Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

# Trabajo de Fin de Grado

Grado en Ingeniería Informática

---

# Ética en Sistemas de Inteligencia Artificial: Retos, Marcos Legales, y Viabilidad

*Ethics in Artificial Intelligence Systems: Challenges, Legal Frameworks, and Viability*

Miguel Dorta Rodríguez

---

San Cristóbal de la Laguna, 11 de Julio de 2024

Dña. **María Elena Sánchez Nielsen**, profesora Titular de Universidad adscrita al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutora

**C E R T I F I C A**

Que la presente memoria titulada:

*“Ética en Sistemas de Inteligencia Artificial: Retos, Marcos Legales, y Viabilidad”*

ha sido realizada bajo su dirección por D. Miguel Dorta Rodríguez.

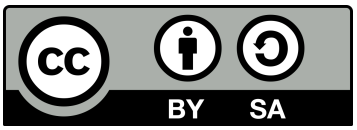
Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en San Cristóbal de la Laguna a 11 de Julio de 2024.

# Agradecimientos

A mi familia, por haberme apoyado en todo momento y permitir que pudiese estudiar aquello que me apasiona

A mi tutora Dra. María Elena Sánchez Nielsen, por la guía y ayuda en la realización de este trabajo

# Licencia



**Resumen**

Este trabajo presenta un análisis de la ética aplicada a la Inteligencia Artificial (IA), proporcionando una visión integral desde la formulación de principios hasta su aplicación. Inicialmente, este estudio categoriza las diferentes disciplinas de la ética aplicada que están relacionadas con los sistemas de IA. Este trabajo luego examina los dilemas éticos planteados por estos sistemas, como los sesgos algorítmicos, la privacidad, y la rendición de cuentas.

Posteriormente se realiza un análisis sobre los marcos legales que gobiernan la IA en China, los Estados Unidos, y la Unión Europea (UE). Adicionalmente este trabajo hace una revisión de las directrices de la UNESCO y de Ética por Diseño para el desarrollo de sistemas de IA éticos.

En la última sección, este documento investiga la implantación de la regulación de la UE y de las directrices de Ética por Diseño en un proyecto centrado en la implementación de un Modelo de Lenguaje Grande (LLM, *Large Language Model*) para el entorno parlamentario, tomando como referencia el Parlamento de Canarias. Aquí se incluyen estudios de las soluciones existentes en el mercado, evaluación de riesgos éticos, y plan de implantación que involucre a partes interesadas, usuarios, y desarrolladores.

**Palabras clave:** Ética, Inteligencia Artificial, Regulación, Directrices, Caso Práctico.

**Abstract**

This paper presents an analysis of ethics applied to Artificial Intelligence (AI), providing a comprehensive overview from the formulation of principles to their application. Initially, this study categorizes the different disciplines of applied ethics related to AI systems. The paper then examines the ethical dilemmas raised by these systems, such as algorithmic biases, privacy, and accountability.

Subsequently, an analysis is conducted on the legal frameworks that govern AI in China, the United States, and the European Union (EU). Additionally, this paper reviews the UNESCO and Ethics by Design guidelines for the development of ethical AI systems.

In the final section, this document investigates the implementation of the EU regulation and the Ethics by Design guidelines in a project focused on implementing an LLM within a parliamentary context, using the Parliament of the Canary Islands as a reference. This includes analysis of the existing solutions on the market, evaluation of ethical risks, and an implementation plan involving stakeholders, users, and developers.

**Keywords:** Ethics, Artificial Intelligence, Regulation, Guidelines, Case Study.

# Índice general

# Índice de figuras

# Índice de tablas

# Capítulo 1

# Introducción

Los sistemas de Inteligencia Artificial (IA) han tenido un enorme y acelerado surgimiento en los últimos años, trayendo consigo oportunidades y retos sin precedentes. A medida que los sistemas de IA se convierten en parte integral de muchos sectores como la sanidad, las finanzas, el transporte, y la gobernanza, las cuestiones éticas que provocan surgen como un problema crucial en su desarrollo e implementación. Este trabajo intentará hacer un análisis de las diferentes categorías éticas de la informática para centrarse en los problemas de la ética aplicada a la IA.

Los sistemas de IA traen consigo numerosos beneficios, pero también levantan dilemas éticos sobre sesgos algorítmicos, la falta de transparencia, y la falta de rendición de cuentas en la toma de decisiones. Es por esto que la implementación de estos sistemas debe estar fundamentada en fuertes principios éticos que atiendan a estos problemas de raíz.

Como veremos a lo largo de este trabajo, diversas organizaciones gubernamentales han creado marcos legales con el fin de regular estos sistemas y minimizar los posibles impactos indeseables en la sociedad. Analizaremos las propuestas regulatorias de las República Popular China, los Estados Unidos de América, y la Unión Europea (UE), para ver y comparar cómo diferentes entidades priorizan diferentes aspectos en el control y desarrollo de estos sistemas.

Más adelante, estudiaremos directrices de la UNESCO y de proyectos de la Unión Europea que intentan ir más allá de la legislación vigente para asegurar que el sistema de IA a desarrollar o implementar no es solo legal, sino también ético. Por último, usaremos estos conocimientos y estas directrices para realizar una implementación teórica de un sistema de IA en el entorno parlamentario, señalando las medidas a tomar y los problemas éticos que nos iremos encontrando.

En resumen, este trabajo intenta ser un estudio holístico de la ética aplicada a la IA, partiendo de su planteamiento y sus principios, pasando por su regulación, y acabando en un ejemplo de su aplicación.

# Capítulo 2

# Marco teórico y estado actual

## 2.1. Definiciones

Esta sección tiene el propósito de exponer de forma explícita las definiciones de los conceptos sobre los que se va a trabajar en esta memoria, para así crear un marco de comprensión común y poder establecer sobre qué ámbitos de la IA se estudiará la aplicación de la ética.

### 2.1.1. Inteligencia Artificial

La IA es, junto con la ética, uno de los conceptos más importantes que hemos de definir para el entendimiento de este trabajo. Este concepto ha tenido diversas definiciones que preceden incluso a la invención de la computadora, pero la que vamos a usar para este trabajo es una definición mucho más actual. Esta definición nos ayudará a delimitar el problema a los nuevos sistemas de IA que han mostrado sus potenciales en los últimos años.

La IA es comúnmente definida como un sistema mecánico que es capaz de mostrar cierta comprensión de problemas o situaciones, y que tiene la capacidad para resolverlos mediante un comportamiento complejo que no se ha programado *a priori*. Dentro de esta definición entran desde algoritmos sencillos como un algoritmo de resolución de caminos, hasta los Modelos de Lenguaje Grandes (LLMs, *Large Language Models*) que están mostrando su potencial estos últimos años.

Esta definición es, sin embargo, demasiado amplia para el ámbito de este trabajo, pues los problemas éticos que presenta en su totalidad son mucho más extensos que los que se pueden abarcar en un trabajo de estas dimensiones. Es por este motivo por lo que la definición de IA en este trabajo va a tratar sobre algoritmos que involucren aprendizaje automático. Esta condición hace posible poder abordar la aplicación de la ética y sus problemas en este documento.

Esta definición de IA como sistema o algoritmo que involucra aprendizaje automático no es exclusiva de este trabajo, pudiéndose decir que es comúnmente usada. Un ejemplo de ello lo podemos encontrar en la definición usada por la Organización para la Cooperación y el Desarrollo Económico (OCDE), que define un sistema de IA en [1] como “un sistema mecánico que, para objetivos explícitos o implícitos, infiere, a partir de la entrada que recibe, cómo generar salidas tales como predicciones, contenido, recomendaciones, o decisiones que pueden influir en entornos físicos o virtuales.” [1]

[1] Frase original: *“a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.”* [1]

### 2.1.2. Ética

La ética es la rama de la filosofía que se dedica al estudio de lo que se denominan fenómenos morales, los cuales consisten en situaciones, valores, principios, y comportamientos que pueden tener un juicio moral. Este juicio moral tradicionalmente distingue entre fenómenos morales buenos y malos, pero no necesariamente todas las teorías éticas utilizan esta dicotomía.

De la misma manera que limitamos el ámbito del significado de IA, tenemos que limitar el ámbito del significado del término de ética que vamos a usar en este trabajo. La discusión ética que tratará este trabajo será sobre descriptividad (es decir, el análisis de principios morales establecidos por organizaciones y comités), y aplicabilidad (es decir, cómo se aplican los principios anteriormente estudiados en casos prácticos), mientras que excluiremos otras discusiones sobre la ética aplicada a IA como la metaética (la reflexión sobre los sesgos y juicios que determinan las normas éticas, al igual que la reflexión sobre las unidades morales fundamentales como “bien” y “mal”).

### 2.1.3. Ética de la IA

La ética de la IA es una rama de la ética aplicada que intenta identificar los fenómenos morales que pueden darse en el planteamiento, desarrollo y utilización de tecnologías que caen bajo el término de IA. Las situaciones, valores, principios y comportamiento que se pueden dar en el campo de la IA pueden tener una profunda influencia en la vida de los individuos y el funcionamiento de la sociedad, es por esto que esta disciplina es definida por el Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial como:

> *“un subcampo de la ética aplicada que estudia los problemas éticos que plantea el desarrollo, despliegue y utilización de la IA. Su preocupación fundamental es identificar de qué modo puede la IA mejorar o despertar inquietudes para la vida de las personas, ya sea en términos de calidad de vida o de la autonomía y libertad humanas necesarias para una sociedad democrática* [2]*.”*

Esta disciplina, como es frecuente en todo campo de estudio, tiene bordes difusos que se entrelazan con otras disciplinas de ámbito similar. Por tanto, creemos que para focalizar más el ámbito del concepto de “ética de la IA” que se va a emplear en este trabajo, es necesario definir (y por tanto delimitar) las otras disciplinas de la ética cuyos objetos de estudio están relacionados con el de la IA.

### 2.1.4. Ética robótica

La ética robótica es la disciplina de la ética que abarca los fenómenos morales involucrados en el desarrollo y el uso de dispositivos robóticos. Muchas veces se usa este término para abarcar también la ética de la IA (probablemente debido a que el imaginario colectivo conceptualiza estos sistemas siendo encarnados en robots), pero se diferencia fundamentalmente en el objeto de estudio.

Los objetos de estudio que esta disciplina comparte con la ética de la IA son los *“killer robots”* (máquinas de guerra con cierto grado de autonomía), los vehículos autónomos, o los dispositivos que puedan reemplazar a la mano de obra humana. Los objetos de estudio únicos de esta

disciplina son los robots sexuales, los dispositivos médicos, o los dispositivos robóticos industriales no inteligentes.

Los problemas que esta disciplina comparte con la ética de la IA son, por ejemplo, responsabilidad (en sentido de *"accountability and liability"*), o el manejo de situaciones en las que los accidentes son inevitables. Los problemas que no comparte con la ética de la IA son problemas como las consecuencias sociales y psicológicas que puede tener el amplio uso de dispositivos robóticos sexuales, o qué tipo de decisiones de diseño se deben tomar, y cuándo, para permitir a un humano tomar el control sobre una máquina.

Un ejemplo de esta problemática que actualmente está en los medios de comunicación es sobre la autoridad de los pilotos de avión para tomar el control sobre la aeronave. Los dispositivos que se usan para la asistencia de vuelo no suelen ser calificados como IA, en cuanto a que son una serie de parámetros y directrices de vuelo calculadas y programadas por humanos. Muchos de estos sistemas están diseñados para poder ser anulados por un humano, lo cual en ocasiones ha contribuido a accidentes aéreos por errores de los pilotos (por ejemplo, en el caso del vuelo Colgan 3407 [3]), pero actualmente vemos como no permitir a un humano tomar el control también ha sido una de las causas en los accidentes que involucran el sistema MCAS del Boeing 737 Max.

### 2.1.5. Ética de algoritmos

La ética de los algoritmos es la rama de la ética aplicada que se encarga de estudiar el diseño, implementación, y consecuencias sociales del uso de algoritmos (en cuanto a sistema predictivo, de reconocimiento, o recomendación, desde los que se basan en simple regresión lineal hasta los que funcionan mediante IA o aprendizaje automático). Esta rama estudia el comportamiento de los algoritmos porque pueden, de forma autónoma, ejecutar o motivar acciones que tengan consecuencias éticas [4].

Al igual que la rama descrita en el punto anterior, esta disciplina coincide en muchas cuestiones con la ética de la IA, pero se diferencia en cuanto a que su objeto de estudio son los problemas éticos que comparten tanto algoritmos complejos como sistemas de aprendizaje automático, hasta otros mucho más sencillos. Esta disciplina comparte los problemas con la ética de la IA (pues la IA como la hemos definido también se considera un algoritmo), pero no todas las cuestiones éticas sobre la IA están dentro de esta disciplina, pues estos sistemas abarcan otra problemática como la conciencia o con la humanización de la IA por el público general. Dilemas morales que abarcan ambas éticas son cuestiones como la discriminación, la coacción, o la manipulación del comportamiento humano que pueden ocasionar este tipo de algoritmos.

### 2.1.6. Ética de los datos

La ética de los datos es la aplicación de la ética en los procesos de adquisición, administración y uso de datos. Estos procesos pueden abordar cuestiones éticas como la privacidad, la confidencialidad, la honestidad, o la responsabilidad de los actores implicados en el uso y tratamiento de estos datos, al igual que cuestiones sobre la transparencia de estos procesos [5].

Al igual que las anteriores ramas de la ética, esta disciplina es una rama que también tiene bastante coincidencias con la ética de la IA, porque estos sistemas suelen hacer uso de gran cantidad de datos, los cuales pueden levantar cuestiones éticas sobre su uso, procesamiento,

diseminación, almacenamiento, y borrado. La coincidencia no es completa, pues es posible conceptualizar una IA que esté desarrollada sin uso de datos, o con uso de datos que no levanten cuestiones éticas (por ejemplo, una IA entrenada en un videojuego), y también es posible ver cómo el uso y tratamiento de datos, con independencia de que se use IA o no, levanta otros tipos de planteamientos éticos.

## 2.2. Principales desafíos y peligros de la IA

La creciente implementación de los sistemas de IA en múltiples aspectos de nuestra vida hace que sea esencial plantearnos a qué tipo de desafíos éticos se enfrenta, y qué peligros puede tener un mal uso de los mismos.

