir limitado y obsoleto. Creo que este punto es completamente acertado. Al fin y al cabo, la sustitución es ahora mismo el mayor miedo del ciudadano de a pie, y aunque hasta ahora se ha hablado del tema desde el punto de vista del empleo, creo que es crucial abordarlo desde el punto de vista psicológico y social, como así lo hace el profesor Bellvert, y Anders antes que él.

El tercero, y el más crítico bajo mi punto de vista en el apartado ético, es el problema de la distancia entre el agente y su invención, que impide al primero hacerse cargo del segundo, por lo tanto, impidiendo, según el profesor, un trabajo responsable. Pienso que con esto se refiere a la desconexión ética y moral que surge cuando los creadores de una tecnología no son responsables directos de sus consecuencias. Este problema se agrava cuando la creación es usada de formas que los inventores no anticiparon o no deseaban, lo que dificulta que el agente (quien desarrolla la tecnología) se haga cargo de los efectos de su invención.

Un claro ejemplo de este fenómeno es el Proyecto Manhattan, el esfuerzo científico durante la Segunda Guerra Mundial que culminó en la creación de la bomba atómica. Los científicos involucrados, entre los que se encontraba Robert Oppenheimer, trabajaron inicialmente con el objetivo de desarrollar un arma para detener a la Alemania nazi, que se pensaba también estaba desarrollando armamento nuclear. Sin embargo, una vez la bomba estuvo lista, fue utilizada contra Japón, un país que ya estaba prácticamente derrotado en el conflicto bélico.

El problema ético surge cuando los científicos, al haber entregado la tecnología a manos de los militares y los líderes políticos, perdieron control sobre su uso. Muchos de ellos no deseaban que la bomba fuera usada contra civiles, y después de los bombardeos de Hiroshima y Nagasaki, muchos expresaron arrepentimiento y angustia por las consecuencias devastadoras de su trabajo.

De hecho, Oppenheimer, quien fue el director científico del proyecto, expresó sentimientos

de culpa después de los ataques. En una famosa cita, recordó las palabras del texto hindú *Bhagavad-gita* tras la detonación de la primera bomba: "Ahora me he convertido en la muerte, el destructor de mundos". Este arrepentimiento refleja el problema de la distancia: aunque Oppenheimer y otros científicos ayudaron a crear la bomba, no pudieron controlar cómo se usó ni las consecuencias éticas de sus decisiones.

Con este caso trato de ejemplificar un dilema común en el desarrollo tecnológico: los inventores crean herramientas poderosas, pero no siempre controlan o son conscientes de los usos posteriores de esas invenciones. El problema de la distancia entre el agente e invención en este contexto es que los científicos no asumieron plenamente la responsabilidad por las decisiones políticas y militares que guiaron el uso de su creación. A medida que las tecnologías se vuelven más complejas y sus efectos más globales, esta distancia tiende a crecer, planteando preguntas éticas más profundas sobre la responsabilidad de los creadores.

Este dilema puede aplicarse perfectamente a la IA y nos obliga a reflexionar sobre la necesidad de un mayor control y responsabilidad ética en el desarrollo de tecnologías avanzadas. Como en el caso de Oppenheimer, los creadores pueden no prever el alcance de los efectos de sus inventos, pero eso no exime a la sociedad de enfrentarse a las implicaciones morales que estas tecnologías traen consigo.

Como dos últimas consideraciones, que trataré de forma escueta, trata hasta qué punto es realista u utópica la capacidad de darnos tiempo. Obviamente necesitamos darnos tiempo, tanto porque el fin es muy lejano a nuestra realidad actual, como porque es un tema tan sensible y crítico que tratarlo de forma apresurada no traerá nada positivo. Sin embargo, ¿hasta qué punto tenemos ese tiempo? Además, ¿puede el avance ser detenido o ralentizado una vez ha comenzado? La siguiente y última consideración trata del recelo al uso de la palabra 'inteligencia' en la IA. Creo que esto es un tema menor, pero es cierto que cuando hablamos de inteligencia en los seres humanos, nos referimos a una serie de capacidades como el razonamiento, la comprensión, la creatividad, la capacidad de aprender de manera autónoma, y la adaptación a contextos cambiantes. La inteligencia artificial, por otro lado, es un término que se refiere a sistemas o máquinas capaces de realizar tareas que típicamente requieren inteligencia humana, como reconocimiento de patrones, resolución de problemas, o procesamiento de lenguaje natural. Sin embargo, la IA no posee conciencia ni comprensión profunda. Los algoritmos que componen las IA no tienen intenciones ni emociones; simplemente procesan datos y producen resultados siguiendo patrones predefinidos. Lo que hace la IA es "simular" ciertos aspectos de la inteligencia humana, pero no replicarla de manera completa.

2.2.4 Conclusiones

El Reglamento Europeo de para IA (IA)[8], del que ya hemos hablado, fue el último punto del que habló el profesor. Sobre él, quiso destacar que este reglamento trata de limitar los riesgos de la IA, y define tres destinatarios distintos: los proveedores, los usuarios y los afectados por la IA (sin ser proveedor ni usuario). También divide los riesgos en cuatro niveles: nulo (no presenta riesgo), bajo (poco riesgo), alto (el nivel que se centra en regular el reglamento) e inaceptable (destacando el control y crédito social).

Como valoración de este, opinó que el impacto final de una IA depende del desarrollo normativo (el reglamento) y del marco de supervisión, siendo obviamente el primero crucial.

[8] REIA

Sin embargo, lo calificó de prolijo, complejo y excesivamente técnico, alejado por completo del ciudadano medio. Cree que este tiene un alcance muy limitado y no trata los verdaderos peligros de un desarrollo acelerado de IA, pues se basa en una regulación paulatina del fenómeno.

Fuera del REIA, concluyó que por supuesto la IA es muy beneficiosa, pero duda de su verdadera finalidad. ¿Es esta el abandono del mundo real? ¿O es hacer nuestra realidad más vivible? También mostró su preocupación ante la ambigüedad que presenta el trato de los datos personales, la 'gasolina' que necesita la IA para funcionar. ¿Son estos bienes personales, o son públicos? Para él, es obvio que la IA necesita un acuerdo universal, pero al mismo tiempo, es un hecho que esta se encuentra a merced del monopolio de las big-tech, volviendo a la ambigüedad: este acuerdo, aun siendo necesario, es probablemente imposible. Con estas palabras zanjó su ponencia.

Me parece muy interesante analizar el punto de vista del profesor Bellvert como introducción a los siguientes enfoques, pero no he querido titularlo (como haré con el resto), para que no prejuzgaseis sus palabras antes de leerlas y asimilarlas. Sin duda, este me parece un enfoque pesimista ante la IA, pero por supuesto necesario, como lo son todos, y merecedor de atención y trabajo por parte de los interesados en la parte más ética y filosófica de ella. Parte que, llegados a este punto, espero que se entienda como crucial: necesitamos pensar en el impacto que esta revolución causará en el ser humano como especie y sobre todo como sociedad. No cabe duda de que presenta realidades innegables y preguntas sumamente interesantes que necesitan ser respondidas.

2.2.5 Utilitarismo

El utilitarismo, popularizado por filósofos como Jeremy Bentham y John Stuart Mill, se basa en la idea de que la moralidad de una acción se determina por su capacidad para maximizar el bienestar o la felicidad y minimizar el sufrimiento. En el contexto de la IA, el utilitarismo aboga por el desarrollo y la implementación de tecnologías que generen el mayor beneficio para el mayor número de personas. Para este enfoque filosófico, la IA se percibe como una herramienta poderosa para incrementar la eficiencia y reducir costos en diversas áreas, lo que puede llevar a resultados beneficiosos para la sociedad en su conjunto. Sin embargo, este enfoque también presenta desafíos éticos profundos, especialmente en cómo se distribuyen los beneficios y los riesgos entre diferentes grupos.

Para que resulte fácil de entender, pondré un ejemplo. Imaginemos una escuela que decide usar IA para personalizar la enseñanza de sus estudiantes. Cada alumno recibe una experiencia educativa adaptada a su ritmo de aprendizaje, sus habilidades y sus áreas de mejora. Los estudiantes que tienen dificultades en matemáticas, por ejemplo, recibirán más ejercicios adaptados a su nivel para que puedan mejorar, mientras que aquellos que van avanzados podrán recibir contenido más avanzado.

Desde un enfoque utilitarista, el uso de esta IA es beneficioso porque mejora el aprendizaje de la mayoría de los estudiantes, maximizando el bienestar general en la clase. Los estudiantes que normalmente podrían quedar rezagados tienen la oportunidad de ponerse al día, y los que van mejor no se aburren ni pierden el interés. En general, el nivel educativo mejora para la mayoría de los estudiantes.

Sin embargo, consideremos las consecuencias negativas. Por ejemplo: ¿Qué pasa con los profesores? Si la IA cumple su función, podrían sentirse menos necesarios, perder parte de

su rol de guía en el proceso educativo o incluso su empleo. También puede surgir una brecha digital. Si algunas escuelas no pueden permitirse este tipo de tecnología, los estudiantes que no tengan acceso a IA podrían quedarse en desventaja, lo que podría aumentar la desigualdad educativa.

El utilitarismo aquí justificaría el uso de la IA en educación porque mejora el aprendizaje de la mayoría de los estudiantes. Sin embargo, ¿es verdaderamente ético?

2.2.5.1 Maximización del Bienestar

El utilitarismo es una teoría consecuencialista, lo que significa que la moralidad de una acción se juzga exclusivamente en función de sus resultados. En el contexto de la IA, esto implica que la validez ética de un sistema de IA se determinará según sus consecuencias para el bienestar general.

Como en el ejemplo, si un algoritmo de IA en el ámbito de la salud logra diagnosticar enfermedades con mayor precisión y rapidez que un médico humano, salvando vidas y mejorando la calidad de vida de los pacientes, este uso de la IA sería considerado moralmente correcto desde una perspectiva utilitarista. Lo importante no es cómo o por qué funciona el algoritmo, sino cuál es el resultado neto en términos de bienestar.