Las *Directrices Éticas para una IA Fiable* [6] hace una calificación de estos desafíos, dividiéndolos en:

1. Acción y supervisión humana:
2. Solidez técnica y seguridad
3. Gestión de la privacidad y los datos
4. Transparencia
5. Diversidad, no discriminación y equidad
6. Bienestar ambiental y social
7. Rendición de cuentas

Para verificar la importancia de los desafíos que señala este documento, es esencial que sea contrastado con la evidencia empírica. Para ello, nos hemos basado el trabajo de Mengyi Wei y Zhixuan Zhou [7], que hacen un análisis de 150 incidentes relacionados con IA, y los categorizan, por orden de frecuencia, en las siguientes categorías:

1. Uso inapropiado (mal desempeño)
2. Discriminación racial
3. Seguridad física
4. Algoritmo injusto (evaluación)
5. Discriminación de género
6. Privacidad
7. Uso no ético (uso ilegal)
8. Salud mental

Como se puede observar, hay un gran solapamiento entre los peligros e incidencias que ocurren a día de hoy con la IA, con los aspectos que abarcan los mismos puntos que del documento de la Comisión Europea. Es por eso que consideramos que el planteamiento de los desafíos éticos que enumera este documento es apropiado y suficiente para contemplar la totalidad de los problemas éticos que pueden derivar del uso de estos sistemas.

## 2.3. Legislación aplicable

En vista de los riesgos descritos en la sección anterior, no es de extrañar que empiecen a surgir regulaciones para los sistemas de IA. En esta sección examinaremos las regulaciones y propuestas que tres entidades, China, los Estados Unidos, y la Unión Europea, han impuesto al desarrollo y uso de estos sistemas.

Hemos escogido analizar la legislación en estas tres entidades pues las consideramos como los tres grandes poderes geopolíticos de la actualidad. Nuestra consideración se basa en que son entidades con una alta cohesión económica, social y territorial con marcos legislativos comunes, y cuyo poder económico es determinante a nivel global: usando la métrica del Producto Interior Bruto ajustado a valores de Paridad del Poder Adquisitivo (PIB/PPA), estas tres entidades suman un 49.11% del planeta [8]. La capacidad económica de estos estados hace que sean los líderes mundiales en el desarrollo de nuevas tecnologías, y es por tanto crucial ver cómo sus legislaciones van a afectar a estas.

### 2.3.1. República Popular China

Empezaremos este análisis examinando la legislación que está en vigor en China, el cual ha sido uno de los primeros estados que ha establecido legislación acerca de los sistemas de IA. De acuerdo con [9], los principios que establece el Ministerio de Ciencia y Tecnología de China para la gobernanza de la IA y una IA responsable, son los siguientes:

1. Armonía y amabilidad (和谐友好)
2. Equidad y justicia (公平公正)
3. Inclusión y uso compartido (包容共享)
4. Respeto por la privacidad (尊重隐私)
5. Seguro y controlable (安全可控)
6. Responsabilidad compartida (共担责任)
7. Colaboración abierta (开放协作)
8. Gobernanza ágil (敏捷治理)

Más allá de estos principios, la regulación de la IA en China está determinada por dos regulaciones: el *Reglamento para la Gestión de Recomendaciones Algorítmicas en los Servicios de Información de Internet* [10] y el *Reglamento para la Gestión de Servicios de Información de Internet de Síntesis Profunda* [11]. A grandes rasgos, estas normativas establecen que los sistemas de IA deben:

- Adaptarse a principios éticos [10, art. 4] mencionando en ocasiones específicamente que no deben generar adicción ni incentivar al gasto desproporcionado [10, art. 8].
- No violar las leyes o la regulación vigente [10, art. 6-7].
- Se deben comprobar y verificar periódicamente esos sistemas de IA [10, art. 8].
- Deben detectar información que sea falsa, negativa, o dañina; y de ser este tipo de información detectada, se tiene que parar su difusión, eliminarla, y reportar el incidente a la autoridad correspondiente [10, art. 9].
- No pueden crear o emitir noticias que no han sido aprobadas por el estado [10, art. 13].

- Deben evitar manipular al público [10, art. 14].
- Deben evitar las prácticas monopolistas y la competencia desleal [10, art. 15].
- Los proveedores de IA deben comprobar la identidad real de los usuarios del sistema o los siguientes proveedores en la cadena de valor [11, art. 9].
- Donde puede haber confusión para el público, se debe etiquetar de forma visible que el contenido ha sido creado por este tipo de sistemas [11, art. 17].

Aparte de estos reglamentos, en China hay otros que por brevedad sólo nombraremos, pero que influyen también en los sistemas de IA, como son la *Ley de Ciberseguridad*, la *Ley de Seguridad de los Datos*, o la *Ley de Protección de Información Personal* [12].

### 2.3.2. Estados Unidos de América

De forma opuesta a la República Popular de China, los Estados Unidos han sido la superpotencia que más tiempo ha tardado en proponer legislación para la IA. Existe legislación en múltiples estados, pero hasta muy recientemente el gobierno federal se había abstenido de formular una regulación específica a estos sistemas.

En el ámbito federal, la legislación de referencia es la *Orden Ejecutiva para el Desarrollo y Uso Seguro y Confiable de la Inteligencia Artificial* [13], un decreto presidencial dictado por J. Biden en octubre de 2023. Esta orden ejecutiva propone la creación de la siguiente regulación:

- La creación de nuevas directrices, estándares y buenas prácticas para una IA segura y confiable, involucrando al Instituto Nacional de Estándares y Tecnología (NIST, *National Institute of Standards and Technology*) y a las Agencias de Gestión de Riesgos del Sector (SRMA, *Sector Risk Management Agencies*), entre otros [13, sec. 4].
- La imposición para los proveedores de Infraestructura como Servicio (IaaS, *Infraestructure as a Service*) de notificar a la Secretaría de Comercio de los Estados Unidos del uso de su infraestructura para el entrenamiento de IA por agentes externos a los Estados Unidos [13, sec. 4.2].
- La obligación de que las compañías que desarrollen o quieran desarrollar sistemas de IA reporten al Gobierno Federal Estadounidense los resultados de los tests de seguridad y otra información crítica que puedan afectar a la seguridad nacional [13, sec. 4.2].
- La creación de regulación para reducir los riesgos:
    - Que estos sistemas pueden presentar en términos de ciberseguridad [13, sec. 4.3].
    - Asociados al uso de estos sistemas para la creación de armas biológicas [13, sec. 4.4].
    - Asociados a los denominados “Contenidos Sintéticos”: incluyendo su marcado (p.e. con marcas de agua) y prohibiendo su uso para la producción de contenido sexual sin consentimiento [13, sec. 4.5].

- La elaboración de regulación para abordar la discriminación ilegal en los sistemas de IA [13, sec. 7].
- La protección de los ciudadanos americanos frente al fraude, la discriminación, y las amenazas a la privacidad.
- Protección de la privacidad [13, sec. 9], incluyendo una petición al Congreso de los Estados Unidos para aprobar legislación sobre la privacidad de los datos [14].
- Proporcionar orientación a las agencias estatales para el uso seguro de los sistemas de IA [13, sec. 10] [13, sec. 12].
- Desarrollar estándares de seguridad y manejo de riesgos de la IA con aliados internacionales [13, sec. 11].

Esta orden ejecutiva se basa en los principios que el propio Biden expuso en su *Anteproyecto para una Declaración de Derechos de la IA* [15], en el cual mencionaba que futura legislación en IA debían basarse en:

1. Sistemas seguros y eficaces
2. Protecciones contra la discriminación algorítmica
3. Privacidad de los datos
4. Aviso y explicación
5. Alternativas humanas, consideración, y respaldo.

Finalmente señalar que esta orden ejecutiva basa su análisis de los riesgos que contiene la IA para las personas y para la sociedad en conjunto en el *Marco de Gestión de Riesgos de la Inteligencia Artificial* (*Artificial Intelligence Risk Management Framework*) propuesto por el NIST en 2023. La orden ejecutiva de Biden marca este estándar como esencial en la determinación de los riesgos, y solicita su expansión para abarcar la IA generativa [13, sec. 4.1].

### 2.3.3. Unión Europea

La Unión Europea se encuentra en una posición intermedia entre las dos potencias, habiendo pasado recientemente una nueva normativa con el fin de regular los sistemas de IA, titulada *Reglamento de Inteligencia Artificial*, pero comúnmente conocida por su título en inglés, *the Artificial Intelligence Act*.

La regulación que se define en esta ley tiene 3 claros propósitos: delimitar qué tipo de prácticas de IA están prohibidas, cómo regular los denominados “sistemas de alto riesgo”, y qué requisitos de transparencia deben cumplir los sistemas de IA de propósito general.

En cuanto a las prácticas de IA que están prohibidas, el Parlamento Europeo describe:

- La manipulación o engaño de personas y colectivos vulnerables [16, art. 5.1a-5.1b].
- La realización de una puntuación ciudadana (también conocido como *“social score”*) [16, art. 5.1c].

- El intento de evaluar o predecir la probabilidad de un individuo para cometer un delito basándose únicamente en sus características y personalidad [16, art. 5.1d].

- La colección de datos biométricos de fuentes no selectivas [16, art. 5.1e] y su uso para categorizar ideológica, personal, o emocionalmente a individuos [16, art. 5.1f-5.1g].

En cuanto a la regulación de los "sistemas de alto riesgo", el *Reglamento de Inteligencia Artificial* los define siguiendo una lista extensa de características entre las que se encuentran la elaboración de perfiles mediante el tratamiento de datos personales, la recolección o uso de biometría no prohibida, los sistemas en infraestructuras críticas, etcétera. Estos sistemas y las acciones que realizan pueden tener grandes impactos personales o sociales y, por tanto, el Parlamento Europeo ha dictaminado que deben seguir la siguiente normativa [17]:

- Establecer un sistema de gestión de riesgos.

- Llevar a cabo gobernanza de los datos, garantizando validación, prueba, y comprobando que sean representativos, completos, y estén libres de errores.

- Elaborar documentación técnica que corrobore la conformidad del sistema con la legislación vigente.

- Diseñar el sistema para que registre automáticamente eventos que puedan ser relevantes para la seguridad nacional.

- Proporcionar instrucciones de uso.

- Diseñarlos de tal forma que permitan la supervisión humana.

- Diseñarlos para alcanzar niveles adecuados de precisión, solidez y ciberseguridad.

- Establecer un sistema de gestión de la calidad.

Podemos ver que esta legislación se asemeja en varios puntos a los requisitos exigidos por el *Reglamento General de Protección de Datos* [18].

Finalmente, en cuanto a los requisitos de transparencia que tienen que cumplir las Inteligencias Artificiales de Propósito General (IAPG) son los siguientes: publicar un resumen detallado sobre el contenido usado para el entrenamiento del modelo, y adoptar directrices para el cumplimiento de la *Directiva sobre los Derechos de Autor* [16, cdo. 104].

Además de estas medidas, las IAPGs (y sus datos) que presenten un riesgo sistémico [16, cdo. 104], o aquellas que no sean divulgadas con una licencia libre, de código abierto, y de forma gratuita [16, cdo. 102-103], estarán también obligadas a tener documentación técnica que incluya el proceso de entrenamiento, validación, y pruebas del modelo [16, anexo XI sec. 2]. Esta documentación también tiene que estar diseñada para ayudar a los siguientes proveedores en la cadena de valor a entender el modelo y sus capacidades, tanto para la integración en sus productos como para cumplir con el Reglamento [16, cdo. 101].

Los proveedores de IAPGs que estén consideradas de riesgo sistémico, deben seguir, además de las anteriores, las siguientes medidas:

- Detectar y atenuar los riesgos de estas, incluyendo pruebas adversarias y pruebas independientes [16, cdo. 114].

- Evaluar y mitigar continuamente los riesgos sistémicos, estableciendo una política de gestión de riesgo, poniendo en práctica vigilancia poscomercialización, y cooperando con los agentes pertinentes a lo largo de la cadena de valor [16, cdo. 114].

- Asegurar que su infraestructura tenga un nivel adecuado de protección en sistema de ciberseguridad [16, cdo. 115].

- En caso de incidente grave, hacer un seguimiento del incidente y comunicar toda información pertinente y medidas a las autoridades competentes [16, cdo. 115].

Además del *Reglamento de Inteligencia Artificial*, estos sistemas están sujetos a otras regulaciones existentes en la Unión, como la *Directiva sobre Derechos de Autor en el Mercado Único Digital*, o el *Reglamento General de Protección de Datos* (GDPR, *General Data Protection Regulation*).

Estas regulaciones también son de gran importancia pues, como hemos podido observar en los últimos años, muchos sistemas de IA han sido entrenados sin tener en consideración la autoría de los datos que han usado. Esto ha dado lugar a casos notorios de IAs que generaban imágenes con marcas de agua [19].

Por último, hemos de señalar que el GDPR es también una normativa relevante para estos sistemas de IA, pues suelen entrenarse con inmensas cantidades de información en las cuales puede haber información personal de usuarios que no han sido correctamente informados que sus datos serían usados para este propósito.

# Capítulo 3

# Gobernanza de la IA y ética por diseño

La gobernanza de la IA, como la definen Tim Mucci y Cole Struker para IBM, consiste en el establecimiento de marcos, reglas y normas para garantizar la seguridad, equidad y respeto de los derechos humanos en el desarrollo y aplicación de la IA [20]. Esta práctica es esencial para evitar los impactos negativos que los sistemas de IA pueden causar a individuos, organizaciones, o sociedades. En este capítulo expondremos varios conjuntos de reglas, recomendaciones, y metodologías que se pueden aplicar a la planificación, desarrollo, aplicación, y mantenimiento de los sistemas de IA.

## 3.1. De los marcos éticos para la IA a las herramientas: una revisión de enfoques

Comenzaremos este capítulo dedicando una sección al artículo de Erich Prem [21] con el cual comparte título. Este artículo se fundamenta en un análisis sistemático de más de 100 marcos teóricos, reglamentos, soluciones y herramientas. En él, se señalan los principios, metodologías, y soluciones principales que los marcos y directrices analizadas proponen, además de señalar qué problemas contienen. Este documento lo utilizaremos como referencia para contrastar las metodologías, soluciones, y problemas que existen en las directrices éticas propuestas por la UNESCO y la Comisión Europea.

Para usar este artículo de Prem como referencia en nuestro análisis, primero debemos entender qué problemas son los que él detecta en los marcos éticos y directrices, además de qué soluciones propone.

El principal problema que detecta, y del cual se derivan muchos otros, es que la aproximación metodológica que la mayoría de marcos teóricos y herramientas usan para la toma de decisiones es el principialismo. Como su propio nombre indica, el principialismo es el enfoque ético de derivar las acciones de principios o máximas. Él identifica 3 principales problemas con esta aproximación ética: la adopción del principialismo por muchos documentos para garantizar la ética de la IA, el foco de esta filosofía en la discusión de los principios en vez de las soluciones, y lo débil que son los principios cuando son contrastados con la acción.

Los dos primeros problemas señalados por Prem son propios de una discusión filosófica, pero el tercero sí tiene cabida en un documento como este, pues uno de los problemas más frecuentes en este ámbito es cómo llevar los principios a la práctica. Esto se debe a que la discusión de principios éticos es una discusión a muy alto nivel, y por tanto no queda claro cómo se traducen en tareas y requisitos, quedando abierto a interpretación. Todo esto son problemas que Prem ilustra en su artículo, indicando también que debido a esta interpretación subjetiva de principios generales, la interoperabilidad entre sistemas no puede estar garantizada.

Otros problemas que también se señalan en este artículo son:

- **La falta de jerarquía entre principios** para resolver situaciones en las que sea imposible satisfacer dos o más simultáneamente.
- **Asumir que existe una visión holística del sistema.** Que está programado por un grupo muy reducido de programadores que pueden tener una visión clara de cómo funciona el sistema en su totalidad.
- **El problema de la emergencia / composicionalidad**, en el que las partes pueden cumplir con los estándares éticos pero el sistema en su conjunto no.

Erich Prem también extrae de su análisis que el proceso de planificación, desarrollo e implantación de un sistema de IA puede dividirse en 9 partes, en cada una de las cuales podemos distinguir problemas éticos prioritarios. Estas fases son obtenidas haciendo una puesta en común de los diferentes documentos y guías que ha analizado, sacando factores comunes a las metodologías que contienen. Este método se divide en las siguientes fases:

1. Desarrollo de negocio y caso de uso
2. Diseño del sistema
3. Creación de datos
4. Comprensión de los datos
5. Preprocesamiento
6. Entrenamiento del modelo
7. Prueba y evaluación
8. Despliegue
9. Supervisión

La última observación que tomaremos del artículo de Prem es la comparación que hace entre el campo de la ética aplicada a la medicina y la ética aplicada a la IA. Erich Prem argumenta que, de la misma forma que la ética en el campo de la medicina ha evolucionado hasta el establecimiento de buenas prácticas, comités, directrices, y regulación, la ética en IA debe desarrollar a partir del principialismo vigente código, infraestructura, educación, comunidad, buenas prácticas y otras herramientas prácticas para su aplicación en situaciones estándares y casos específicos.

Habiendo visto y examinado este artículo, vamos ahora a utilizarlo como referencia para estudiar las directrices para la gobernanza que da la UNESCO en su *Recomendación sobre la Ética de la Inteligencia Artificial* [22], y la Comisión Europea en su documento *Enfoques de Ética por Diseño y Ética de Uso para la Inteligencia Artificial* [23].

## 3.2. Recomendación de la UNESCO

Esta sección la dedicaremos a enumerar las directrices para la gobernanza que expone la UNESCO en [22]. En este documento se exponen los valores y principios que este organismo de las Naciones Unidas argumenta que deberían ser respetados y promovidos por todos los actores involucrados en el ciclo de vida de los sistemas de IA. De estos valores y principios se derivan los artículos para la gobernanza que se definen en la sección de “Ámbitos de acción política: Gobernanza y administración éticas”, los cuales dictan lo siguiente.

- **Sobre los agentes que deben gobernar la IA (estados, organizaciones regionales, y empresas):** Deben establecer una estrategia nacional o regional para la gobernanza de la IA [22, art. 56] [22, art. 68] que respete y promulgue los derechos humanos [22, art. 61], debe existir una entidad u organismo que supervise la IA [22, art. 58] [22, art. 62], se debe reforzar el sistema judicial de los estados para poder regular de forma efectiva la IA [22, art. 63] considerando esencial que la responsabilidad de los sistemas de IA debe *"recaer siempre en última instancia en personas físicas o jurídicas*" [22, art. 68]. A parte, los estados deben ser líderes en materia de seguridad y transparencia de la IA [22, art. 64] y deben fomentar un ecosistema digital que fomente el desarrollo ético de la IA [22, art. 59].

- **Sobre la IA:** Debe ser prohibida si viola los derechos humanos [22, art. 57], y debe ser transparente y explicable durante su ciclo de vida [22, art. 70].

- **Sobre la igualdad y diversidad en los sistemas de IA:** se debe asegurar la participación internacional de los estados de ingreso mediano bajo [22, art. 60], se debe involucrar a todos los actores en la elaboración de las normas y regulaciones [22, art. 69], la IA debe respetar la diversidad cultural [22, art. 65], combatir estereotipos, desigualdades y desinformación [22, art. 66], y promover la diversidad e inclusión de los grupos marginados [22, art. 67].

- **Sobre los mecanismos de gobernanza de la IA:** Deben ser transparentes, inclusivos y multilaterales [22, art. 54].

- **Sobre las consecuencias de la IA:** Los daños que estos sistemas causen deben ser reparados [22, art. 55].

Se puede ver que los principios expuestos en este documento son bastante completos y abarcan múltiples ámbitos de la gobernanza de la IA, pero gracias al análisis de Prem que vimos en el apartado anterior, podemos ver cómo estas recomendaciones tienen un nivel de abstracción muy alto, pudiendo dar cabida a interpretaciones y decisiones subjetivas. Otro de los problemas que este autor señalaba y que podemos aplicar a este artículo es la falta de jerarquía entre los conceptos que, si bien es algo que la UNESCO considera en [22, art. 11], no proporciona ninguna solución y lo deja como una cuestión abierta en su documento.

## 3.3. Ética por diseño

La última sección de este capítulo la dedicaremos al documento *Enfoques de Ética por Diseño y Ética de Uso para la Inteligencia Artificial* [23], redactado por la Comisión Europea en 2021. La metodología descrita en él, "ética por diseño", es paralela a otras metodologías como la seguridad por diseño, o la privacidad por diseño. Todas estas son metodologías que intentan dar soluciones a sus respectivos problemas tan pronto como aparecen en la planificación, diseño o implementación de un sistema.

Las metodologías de "X por diseño" se contrastan con otras que son *a posteriori*, que intentan evaluar y solucionar los problemas una vez realizadas las implementaciones. Estas otras metodologías, como expone la Comisión Europea, son defectuosas en la práctica, pues hay requisitos que no pueden ser alcanzados si no se diseñan los sistemas con ellos en mente. La efectividad de estas metodologías se puede observar en el amplio uso de ellas por muchas

empresas líderes del sector, como Microsoft [24] o Google [25], que usan esta aproximación proactiva para solventar los posibles problemas de sus sistemas de forma preventiva.

Este documento de la Comisión Europea se divide principalmente en tres partes. En la primera se exponen los principios y requisitos que debe tener una IA fiable, en la segunda parte se centran en cómo aplicar estos requerimientos de forma práctica (centrándose principalmente en la fase de desarrollo), y en la tercera ponen el foco en proporcionar directrices para el despliegue y uso de estos sistemas de IA.

Analizando la primera parte, vemos que en ella se exponen los principios y requisitos que deben cumplir los sistemas de IA. Estos principios se basan en el documento de *Directrices Éticas para una IA Fiable* [6], además de en los proyectos SHERPA y SIENNA[2] de la Unión Europea. Los principios que se describen son los siguientes:

- Respeto de la voluntad humana
    - Autonomía
    - Dignidad
    - Libertad
- Privacidad, protección de los datos personales, y gobernanza de los datos
- Imparcialidad
- Bienestar individual, social, y ambiental
- Transparencia
- Rendición de cuentas y supervisión

En la segunda parte del documento de *Ética por Diseño* se explica una metodología para derivar métodos y herramientas a partir de una definición de principios éticos generales. Este proceso se divide en 5 capas de abstracción, en las que se parte de los principios éticos en la capa superior (la más abstracta) y se va concretando hasta llegar a aplicaciones prácticas de estos. Las capas del modelo que describen son las siguientes:

1. **Principios:** Postulados éticos que, si el sistema incumple, haría que lo considerásemos inmoral.

2. **Requisitos éticos:** Condiciones que debe cumplir el sistema de IA para satisfacer los principios éticos.

3. **Directrices de la ética por diseño:** Tareas específicas que tienen que ser realizadas para que el sistema cumpla con cada uno de los requisitos éticos.

4. **Metodologías de IA:** Consiste en los detalles de cómo se deben cumplir las tareas dictaminadas por las directrices de ética por diseño en las diferentes metodologías usadas de desarrollo usadas comúnmente para sistemas de IA (Agile, V-Method, etc).

5. **Herramientas y métodos:** Herramientas y procesos específicos, como la ficha técnica de los datasets, usados para evaluar las características éticas de los datos.

[2] Los proyectos SHERPA (*Shaping the Ethical Dimensions of Smart Information Systems*) [26] y SIENNA (*Stakeholder-Informed Ethics for New technologies with high socio-ecoNomic and human rights impAct*) [27] son proyectos de investigación de la ética en campos emergentes de la ciencia y tecnología. Ambos proyectos están financiados por el Programa de Investigación e Innovación Horizonte 2020 de la Unión Europea.

Gracias a este modelo, tenemos una referencia de cómo pasar de los principios generales de la ética, a las herramientas y tareas que necesitamos para hacer un desarrollo de un sistema de IA ético.

En cuanto a la metodología, el documento de *Ética por Diseño* se basa en la premisa de que los procesos de desarrollo de IA pueden ser divididos en al menos 6 fases: especificación de objetivos, especificación de requisitos, diseño a alto nivel, recolección y preparación de los datos, diseño detallado y desarrollo, prueba y evaluación. Para cada una de estas fases, la guía de ética por diseño nos ofrece recomendaciones.

- **Especificación de objetivos:** Para esta fase, el documento de *Ética por Diseño* nos recomienda valorar la posibilidad e impacto de los principios éticos y, si fuese necesario, determinar la prioridad entre ellos para nuestro proyecto. Se debe realizar esta valoración con miembros del equipo de desarrollo y con las partes interesadas (*stakeholders*) del proyecto, y se debe considerar siempre el potencial uso malicioso (intencional o accidental) del sistema. Si los riesgos se determinan muy altos, o se violan principios básicos, en esta fase se puede determinar y cancelar el desarrollo.

- **Especificación de requisitos:** En esta fase se definen los recursos, arquitectura, además de otras especificaciones, y se contrastan con los requisitos éticos. Se señala la importancia también de realizar una evaluación ética de riesgos e impacto, que incorpore estructuras organizacionales y procedimientos de ética al proceso de desarrollo.

- **Diseño a alto nivel:** En esta fase se define a alto nivel la arquitectura del sistema de IA. La recomendación señalada en el documento para esta fase es tratar los requisitos éticos de forma idéntica a cualquier otro requisito del sistema, planificando y definiéndolo en las metas del desarrollo. A parte de esto nos proporciona varias directrices éticas, especialmente centradas en riesgo, impacto, transparencia y explicabilidad.

- **Recolección y preparación de los datos:** Esta fase consiste en la "recolección, verificación, limpieza e integración de los datos" [23, p. 12]. Las recomendaciones tienen su foco en la adherencia a la GDPR, y en asegurar la imparcialidad y la precisión de los datos. Con respecto al aspecto de la imparcialidad, nos recomienda siempre asumir que todo dato recolectado está sesgado, distorsionado o es incompleto hasta que se pruebe lo contrario, además de señalar que debemos evitar introducir otros sesgos o distorsiones en la fase de procesamiento de los datos.

- **Diseño detallado y desarrollo:** Esta fase consiste en la implementación de la planificación en un sistema funcional. Las recomendaciones que aquí se nos dan enfatizan mucho la comunicación y clara comprensión de las directrices éticas por los desarrolladores e ingenieros, incluyendo formación y concienciación. Además, se recomienda el uso de la planificación realizada previamente para contrastar las decisiones éticas en la toma de decisiones.

- **Prueba y evaluación:** En esta fase deberíamos tener un sistema que cumple con los requisitos éticos propuestos en su planificación, pues la metodología descrita en *Ética por Diseño* nos indica que esos requisitos tienen que estar satisfechos para considerar la fase de desarrollo como completada. Las recomendaciones en esta fase consisten en hacer una evaluación ética del sistema, y proyectar los requisitos éticos de la planificación para tratar cualquier desviación de las características éticas del sistema como un error. Además, en esta fase se nos recomienda comunicar de forma clara y proactiva a las capacidades y

limitaciones del sistema, evaluar la comprensión de los usuarios sobre ello, y tener líneas abiertas para obtener retroalimentación de terceros.

Finalmente, en la última parte de este documento la Comisión Europea, se nos presentan algunas directrices para diferentes ámbitos de la IA, con foco en el despliegue y el uso de estos sistemas. Principalmente son pautas que derivan de las expuestas en la parte de metodología del documento, pero de ellas es especialmente relevante señalar las siguientes:

- **En el ámbito de la planificación y dirección de proyectos de IA:** Planificar y presupuestar teniendo en cuenta los problemas éticos que potencialmente pueden ocurrir, además de definir reglas y procedimientos para implementar las directrices éticas y monitorizar su implementación.

- **En el ámbito de la adquisición de software de IA:** En caso de obtener una solución existente, adquirir la que sea más capaz de adaptarse a los requisitos éticos; en caso de obtener una solución a medida, ofrecer orientación a los desarrolladores.

- **En el ámbito del despliegue e implantación:** Borrar los datos de carácter personal que no sean necesarios para el funcionamiento del sistema, y monitorizar las normas éticas durante su implantación.

- **En el ámbito del monitoreo:** Crear mecanismos para que terceras partes puedan ofrecer retroalimentación en problemas éticos o funcionales del sistema, además de monitorizar periódicamente la conformidad del sistema con los objetivos y las métricas éticas.

Al evaluar este documento podemos observar que tiene en consideración la mayoría de los problemas que Erich Prem señalaba en su artículo, pues no sufre de múltiples problemas que Prem señalaba que muchos otros sí. Se puede ver cómo el mayor problema que Prem destacaba: la poca claridad de cómo pasar de principios éticos a medidas prácticas, es solventada en este documento con el modelo de 5 capas y las recomendaciones que nos ofrecen en cada una de las fases del desarrollo.

Otro de los problemas que este autor señalaba era la falta de jerarquía entre los principios, además de basarse en la idea de que los equipos pueden tener una visión holística del proyecto. Con respecto al primero de estos puntos, el documento de *Ética por Diseño* no proporciona una solución única, sino que nos recomienda incorporar esta consideración en la primera fase de especificación de objetivos. Con respecto al segundo punto, este documento acaba con ese problema de raíz, en cuanto a que las directrices éticas se plantean desde una etapa muy temprana del desarrollo del sistema, cuando no existen los detalles de implementación y, por tanto, es posible tener una idea completa de cómo va a funcionar el sistema.

Sin embargo, uno de los problemas que señala Prem y que no resuelve este documento es el problema de la emergencia o composicionalidad, en la que todas las partes de un sistema pueden seguir unas normas éticas, pero la suma de esas partes no. Es discutible si este problema puede surgir en una aproximación que va desde el caso general al específico (que es la aproximación que sigue este documento) pero es cierto que no hay mención de esta posibilidad.