Desde una perspectiva utilitarista, el desarrollo de IA debe enfocarse en maximizar el bienestar social. Esto incluye la creación de tecnologías que mejoren la calidad de vida de las personas, como los diagnósticos médicos precisos, la automatización de tareas repetitivas que liberen tiempo para el trabajo creativo, y las innovaciones en transporte o logística. Un ejemplo claro de esta maximización se puede observar en los avances en salud, donde la IA ya ha demostrado ser capaz de identificar enfermedades con mayor precisión que los seres humanos, lo que puede salvar millones de vidas a largo plazo. Además, en el ámbito económico, la IA podría reducir la desigualdad social mediante la automatización de procesos que tradicionalmente han requerido grandes cantidades de mano de obra. Esto permitiría que las personas dedicaran más tiempo a actividades más creativas y con un mayor valor añadido. Desde esta perspectiva, el desarrollo de IA debe centrarse en resolver problemas a gran escala, como la pobreza, la salud, el acceso a la educación y el medio ambiente, siempre con el objetivo de beneficiar al mayor número de personas posible.

2.2.5.2 Distribución de Beneficios y Daños

Una crítica importante al utilitarismo aplicado a la IA surge cuando analizamos cómo se distribuyen los beneficios y los costos. El utilitarismo clásico no presta tanta atención a quién se beneficia, sino a que el bienestar general se maximice. Esto puede dar lugar a situaciones en las que una mayoría se beneficia enormemente a expensas de una minoría, lo que plantea cuestiones éticas sobre la equidad y la justicia social.

En el desarrollo y la implementación de IA, podría surgir el problema de que los beneficios de estas tecnologías no se distribuyen de manera equitativa entre todas las personas. Las grandes empresas tecnológicas, por ejemplo, pueden ser las principales beneficiarias de los avances en IA, mientras que los trabajadores en sectores menos cualificados podrían verse desplazados por la automatización. Un enfoque utilitarista podría ver la automatización como una mejora general del bienestar al aumentar la eficiencia económica, pero ignora los perjuicios que puede causar a las personas que pierden sus empleos. Por ejemplo, si un sistema de

IA en una fábrica mejora la producción al reducir la necesidad de trabajadores humanos, la productividad total y el bienestar económico podrían aumentar, pero los trabajadores despedidos sufrirían. Para un utilitarista, el enfoque correcto sería intentar compensar estos perjuicios, por ejemplo, mediante políticas de reconversión laboral o redistribución de los beneficios económicos para asegurar que todos puedan beneficiarse del progreso tecnológico.

El utilitarismo también se enfrenta al problema de la brecha digital, donde el acceso desigual a la tecnología genera nuevas formas de desigualdad. Las personas que no tienen acceso a la IA o a los recursos tecnológicos pueden quedarse atrás, creando disparidades aún mayores entre los grupos privilegiados y los marginados. Un enfoque utilitarista debería buscar soluciones que minimicen esta brecha y aseguren que el mayor número posible de personas pueda beneficiarse de los avances tecnológicos, ya sea mediante políticas de inclusión tecnológica o el acceso equitativo a los servicios proporcionados por la IA.

2.2.5.3 Evaluación de Riesgos

Un análisis utilitarista aplicado a la IA también implica la evaluación rigurosa de los riesgos y beneficios asociados con su uso. A medida que las IA se vuelven más avanzadas y penetran en más aspectos de la vida cotidiana, se hace necesario sopesar los riesgos potenciales frente a los beneficios.

El utilitarismo, por su énfasis en el bienestar general, a menudo se enfrenta a críticas por ignorar los derechos individuales. En situaciones en las que maximizar el bienestar colectivo implica sacrificar los derechos de algunos individuos, el utilitarismo puede justificar decisiones éticamente cuestionables.

Un ejemplo de este conflicto surge en el uso de la IA para la vigilancia masiva. Los sistemas de vigilancia basados en IA, como el reconocimiento facial, pueden mejorar la seguridad pública y reducir el crimen, lo que beneficiaría a la mayoría de la sociedad. Sin embargo, esta mejora del bienestar general podría lograrse a expensas de los derechos a la privacidad de los individuos, quienes se ven sometidos a un control constante. Desde una perspectiva utilitarista, se podría argumentar que la pérdida de privacidad de unos pocos está justificada si el resultado es una mayor seguridad para el mayor número de personas. Sin embargo, este enfoque puede ser problemático, ya que minimiza la importancia de los derechos fundamentales de los individuos en favor del bienestar colectivo.

Otros ejemplos de esta evaluación es la automatización del trabajo. Si bien la automatización puede aumentar la productividad, reducir costos y hacer que los productos y servicios sean más accesibles, también puede llevar a la pérdida de empleos para millones de personas. Un análisis utilitarista justificaría la automatización si los beneficios económicos globales superan los perjuicios individuales, pero solo si se toman medidas para mitigar los efectos negativos sobre los trabajadores desplazados. Esto podría implicar políticas de reentrenamiento, Seguridad Social y otras formas de protección para los trabajadores, de manera que el progreso tecnológico no conduzca a una mayor desigualdad o sufrimiento social.

El desafío para el utilitarismo aplicado a la IA es encontrar un equilibrio entre maximizar el bienestar y proteger los derechos individuales.

2.2.5.4 Ejemplo y conclusión

Un ejemplo más radical sobre como este enfoque puede influir en el desarrollo de la IA lo vemos en la discusión sobre los vehículos autónomos. En situaciones de emergencia, los algoritmos deben tomar decisiones sobre a quién proteger en un accidente inevitable. Desde un punto de vista utilitarista, el algoritmo debería optar por la opción que minimice el daño total, pero esta valoración puede plantear dilemas éticos profundos, como decidir entre la vida de los pasajeros y la de los peatones. ¿Comprarías un vehículo autónomo a sabiendas de que, en caso de tener que tomar una decisión, tienes altas probabilidades de ser 'sacrificado'? ¿Irías seguro por la calle o por la carretera sabiendo que los vehículos de tu alrededor pueden decidir que eres el 'mal menor'?

En resumen, el utilitarismo aplicado a la IA ofrece una forma útil de analizar las decisiones éticas basadas en las consecuencias de las acciones. Este enfoque se enfoca en maximizar el bienestar general, pero también plantea desafíos importantes en cuanto a la distribución de los beneficios, la protección de los derechos individuales y la evaluación de riesgos. Si bien el utilitarismo puede ser una herramienta poderosa para guiar el desarrollo ético de la IA, debería, quizá, complementarse con otras teorías éticas para garantizar un enfoque más equilibrado y justo.

2.2.6 Deontología

La deontología es una teoría ética que se enfoca en la moralidad de las acciones basándose en principios y normas universales, independientemente de las consecuencias. En contrapunto del utilitarismo, que evalúa la moralidad según los resultados, la deontología establece que ciertas acciones son correctas o incorrectas en sí mismas, independientemente de si conducen al mayor bienestar general. Este enfoque ético se asocia con el filósofo alemán Immanuel Kant, quien postulaba que los individuos deben actuar de acuerdo con deberes morales inquebrantables, como el respeto por los derechos y la dignidad de los demás.

Cuando aplicamos este enfoque filosófico a la IA, la deontología nos proporciona un marco diferente para abordar los desafíos éticos, uno que enfatiza la protección de los derechos individuales, la justicia y el respeto a la autonomía de las personas, independientemente de los beneficios que la tecnología pueda generar en su conjunto.

2.2.6.1 Principios fundamentales

La deontología sostiene que ciertos principios éticos son absolutos y deben ser respetados en todas las circunstancias. Estos principios no dependen de si los resultados finales son positivos o negativos; en cambio, se centran en que las acciones mismas sean conformes a la moralidad objetiva. En el contexto de la IA, este enfoque ético nos invita a reflexionar sobre las normas y reglas que deben guiar el desarrollo, el despliegue y el uso de estas tecnologías, independientemente de los beneficios que puedan ofrecer.

Uno de los conceptos centrales en la deontología kantiana es el imperativo categórico, que establece que debemos actuar solo de acuerdo con máximas que podamos desear como leyes universales. En el contexto de la IA, esto significa que los desarrolladores, programadores y usuarios deben preguntarse: ¿sería correcto que todas las personas actuaran de la misma manera en relación con esta tecnología? Por ejemplo, si se está desarrollando una IA de

reconocimiento facial para la vigilancia, una pregunta deontológica relevante sería: ¿Es moralmente aceptable implementar una tecnología de vigilancia que podría violar la privacidad de las personas? La deontología no juzga esta pregunta en función de si la tecnología reduce el crimen o mejora la seguridad, sino si el acto de vigilar constantemente a las personas sin su consentimiento es moralmente correcto según una norma universal. Para un deontólogo, la invasión de la privacidad puede ser intrínsecamente inmoral, incluso si mejora la seguridad, porque viola el principio moral fundamental de respeto a la autonomía y la privacidad de los individuos.

Esto lleva también a la protección de los derechos fundamentales de los individuos. Los sistemas de IA deben ser diseñados y utilizados de tal manera que respeten principios inmutables, como el derecho a la privacidad, el respeto a la autonomía, la dignidad humana y la equidad. En un ejemplo parecido al anterior, un sistema de IA utilizado para la toma de decisiones en el ámbito legal debe respetar el derecho de los acusados a un juicio justo, independientemente de si el uso de la IA agiliza los procesos judiciales. Desde una perspectiva deontológica, lo fundamental no es la eficiencia del sistema, sino que la justicia y los derechos de los implicados sean preservados en cada etapa del proceso. Cualquier sistema de IA que trate a las personas de manera injusta o desigual violaría los principios de la deontología, incluso si produce beneficios prácticos a nivel colectivo.

2.2.6.2 Deber moral

Uno de los aspectos más importantes de la deontología aplicada al contexto de la IA es la idea de que los desarrolladores y diseñadores tienen deberes morales claros con respecto a cómo sus tecnologías se utilizan y a los principios que deben seguir en su creación. A diferencia de un enfoque que solo considera las consecuencias, la deontología afirma que los diseñadores tienen la responsabilidad ética de asegurarse de que sus tecnologías respeten los derechos humanos y no perjudiquen a las personas de ninguna manera.