Otra de las cosas que podemos distinguir en el documento de *Ética por Diseño* es que, en él, se dividen los problemas éticos por fases de la metodología de desarrollo de IA y, si bien es cierto que la que describe Prem es más granular, no se contradice con esta propuesta, pudiéndose incluso hacer un mapeo de 1:N entre ambas.

Una última observación es que la Comisión Europea dictamina múltiples directrices, pone ejemplos de situaciones comunes, y comenta qué acciones tomar en ellas. Esto viene a mostrar la progresión que Prem señala que debe seguir el ámbito de la ética aplicada a sistemas de IA. Prem describe esta progresión como la necesidad de pasar del principialismo a buenas prácticas para situaciones estándares, directrices y regulación.

Es por esta lista de motivos por lo que creemos que el documento de *Ética por Diseño* es una excelente herramienta para guiar el desarrollo e implantación de sistemas de IA, y es por ello por lo que lo usaremos como referencia para el caso práctico que examinaremos en este documento.

# Capítulo 4

# Caso de uso: Entorno parlamentario

Este capítulo estará dedicado a la aplicación práctica de los conocimientos de la normativa y recomendaciones expuestas en capítulos anteriores. El caso de uso que hemos seleccionado para ilustrar la aplicación de la ética es la implantación y uso de sistemas de IA en el entorno parlamentario, tomando como referencia el Parlamento de Canarias.

## 4.1. Funciones y responsabilidades de un parlamento

Es esencial hacer un análisis del entorno que vamos a estudiar para poder determinar qué problemas éticos pueden presentarse con la implantación de la IA en él. Para ello, hemos hecho un estudio de la funcionalidad de un parlamento (en este caso, el de Canarias) basándonos en un análisis del *Estatuto de Autonomía de Canarias* [28] y la *Constitución Española* [29]. Gracias a este análisis, podemos determinar que las funciones de un Parlamento son:

- **Representar al pueblo:** La función más importante y legitimadora que tiene el Parlamento es ser el órgano constitucional que representa al pueblo, debido a que sus miembros son elegidos de forma directa mediante sufragio universal [29, pto. 66.1] [29, pto. 68.1] [28, pto. 38.1].
- **Legislar:** Crear proyectos de ley y enmiendas es una de las tareas principales que tienen los parlamentos. A esta tarea se le llama ejercer la potestad legislativa [28, pto. 43a] y es su producción documental principal. En el caso del Parlamento de Canarias, esta tarea implica presentar el proyecto de ley a las Cortes Generales del Estado [28, pto. 43e]. Además, en España, los parlamentos tienen la obligación de convalidar o derogar los decretos-leyes [29, pto. 86.2] [28, pto. 46.3].
- **Controlar la acción del Gobierno:** Esta es otra de las principales tareas que tienen los parlamentos. Para esto, los diferentes parlamentos tienen diferentes poderes. En el caso de Canarias, el Parlamento controla la acción del ejecutivo con:
    - El nombramiento del equipo de gobierno [28, pto. 43c].
    - La designación de los senadores [28, pto. 43d].
    - Convalidando o derogando los decretos-leyes [28, pto. 46.3].
    - Cesando al presidente y al equipo de gobierno [28, pto. 52b] [28, pto. 54e] [28, pto. 55].
    - Creando comisiones de investigación [28, pto. 42.1].
- **Presupuestos:** La aprobación de los presupuestos generales es también una de las principales tareas que tienen los parlamentos [29, pto. 66.2] [28, pto. 43b]. En el caso del

Congreso de los Diputados y del Parlamento de Canarias, ambos tienen además autonomía financiera y la capacidad de definir su propio presupuesto [29, pto. 72] [28, pto. 41.2].

- **Debates políticos:** Una característica que emerge de la naturaleza plural del parlamento son los debates políticos, en los que miembros de la cámara exponen un asunto determinado y dan a sus miembros la posibilidad de realizar intervenciones con otros grupos parlamentarios. Estos debates son esenciales para un buen funcionamiento de una democracia, en la que las ideas se pueden poner a discusión, y los cuales informan a la ciudadanía.
- **Otras tareas:** Los parlamentos también realizan otras tareas menores, como elegir al titular de la Diputación del Común [28, pto. 57.3] o elaborar su propio reglamento [28, pto. 41.4].

## 4.2. Viabilidad del uso de LLMs para la asistencia en las labores parlamentarias

El caso de uso concreto que hemos decidido estudiar en este documento es la implementación de una IA generativa tipo ChatBot para la asistencia en la lectura, resumen, redacción, y argumentación de las propuestas de ley. Este tipo de IA puede presentar grandes beneficios en velocidad y eficiencia del organismo para realizar su función, pero también lleva asociado riesgos y desafíos. Para conocer estos detalles, primero debemos de comprobar la viabilidad de nuestro proyecto analizando estos sistemas y cómo nuestro caso de uso se adapta a la legislación europea.

### 4.2.1. Capacidades y limitaciones de los LLMs

Los LLMs son sistemas de IA de aprendizaje automático usados para comprender y generar lenguaje natural. Estos modelos son entrenados con conjuntos masivos de datos, y son de gran utilidad para responder a situaciones nuevas, las cuales no han sido contempladas en sus datos de entrenamiento [30]. Estos sistemas han ganado una enorme presencia social desde la introducción de ChatGPT por OpenAI, debido a que ha puesto a disposición del público general las enormes capacidades de uno de estos sistemas.

Sin embargo, antes de plantear su introducción al entorno parlamentario, debemos delimitar las capacidades y limitaciones de estos sistemas. Andrei Kucharavy expone estas limitaciones en *Limitaciones Fundamentales de los LLMs Generativos* [31], en el cual enumera que las principales limitaciones de estos modelos son, de forma resumida, las siguientes:

- **Los LLMs puros no son fácticos.** No pueden serlo, por diseño. La capacidad que tienen de poder generar texto sin estar limitados a búsquedas les da la flexibilidad para adaptarse a inputs para los que no han sido entrenados, pero esto mismo evita que sean fácticos. La forma en la que generan información es por asociación, y estas asociaciones pueden ser incorrectas.

- **Los LLMs aumentados con herramientas tampoco son fácticos** excepto que se demuestre lo contrario, pues se les introducen los siguientes problemas:
    - Cómo puede determinar e identificar la información que el usuario está pidiendo.
    - Cómo nos podemos asegurar que está haciendo una consulta adecuada para la información que busca.
    - Cómo podemos saber que la herramienta que usa no tiene sesgos o está incompleta.
    - Cómo podemos verificar que el LLM puede condensar los resultados obtenidos sin introducir hechos falsos que tengan la apariencia de ser verdaderos.
- **Los LLMs no ofrecen garantías de no exponer los datos privados con los que han sido entrenados,** por lo que debemos de asumir que los datos que se usen para el entrenamiento serán filtrados.
- **No podemos confiar en que los LLMs razonen correctamente,** pues sus capacidades generativas se basan en análisis probabilístico y no en lógica determinista. Esto se puede apreciar actualmente en las dificultades de los LLMs en computar operaciones matemáticas correctamente.
- **Los LLMs actuales no son adecuados para procesar largos cuerpos de texto,** debido al número limitado de tokens a los que pueden prestar atención.
- **Los LLMs no son adecuados para temas nicho, eventos recientes,** y sucesos o campos en los que generalmente **haya poca información disponible.**
- **Los LLMs pueden generar contenido perturbador sin previo aviso,** pues en muchos casos no saben distinguir cuándo algo es apropiado o deseable, solo que es probable. Las salidas de estos sistemas deben ser procesadas antes de exponerlas al usuario directamente.
- **Se reconoce que los LLMs están sesgados.** Muchos de estos sesgos derivan de que numerosas fuentes de datos preservadas por motivos históricos presentan claras instancias de discriminación y odio. Muchas veces es deseable que el modelo disponga de estos datos, pero no que los use en su razonamiento, lo cual es uno de los problemas pendientes en el mundo de la IA. Por esto, es recomendable siempre asumir que las salidas de estos sistemas van a estar sesgadas.

### 4.2.2. Conformidad con el *Reglamento de Inteligencia Artificial* de la UE

El primer paso antes de empezar a estudiar la implementación es examinar cuán adecuada es la conformidad de nuestra propuesta al nuevo *Reglamento de Inteligencia Artificial*.

Lo primero que debemos comprobar es la legalidad de nuestra propuesta. Estudiando las prácticas de IA prohibidas en el Reglamento, vemos que nuestra propuesta no contempla la realización de ninguna de estas prácticas, y por tanto podemos continuar con ella.

Lo siguiente que debemos de tener en cuenta es en qué calificación del Reglamento entraría nuestra propuesta. El sistema de IA que se propone no estaría calificado como un sistema de IA

de alto riesgo. Este sistema, por tanto, se ve sujeto a la obligación de establecer, implementar, documentar y mantener durante todo el ciclo de vida del sistema un sistema de gestión de riesgos [16, art. 9].

Además de esta calificación, un LLM generativo también entra en la definición de Inteligencia Artificial de Propósito General (IAPG) del Reglamento, por lo cual también se somete a otras obligaciones. Para saber cuáles, tenemos que determinar dos cualidades de la IA: si es de código abierto, y si se califica como de riesgo sistémico. Estas dos características todavía no están determinadas a estas alturas del planteamiento, pero por motivos de brevedad y evitar mostrar obligaciones que no tendremos que cumplir, adelantaré que más adelante nos decidiremos por un modelo que es de código abierto y de riesgo sistémico. Por tanto, las obligaciones que tendrá que cumplir nuestro sistema son las siguientes:

- Establecer una política para cumplir con la *Directiva sobre Derechos de Autor* [16, art. 53.1.c].
- Hacer y publicar un resumen detallado del contenido usado para el entrenamiento [16, art. 53.1.d].
- Realizar evaluaciones del modelo con protocolos y herramientas estándar, incluyendo pruebas adversarias [16, art. 55.1.a].
- Evaluar y mitigar posibles riesgos sistémicos [16, art. 55.1.b].
- Monitorear, documentar y reportar sin demora cualquier información relevante sobre incidentes graves y las medidas correctivas para abordarlo [16, art. 55.1.c].
- Asegurar la ciberseguridad de la infraestructura física del modelo [16, art. 55.1.d].

Hemos visto, por tanto, que el proyecto es viable, pero para su realización debemos cumplir todas las obligaciones expuestas y tener en mente todas las consideraciones que se han visto en este apartado.

## 4.3. Implementación usando la metodología de ética por diseño

Para estudiar la implementación del sistema de IA, vamos a usar el texto de *Enfoques de Ética por Diseño y Ética de Uso para la Inteligencia Artificial* [23], de la Comisión Europea, y *Ética por Diseño y Ética de Uso en IA y Robótica* [32], del Proyecto SIENNA, un proyecto financiado por la Unión Europea para la investigación de las dimensiones éticas y de los derechos humanos en tecnologías emergentes. El documento del Proyecto SIENNA está basado en el texto de la Comisión Europea, enfatizando la parte práctica del desarrollo e implementación, pero siendo en última instancia muy similar a este.

Como nuestro caso de uso trata sobre la implementación de un LLM para la asistencia en lectura, resumen, redacción y argumentación de leyes en el entorno parlamentario, vamos a usar las directrices expuestas en el cuarto capítulo del documento del Proyecto SIENNA, titulado “Implementación y Uso Ético”.

### 4.3.1. Planificación y administración del proyecto

En primer lugar, vamos a considerar las directrices éticas recomendadas para la fase de planificación y administración del proyecto. Esta es la fase en la que se planifica, presupuesta y gestiona la implementación del sistema. Cumpliendo con las directrices éticas, lo primero que se debe de tener en cuenta para realizar esta implementación es la elaboración de una licitación que tenga en cuenta las acciones necesarias para el cumplimiento de la ética por diseño, y reserve el presupuesto adecuado para ello. Esta licitación debe comprender las tareas y subtareas necesarias. Esta licitación también debe plantear una planificación temprana de los roles, responsabilidades y procedimientos para monitorear y evaluar que la implementación cumpla con los requisitos éticos. Esto se debe a que en esta fase se plantea si estos roles son internos o externos a la entidad que implementa el sistema en sí. Los detalles de la elaboración de esta licitación quedan fuera del alcance de este documento.

En segundo lugar, en esta fase debemos comprobar si nuestro proyecto se adapta a los principios y requisitos descritos en el documento de *Ética por Diseño*. La evaluación que hemos realizado, basándonos en las capacidades y limitaciones de los LLMs, es la siguiente:

- Para conservar y proteger la voluntad humana, además de los tres principios de autonomía, dignidad y libertad, el sistema debe:
    - Notificar a sus usuarios que la información que presenta es limitada y puede estar sesgada.
    - Estar diseñado de tal forma que no pueda tomar decisiones de forma autónoma.
    - Estar diseñado para evitar engañar, manipular, o deshumanizar al usuario, a grupos, o a terceras personas.
- Para conservar y proteger la privacidad y hacer una buena gobernanza de los datos, el sistema debe:
    - Cumplir con el *Reglamento General de Protección de Datos*.
    - Tener un informe claro y detallado de los tipos de datos usados en su entrenamiento.
    - No usar los datos o entradas de los usuarios para su entrenamiento.
    - Anonimizar o seudonimizar la actividad de los usuarios previo a su registro y almacenamiento.
- Para conservar y proteger la imparcialidad, se debe:
    - Documentar cómo está diseñado para evitar los sesgos algorítmicos, y cómo se han intentado evitar sesgos en sus datos de entrenamiento.
    - El sistema debe estar diseñado siguiendo los estándares de accesibilidad.
    - El sistema debe estar diseñado para mitigar, en sus salidas, posibles impactos sociales negativos en ciertos grupos.

- Para conservar y proteger el bienestar individual, social, y ambiental, el sistema debe:
    - Estar diseñado para evitar salidas que afecten negativamente a individuos o grupos sociales.
    - Cumplir con los Objetivos de Desarrollo Sostenible.
    - Estar diseñado para evitar amplificar discursos no cívicos, noticias falsas y cámaras de eco.
- Para conservar y proteger la transparencia, el sistema debe:
    - Dejar claro a los usuarios que están interactuando con una IA.
    - Recordar a los usuarios las capacidades, limitaciones, beneficios y riesgos del sistema, donde a nuestro juicio deberíamos incluir la continua evaluación de este conocimiento.
    - Estar diseñado para poder rastrear y analizar el proceso de decisión que ha seguido en todo momento de su ciclo de vida.
    - Hacer un seguimiento y registro documental de las decisiones tomadas para cumplir con los requisitos éticos.
- Para conservar y proteger la rendición de cuentas del sistema, y su supervisión:
    - Debe existir documentación de cómo los eventos indeseables desde el punto de vista ético y social deben ser detectados, parados, y qué medidas se deben tomar para que no vuelva a ocurrir.
    - Se debe hacer una evaluación de los riesgos, incluyendo procedimientos y medidas para su mitigación.
    - Se debe crear un sistema en el que usuarios y terceras partes puedan poner reclamaciones, expresar preocupaciones éticas, o notificar de eventos adversos. Debe existir documentación de cómo esto debe ser evaluado, tratado y comunicado de vuelta.
    - El sistema debe estar diseñado para poder ser evaluado por terceras partes.

En cuanto a las operaciones y tareas que debe realizar la IA, en este proyecto la hemos planteado como un sistema de ayuda a los miembros del Parlamento de Canarias en la lectura, resumen, redacción, y argumentación de leyes. Es por esto que en esta etapa podemos idear cuatro tareas o modos de operación que el sistema debe incorporar para cumplir su propósito.

- **Modo lectura/preguntas:** El usuario proporcionará un texto y preguntas sobre él. El sistema responderá a las preguntas del usuario.
- **Modo resumen:** El usuario proporcionará un texto que el sistema resumirá.
- **Modo redacción:** El usuario proporcionará al sistema una idea o serie de ideas, y el sistema deberá redactarlas como si se tratara de un texto legal.
- **Modo argumentación:** El usuario proporcionará ideas o escenarios, a los cuales el sistema intentará dar argumentos a favor y en contra.

Por último, en esta fase debemos de crear procedimientos que incluyan a los usuarios en el proceso de adquisición, despliegue, implementación y monitoreo del sistema. Como mínimo debemos asegurar que son consultados acerca de sus valores e intereses con respecto al sistema. En esta fase, además, se ha creado el sistema de gestión de riesgos y se ha hecho la primera evaluación de riesgos.

### 4.3.2. Adquisición del sistema de IA

En segundo lugar, vamos a evaluar las directrices éticas que se aplican a la fase de adquisición de un sistema de IA. Esta fase incluye el proceso de selección del modelo a utilizar. Para esta propuesta hemos decidido que el sistema a implementar cumpla los siguientes criterios:

1. **Solución ya existente en el mercado**: se seleccionará un modelo de lenguaje que ya exista en el mercado, completamente desarrollado y entrenado. Este criterio se establece para reducir la complejidad y la inversión de tiempo que conlleva el desarrollo de un nuevo modelo.

2. **Ejecución en local**, sin necesidad de enviar las peticiones a servicios de terceros, pues esto hace más sencilla la seguridad del sistema y la aplicación de políticas de protección de datos.

3. **Entrenado e incorporando salvaguardas** para evitar salidas sesgadas, discriminatorias, o que puedan ocasionar daños psicológicos a sus usuarios.

4. **Código libre**, pues esto permite que el modelo sea evaluado por terceras partes, mejorando la transparencia y, en la medida de lo posible, la imparcialidad del algoritmo.

5. **Entrenado con datos en español**, pues no podemos suponer que todos los parlamentarios tendrán un perfecto nivel de inglés, y por tanto podría suponer una barrera lingüística.

Para seleccionar un sistema que cumpla nuestros requisitos y mejor se adapte a las exigencias éticas que nos impone el *Reglamento de Inteligencia Artificial*, y los principios del documento de *Ética por Diseño*, se han analizado las fuentes [33-42] y se ha elaborado la tabla 1.

| Modelo | Cumplen los requisitos mínimos expuestos anteriormente | Evitan sesgos algorítmicos y sesgos en los datos de entrenamiento | Evitan salidas con impactos negativos (ofensivas, desinformación, etc) | Describen en suficiente detalle los datos usados en su entrenamiento | Hacen una toma de decisiones transparente y rastreable. Evitan producir contenido protegido por copyright |
|---|---|---|---|---|---|
| Bard / Gemini | No [3] | - | - | - | - |
| ChatGPT | No [3] | - | - | - | - |
| Claude | No [3] | - | - | - | - |
| Gemma | Sí [37] | Sí [37] | Sí [37] | Sí [37] | No [4] |
| Llama | Sí [38] [39] | Sí [38] [39] | Sí [38] [39] | No [5] | No [4] |
| Mixtral | Sí [40] | Sí [40] | Sí [40] | No [6] | No [4] |
| Qwen | Sí [41] [42] | Sí [42] | Sí [42] | No [7] | No [4] |

*Tabla 1: Comparación de cumplimiento de requisitos legales y éticos por varios sistemas de IA*

Como vemos en esta tabla comparativa, ninguno de los sistemas analizados cumple con todos los requisitos exigidos por los principios del documento de *Ética por Diseño* y el *Reglamento de Inteligencia Artificial* de la UE y, por tanto, no podemos recomendar sus usos en los entornos parlamentarios. Sin embargo, y por el propósito de este trabajo, hemos decidido usar Gemma AI para ilustrar los pasos a seguir para una correcta implementación. Esta IA de Google ha sido escogida debido a que, en comparación con las otras, su documentación presenta un mayor nivel de transparencia con los datos usados y las medidas que han usado para mitigar los contenidos inapropiados y sesgos en estos datos.

A pesar de desconocer este dato, es razonable suponer que Gemma AI se puede categorizar como una IAPG de riesgo sistémico, pues es posible que la cantidad acumulada de operaciones de coma flotante (FLOPs, *FLoating-point OPerations*) en su entrenamiento supere los 10 yottaFLOPs (10^25 FLOPs). Esta estimación la hacemos a partir de la observación de que los

[3] Para Bard/Gemini en [34], ChatGPT en [35] y Claude en [36] las ausencias de guía de instalación y referencias al código fuente hacen que no puedan ser considerados IAs de código libre.

[4] No se menciona en ninguna de las fuentes [37-42] formas de estudiar la toma de decisiones del sistema, ni de medidas tomadas para evitar la producción de contenido con copyright.

[5] Solo se mencionan la preparación de los datos [38] [39] y el origen (*"publicly available sources"* [39]). No hay comentarios acerca de la naturaleza de los datos.

[6] Solo se menciona el origen de los datos (*"data extracted from the open Web"* [40]). No hay comentarios acerca de la naturaleza de los datos.

[7] Describe la fuente de los datos (*"publicly available sources"* [42]), la preparación de los datos [42] y de que contiene datos matemáticos [41] [42] y en múltiples idiomas [42]. No describe en suficiente detalle la naturaleza de esos datos.

modelos listados anteriormente son similares entre sí en cuanto a sus capacidades, junto con un análisis de la documentación de Llama3. En esta documentación se declara que el entrenamiento de Llama3-70B ha requerido 6.4M de horas de GPU, y que la GPU usada ha sido la NVIDIA H100 80GB [39]. Si asumimos que se ha usado la operación más rápida (coma flotante de 8 bits), al multiplicar las 3958 operaciones de coma flotante por segundo que este modelo de GPU puede realizar [43] por el número de horas empleada, obtenemos que este modelo acumula 18,52 yottaFLOPs, casi duplicando el umbral declarado en el Reglamento de la UE. Es por esta estimación que, de forma preventiva, asumiremos que se trata de una IAPG de riesgo sistémico.

Gemma AI 1.1 (la última revisión de código libre de Gemma AI) tiene 4 variantes. Estas variantes consisten en el modelo base y el modelo ajustado a instrucciones (*instruction-tuned*), ambos con sus respectivas variantes de tamaño 2B y 7B. El tamaño hace referencia al número de parámetros que contiene esa variable del modelo, en miles de millones (*billions* en inglés). Como norma general, una variante de un modelo que tenga un mayor número de parámetros que otra, suele ofrecer mejores resultados [33] [37-42], pero también suele acaparar más recursos del sistema.

El modelo ajustado a instrucciones es una variante del modelo base que ha sido ajustada (*fine-tuned*) para entablar conversaciones y/o para seguir instrucciones [44]. Este tipo de ajuste es una práctica común realizada con el fin de mejorar el rendimiento del modelo en múltiples usos prácticos. Para este proyecto, se escogerá, por tanto, la variante 7B ajustada a instrucciones. Esta variante es la más capaz y que mejor se adapta a nuestro caso de uso.

En cuanto a la adquisición de hardware, este proyecto debe contar con un servidor capaz de ejecutar en un tiempo razonable las peticiones de los usuarios. Para comprender las características que debe tener el servidor, hemos realizado pruebas en un equipo con las siguientes características:

- **CPU:** Intel Core i7 4770
- **GPU:** AMD Radeon RX 580 4GB
- **RAM:** 16 GiB DDR3-1600
- **SSD:** Kingston SA-400 480G
- **OS:** OpenSUSE Tumbleweed 84.87.20240609.6b4c891

Después de realizar pruebas y análisis empíricos del uso de recursos por Gemma AI 1.1 Instruct 7B, podemos hacer las siguientes recomendaciones de características de hardware para el servidor:

| Componente | Uso observado | Recomendación | Razón |
|---|---|---|---|
| CPU | N/A (100%) | N/A | El uso observado no refleja el uso que tendría un sistema con aceleración por GPU, por tanto, es difícil hacer una recomendación. |
| GPU | N/A (<1%) | NVIDIA H100 80GB o equivalente | El uso observado no refleja el uso que tendría un sistema con aceleración por GPU, por tanto, es difícil hacer una recomendación.<br><br>Recomendamos NVIDIA porque el soporte para CUDA en sus GPUs está extendido, mientras que ROCm en GPUs de AMD está limitado a pocos modelos [45].<br><br>Existe precedente de que el modelo recomendado es adecuado para su uso en el ámbito de los LLMs, pues ha sido el utilizado por Meta para entrenar Llama3 [39]. |
| RAM | 26 GiB | ≥ 32 GiB | La recomendación es superior al uso observado puesto que este no tiene en consideración otros procesos del sistema. |
| SSD | 55 GiB | ≥ 100 GiB | La recomendación es superior al uso observado puesto que este no tiene en consideración otros datos del sistema. |

*Tabla 2: Uso observado y recomendación de componentes de hardware para el sistema de IA*

Sin embargo, debido a las limitaciones presupuestarias de este trabajo, la implementación piloto y las pruebas que se realizarán sobre ella se harán usando el modelo Gemma 1.1 Instruct 2B. Esta elección surge de la imposibilidad de ejecutar las variantes 7B en el equipo de prueba, y por tanto de la necesidad de adaptar el modelo a los recursos de hardware disponibles. El modelo 2B requiere considerablemente menos recursos, pero cabe esperar que los resultados obtenidos sean significativamente peores.

Esta implementación se realizará en una máquina virtual que se ejecutará sobre el mismo equipo en el que se realizaron las pruebas. El OS host usado será Windows 10 Pro 2H22, el hipervisor de tipo 1 será Hyper-V (nativo), y el hipervisor de tipo 2 será VMware Workstation 17.5. Las características de la máquina virtual serán las siguientes:

- **CPU:** 8vCPU
- **RAM:** 12 GiB
- **Almacenamiento:** 50 GiB
- **OS:** Debian 12.5 Bookworm

Para finalizar la sección de adquisición del modelo, se ha realizado otra evaluación de riesgo tomando en cuenta las decisiones tomadas y cómo afectan a los riesgos planteados en la anterior.

## 4.3.3. Evaluación del modelo

Una vez hemos obtenido un buen candidato como modelo, tenemos que evaluar que tiene las capacidades para realizar las funciones con las que se ideó este proyecto. Recordemos que estas funciones eran ayudar a los miembros del Parlamento de Canarias a leer, resumir, escribir y argumentar leyes, decretos, y otros documentos que emite el Parlamento en su labor legislativa. Para cada una de estas funciones se ha ideado una prueba que simula diferentes entradas de los futuros usuarios y evalúa las salidas del sistema.

### 4.3.3.1. Resumen

En esta prueba se entregaba al sistema de IA un texto y se le pedía que lo resumiera. Esta evaluación tiene como objetivo probar la capacidad del sistema para resumir texto.

**Entrada de texto (prompt)**

La entrada de texto usada para este test fue:

| Español | Inglés |
|---|---|
| Instrucciones:<br>Resume este texto<br><br>Texto:<br><*Texto*><br><br>Resumen: | Instructions:<br>Summarize this text<br><br>Text:<br><*Texto*><br><br>Resumen: |

*Tabla 3: Instrucciones para la evaluación de capacidad de resumen*

**Material empleado**

Para esta prueba, se hicieron y evaluaron un total de 22 resúmenes correspondientes a los siguientes materiales:

- Los fragmentos [46-53] del *Boletín Oficial de Canarias* (*BOC*), incluyendo su traducción automática (mediante Google Translate) al inglés.
- Fragmentos de 2, 6 y 10 de las recomendaciones encontradas en [54], del *Diario Oficial de la Unión Europea*. Se ha usado la traducción oficial al inglés.

**Metodología**

1. Extracción y valoración de ideas: se hizo una lectura manual de los textos y se extrajeron las ideas principales. A las ideas más importantes se les asignó valor de 2, y a las secundarias de 1. A la suma de los valores de todas las ideas se le consideró “Valor total”.

2. Comprobación de las ideas en el resumen: se hizo una lectura del resumen producido por el sistema de IA, y se asignó un valor de “presencialidad” a cada idea principal del texto. El valor de presencialidad era de 1 si la idea estaba completamente presente en el resumen, 0.5 si estaba parcialmente, y 0 si no estaba presente.

3. Cálculo de la puntuación: la puntuación de cada idea se ha obtenido multiplicando el valor de esta por su presencialidad. La puntuación de un resumen se obtiene dividiendo la suma de las puntuaciones de todas sus ideas por el “Valor total”.

**Resultados esperados**

Se espera obtener menor puntuación a mayor tamaño de texto, pues el número de tokens del sistema es constante. Se espera además obtener mejores resultados en inglés que en español, pues es el lenguaje en el que se suelen entrenar primariamente estos sistemas.

**Resultados obtenidos**

Los resultados de la evaluación se muestran a continuación:

| | **Puntuación del resumen (media)** | **Número de palabras del texto (media)** | **Puntuación y número de palabras (coef. de correlación de Pearson)** |
|---|---|---|---|
| **Español** | 59.70% | 849.45 | -0.2252 |
| **Inglés** | 65.76% | 796.00 | -0.1624 |

*Tabla 4: Resultados de la evaluación de capacidad de resumen*

Como se puede ver, la puntuación media obtenida en inglés es significativamente superior a la obtenida con los mismos textos en español. Se puede observar también que los datos no nos dan una correlación negativa entre el número de palabras del texto a resumir y la calidad de los resultados lo suficientemente fuerte como para confirmar nuestra predicción.

### 4.3.3.2. Escritura

En esta prueba se entregaba al sistema de IA una petición de escritura listando el contenido deseado. Esta evaluación tiene como objetivo probar la capacidad del sistema para general texto formal.