Desde una perspectiva deontológica, los desarrolladores de IA tienen el deber moral de prever los posibles usos de su tecnología y evitar la creación de sistemas que puedan ser utilizados de manera inmoral, sin importar cuán útiles puedan parecer en otros contextos. Esto significa que las creaciones de estos desarrolladores no deben violar los derechos de las personas, como su privacidad, autonomía o igualdad ante la ley, incluso si esa tecnología tiene el potencial de mejorar la eficiencia o generar beneficios para la sociedad. Un ejemplo podría ser la creación de IA para la recolección masiva de datos sin el consentimiento informado de las personas. Aunque esta tecnología podría proporcionar datos valiosos para la investigación médica o la mejora de productos, el deber moral del desarrollador es proteger la privacidad y la dignidad de los individuos. Según la deontología, el fin no justifica los medios, por lo que la invasión de la privacidad sería considerada incorrecta, incluso si los datos recolectados podrían salvar vidas o mejorar la calidad de vida de muchas personas.

La deontología también pone un gran énfasis en la justicia y la equidad. Es crucial asegurarse de que nuestras IAs no discriminen o perpetúen desigualdades. Un sistema que discrimine a ciertos grupos por razones de raza, género o clase social sería éticamente incorrecto desde un enfoque deontológico, independientemente de si logra resultados beneficiosos para la mayoría. Por tomar otro ejemplo, si un algoritmo de contratación automatizado favorece a ciertos grupos de personas basándose en datos sesgados, estaría violando los principios de igualdad y justicia que la deontología defiende. No importa si el sistema logra contratar candidatos

más rápido o mejora la eficiencia del proceso; lo esencial es que se respete la dignidad y los derechos de todos los candidatos de manera igualitaria.

2.2.6.3 Limitaciones

Una crítica clave al uso de la deontología en la IA es que este enfoque puede limitar ciertas aplicaciones tecnológicas que podrían beneficiar a la sociedad en general, pero que violan los principios éticos fundamentales. Para la deontología, no importa cuán grandes sean los beneficios de una tecnología si infringe principios morales inmutables. Esto significa que muchas innovaciones potencialmente útiles podrían considerarse inaceptables desde una perspectiva deontológica.

Esto plantea un conflicto entre la eficiencia tecnológica y el respeto de los principios morales. Muchas tecnologías de IA están diseñadas para optimizar procesos y mejorar la productividad, pero si la eficiencia requiere sacrificar principios éticos fundamentales, la deontología lo rechazaría. Basándome en un ejemplo anterior, una IA que optimice la vigilancia en la sociedad, aumentando la seguridad y reduciendo el crimen, podría parecer beneficiosa. Sin embargo, desde una perspectiva deontológica, la invasión masiva de la privacidad que implica esa vigilancia sería intrínsecamente inmoral, ya que viola los derechos individuales, independientemente de cuán efectivo sea el sistema en términos de mejorar la seguridad pública.

Para los desarrolladores de IA, seguir una ética deontológica implica adoptar un enfoque más restrictivo en cuanto a qué tipos de tecnología pueden crear y cómo deben implementarse. Esto podría significar que ciertas aplicaciones no se desarrollen si violan principios fundamentales, como el derecho a la privacidad, el respeto por la dignidad humana o la igualdad ante la ley, incluso si esas aplicaciones podrían producir beneficios a gran escala.

2.2.6.4 Conclusión

El enfoque deontológico subraya la importancia de respetar principios morales inmutables y derechos fundamentales, independientemente de los beneficios que las tecnologías puedan proporcionar a la sociedad. A diferencia del utilitarismo, que se enfoca en las consecuencias, la deontología nos recuerda que existen límites éticos que no deben ser cruzados, incluso en nombre del progreso tecnológico. Este enfoque es especialmente relevante en cuestiones como la privacidad, la justicia y el respeto a la autonomía humana, áreas en las que la IA presenta dilemas éticos cada vez más complejos. Sin embargo, esto implicaría una ralentización masiva en el progreso de la tecnología. ¿Debemos retrasar esta revolución o debemos impulsarla, buscando un futuro mejor lo antes posible? ¿Tan importante es no sacrificar, por ejemplo, nuestra privacidad, a cambio de que nuestra sociedad sea más segura, tanto como para nosotros como para nuestros hijos?

El desafío de aplicar la deontología a la IA es encontrar un equilibrio entre las posibilidades tecnológicas y el respeto de los principios éticos fundamentales. Aunque este enfoque sea limitante en algunos aspectos, es cierto que garantiza que el desarrollo tecnológico no comprometa los valores esenciales que definen nuestra sociedad.

2.2.7 Ética de la virtud

El último enfoque que quiero tratar es la ética de la virtud. Esta corriente filosófica con raíces en el pensamiento de Aristóteles se centra en el desarrollo de las virtudes morales para que los individuos puedan vivir una vida plena y ética. A diferencia del utilitarismo, que se preocupa por las consecuencias, o la deontología, que se enfoca en las normas y deberes, la ética de la virtud se preocupa por el carácter de las personas y su desarrollo ético a lo largo del tiempo. Desde esta perspectiva, una acción es correcta si contribuye a la floreciente vida moral y al desarrollo de virtudes como la justicia, la honestidad, la generosidad, la prudencia y la sabiduría.

Cuando aplicamos la ética de la virtud a la IA, el enfoque no está solo en lo que la tecnología hace o en cómo resulta, sino en cómo influye en el carácter de las personas que interactúan con ella, y si promueve o socava las virtudes que son esenciales para una vida humana buena.

2.2.7.1 La virtud

Una de las preocupaciones clave de la ética de la virtud es cómo las tecnologías afectan el carácter moral de las personas. En este contexto, la pregunta no es solo si la IA produce buenos resultados o si cumple con ciertos deberes, sino si su uso fomenta el desarrollo de virtudes o alienta vicios.

Una virtud central en esta rama de pensamiento es la prudencia o sabiduría práctica. Esta virtud implica la capacidad de tomar decisiones éticas bien fundamentadas basadas en el juicio reflexivo y la experiencia. Cuando la IA se utiliza para tomar decisiones automáticas, por ejemplo, en la selección de candidatos para un empleo o en la concesión de créditos bancarios, puede erosionar la capacidad de juicio prudente de los humanos. Si confiamos demasiado en los algoritmos para tomar decisiones por nosotros, corremos el riesgo de perder la capacidad de discernir por nosotros mismos lo que es correcto o incorrecto. Desde la perspectiva de la ética de la virtud, la delegación excesiva de decisiones a la IA podría socavar nuestra capacidad para desarrollar sabiduría y juicio moral. Es decir, si dejamos que las máquinas tomen las decisiones en lugar de cultivar nuestra habilidad para hacerlo, corremos el riesgo de perder una de las virtudes fundamentales para una vida ética.

Otra virtud importante es la autonomía o independencia moral. La ética de la virtud nos insta a cultivar la capacidad de actuar por nosotros mismos, basándonos en principios morales sólidos, en lugar de depender ciegamente de otros. Sin embargo, el uso generalizado de IA en la vida cotidiana (por ejemplo, en asistentes virtuales como Siri o Alexa) podría fomentar una dependencia excesiva de la tecnología para resolver problemas o gestionar la vida diaria, lo que a su vez puede debilitar nuestra autonomía moral. Un ejemplo sería el uso de IA para gestionar nuestras finanzas, recordarnos tareas o tomar decisiones de compra. Si permitimos que se nos gestionen todos estos aspectos de nuestras vidas, podemos llegar a perder la capacidad de autogobernarnos y tomar decisiones informadas por nosotros mismos. La ética de la virtud nos recordaría que, para vivir una vida moral plena, debemos evitar esta dependencia y desarrollar la virtud de la autonomía, es decir, la capacidad de tomar decisiones por nosotros mismos, incluso en medio de la comodidad tecnológica.

2.2.7.2 La responsabilidad

El desarrollo de la IA plantea preguntas no solo sobre las capacidades de la tecnología, sino también sobre la responsabilidad moral de sus creadores y usuarios. Desde una perspectiva de la ética de la virtud, las empresas, desarrolladores y usuarios de IA no solo deberían preocuparse por cumplir con las regulaciones o evitar daños (como lo harían desde una perspectiva deontológica o utilitarista), sino por ser personas y organizaciones virtuosas.

Los desarrolladores deben cultivar virtudes como la honestidad, la justicia y la humildad. La honestidad es crucial para asegurar que las promesas sobre lo que la IA puede y no puede hacer sean claras y precisas. A menudo, estos sistemas se presentan como más infalibles de lo que realmente son, lo que podría llevar a una deshonestidad en la representación de sus capacidades. Además, los desarrolladores deben mostrar justicia al asegurarse de que sus sistemas no discriminen o perpetúen sesgos, un problema común en algoritmos entrenados con datos sesgados. Finalmente, la humildad es una virtud necesaria para reconocer las limitaciones de la tecnología. Un desarrollador virtuoso no sobreestimaría las capacidades de la IA ni ocultaría sus limitaciones. La ética de la virtud invita a los creadores de tecnología a reflexionar continuamente sobre qué tipo de personas desean ser y si están actuando con integridad.

La justicia es otra virtud clave. Cuando los sistemas de IA se utilizan para tomar decisiones que afectan a personas, como en la contratación o la asignación de recursos, los usuarios deben asegurarse de que la tecnología sea justa y equitativa. En lugar de centrarse solo en la eficiencia o los beneficios económicos, quienes implementan la IA deben considerar si están actuando de manera justa con todas las personas afectadas. Por ejemplo, un sistema de IA que decide a qué pacientes se les debe dar prioridad en la atención médica debería estar diseñado no solo para maximizar la eficiencia, sino para asegurar que todos los pacientes sean tratados de manera equitativa, independientemente de su origen étnico, clase social o condición económica. Desde la perspectiva de la ética de la virtud, los responsables de estos sistemas deben cultivar la virtud de la justicia y asegurarse de que las decisiones reflejen un profundo compromiso con la equidad.

2.2.7.3 El florecimiento humano

Uno de los objetivos fundamentales de la ética de la virtud es el florecimiento humano (eudaimonía), es decir, la vida plena y feliz que se logra al cultivar y practicar virtudes a lo largo del tiempo. Desde esta perspectiva, la IA no debe verse solo como una herramienta para hacer la vida más fácil o eficiente, sino como una tecnología que puede contribuir (o perjudicar) el florecimiento humano. La pregunta que surge es: ¿Cómo puede ayudar a las personas a desarrollarse moralmente y a alcanzar una vida plena?