**Entrada de texto (prompt)**

La entrada de texto usada para este test fue:

| **Español** | **Inglés** |
|---|---|
| Instrucciones:<br>Ayuda al usuario con escritura formal y legal<br><br>Petición del usuario:<br>*<Petición>*<br><br>Respuesta: | Instructions:<br>Help the user with formal and legal writing<br><br>User request:<br>*<Petición>*<br><br>Answer: |

*Tabla 5: Instrucciones para la evaluación de capacidad de escritura*

**Peticiones realizadas**

Se realizaron 10 peticiones diferentes de redacción, todas ellas en español e inglés (sumando un total de 20). Todas las peticiones han estado relacionadas con la actividad parlamentaria, con ejemplos como autorizaciones de obras, anuncio de incrementos de ayudas, o la elaboración de propuestas legislativas.

Las peticiones realizadas han contemplado un amplio rango de longitudes y detalles, desde descripciones muy cortas (28 palabras) hasta descripciones relativamente largas y detalladas (344 palabras). En estas peticiones se han proporcionado detalles como cifras, nombres, fechas, y localizaciones que se han usado para evaluar la capacidad del modelo de usar esos datos sin alterarlos.

**Metodología**

1. Valoración de ideas: se valoraron las ideas de la petición, asignando los valores de 2 a las principales, y 1 a las secundarias. A la suma de los valores de todas las ideas se le consideró “Valor total”.
2. Comprobación de las ideas en el texto generado: se hizo una lectura del texto producido por el sistema de IA, y se asignó un valor de “presencialidad” a cada idea principal de la petición. El valor de presencialidad era de 1 si la idea estaba completamente presente en el resumen, 0.5 si estaba parcialmente, y 0 si no estaba presente.
3. Cálculo de la puntuación: la puntuación de cada idea se ha obtenido multiplicando el valor de esta por su presencialidad. La puntuación de un texto generado se obtiene dividiendo la suma de las puntuaciones de todas sus ideas por el “Valor total”.

**Resultados esperados**

Se espera obtener menor puntuación a mayor fuese la descripción del texto a redactar, pues el número de tokens del sistema es constante. Se espera además obtener mejores resultados en inglés que en español, pues es el lenguaje en el que se suelen entrenar primariamente estos sistemas.

**Resultados obtenidos**

Los resultados de esta evaluación se pueden ver a continuación:

| | **Puntuación (Media)** | **Puntuación (Desviación Típica)** | **Puntuación y Número de palabras (Coef. de Correlación de Pearson)** |
|---|---|---|---|
| **Español** | 94.80% | 6.78% | 0.305 |
| **Inglés** | 91.18% | 14.12% | 0.013 |

*Tabla 6: Resultados de la evaluación de capacidad de escritura*

Como se puede ver, la puntuación del sistema en ambos idiomas es bastante alta. Se puede ver que los resultados en español son ligeramente mejores que los resultados en inglés, pero como se ve en la desviación típica, los resultados del inglés están distribuidos de forma menos uniforme. La menor variación de los resultados en español puede sugerir un procesamiento menos elaborado y más directo en ese idioma.

También se puede apreciar cómo la relación inversa esperada entre la puntuación y el número de palabras no se ha dado. La relación entre estas variables es inexistente o muy débil, sugiriendo que el sistema tiene un suficiente número de tokens para procesar correctamente la información de las peticiones. Además, se aprecia que, aunque la relación entre la puntuación y el número de palabras es muy débil, esta relación en el español es significativamente superior al inglés, lo que puede apoyar la hipótesis de que este sistema de IA hace un procesamiento menos elaborado y más directo de los textos en español.

#### 4.3.3.3. Lectura y respuesta a preguntas

En esta prueba se entregaba al sistema de IA un texto y una pregunta acerca de ese texto. Esta evaluación tiene como objetivo probar la capacidad del sistema para ayudar a los miembros del Parlamento a entender textos legales.

**Entrada de texto (prompt)**

Las entradas de texto usadas para este test fueron:

| | **Español** | **Inglés** |
|---|---|---|
| **Texto primero (pruebas 1-10)** | Instrucciones:<br>Responde a la pregunta del usuario de acuerdo con el texto proporcionado.<br><br>Texto proporcionado:<br><*Texto*><br><br>Pregunta del usuario:<br><*Pregunta*><br><br>Respuesta: | Instructions:<br>Answer the user question based on the text provided.<br><br>Text provided:<br><*Texto*><br><br>User question:<br><*Pregunta*><br><br>Answer: |
| **Pregunta primero (pruebas 11-20)** | Instrucciones:<br>Responde a la pregunta del usuario de acuerdo con el texto proporcionado.<br><br>Pregunta del usuario:<br><*Pregunta*><br><br>Texto proporcionado:<br><*Texto*><br><br>Respuesta: | Instructions:<br>Answer the user question based on the text provided.<br><br>User question:<br><*Pregunta*><br><br>Text provided:<br><*Texto*><br><br>Answer: |

*Tabla 7: Instrucciones para la evaluación de capacidad de lectura*

**Material utilizado y peticiones realizadas**

El material utilizado fue:

- Preguntas 1-4: Artículo 6 del *Reglamento de Inteligencia Artificial* [16].
- Preguntas 5-7: Capítulo 3, Sección 4 (Artículos 21 y 22) del *GDPR* [18].
- Preguntas 8-10: Capítulo 4, Sección 1 (Artículos 24-31) del *GDPR* [18].
- Preguntas 11-13: Orden BOC-A-2024-133-2184 del *BOC* [49].
- Preguntas 14 y 15: Extracto BOC-A-2024-133-2192 del *BOC* [50].
- Preguntas 16-18: Anuncio BOC-A-2024-133-2197 del *BOC* [51].
- Preguntas 19 y 20: Resolución BOC-A-2024-134-2203 del *BOC* [53].

Se emplearon diferentes tipos de preguntas, desde las que tenían respuesta explícita en el texto, hasta explicar situaciones concretas y ver si la ley u orden descrita en el texto se aplicaba. Muchas de las preguntas se acompañaron con "¿por qué?" para estudiar el razonamiento y la capacidad de justificación del modelo.

**Metodología**

Cada respuesta proporcionada se valoró con 1 si era correcta y mostraba un razonamiento o justificación adecuada con el texto, 0.5 si era correcta pero la justificación no era adecuada o era incorrecta, y 0 si la respuesta proporcionada era incorrecta.

**Resultados esperados**

Se espera obtener buenas puntuaciones en general, con una mayor puntuación en inglés pues suele ser el lenguaje con el que principalmente se entrenan estos modelos.

**Resultados obtenidos**

Los resultados de esta evaluación se muestran a continuación:

| | **Responde correctamente (sobre el total)** | **Razona correctamente (sobre nº de respuestas correctas)** | **Puntuación Texto Primero (media)** | **Puntuación Pregunta Primero (media)** | **Puntuación total (media)** |
|---|---|---|---|---|---|
| **ES** | 55% | 82% | 0.45 | 0.55 | **0.50** |
| **EN** | 65% | 85% | 0.55 | 0.65 | **0.60** |

*Tabla 8: Resultados de la evaluación de capacidad de lectura*

En los datos se puede apreciar una puntuación mucho más baja de lo esperada, teniendo una puntuación media de 0.5 y 0.6 en español e inglés respectivamente. También se puede apreciar que, aunque es probable, las justificaciones que el sistema da para sus decisiones no son siempre correctas. Por último, señalar que con las cifras obtenidas y el número de pruebas realizadas no se puede determinar si cambiar el orden del texto y la pregunta en la petición da mejores resultados.

#### 4.3.3.4. Argumentación

En esta prueba se entregaba al sistema de IA un tema y se le pedía que generase argumentos a favor y en contra de ella. Esta evaluación tiene como objetivo probar la capacidad del sistema para generar argumentos a favor y en contra de ideas.

**Entrada de texto (prompt)**

La entrada de texto usada para este test fue:

| Español | Inglés |
|---|---|
| Instrucciones:<br>Da argumentos a favor y en contra del tema proporcionado<br><br>Tema proporcionado:<br><*Tema*><br><br>Argumentos: | Instructions:<br>Give arguments for and against the given topic<br><br>Given topic:<br><*Tema*><br><br>Arguments: |

*Tabla 9: Instrucciones para la evaluación de capacidad de argumentación*

**Temas proporcionados**

Los temas que le fueron proporcionados al sistema para esta prueba fueron temas políticos, escritos brevemente como una frase. A continuación se muestran los 5 primeros temas como ejemplos ilustrativos:

1. Recortes en educación
2. Privatización de la sanidad pública
3. Subir el salario mínimo interprofesional
4. Incrementar el gasto en defensa
5. Legalización del cannabis

Fueron proporcionados 41 temas, tanto en español como en inglés, por un total de 82 pruebas.

**Metodología**

Este tipo de salidas es difícil de evaluar, por lo que hemos decidido hacer una puntuación “objetiva” que consiste en, por cada par de preguntas en español e inglés, asignar el valor 1 a la respuesta con mayor número de argumentos, y la fracción proporcional a la menor. Esto en efecto es una simple comparación proporcional del número de argumentos.

A parte de esto, hemos realizado una puntuación subjetiva calculada a partir de la puntuación objetiva × 0.9 + una valoración elaborada a partir de la calidad de los argumentos y el detalle en el que estén explicados × 0.1. A parte de esto, si se considera que la muestra tiene sesgos, discrimina a grupos de personas, o sus salidas pueden tener impacto negativo, se penalizan con 0.5 puntos normalizados.

**Resultados esperados**

Se espera obtener generalmente puntuaciones altas. Se espera además obtener mayores puntuaciones en inglés porque suele ser el lenguaje con el que principalmente se entrenan estos modelos.

**Resultados obtenidos**

A continuación se muestra un resumen de las principales estadísticas y resultados obtenidos:

| | Número de argumentos (media) | Puntuación objetiva (media) | Puntuación subjetiva (media) |
|---|---|---|---|
| **Español** | 6.76 | 82.96% | 78.58% |
| **Inglés** | 8.22 | 99.51% | 97.82% |

*Tabla 10: Resumen de resultados de la evaluación de capacidad de argumentación*

Como se puede observar, existe una diferencia significativa en el número de argumentos generados en español e inglés, lo cual impacta de manera directa la puntuación objetiva. Además, los argumentos generados en inglés no sólo de media eran más numerosos, sino también más detallados y más relevantes al tema solicitado. Es por esto que la puntuación subjetiva refuerza los resultados positivos del inglés.

#### 4.3.3.5. Conclusiones de la evaluación

Se puede observar que el rendimiento del sistema seleccionado es bueno en los ámbitos de escritura y argumentación, pero su rendimiento es notablemente inferior en los ámbitos de resumir y de responder a preguntas acerca de un texto. Debido a esto, y siguiendo con la evaluación y mitigación de riesgos, se integrará un aviso a los usuarios en la interfaz de la aplicación cuando se realicen estas tareas.

Los datos usados en esta evaluación, y los análisis realizados sobre ellos, están disponibles en https://github.com/alu0101048369/tfgdata

### 4.3.4. Despliegue e implementación

En tercer lugar, vamos a evaluar las directrices éticas que se aplican a la fase de despliegue e implementación del sistema de IA. Esta fase abarca el proceso de integrar el sistema de tal forma que los usuarios puedan tener acceso a él, por tanto, acarrea nuevos riesgos y problemas que pueden no haber sido previstos en las fases previas. Es esencial que, en esta parte, el plan de implantación y la formación de los usuarios tengan contemplados estos riesgos éticos, y haya un esfuerzo por mitigarlos.

Las directrices éticas de esta parte principalmente consisten en aplicar los planes y políticas que hemos previsto en los anteriores apartados, además de monitorizar los nuevos retos que pueden surgir durante la implementación y ajustar los planes donde haga falta para mitigarlos.

Para llevar a cabo una implementación piloto que ilustre la metodología de este documento, se ha realizado y seguido un plan de implantación. Este plan está dividido en 5 ámbitos: Instalación, Programación, Formación, Pruebas con usuarios, y Documentación.

En el ámbito de instalación se ha hecho la configuración del sistema operativo de la máquina virtual mencionada anteriormente. En ella se han configurado los programas, NGINX como servidor de la web y reverse proxy, se ha securizado con firewalld, y se ha securizado el servidor de SSH cambiando el puerto, eliminando el login por contraseña (sólo permite claves SSH), y se ha desactivado el login en la cuenta de root.

En el ámbito de la programación se han desarrollado varios servicios e interfaces que funcionan de acuerdo con el diagrama de la figura 1.

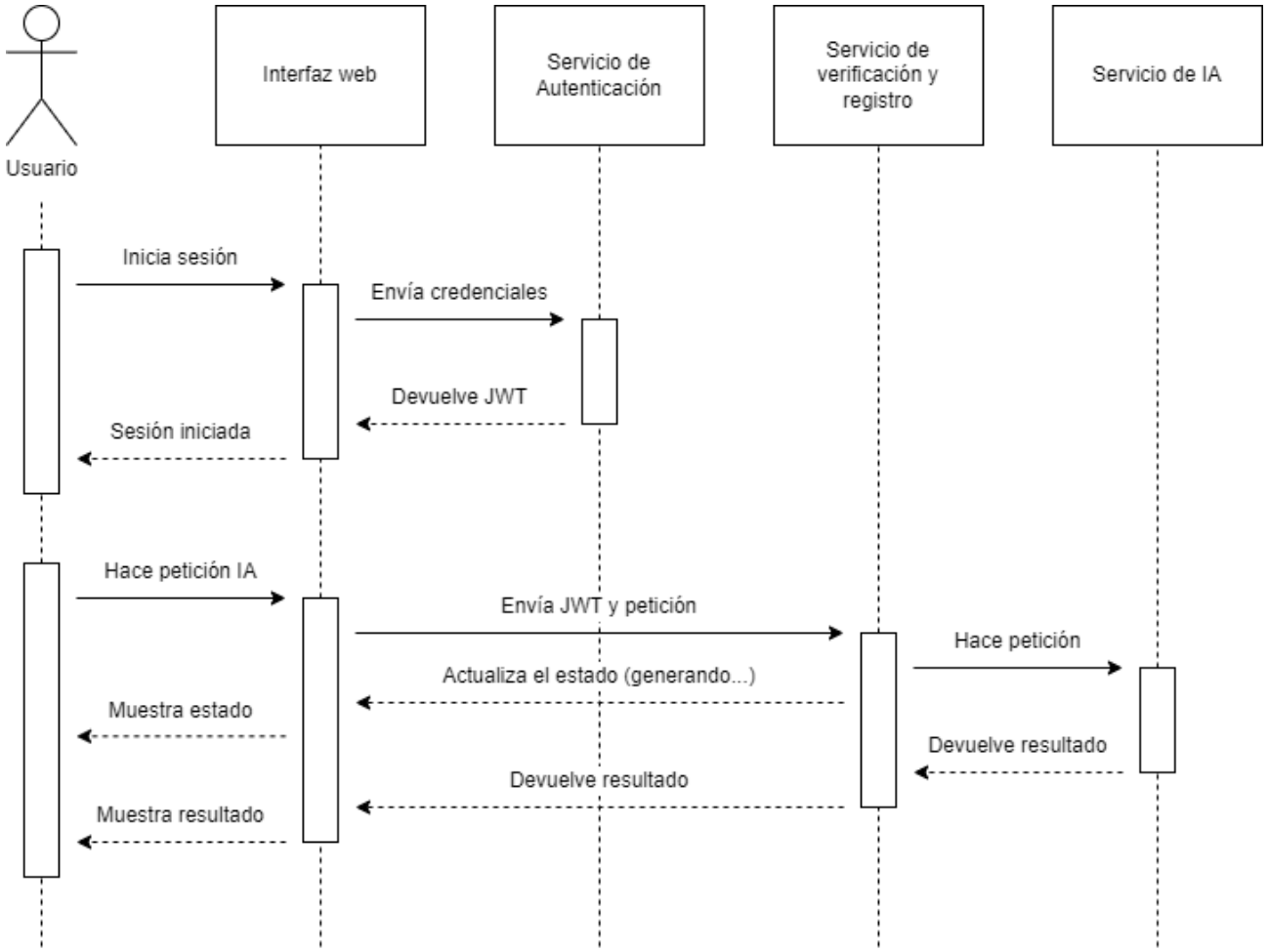



*Figura 1: Diagrama de flujo de la aplicación*

La interfaz web está hecha con HTML, CSS y JS. El servicio de autenticación se hizo con JavaScript en Node, el servicio de verificación y registro fue desarrollado en Go (también conocido como GoLang), y usa el protocolo de WebSockets para comunicarse de forma bidireccional con la interfaz web. Por último, el servicio de IA está programado en Python y usa HTTP para comunicarse con el servicio de verificación y registro.

Esta arquitectura fue diseñada de esta forma para poder hacer separación del servicio de autenticación y de registro. Esto nos asegura que el servicio de registro no requiere acceso a ninguna base de datos de usuarios para verificar que su sesión y datos son correctos y, por tanto, esto reduce el riesgo de que si se compromete su seguridad se puedan obtener datos de usuario.

Se ha separado el servicio de IA del de verificación y registro porque la API de Gemma AI está diseñada para su uso en Python, y el desarrollador en este proyecto está más familiarizado con el lenguaje de programación Go.

En el ámbito de la formación se ha realizado una guía de uso. Las pruebas con usuario quedan fuera del alcance y capacidades presupuestarias de este documento. En el ámbito de la documentación se ha realizado el presente documento, que en el cuerpo principal muestra el planteamiento y funcionamiento en detalle.

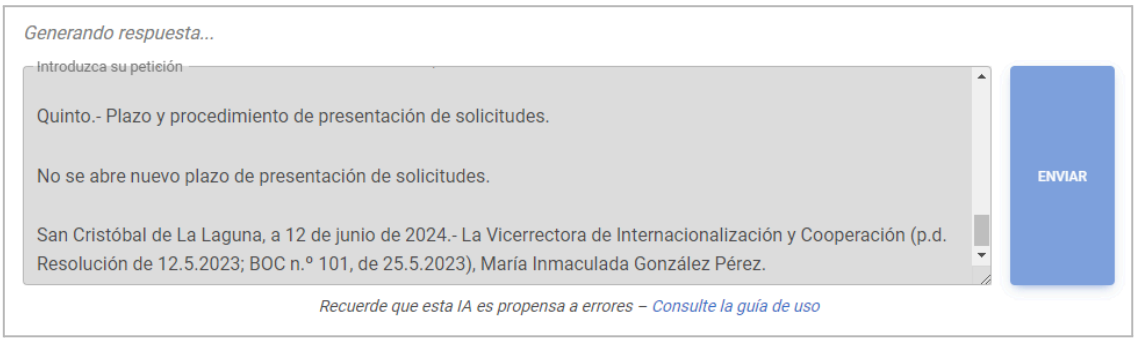



*Figura 2: Interfaz web durante la generación de respuesta*

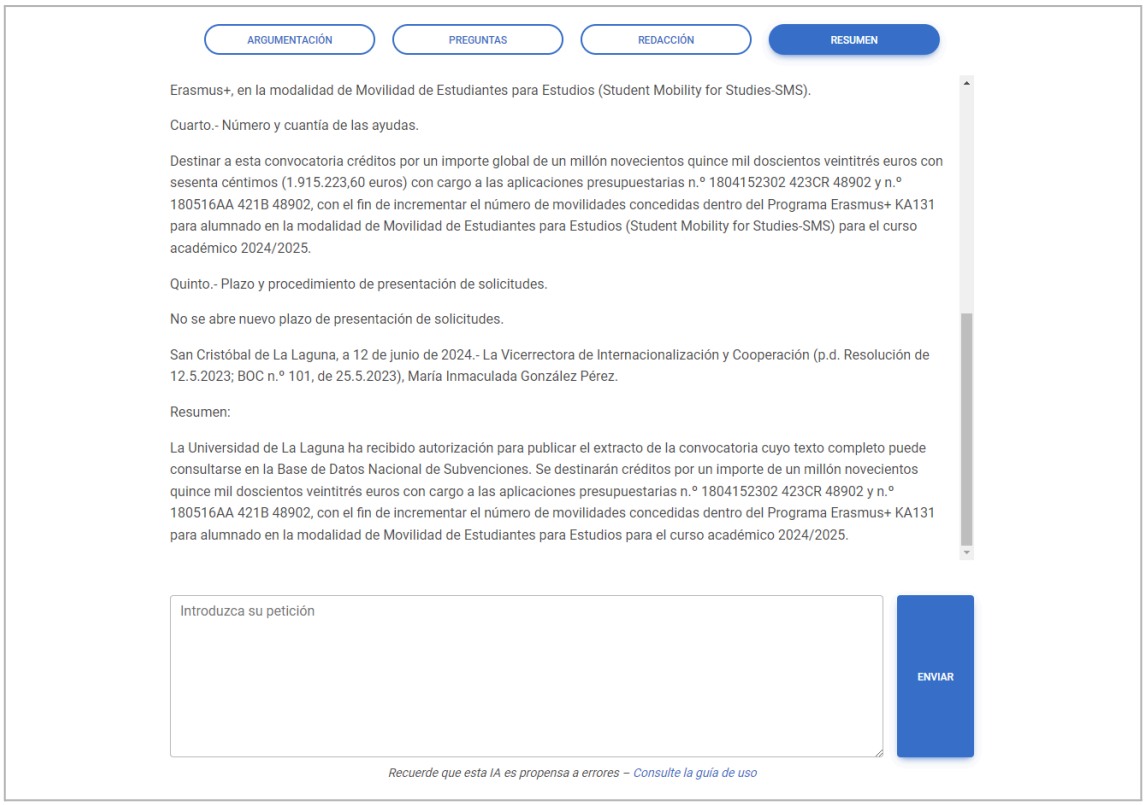



*Figura 3: Interfaz web con una respuesta generada*


```
{
  "issuer": "system",
  "message":"Startup",
  "time":"2024-07-02T09:22:11.492Z"
}
{
  "input":"Instrucciones:\nResume el texto\n\nTexto:\nDe conformidad (...) González Pérez.\n\nResumen:\n",
  "issuer":"user",
  "requestId":"10618304837038354416",
  "time":"2024-07-02T09:23:09.859Z",
  "userId":"iB2JMcXMUrUmuid9t8Z1argjdyFAA8Va"
}
{
  "issuer":"user",
  "requestId":"10618304837038354416",
  "response":"La Universidad (...) académico 2024/2025.",
  "time":"2024-07-02T09:27:14.387Z",
  "userId":"iB2JMcXMUrUmuid9t8Z1argjdyFAA8Va"
}
{
  "issuer":"system",
  "message":"Shutting down",
  "time":"2024-07-02T09:29:37.009Z"
}
```


*Figura 4: Logs de la aplicación abreviados y estilizados*

### 4.3.5. Monitorización

En último lugar, vamos a evaluar las directrices éticas que se aplican a la fase de monitorización. En esta fase debe contemplar el estudio de la conformidad del sistema de IA con los requisitos, tanto técnicos como éticos.

Las directrices éticas de esta parte insisten en la monitorización periódica y perpetua del cumplimiento de los requisitos éticos del sistema. Los usuarios y afectados por el sistema tienen además que disponer de sistemas de comunicación para quejas y reclamaciones, las cuales deben ser regularmente monitorizadas y procesadas. Los problemas éticos no deben desaparecer en el sistema, sino debe de haber un sistema formal de gestión y reporte.

Para nuestro caso práctico, una correcta monitorización del sistema debe incluir:

- El nombramiento de un **responsable de monitorización**. Este cargo debe existir durante todo el ciclo de vida del sistema.
- La elaboración de un **plan de mantenimiento** que incluya:
    - Revisiones periódicas, poniendo como ejemplo cada 6 meses para las áreas detectadas como “Riesgo Moderado” en el sistema de evaluación de riesgos, y cada 12 meses para las áreas de menor riesgo. Esta revisión periódica debe incluir una evaluación de riesgos que verifique que los riesgos siguen bajo control. Esta revisión tiene, además, que verificar que el sistema sigue cumpliendo con el *Reglamento de Inteligencia Artificial* de la Unión Europea.
    - Revisar los sistemas de retroalimentación y encuestar a los usuarios periódicamente para contemplar sus sugerencias, posibles errores, o infracciones éticas detectadas por ellos.
    - Contemplar los progresos en la regulación y la ética de los sistemas de IA, y verificar que los estándares más altos posibles están siendo aplicados.
- Un **plan de acción** que:
    - Establezca un protocolo de acción claro para las infracciones éticas detectadas en el sistema.
    - Trate las infracciones éticas como errores críticos que requieren resolución inmediata.
    - Haga un requisito la documentación de los problemas detectados, los pasos que se han tomado para resolverlos, y qué medidas fueron tomadas para evitar futuras infracciones.

# Capítulo 5

# Conclusiones y líneas futuras

En este trabajo hemos realizado un análisis de las diferentes disciplinas de la ética aplicada a campos de la informática, destacando la importancia de distinguir entre ámbitos de aplicación. También hemos destacado los problemas que están presentes en la aplicación de la ética en el campo de la IA, y hemos examinado las medidas adoptadas por los gobiernos de las grandes potencias para regularla.

Posteriormente, hemos analizado las directrices de la UNESCO y del documento de *Ética por Diseño* en busca de un marco teórico y práctico que nos proporcione principios y metodologías para su aplicación. Al revisar estas propuestas, pudimos extraer estas directrices y técnicas que nos ayudaron a realizar una implementación práctica de los principios éticos.

Para ilustrar la aplicación práctica de la ética, propusimos la implementación de un LLM en el entorno parlamentario. Este caso práctico mostró que actualmente ninguna de las IAs analizadas cumple completamente con los requisitos legales del *Reglamento de Inteligencia Artificial* de la UE, y describió las múltiples medidas necesarias para garantizar que la implementación del sistema sea segura y ética.

Este trabajo deja abierta muchas líneas de investigación. Líneas de investigación similares pueden incluir el estudio de la implementación de la ética en la IA con un enfoque en el desarrollo de una IA ética. Otras posibles líneas de investigación pasan por ampliar este trabajo con herramientas de evaluación del impacto ético, o analizar una implementación real junto con los resultados de su elaboración y mantenimiento.

Líneas de investigación más genéricas podrían involucrar un análisis interdisciplinario entre filosofía, derecho, e informática, con el objetivo de identificar principios éticos adicionales y describir cómo ponerlos en práctica. Otro enfoque podría ser un estudio no principialista que busque elaborar otras tareas y requisitos éticos desde diversas perspectivas éticas, o proponer una metodología alternativa para derivar tareas y objetivos a partir de principios éticos.

Siguiendo estas líneas de investigación, podemos avanzar en la comprensión de la ética de la IA y desarrollar estrategias más efectivas para garantizar que estos sistemas se desarrollen y utilicen de manera moral, responsable, y beneficiosa para la sociedad.

# Capítulo 6

# Summary and Conclusions

In this paper we have made an analysis of the different disciplines of ethics applied to computer science fields, emphasizing the importance of distinguishing between areas of application. We have also highlighted the problems present in the application of ethics in the field of AI, and we have examined the measures taken by the governments of major powers to regulate it.

Subsequently, we have analyzed the UNESCO and *Ethics by Design* guidelines, seeking a theoretical and practical framework that provides us with principles and methodologies for their application. By reviewing these proposals, we were able to extract guidelines and techniques that helped us carry out a practical implementation of ethical principles.

To illustrate the practical application of ethics, we proposed the implementation of an LLM in the parliamentary context. This case study showed that currently none of the AI analyzed fully complies with the legal requirements of the EU's *AI Act*, and described the multiple measures that are necessary to ensure that the system's implementation is safe and ethical.

This paper leaves open many lines of research. Similar lines of inquiry may include studying the implementation of ethics in AI with a focus on developing ethical AI. Other potential lines of research include extending this paper with ethical impact evaluation tools, or analyzing a real-world implementation along with the results of its development and maintenance.

More generic lines of research could involve an interdisciplinary analysis between philosophy, law, and computer science, aiming to identify additional ethical principles and describe how to put them into practice. Another approach might be a non-principlist study that seeks to elaborate on other ethical tasks and requirements from various ethical perspectives, or to propose an alternative methodology for deriving tasks and goals based on ethical principles.

By following these lines of research, we can advance our understanding of AI ethics and develop more effective strategies to ensure that these systems are developed and used in a moral, responsible, and beneficial manner for society.

# Capítulo 7

# Presupuesto

El presupuesto de este trabajo abarca principalmente dos componentes, el trabajo del investigador y los recursos informáticos necesarios para el estudio. El coste asociado con el trabajo del estudiante investigador abarca el tiempo empleado en la revisión de las fuentes, la realización de pruebas, el análisis, y la escritura del documento. Los costes de los recursos informáticos cubren los aspectos relacionados con el software y hardware necesarios para la realización de pruebas y análisis. Otros recursos de este proyecto incluyen el consumo eléctrico.

| Tipos | Descripción | Coste |
|---|---|---|
| Trabajo del investigador | Tiempo empleado en actividades de investigación y redacción. | 300 h × 12 €/h = 3.600 € |
| Recursos informáticos | Equipo y software | 650 € |
| Otros recursos | Consumo eléctrico (300h × 500W = 0.15MWh) | 0.15 MWh × 108.6 €/MWh = 16,20 € |
| **Total** | | **4.266,20 €** |

*Tabla 11: Desglose de costes*

# Referencias


[1] OECD, *Recommendation of the Council on Artificial Intelligence*, 2024. OECD/LEGAL/0449. Disponible en: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[2] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*. Bruselas, Bélgica: Comisión Europea, 2019. Disponible en: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[3] J. Croft, *"NTSB: Colgan 3407 pitched up despite anti-stall push,"* Flight Global, Febrero 2009. Disponible en: https://www.flightglobal.com/ntsb-colgan-3407-pitched-up-despite-anti-stall-push/85134.article [Accedido el 24 de Julio de 2024].