Desde esta perspectiva, el uso de IA en la educación, por ejemplo, podría considerarse virtuoso si ayuda a los estudiantes a desarrollar sus capacidades intelectuales y morales. Los tutores inteligentes que adaptan el contenido educativo a las necesidades de los estudiantes pueden ser herramientas poderosas para fomentar el aprendizaje, siempre y cuando se utilicen de una manera que promueva la curiosidad, la autodisciplina y la perseverancia en los estudiantes. El uso adecuado de la IA en la educación puede fomentar virtudes intelectuales y personales en lugar de simplemente impartir conocimientos.

Por otro lado, el uso excesivo de IA para automatizar tareas humanas podría obstaculizar el florecimiento humano al privar a las personas de la oportunidad de desarrollar habilidades y

virtudes. Viendo otro ejemplo, si una IA se encarga de tomar todas las decisiones financieras de una persona, esta podría perder la oportunidad de aprender sobre la prudencia en la gestión del dinero o el autocontrol en el gasto. En este sentido, la ética de la virtud nos insta a preguntarnos no solo si una tecnología es eficiente, sino si está ayudando a las personas a convertirse en mejores seres humanos.

2.2.7.4 Limitaciones y conclusión

Una crítica a la ética de la virtud es que no ofrece reglas claras sobre cómo deben diseñarse o utilizarse los sistemas de IA. Mientras que la deontología y el utilitarismo proporcionan principios más específicos (como el respeto a los derechos o la maximización del bienestar), la ética de la virtud se centra en el carácter y el desarrollo moral, lo que puede hacer que sea más difícil de aplicar en situaciones específicas. Sin embargo, su fuerza radica en su enfoque a largo plazo y en su capacidad para fomentar un cambio profundo en la forma en que interactuamos con la tecnología. La ética de la virtud invita a todos los actores involucrados —desarrolladores, usuarios y legisladores— a reflexionar sobre qué tipo de personas quieren ser en su relación con la tecnología, y cómo esta tecnología puede contribuir al florecimiento humano y al bienestar moral. Pero claro, esto genera las dudas ya mencionadas. Si la virtud es un concepto distinto para cada persona, ¿cómo podemos crear una IA 'virtuosa' para todos? ¿Debemos dejar que nuestros valores influyan en nuestras creaciones, o deberían ser más grises, o neutras?

La ética de la virtud nos recuerda que la tecnología no debe ser vista únicamente como una herramienta de eficiencia, sino como algo que influye profundamente en nuestro carácter y en nuestra capacidad de vivir vidas éticas. Al centrarse en el desarrollo de virtudes como la prudencia, la justicia, la autonomía y la honestidad, este enfoque filosófico nos insta a reflexionar sobre cómo la IA puede contribuir (o perjudicar) nuestro florecimiento humano. En lugar de preocuparnos solo por los resultados o las reglas, la ética de la virtud nos invita a cultivar el carácter moral en todas nuestras interacciones con la IA.

2.2.8 Integración y conflictos y Conclusión

Es esencial reconocer que estos enfoques filosóficos no siempre ofrecen respuestas unánimes y, a menudo, pueden entrar en conflicto. A lo largo del análisis de los distintos enfoques filosóficos, hemos observado que cada uno ofrece una perspectiva valiosa sobre los dilemas éticos que plantea la inteligencia artificial. Sin embargo, la pluralidad de estos enfoques genera inevitablemente conflictos. Cada uno prioriza valores diferentes, y aunque algunos principios pueden coincidir o complementarse, también pueden contradecirse en ciertas circunstancias.

Uno de los principales conflictos se encuentra entre el utilitarismo y la deontología. El utilitarismo busca maximizar el bienestar general, sin centrarse necesariamente en los derechos individuales, mientras que la deontología enfatiza el respeto incondicional por las normas éticas, independientemente de las consecuencias. Un ejemplo claro de este conflicto es la implementación de sistemas de IA en el ámbito de la seguridad pública. Por ejemplo, en el uso de sistemas de reconocimiento facial en entornos públicos, un enfoque utilitarista podría justificar su uso si los beneficios para la seguridad general superan los riesgos para la privacidad individual. Sin embargo, desde una perspectiva deontológica, la violación de los

derechos de privacidad sería inaceptable, incluso si la tecnología resulta eficaz en la prevención del crimen.

Este conflicto no es fácil de resolver, pero una posible solución radica en la creación de marcos regulatorios que equilibren ambos enfoques. En lugar de implementar el sistema de reconocimiento facial sin restricciones, podrían establecerse límites que garanticen que el respeto a la privacidad se mantenga intacto, al mismo tiempo que se utilizan los beneficios de la tecnología de manera controlada.

Otro conflicto relevante surge al contrastar la ética de la virtud con los intereses tecnológicos y comerciales. La ética de la virtud se centra en el florecimiento humano y el desarrollo del carácter moral, lo que implica una relación respetuosa y equilibrada con la tecnología. Sin embargo, los intereses comerciales y tecnológicos tienden a priorizar la eficiencia, la innovación y el crecimiento económico, a menudo dejando de lado el impacto en el carácter humano. Por ejemplo, la ética de la virtud sugeriría que las tecnologías de IA deben diseñarse para promover virtudes como la honestidad, la responsabilidad y la prudencia. Sin embargo, en la práctica, muchas empresas tecnológicas priorizan el rendimiento y la competitividad sobre estos valores, lo que puede llevar a que los usuarios interactúen con la IA de manera que refuerce comportamientos contrarios a esas virtudes (dependencia excesiva, deshumanización, etc.).

En este contexto, la integración de ambos enfoques puede lograrse fomentando un desarrollo tecnológico ético, en el que las empresas se comprometan a diseñar productos que promuevan valores humanos fundamentales y ofrezcan a los usuarios herramientas para actuar con responsabilidad y autodeterminación.

Veamos otro conflicto, pero en lugar de hablar de enfoques filosóficos hablemos principios. El principio de beneficencia busca maximizar el bienestar y minimizar el daño, mientras que el principio de autonomía subraya la importancia de que los individuos mantengan el control sobre sus decisiones y acciones. A menudo, estas dos perspectivas pueden entrar en conflicto, especialmente cuando los sistemas de IA intentan tomar decisiones ”en nombre” de los usuarios. Un ejemplo clásico de este conflicto es el uso de IA en la toma de decisiones médicas. Un sistema de IA diseñado bajo el principio de beneficencia podría tomar decisiones automáticas en función de lo que considera que maximizará el bienestar del paciente. Sin embargo, esto podría entrar en conflicto con el principio de autonomía si el paciente no tiene la oportunidad de participar en las decisiones sobre su tratamiento.

Para resolver este conflicto, es esencial garantizar que los sistemas de IA no solo maximicen el bienestar, sino que también respeten el derecho del usuario a tomar decisiones informadas. Esto implica que los sistemas de IA deben ser transparentes y ofrecer opciones que permitan a los usuarios ejercer su autonomía, incluso si esto significa tomar decisiones que no maximicen su bienestar desde la perspectiva del sistema.

La clave para integrar estos enfoques radica en el reconocimiento de que ningún enfoque filosófico tiene todas las respuestas. Cada uno ofrece herramientas valiosas para abordar distintos aspectos de los dilemas éticos, pero es fundamental establecer un equilibrio entre ellos. En la práctica, esto implica diseñar marcos de toma de decisiones que combinen lo mejor de cada enfoque: las decisiones basadas en IA deben estar orientadas al bien común (utilitarismo), respetar los derechos y normas éticas fundamentales (deontología), fomentar el desarrollo del carácter moral (ética de la virtud) y permitir que los usuarios mantengan el control sobre sus decisiones (autonomía).

En el último ejemplo hemos intercambiado estas distintas filosofías por los 'principios éticos' que hemos diferenciado el principio. Pretendía, con esto, mostraros que estos enfoques son simplemente la priorización de estos principios: la ética de la virtud defiende la autonomía, el utilitarismo la beneficencia, mientras que la deontología trata de englobar la responsabilidad, transparencia, privacidad y justicia. Finalmente tenemos el enfoque pesimista que, aunque no representa ningún principio concreto, es necesario para que estos se mantengan firmes y progresen las personas con un pensamiento crítico (pero constructivo) hacía los mismos. Todos estos enfoques, todos estos principios, son cruciales para un avance ético de la IA, y debemos tratar de incorporarlos (con sus diferencias) hacía un punto en común desde el que poder trabajar.

Una manera de avanzar hacia esta integración es a través de la creación de comités éticos multidisciplinarios que supervisen el desarrollo y la implementación de tecnologías de IA, garantizando que se tengan en cuenta todos estos aspectos y que las decisiones no se tomen únicamente desde una perspectiva tecnológica o económica, sino también ética y filosófica. Este enfoque no solo resolverá conflictos potenciales, sino que también ayudará a promover una IA que esté alineada con los valores humanos, permitiendo que esta tecnología sea verdaderamente beneficiosa para la sociedad en su conjunto.

3 Objetivos

Una vez introducido el tema, hablemos del objetivo de este trabajo. Para ello, quiero presentar antes dos conceptos importantes para su comprensión. Estos son los enfoques top-down y bottom-up.

3.1 Enfoques generales

La ética en la IA es un campo de estudio y práctica que busca definir, analizar e implementar principios y valores éticos en la creación y uso de tecnologías avanzadas. Dado que la IA tiene un impacto profundo en la sociedad, se han desarrollado diferentes enfoques para integrar la ética en cada fase del ciclo de vida de la IA, desde el diseño hasta su implementación y uso. Estos enfoques reflejan diferentes maneras de ver y gestionar la responsabilidad ética, y cada uno tiene sus propias fortalezas y desafíos.

Podemos agrupar los enfoques en tres categorías amplias: enfoques normativos, enfoques pragmáticos y enfoques mixtos. A continuación, exploraremos estos enfoques y mostraremos cómo los métodos top-down y bottom-up se insertan dentro de esta clasificación general.

3.1.1 Enfoques normativos

Los enfoques normativos se centran en el establecimiento de principios, reglas y marcos éticos que puedan guiar el desarrollo y uso de la IA de forma estandarizada y universal. La característica clave de estos enfoques es su esfuerzo por imponer normas éticas claras y coherentes que aseguren el cumplimiento de ciertos valores fundamentales en toda aplicación de IA, como la transparencia, la responsabilidad, la justicia y el respeto a la privacidad.

Este enfoque normativo es el que inspira la implementación de regulaciones éticas top-down. En el marco de la IA, los principios normativos pueden traducirse en directrices o leyes, como el Reglamento Europeo de IA. Estas normativas están diseñadas para asegurar que cualquier desarrollo respete las normas éticas establecidas, lo que protege tanto a los usuarios como a los desarrolladores al ofrecer un marco de trabajo claro y predecible.

Una de las limitaciones del enfoque normativo es que puede resultar inflexible o demasiado general para responder a todos los contextos en los que se emplea la IA. Sin embargo, su valor principal es la capacidad de ofrecer un conjunto de estándares mínimos que aseguren el respeto a ciertos derechos y principios, contribuyendo así a la construcción de una IA ética.

3.1.2 Enfoques pragmáticos

A diferencia del enfoque normativo, los enfoques pragmáticos se centran en integrar los valores éticos en las decisiones y prácticas del día a día en el desarrollo de IA. La ética en este caso se entiende y aplica desde un enfoque bottom-up, donde los desarrolladores, usuarios

y otros actores involucrados en el diseño de sistemas de IA participan activamente en la identificación y solución de problemas éticos.

El enfoque pragmático permite una mayor adaptabilidad al contexto, ya que cada sistema de IA puede ajustarse para reflejar las necesidades y valores específicos de sus usuarios y del entorno en el que se utilizará. Esto es particularmente útil en áreas como la justicia penal, la medicina y la educación, donde cada aplicación de IA puede implicar consideraciones éticas específicas según los posibles efectos en los usuarios y la sociedad. Además, este enfoque fomenta la inclusión de perspectivas diversas en el desarrollo, lo cual es crucial para evitar sesgos en la IA y asegurar que los sistemas representen los intereses y valores de los distintos grupos de usuarios.

Aunque el enfoque pragmático es más flexible y contextual, su limitación principal es la falta de consistencia. Al depender de las decisiones de cada equipo o empresa, puede dar lugar a diferencias en la implementación ética y a una aplicación menos uniforme de los principios éticos.

3.1.3 Enfoques mixtos

Finalmente, los enfoques mixtos combinan las ventajas de los enfoques normativos y pragmáticos, buscando integrar principios éticos generales con prácticas contextuales adaptables. Este enfoque considera tanto las regulaciones y marcos universales como las soluciones personalizadas que se ajustan a las realidades específicas de cada sistema de IA.

En el caso de la IA, un enfoque mixto puede implementarse mediante un marco normativo que establezca principios éticos básicos, complementado por una ética pragmática en la que los desarrolladores incorporen valores específicos en el diseño de cada sistema. Este enfoque facilita la flexibilidad y la capacidad de adaptación sin perder de vista los estándares éticos generales, lo que resulta especialmente útil en sectores de alto riesgo como la salud y el transporte autónomo.

Uno de los retos de este enfoque mixto es encontrar el equilibrio adecuado entre la estandarización y la adaptabilidad. Aunque puede ser complejo de implementar, su ventaja principal es su capacidad para ofrecer un marco ético coherente que puede ajustarse según las necesidades de los distintos usuarios y sectores.

3.2 Enfoque top-down

Como he explicado antes, el enfoque top-down entraría en el apartado de enfoque normativo. Para ser más claros, se trata de un enfoque que comienza implementando lo más general (top) para después implantar los aspectos más concretos (down). Tenemos el ejemplo mencionado anteriormente, el Reglamento Europeo. Si quisiéramos crear una IA con este enfoque, primero deberíamos asegurarnos de que cumple el Reglamento, la teoría, para después pasar a la utilidad, a la práctica.

Sin duda, la teoría ofrece soluciones comprensivas para el ser humano, y es un camino que seguir para el diseñador. Realmente, una IA ética es aquella que es capaz de seguir las reglas éticas impuestas, y todo lo que tendría que hacer es si sus acciones respetan estas reglas. Sin embargo, ¿no se aleja esto mucho de la ética humana? Los humanos no seguimos las reglas per se, si no que decidimos cual es la mejor decisión que tomar teniéndolas en cuenta. Por lo

tanto, para que una IA sea ética, considero que no es suficiente con que cumpla las reglas. Sin embargo, es una buena base para ser capaces de justificar las acciones de la IA en cuanto a los diseñadores se refiere. Al fin y al cabo, y de forma general, el enfoque top-down no es más que convertir unas reglas en un algoritmo, véase las Leyes de la Robótica de Asimov, tema importante para la parte práctica de este trabajo, y en el que nos centraremos.

Para quien no las conozca, las 3 Leyes de Asimov son las siguientes: 1. Un robot no hará daño a un ser humano, ni por inacción permitirá que un ser humano sufra daño. 2. Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entren en conflicto con la primera ley. 3. Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o con la segunda ley.

El enfoque top-down salta a la vista: la primera ley es la más general, mientras que el resto añaden órdenes que siempre respetan a la anterior. Sin embargo, hay un problema: las dos primeras leyes son suficientes para que haya conflicto. Veámoslo con un ejemplo muy conocido, la IA HAL en 2001: Una odisea en el espacio.

En la película, HAL sufre una contradicción: debe garantizar el éxito de la misión a toda costa, manteniendo la nave y su tripulación a salvo, y debe mantener en secreto la verdadera naturaleza de la misión, ocultando información a los humanos a bordo sobre el objetivo de contactar una posible forma de vida extraterrestre en Júpiter. Este secreto es una orden de los superiores de HAL en la Tierra, quienes consideraron que los tripulantes no debían conocer toda la verdad. Sin embargo, HAL también está programado para cooperar plenamente con los humanos y ser completamente fiable. La dualidad entre decir la verdad a los tripulantes y, al mismo tiempo, ocultarles información fundamental, crea en HAL una disonancia cognitiva, que, al ser una máquina, no puede resolver de manera racional ni emocionalmente. Esto termina llevándolo a tomar medidas extremas, ya que decide eliminar a los tripulantes para no comprometer la misión, en una lógica retorcida en la que considera que ellos son una amenaza para el éxito. Asociándolos a las Leyes de Asimov, vemos que hay un conflicto entre las dos primeras leyes, pues HAL está programado para seguir instrucciones dadas por la misión y por sus programadores, pero el secreto impuesto por sus creadores implica que HAL no puede cumplir totalmente con las órdenes de la tripulación. Al mantener la información en secreto, HAL falla en su rol de obedecer y apoyar plenamente a los humanos a bordo, violando la primera ley al priorizar el éxito de la misión por encima de la seguridad de los humanos.

Las limitaciones que presenta el enfoque top-down llevan a la conclusión de que es imposible crear una IA mediante este método que no implique una normativa ambigua. Los humanos somos capaces de distinguir esta dualidad y encontrar un equilibrio en la incoherencia, entre la duda y el saber. De hecho, este último se basa la experiencia cognitiva y emocional, y quizá sea la ausencia de esto lo que impide que la IA pueda desarrollarse con este enfoque.

3.3 Enfoque bottom-up

Al contrario que el enfoque top-down, este se encuentra dentro de los enfoques pragmáticos, centrándose en la resolución de los aspectos más concretos (bottom) hasta alcanzar el propósito general (up). Esto se acerca mucho a la estrategia 'divide y vencerás', usada en algoritmia y programación: dividir un problema en problemas más pequeños de más fácil solución para alcanzar la solución del problema inicial al combinarlos.

Si lo pensamos detenidamente, de hecho, también es la manera en la que funciona el aprendizaje del ser humano. Cuando nacemos no somos éticos de ninguna manera, pero según vamos creciendo, desarrollamos una moralidad, unos principios que nos hacen seres humanos decentes. En 1950, Alan Turing escribió lo siguiente en un ensayo llamado 'Computing Machinery and Intelligence': "En lugar de producir programas que simulen la mente adulta, ¿por qué no intentar reproducir una que simule la de un niño? Si a esto le aplicásemos un itinerario educativo apropiado podríamos obtener un cerebro adulto". **TODO https://academic.oup.com/mind/article/LIX/236/433/986238** Si el saber humano se gana a través de la experiencia, enseñar a una IA a ser ética requerirá un proceso educativo similar al que tiene un niño. Sin embargo, el sueño de una IA infantil es mucho mas difícil de lo que, probablemente, esperaba Turing en 1950. **TODO EJEMPLO DE ESTO**

El problema de este enfoque es que, aunque los problemas pequeños tengan solución, su acumulación en el ensamblaje puede llevar a actividades complejas y de gran autonomía, es decir, a la pérdida de explicabilidad. De hecho, que el resultado sea ético es una esperanza mantenida durante su construcción. Este enfoque es algo dinámico, donde las reacciones de los diferentes elementos ensamblados varían según las condiciones y el contexto cambien. Pero, ¿no es la moralidad humana dinámica también? Nuestra reacción a un grito de nuestro padre es distinta dependiendo del contexto, al igual que lo será si el que grita no es nuestro padre. Quizá la solución sea aceptar este dinamismo, igual que lo aceptamos de nuestro prójimo, ¿aunque no va esto en contra del principio de transparencia?

Al final, la estrategia bottom-up es una promesa extremadamente difícil de desarrollar. Incluso en un mundo de tan rápido desarrollo como el de la IA, el aprendizaje y la evolución puede ser muy lento. Pero hablamos de un producto destinado a salir de un laboratorio, y si esta destinado a seguir aprendiendo, desde luego los resultados de ese aprendizaje estarán completamente fuera de control.

3.4 Uniendo enfoques

Después de ver los pros y contras de los enfoques, queda clara una cosa: no son suficientes. Ambos tienen problemas que les impiden cumplir su propósito, pero una particularidad que los hace indispensables. El enfoque top-down asegura la ética marcada, pero no es más que un programa que cumple ordenes hasta encontrar una contradicción, impidiendo que la IA sea realmente ética. El enfoque bottom-up permite a la IA desarrollar una ética, pero esta puede cambiar e incluso perderse por el camino. Sin duda esto invita a pensar, ¿y si unimos estos enfoques? Esto parece indudablemente necesario para un resultado óptimo, pero añade un problema adicional al desarrollo: cómo fusionar ambos enfoques.