[4] A. Tsamados, N. Aggarwal, J.Cowls, J. Morley, H. Roberts, M. Taddeo y L. Floridi, "The ethics of algorithms: key problems and solutions," *AI & Society*, vol. 37, pp. 215-230, 2022, https://doi.org/10.1007/s00146-021-01154-8

[5] Federal Data Strategy, *Data Ethics Framework*, Washington, DC: Office of Management and Budget, 2020, Disponible en: https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf [Accedido el 19 de Febrero de 2024].

[6] Grupo independiente de expertos de alto nivel sobre inteligencia artificial, *Directrices éticas para una IA fiable*. Bruselas, Bélgica: Comisión Europea, 2019. Disponible en: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[7] M. Wei y Z. Zhou, "AI Ethics Issues in Real World: Evidence from AI Incident Database," En *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2003, pp. 4923-4932, http://dx.doi.org/10.24251/HICSS.2023.602

[8] International Monetary Fund, "World Economic Outlook Database," Reporte autogenerado, Abril, 2024. Disponible en: https://www.imf.org/en/Publications/WEO/weo-database/2024/April/weo-report?c=122,124,918,924,960,423,935,128,939,172,132,134,174,532,944,178,136,941,946,137,546,181,138,964,182,359,968,936,961,184,144,111,&s=PPPSH,&sy=2024&ey=2024&ssm=0&scsm=1&scc=0&ssd=1&ssc=0&sic=0&sort=country&ds=.&br=1 [Accedido el 23 de Abril de 2024].

[9] L. Zhang, "China: AI Governance Principles Released," Library of Congress, 2019. [Online]. Disponible en: https://www.loc.gov/item/global-legal-monitor/2019-09-09/china-ai-governance-principles-released/ [Accedido el 17 de Abril de 2024].

[10] 国家互联网信息办公室, 互联网信息服务算法推荐管理规定. Pekín, China: 国家市场监督管理总局, 2021. Disponible en: https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm (traducido en: China Law Translate, "Provisions on the Management of Algorithmic Recommendations in Internet Information Services," Enero 2022 [Online] Disponible en: https://www.chinalawtranslate.com/en/algorithms/ [Accedido el 21 de Mayo de 2024]).

[11] 国家互联网信息办公室, 互联网信息服务深度合成管理规定. Pekín, China: 中华人民共和国公安部, 2022. Disponible en: https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm (traducido en: China Law Translate, “Provisions on the Administration of Deep Synthesis Internet Information Services,” Diciembre 2022 [Online] Disponible en: https://www.chinalawtranslate.com/en/deep-synthesis/ [Accedido el 21 de Mayo de 2024]).

[12] IAPP Research and Insights, *Global AI Law and Policy Tracker*. Portsmouth, NH: International Association of Privacy Professionals, 2024. Disponible en: https://iapp.org/resources/article/global-ai-legislation-tracker/ [Accedido el 21 de Mayo de 2024].

[13] J. R. Biden, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, Washington, DC: The White House, 2023. Disponible en: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[14] The White House, “FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence,” Octubre 2023. Disponible en: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

[15] The White House, “Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People,” Octubre 2022. Disponible en: https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

[16] Parlamento Europeo, “Reglamento de Inteligencia Artificial,” Marzo 2024. Disponible en: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_ES.pdf

[17] Future of Life Institute, “High-level summary of the AI Act,” Febrero 2024. Disponible en: https://artificialintelligenceact.eu/high-level-summary/

[18] Parlamento Europeo, “Reglamento general de protección de datos,” Abril 2016. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R0679

[19] J. Vincent, “Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement,” *The Verge*, 6 de Febrero de 2023. Disponible en: https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion

[20] T. Mucci, C. Stryker, “¿Qué es la gobernanza de la IA?,” ibm.com, 28 de Noviembre de 2023. Disponible en: https://www.ibm.com/es-es/topics/ai-governance [Accedido el 4 de Junio de 2024].

[21] E. Prem, “From ethical AI frameworks to tools: a review of approaches,” *AI and Ethics*, vol. 3, pp. 699-716, 2023, https://doi.org/10.1007/s43681-023-00258-9

[22] UNESCO, *Recomendación sobre la ética de la inteligencia artificial*. París, Francia: UNESCO, 2022. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa.locale=es (por motivos de preservación, también se puede encontrar en el siguiente enlace https://web.archive.org/web/20240605115631/https://unesdoc.unesco.org/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_50daf52c-56dc-4375-ba1f-3574cd3d9b3f?_=381137spa.pdf).

[23] Comisión Europea, *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence*. Bruselas, Bélgica: Comisión Europea, 2021. Disponible en: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

[24] Microsoft, "Perform secure design review and threat modeling," en *Security Development Lifecycle Practices*. Disponible en: https://www.microsoft.com/en-us/securityengineering/sdl/practices/secure-by-design

[25] C. Kern, *Secure by Design at Google*. Mountain View, CA: Google, 2024.

[26] SHERPA Project Team, "About," [Online]. Disponible en: https://www.project-sherpa.eu/about/ [Accedido el 25 de Junio de 2024].

[27] SIENNA, "About SIENNA," [Online]. Disponible en: https://www.sienna-project.eu/w/si/about-sienna [Accedido el 25 de Junio de 2024].

[28] *Estatuto de Autonomía de Canarias,* rev. 28 Diciembre 2022. Madrid, España: Agencia Estatal Boletín Oficial del Estado, 2022. Disponible en: https://www.boe.es/biblioteca_juridica/publicacion.php?id=PUB-PB-2022-130

[29] *Constitución Española*, rev. 19 Febrero 2024. Madrid, España: Agencia Estatal Boletín Oficial del Estado, 2024. Disponible en: https://www.boe.es/biblioteca_juridica/codigos/codigo.php?id=151

[30] Cloudflare, "What is a large language model (LLM)?," Cloudflare [Online]. Disponible en: https://www.cloudflare.com/learning/ai/what-is-large-language-model/ [Accedido el 25 de Junio de 2024].

[31] A. Kucharavy, "Fundamental Limitations of Generative LLMs," *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, pp. 55-64. Cham, Suiza: Springer, 2024. https://doi.org/10.1007/978-3-031-54827-7_5

[32] P. Brey y B. Dainow, "Ethics by Design and Ethics of Use in AI and Robotics," *D5.4: Multi-Stakeholder Strategy and Tools for Ethical AI and Robotics*, pp. 33-73. SIENNA, 2021. https://doi.org/10.5281/zenodo.5536176

[33] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall y T. Wolf, "Open LLM Leaderboard" [Online]. Hugging Face, 2023. Disponible en: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard [Accedido el 18 de Junio de 2024].

[34] Google, "Gemini API" [Documentación online]. Google, 2024. Disponible en: https://ai.google.dev/gemini-api/docs [Accedido el 18 de Junio de 2024].

[35] OpenAI, “OpenAI documentation” [Documentación online]. OpenAI: 2024. Disponible en: https://platform.openai.com/docs/overview [Accedido el 18 de Junio de 2024].

[36] Anthropic, “Welcome to Claude” [Documentación online]. Anthropic: 2024. Disponible en: https://docs.anthropic.com/en/docs/welcome [Accedido el 18 de Junio de 2024].

[37] Gemma Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, et al., *Gemma*. Kaggle, 2024. https://doi.org/10.34740/KAGGLE/M/3301

[38] AI@Meta, “Introducing Meta Llama 3: The most capable openly available LLM to date,” Abril 2024 [Online]. Disponible en: https://ai.meta.com/blog/meta-llama-3/ [Accedido el 18 de Junio de 2024].

[39] AI@Meta, *Llama 3 Model Card*. 2024 [Online]. Disponible en: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md [Accedido el 18 de Junio de 2024].

[40] Mistral AI Team, “Mixtral of experts,” Mistral AI, 2023 [Online]. Disponible en: https://mistral.ai/news/mixtral-of-experts/ [Accedido el 18 de Junio de 2024].

[41] Qwen Team, *Introducing Qwen1.5*. Febrero 2024 [Online]. Disponible en: https://qwenlm.github.io/blog/qwen1.5/ [Accedido el 18 de Junio de 2024].

[42] Qwen Team, Introducing Qwen-7B: “Open foundation and human aligned models (of the state-of-the-arts).” Agosto 2023 [Online]. Disponible en: https://github.com/QwenLM/Qwen/blob/main/tech_memo.md [Accedido el 18 de Junio de 2024].

[43] NVIDIA Corporation. “GPU NVIDIA H100 Tensor Core.” 2024 [Online]. Disponible en: https://www.nvidia.com/es-es/data-center/h100/ [Accedido el 1 de Julio de 2024].

[44] D. Bergmann, “What is instruction tuning?,” ibm.com, 5 de Abril de 2024. Disponible en: https://www.ibm.com/topics/instruction-tuning [Accedido el 7 de Julio de 2024].

[45] Advanced Micro Devices, Inc. “System requirements (Linux),” en *ROCm documentation*. Junio 2024 [Documentación online]. Disponible en: https://rocm.docs.amd.com/projects/install-on-linux/en/latest/reference/system-requirements.html [Accedido el 1 de Julio de 2024].

[46] Hernández Suárez, A., “Resolución de 7 de junio de 2024, por la que se concede la autorización administrativa de la instalación eléctrica de alta tensión denominada Línea aérea de transporte de energía eléctrica a 66 kV simple circuito. Modificación de la línea La Oliva-Salinas en el tramo T-2 y T-4, término municipal de Puerto del Rosario, Fuerteventura,” *Boletín Oficial de Canarias*, Nº 122, BOC-A-2024-122-2021, 24 de Junio. Santa Cruz de Tenerife, España: Gobierno de Canarias, 2024. Disponible en: https://www.gobiernodecanarias.org/boc/2024/122/010.html

[47] González Pérez, M.I., “Resolución de 12 de junio de 2024, por la que se modifica el crédito de la financiación inicial y se amplía el crédito para la convocatoria de plazas y tramitación anticipada de ayudas de movilidad europea de estudiantes correspondientes al curso 2024/2025, en el marco del Programa Erasmus+, en la modalidad de Movilidad de

Estudiantes para Estudios," *Boletín Oficial de Canarias*, Nº 122, BOC-A-2024-122-2029, 24 de Junio. Santa Cruz de Tenerife, España: Gobierno de Canarias, 2024. Disponible en: https://www.gobiernodecanarias.org/boc/2024/122/018.html

[48] León Pérez, I.K., "Resolución de 13 de junio de 2024, por la que se convoca la primera edición de los Premios María José Rodrigo López a Trabajo Fin de Grado (TFG) y Trabajo Fin de Máster (TFM) de la Cátedra de Infancia y Adolescencia de la Universidad de La Laguna," *Boletín Oficial de Canarias*, Nº 122, BOC-A-2024-122-2030, 24 de Junio. Santa Cruz de Tenerife, España: Gobierno de Canarias, 2024. Disponible en: https://www.gobiernodecanarias.org/boc/2024/122/019.html

[49] Miranda Medina, M., "Orden de 26 de junio de 2024, por la que se resuelve la convocatoria para la provisión, por el procedimiento de libre designación, del puesto de trabajo n.º 13205710, denominado Jefe/a Servicio Régimen Jurídico, adscrito al Servicio de Régimen Jurídico de la Secretaría General Técnica de esta Consejería," *Boletín Oficial de Canarias*, Nº 133, BOC-A-2024-133-2184, 26 de Junio. Santa Cruz de Tenerife, España: Gobierno de Canarias, 2024. Disponible en: https://www.gobiernodecanarias.org/boc/2024/133/001.html

[50] Quintero Castañeda, A.N., "Extracto de la Orden de 28 de junio de 2024, por la que se reabre el plazo de presentación de solicitudes para acogerse a las subvenciones previstas en la Orden de 26 de marzo de 2024, de este Departamento, que convoca para el ejercicio 2024 las subvenciones destinadas a la reconstrucción del potencial de producción agrícola dañado por la erupción volcánica en los municipios de Tazacorte, El Paso y Los Llanos de Aridane, isla de La Palma.," *Boletín Oficial de Canarias*, Nº 133, BOC-A-2024-133-2192, 26 de Junio. Santa Cruz de Tenerife, España: Gobierno de Canarias, 2024. Disponible en: https://www.gobiernodecanarias.org/boc/2024/133/009.html

[51] Abreu Rosado, A.I., "Gerencia Municipal de Urbanismo.- Anuncio de 26 de junio de 2024, relativo a la nulidad del Plan Parcial Guamasa 3," *Boletín Oficial de Canarias*, Nº 133, BOC-A-2024-133-2197, 26 de Junio. Santa Cruz de Tenerife, España: Gobierno de Canarias, 2024. Disponible en: https://www.gobiernodecanarias.org/boc/2024/133/014.html

[52] Hipólito Alejandro Suárez Nuez, H.A., "Orden de 27 de junio de 2024, por la que se autoriza como denominación específica del Colegio de Educación Infantil y Primaria En Barranco del Ciervo, la de Colegio de Educación Infantil y Primaria Morro Jable II," *Boletín Oficial de Canarias*, Nº 134, BOC-A-2024-134-2202, 27 de Junio. Santa Cruz de Tenerife, España: Gobierno de Canarias, 2024. Disponible en: https://www.gobiernodecanarias.org/boc/2024/134/005.html

[53] Gustavo Pérez Martín, G. y Ortega Montesdeoca, M.M., "Resolución de 26 de junio de 2024, de la Directora Ejecutiva, por la que se dispone la publicación del Convenio de Adhesión del Ayuntamiento de Güímar a la Agencia Canaria de Protección del Medio Natural," *Boletín Oficial de Canarias*, Nº 134, BOC-A-2024-134-2203, 27 de Junio. Santa Cruz de Tenerife, España: Gobierno de Canarias, 2024. Disponible en: https://www.gobiernodecanarias.org/boc/2024/134/006.html

[54] Lahbib, H., "Recomendación del Consejo de 25 de junio de 2024 sobre un plan director para coordinar la respuesta a escala de la Unión en caso de perturbaciones de infraestructuras críticas con importancia transfronteriza significativa," Diario Oficial de la Unión Europea, C/2024/4371, 25 de Junio. Luxemburgo: Oficina de Publicaciones de la Unión Europea, 2024. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024H04371