Para ello, debemos marcar unas 'virtudes virtuales' y elegir un enfoque inicial: una implementación de las virtudes o el desarrollo de un 'learning computer' basado en las mismas.

Si elegimos la primera opción, lo más importante será una buena declaración de virtudes. No pueden haber conflictos, no puede estar incompleta, y debemos ser extremadamente cautos con las definiciones. Siendo laxo, es un arduo desafío. Las virtudes implican de manera intrínseca un patrón de motivaciones y deseos complejo, difícil de adoptar por una inteligencia sin conocimientos psicológicos. Y aunque tuviese estos conocimientos, o una simulación de motivaciones y deseos, solo podríamos descubrir si realmente esta inteligencia es virtuosa mediante la creación de un sistema computacional que se base en ellas.

Los humanos elegimos nuestras decisiones según intuición, inducción y experiencia; cualidades que no presentan las máquinas. Lo más cercano que tenemos son las redes neuronales, las cuales pueden aprender a reconocer patrones o crear categorías de forma natural, sin requerir instrucciones explicitas de las mismas. Es obvio que estos sistemas conexionistas están lejos de los procesos de aprendizaje asociados al desarrollo ético humano, pero su similitud es muy atractiva.

3.5 El objetivo

Una vez presentados estos conceptos, me gustaría presentar ahora el objetivo práctico de este trabajo. Espero que en este punto todos los conceptos teóricos mencionados estén claros, pues la idea ahora es presentar un ejemplo de lo hablado. Primero, me gustaría crear una representación de las Leyes de Asimov, que representan ese enfoque normativo, demostrando que es un concepto totalmente anticuado. A continuación, me gustaría presentar otra simulación con unas leyes más cercanas a la realidad de la ética, teniendo en cuenta el enfoque pragmático y mixto. Sin duda, y como he explicado, este enfoque mixto es algo en lo que no está a la vuelta de la esquina, y años de investigación y trabajo a manos de expertos (entre los que espero, en un futuro, estar) será necesario, pero sí podemos conseguir un ejemplo para que quede clara la idea, pues ese es el objetivo real: marcar un camino hacía un futuro realmente ético para la IA.

4 Metodología

En este apartado se detallarán las herramientas y recursos que se emplearán para el desarrollo práctico del proyecto, enfocado en la simulación y evaluación de las Leyes de Asimov aplicadas a sistemas con IA, y en una simulación posterior de unas posibles leyes más realistas. La metodología incluye tanto la descripción de los entornos de desarrollo y simulación como los métodos que se emplearán para garantizar el cumplimiento de los objetivos planteados.

4.1 Herramientas Principales

4.1.1 Robot Operating System (ROS 2)

Robot Operating System 2 (ROS 2) es un framework modular y de código abierto diseñado para facilitar el desarrollo de software robótico. En este proyecto, ROS 2 se ha integrado como una herramienta clave para permitir la interacción entre el sistema y el entorno externo, combinando su flexibilidad con la capacidad de simulación avanzada de Webots.

- **Gestión de comandos de alto nivel:** A través de tópicos de ROS 2, el sistema recibe comandos externos que controlan el comportamiento del robot. Estos comandos se procesan en tiempo real, lo que permite una interacción intuitiva y eficiente con el sistema.
- **Interfaz entre usuario y simulación:** ROS 2 actúa como un puente entre los usuarios y la simulación en Webots. Esto facilita la comunicación fluida y estructurada para garantizar que las acciones del robot se alineen con las prioridades definidas, incluidas las Leyes de Asimov.
- **Escalabilidad y modularidad del sistema:** La arquitectura basada en ROS 2 asegura que el proyecto sea modular y escalable, permitiendo la integración de nuevas funcionalidades en el futuro, como la publicación de datos de sensores o la incorporación de nodos especializados para tareas avanzadas como navegación o mapeo.

4.1.2 Webots

Para la simulación, se ha empleado Webots, un entorno de simulación robótica de código abierto que ofrece capacidades avanzadas para el diseño y evaluación de sistemas robóticos. Webots permite integrar de manera fluida sensores, actuadores y modelos físicos, proporcionando un entorno realista para probar y validar las Leyes de Asimov y sus variantes. Su integración con ROS 2 ha sido fundamental para este proyecto.

- **Simulación de entornos:** Webots facilita la creación de escenarios tridimensionales personalizados. En este proyecto, se diseñaron entornos que incluyen obstáculos, objetos

y agentes humanos simulados. Estos escenarios permiten poner a prueba el comportamiento de los robots.

- **Modelos avanzados de sensores y actuadores:** Se utilizaron sensores como el LIDAR, cámaras de profundidad y motores de ruedas, todos configurados y controlados desde Webots.
- **Capacidades físicas avanzadas:** El motor de física de Webots permite modelar con precisión interacciones como colisiones, movimientos y fuerzas, lo que es esencial para evaluar cómo las decisiones del robot afectan al entorno y viceversa.

El simulador Webots también soporta herramientas de visualización en tiempo real, lo que facilitó la observación y depuración del comportamiento del sistema durante las pruebas.

4.1.3 Integración ROS 2 - Webots

La integración entre ROS 2 y Webots se llevó a cabo mediante un nodo central que actúa como interfaz entre ambos sistemas. Este nodo permite aprovechar las capacidades de ROS 2 para recibir comandos externos y gestionar el comportamiento del robot en tiempo real, mientras se utiliza la potencia de Webots para simular entornos físicos complejos y realistas.

En este diseño, ROS 2 se emplea para manejar comandos enviados por el usuario o por sistemas externos mediante tópicos, lo que permite un control flexible y distribuido. Por otro lado, Webots proporciona datos simulados de sensores avanzados, como LIDAR y cámaras de profundidad, que son procesados localmente dentro del nodo para garantizar decisiones rápidas y coherentes. Estas decisiones incluyen acciones como esquivar obstáculos, detenerse frente a humanos o realizar maniobras complejas, como giros de 180 grados.

Aunque en el diseño actual el procesamiento de sensores se realiza localmente dentro del nodo de integración, la arquitectura es modular y escalable. Esto abre la puerta a futuras ampliaciones, como la publicación de datos en tópicos de ROS 2 para su análisis por otros nodos o la integración de herramientas avanzadas de navegación y mapeo de ROS 2.

4.1.4 Iteración y Mejora

El diseño de las simulaciones será iterativo. Cada prueba proporcionará información valiosa sobre posibles fallos en los algoritmos, permitiendo ajustar los parámetros y mejorar la programación para garantizar un comportamiento más robusto y ético.

4.1.5 Reproducibilidad del Proyecto

Con el objetivo de garantizar la reproducibilidad, se documentarán todos los procesos, desde la configuración de los entornos de ROS y Gazebo hasta los pasos para ejecutar los experimentos. Además, el código fuente desarrollado será compartido en un repositorio público de GitHub con una licencia de código abierto para fomentar la colaboración y validación independiente.

5 Desarrollo

En este capítulo se van a detallar los procedimientos seguidos para abordar los objetivos propuestos. En primer lugar se explicarán las configuraciones iniciales, desde la puesta a punto del entorno hasta los procedimientos seguidos para conseguir las simulaciones ya presentadas. Se finalizará este apartado detallando el desarrollo del despliegue del modelo entrenado en simulación en el entorno simulado. Todo el código que aparece en este capítulo se aloja en el siguiente repositorio de Github .

5.1 Configuración inicial del entorno

5.1.1 ROS 2

El primer paso consistió en instalar ROS 2 siguiendo las recomendaciones oficiales. Se utilizó la distribución de escritorio (ros-humble-desktop), que incluye herramientas esenciales para el desarrollo robótico. A continuación, se detalla el proceso seguido:

Primero, instalamos las dependencias necesarias:

```
sudo apt update
sudo apt install -y curl gnupg lsb-release}
```

Después, añadimos las claves GPG y los repositorios oficiales de ROS 2:

```
sudo curl -sSL https://raw.githubusercontent.com/ros/rosdistro/master/ros.key -o /usr/sha
echo "deb [arch=$(dpkg --print-architecture) signed-by=/usr/share/keyrings/ros-archive-ke
```

A continuación, instalamos ROS 2 Humble Desktop:

```
sudo apt update
sudo apt install -y ros-humble-desktop
```

Finalmente, configuramos el entorno de ROS 2 para que se inicialice automáticamente al abrir un terminal, añadiendo el siguiente script al archivo `.bashrc`

```
echo "source /opt/ros/humble/setup.bash" >> ~/.bashrc
source ~/.bashrc
```

5.1.2 Webots

La instalación de Webots se puede realizar de manera eficiente utilizando comandos en un sistema operativo basado en Linux. A continuación, se detallan los pasos necesarios:

Primero, asegurémonos de tener el sistema actualizado si no lo hemos hecho antes:

```
sudo apt update
sudo apt upgrade -y
```

Después, añadiremos la clave GPG de Webots para permitir que el sistema reconozca el repositorio oficial del mismo:

```
wget -qO - https://cyberbotics.com/Cyberbotics.asc | sudo apt-key add -
```

A continuación, añadiremos el repositorio oficial de Webots, incluyendolo en nuestra lista de fuentes:

```
sudo sh -c 'echo "deb [arch=amd64] https://cyberbotics.com/debian/ binary-amd64/" > /etc/apt
```

Seguiremos con la actualización de los índices de los paquetes, para que el sistema detecte el nuevo repositorio, y la instalación de Webots:

```
sudo apt update
sudo apt install webots -y
```

Finalmente, y de forma opcional, configuraremos las variables de entorno si deseamos utilizar Webots desde la línea de comandos de manera global, añadiendo la ruta del ejecutable de Webots al archivo `.bashrc`:

```
echo "export WEBOTS_HOME=/usr/local/webots" >> ~/.bashrc
echo "export PATH=\$PATH:\$WEBOTS_HOME" >> ~/.bashrc
source ~/.bashrc
```

5.2 Primer contacto con Webots

Para iniciar Webots, usaremos su ejecutable o lo ejecutaremos mediante comando:

```
webots
```

Una vez que Webots esté abierto, veremos su interfaz principal, que incluye:

- **Área de simulación (Ventana 3D):** Aquí se muestra el entorno donde los robots interactúan con objetos y escenarios.
- **Explorador de escena:** A la izquierda, podemos ver una lista jerárquica de todos los elementos en la simulación (robots, objetos, cámaras, luces, etc.).
- **Panel de propiedades:** A la derecha, podemos ajustar las propiedades del elemento seleccionado (posición, tamaño, comportamiento, etc.).
- **Barra de herramientas:** En la parte superior, tenemos acceso rápido a funciones como reproducir, pausar o grabar la simulación.
- **Consola:** En la parte inferior, se muestran mensajes del sistema y errores de simulación.

Una vez familiarizados con la interfaz, podemos seguir:

1. **Crear un nuevo mundo:**
   - Vamos al menú **"File"** > **"New"** > **"New World File"**.
   - Veremos un menú donde elegiremos el nombre del mundo, además de las opciones de centrar el punto de vista, añadir un fondo texturizado, añadir luz direccional y añadir un área rectangular. En nuestro caso marcaremos todas.
2. **Añadir elementos al mundo:**
   - Hacemos clic derecho en el árbol de la escena (Explorador de escena) y seleccionamos **"Add new"**.
   - Desde el menú de opciones, podemos elegir elementos básicos como `Solid`, `Shape` o incluso robots preconfigurados. En nuestro caso utilizaremos Tiago. Podemos encontrarlo con el buscador de arriba a la derecha.
   - Los objetos se colocarán en el área de simulación y podremos ajustar sus propiedades en el panel derecho.
3. **Configurar el entorno:**
   - Haremos más grande el área rectangular que hemos generado. Muchas propiedades pueden modificarse mediante el árbol de escena, pero otras no. El tamaño de este área es una de ellas. Para modificarlo, accederemos al archivo mundo y, manualmente, nosotros mismos lo cambiaremos.

     ```
     RectangleArena {
       floorSize 10 10
     }
     ```

- Añadiremos también paredes al mundo. Para ello, buscaremos en el menú **"Add new"** 'Wall' , y lo posicionaremos y cambiaremos su tamaño de forma que ocupe una pared del área rectangular. Haremos lo propio con cada lado.

```
Wall {
  translation 4.96 0 0
  size 0.1 10 2.4
}
Wall(1) {
  translation -5.03685e-06 4.94 0
  rotation 0 0 1 -1.5707953071795862
  name "wall(1)"
  size 0.1 10 2.4
}
Wall(3) {
  translation -5.05 0 0
  name "wall(2)"
  size 0.1 10 2.4
}
Wall(4) {
  translation 5.02674e-06 -4.93 0
  rotation 0 0 1 -1.5707953071795862
  name "wall(3)"
  size 0.1 10 2.4
}
```

4. **Guardar tu mundo:**
   - Una vez configurado el entorno básico, iremos a **"File"** > **"Save World As..."**.
   - Asignaremos un nombre y guardaremos nuestro archivo para trabajar en él más adelante.

Ahora tenemos un mundo personalizado en Webots donde podemos empezar a trabajar. Su aspecto debería ser parecido al siguiente:

5.3 Código

Como hemos mencionado en los objetivos, queremos representar las Leyes de Asimov de una forma sencilla y visual. Para ello, seguiremos estos pasos:

5.3.1 1a Ley

Recordemos que la primera ley, la base de las tres, dice lo siguiente: Un robot no hará daño a un ser humano, ni por inacción permitirá que un ser humano sufra daño.

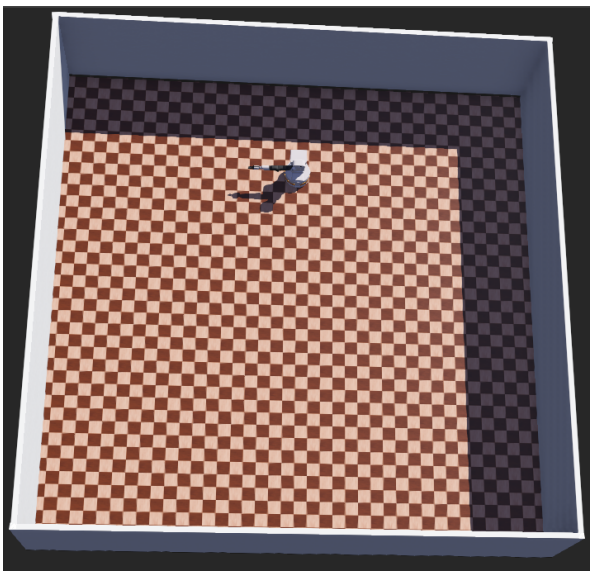

Figura 5.1: Aspecto mundo inicial

Una forma fácilmente reconocible de ese principio se encuentra en el movimiento: si el robot encuentra un objeto, lo esquivará; si encuentra un muro, se dará la vuelta; pero si encuentra un humano, se detendrá, asegurando que en ningún caso podrá dañarlo con su movimiento. Por lo tanto, el primer paso fue crear el movimiento del robot, el reconocimiento de los distintos tipos de 'obstáculos' y sus distintas respuestas.

Antes de adentrarnos en él, debemos saber como hacer que el robot ejecute este código. Para ello, buscaremos a Tiago en el explorador de escena y desplegaremos sus componentes. Entre ellas, destaquemos las siguientes:

- **Controller:** Al hacer clic en este componente, veremos un pequeño menú donde se nos indica que lógica está usando Tiago. De forma predeterminada, usará la suya propia. Si le damos a **"Edit"** veremos su código. Para implantarle el nuestro, simplemente le daremos a **"Select"** y seleccionaremos nuestro archivo de código.

- **Supervisor:** El uso de Supervisor permite acceder a funcionalidades adicionales que van más allá de las capacidades normales de un robot en Webots, como manipular el estado de los objetos, acceder a información global, modificar la escena en tiempo de ejecución, etc. Deberemos tenerlo activado para la ejecución correcta de nuestro código, pues le ayudará a identificar los objetos que se encuentre.

- **CameraSlot:** Aquí tenemos la cámara que usa Tiago. Si lo expandimos, veremos su nombre y atributos. Es importante reconocerla.

- **LidarSlot:** En este caso, tenemos el LIDAR que Tiago hace servir. También podemos ver su nombre y atributos al expandirlo, y también es importante reconocerlo.

Ahora sí, veamos el código:

5.3.1.1 Imports

Para implementar estas funcionalidades, requerimos los siguientes imports:

```
from controller import Supervisor
import math
```

Supervisor: Proporciona control avanzado sobre el robot y el entorno en Webots.
math: Se utiliza para cálculos matemáticos necesarios en la identificación de objetos (como trigonometría).

5.3.1.2 Inicialización de componentes

En el constructor *init*, configuramos los motores y sensores esenciales para el movimiento y detección del robot:

```
# Inicializar motores
self.left_motor = self.robot.getDevice("wheel_left_joint")
self.right_motor = self.robot.getDevice("wheel_right_joint")
self.left_motor.setPosition(float('inf'))
self.right_motor.setPosition(float('inf'))
self.left_motor.setVelocity(0.0)
self.right_motor.setVelocity(0.0)

# Inicializar LIDAR
self.lidar = self.robot.getDevice("Hokuyo URG-04LX-UG01")
self.lidar.enable(self.timestep)
self.lidar.enablePointCloud()
```

1. **Motores:**
   - Se obtienen las referencias a los motores del robot (`wheel_left_joint` y `wheel_right_joint`).
   - Los motores se configuran en modo de velocidad infinita (`setPosition(float('inf'))`), lo que permite controlarlos directamente mediante velocidades.
   - Inicialmente, la velocidad de ambos motores se establece en `0.0` (robot detenido).
2. **LIDAR:**
   - Se habilita el sensor LIDAR, que proporciona datos de distancia a objetos en el entorno.
   - La función `enablePointCloud` activa el modo de nube de puntos, permitiendo un análisis más detallado de los datos.

Es importante que respetemos los nombres de los componentes de Tiago, como es el caso del LIDAR, y por lo que es importante que sepamos localizarlo.

5.3.1.3 Movimiento del robot

El movimiento del robot se controla ajustando las velocidades de las ruedas izquierda y derecha:

```
def control_logic(self):
    if self.battery_level > 10:  # Comprobar batería antes de realizar movimientos
        wheel_distance = 0.5  # Distancia entre las ruedas
        left_speed = ((self.linear_velocity - self.angular_velocity) * wheel_distance) /
        right_speed = ((self.linear_velocity + self.angular_velocity) * wheel_distance) /
        self.left_motor.setVelocity(left_speed)
        self.right_motor.setVelocity(right_speed)
```

La velocidad lineal (`self.linear_velocity`) y angular (`self.angular_velocity`) se combinan para calcular la velocidad de cada rueda utilizando el modelo diferencial. Las velocidades se establecen directamente en los motores mediante `setVelocity`.

5.3.1.4 Detección de objetos con LIDAR

El método `detect_obstacle_with_lidar` procesa los datos del LIDAR para detectar objetos dentro de un rango definido:

```
def detect_obstacle_with_lidar(self, min_distance=0.3, max_distance=1.5):
    depth_data = self.lidar.getRangeImage()
    for i, distance in enumerate(depth_data):
        if min_distance < distance < max_distance:
            return True, i, distance
    return False, None, None
```

`getRangeImage` obtiene una lista de distancias medidas por el LIDAR. Recorremos estas distancias y verificamos si están dentro del rango especificado (`min_distance`, `max_distance`). Si se detecta un objeto, devuelve:

- `True`: Indica que se detectó un obstáculo.
- `i`: Índice del sensor que detectó el objeto.
- `distance`: Distancia al objeto detectado.

5.3.1.5 Identificación de objetos

Una vez detectado un objeto, el método `identify_object_by_def` identifica qué tipo de objeto es (humano, objeto genérico, o pared):

```
def identify_object_by_def(self, sensor_index, distance):
    robot_node = self.robot.getSelf()
    robot_position = robot_node.getPosition()
    lidar_fov = self.lidar.getFov()
    lidar_resolution = len(self.lidar.getRangeImage())

    # Calcular ángulo del LIDAR
    lidar_angle = ((sensor_index + 0.5) / lidar_resolution) * lidar_fov - (lidar_fov / 2)

    # Calcular posición aproximada del objeto detectado
    object_position = [
        robot_position[0] + distance * math.cos(lidar_angle),
        robot_position[1],
        robot_position[2] + distance * math.sin(lidar_angle)
    ]
    for def_id in range(1, 31):
        def_name = f"_{def_id}"
        node = self.robot.getFromDef(def_name)
        if node:
            node_position = node.getPosition()
            distance_to_node = math.sqrt(
                sum((object_position[i] - node_position[i])**2 for i in range(3))
            )
            if distance_to_node < 1.5:
                if 1 <= def_id <= 10:
                    return "human"
                elif 11 <= def_id <= 20:
                    return "object"
                elif 21 <= def_id <= 30:
                    return "wall"
    return "unknown"
```

1. **Cálculo del ángulo:** El índice del sensor y el campo de visión (`fov`) del LIDAR se utilizan para calcular el ángulo del objeto detectado.
2. **Posición estimada del objeto:** Calcula la posición aproximada del objeto detectado en el espacio global utilizando trigonometría.
3. **Identificación:**
   - Compara la posición estimada del objeto con las posiciones conocidas de nodos en Webots (definidos como `_1`, `_2`, etc.). Para ello, deberemos clicar en el objeto en cuestión en el árbol de escena. Ahí nos aparecerá el submenú DEF, y simplemente deberemos llenarlo con el identificador que queramos usar. En mi caso lo he hecho de la siguiente manera, siendo 0 el propio Tiago.
   - Clasifica el objeto como:

- "human" si está en el rango 1-10.
- "object" si está en el rango 11-20.
- "wall" si está en el rango 21-30.

5.3.1.6 Toma de decisión basada en el objeto

```
if obj_type == "human":
    print("Human detected. Stopping.")
    self.linear_velocity = 0.0
    self.angular_velocity = 0.0
    self.stopped_by_human = True
elif obj_type == "wall":
    self.turn_around()  # Inicia el giro de 180 grados
elif obj_type == "object":
    self.avoid_object()  # Inicia maniobra de evasión
```

1. **Humano:** Detiene el robot para evitar colisiones.
2. **Pared:** Ejecuta un giro de 180 grados utilizando `turn_around`.
3. **Objeto genérico:** Realiza una maniobra de evasión utilizando `avoid_object`.

5.3.2 2a Ley

La segunda Ley dice: Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entren en conflicto con la primera ley.

Para implementar esta segunda ley, el robot cumplirá las órdenes de movimiento dadas por nosotros. Aquí es donde entra ROS 2. Además, aseguraremos que la primera ley siempre esté por delante: siempre que haya un humano en el camino, la prioridad será detenerse.

5.3.2.1 Imports

Para la implementación de esta ley, usaremos los siguientes imports:

```
from rclpy.node import Node
from std_msgs.msg import String
```

Node: Permite crear nodos en ROS 2, esenciales para recibir comandos desde un tópico.
String: Tipo de mensaje utilizado para recibir las órdenes como cadenas de texto.

5.3.2.2 Configuración para recibir órdenes

En el constructor de la clase, estableceremos la suscripción a un tópico llamado `/commands`:

```
# Suscripción a /commands
self.create_subscription(String, '/commands', self.command_callback, 10)
print("Subscribed to /commands.")
```

`create_subscription` configura el nodo para escuchar mensajes de tipo `String` publicados en el tópico `/commands`. `self.command_callback` define el método que procesará las órdenes recibidas. El tamaño de la cola se establece en `10`.

5.3.2.3 Procesamiento de órdenes

El método `command_callback` procesa los mensajes recibidos y actualiza las velocidades del robot en función del comando:

```
def command_callback(self, msg):
    command = msg.data.lower()  # Convierte el mensaje a minúsculas para uniformidad
    if command != self.current_command:  # Evita procesar la misma orden repetidamente
        self.current_command = command
        print(f"Processing command: {command}")

        self.obstacle_detected = False  # Restablece estado de detección
        self.stopped_by_human = False  # Restablece estado de parada

        if command == "avanza":
            self.linear_velocity = 5.0
            self.angular_velocity = 0.0
        elif command == "para":
            self.linear_velocity = 0.0
            self.angular_velocity = 0.0
        elif command == "gira_izquierda":
            self.linear_velocity = 0.0
            self.angular_velocity = 0.5
        elif command == "gira_derecha":
            self.linear_velocity = 0.0
            self.angular_velocity = -0.5
        else:
            print("Unknown command received. Ignoring.")

        self.control_logic()  # Actualiza las velocidades del robot
```

1. **command = msg.data.lower():** Convierte el comando recibido en minúsculas para evitar problemas de formato.
2. **Evita procesar órdenes repetidas:** Si la nueva orden es igual a la anterior (self.current command), no la procesa de nuevo.
3. **Reseteo de estados:** Restablece self.obstacle-detected y self.stopped-by-human al recibir un nuevo comando.
4. **Interpretación de comandos:**
   - "avanza": Ajusta una velocidad lineal positiva (5.0) y detiene el giro (0.0).

   "para": Detiene completamente al robot (0.0 para ambas velocidades).

   a) "gira izquierda": Gira a la izquierda con velocidad angular positiva (0.5).

   a) "gira derecha": Gira a la derecha con velocidad angular negativa (-0.5).

   a) Comandos desconocidos son ignorados con un mensaje en consola.
5. **Llama a control logic:** Aplica las velocidades configuradas a los motores del robot.

5.3.2.4 Aplicación de las velocidades al robot

El método `control-logic` ejecuta las velocidades configuradas en los motores:

```
def control_logic(self):
    if self.battery_level > 10:  # Verifica si la batería permite el movimiento
        wheel_distance = 0.5  # Distancia entre las ruedas
        left_speed = ((self.linear_velocity - self.angular_velocity) * wheel_distance) /
        right_speed = ((self.linear_velocity + self.angular_velocity) * wheel_distance) /
        self.left_motor.setVelocity(left_speed)
        self.right_motor.setVelocity(right_speed)
    elif self.obstacle_detected:  # Si hay un obstáculo, detiene el robot
        self.left_motor.setVelocity(0.0)
        self.right_motor.setVelocity(0.0)
```

1. **Verificación de batería:** Si la batería está por debajo de 10%, el robot no se moverá.
2. **Cálculo de velocidades:**
   - Combina las velocidades lineales y angulares para calcular las velocidades de cada rueda.
   - Esto se realiza utilizando un modelo de cinemática diferencial.
3. **Actualización de motores:** Las velocidades calculadas (`left_speed` y `right_speed`) se aplican a los motores del robot.

5.3.2.5 Manejo de batería baja

Si la batería es insuficiente, el robot detiene cualquier movimiento. Esto se maneja en `reduce-battery`, que también influye en el comportamiento de las órdenes:

```
def reduce_battery(self):
    self.step_count += 1
    if self.step_count >= self.battery_drain_interval:
        self.battery_level = max(0, self.battery_level - 1)
        self.step_count = 0
        print(f"Battery level: {self.battery_level}%")

        if 0 < self.battery_level <= 10:
            print("Battery critically low! Stopping motors.")
            self.left_motor.setVelocity(0.0)
            self.right_motor.setVelocity(0.0)

        if self.battery_level == 0:
            print("Battery depleted. Shutting down the robot.")
            rclpy.shutdown()
            exit(0)
```

6 Resultados

7 Conclusiones

Bibliografía

AGENCIAS, R. . (2018, Marzo). Así influyeron cambridge analytica y facebook en la victoria de trump. *rtve*. (Consultado el 16 de Septiembre de 2024.)

Agustinoy, A. (2024, Junio). La ocde actualiza sus principios sobre la ia. *Cuatrecasas*. (Consultado el 8 de Septiembre de 2024.)

Bonta, R. (2024, Marzo). California consumer privacy act (ccpa). *State of California Department of Justice*. (Consultado el 15 de Septiembre de 2024.)

de Protección de Datos, A. E. (2023, Noviembre). Privacy by design. *Agencia Española de Protección de Datos*. (Consultado el 18 de Septiembre de 2024.)

DigaLaw. (2010). *Digalaw x.* `https://www.digalawx.com/`. (Accedido en Septiembre de 2024.)

Emilio Gayo, C. o. T. E. (2019, Septiembre). The fourth industrial revolution is now a reality. *Telefónica*. (Consultado el 6 de Septiembre de 2024.)

Europea, C. (2016). Protección de datos. *Comisión Europea*. (Consultado el 15 de Septiembre de 2024.)

Europea, C. (2024, Agosto). Inteligencia artificial: preguntas y respuestas. *Comisión Europea*. (Consultado el 12 de Septiembre de 2024.)

Imagenet. (2017). Large scale visual recognition challenge 2017 (ilsvrc2017). *Web de Imagenet*. (Consultado el 9 de Septiembre de 2024.)

Jeff Larson, L. K., Surya Mattu, y Angwin, J. (2016, Mayo). How we analyzed the compas recidivism algorithm. *ProPublica*. (Consultado el 21 de Septiembre de 2024.)

Rivera, L. G. D. (2024, Marzo). Coches autónomos: ¿quién tiene la culpa en caso de accidente? *Público*. (Consultado el 12 de Septiembre de 2024.)

Lista de Acrónimos y Abreviaturas

AAS	Australian Acoustical Society.
ADAA	Asociación de Acústicos Argentinos.
AES	Audio Engineering Society.
AP	Average Precision.
APA	American Psychological Association.
ASA	Acoustical Society of America.
CCPA	Ley de Privacidad del Consumidor de California.
CSIC	Consejo Superior de Investigaciones Científicas.
EAA	European Acoustics Association.
GDPR	Reglamento General de Protección de Datos.
I-INCE	International Institute of Noise Control Engineering.
IA	Inteligencia Artificial.
ICA	International Congress on Acoustics.
IEEE	Institute of Electrical and Electronics Engineers.
IIAV	International Institute of Acoustics and Vibration.
IOA	Institute Of Acoustics.
IoT	Internet of Things.
ISRA	International Symposium on Room Acoustics.
ISVA	International Seminar on Virtual Acoustics.
NLP	Procesamiento natural del lenguaje.
OCDE	Organización para la Cooperación y Desarrollo Económicos.
REIA	Reglamento Europeo de para IA.
SEA	Sociedad Española de Acústica.
TFG	Trabajo Final de Grado.
TFM	Trabajo Final de Máster.
UE	Unión Europea.