SOCIOLOGÍA & TECNOCIENCIA

Ed UVa EDICIONES Universidad Valladolid

Acerca de la idea de una Inteligencia Artificial Responsable*

About the idea of a Responsible Artificial Intelligence

RAÚL TABARÉS GUTIÉRREZ
TECNALIA, Basque Research and Technology Alliance. Parque Científico-Tecnológico de Bizkaia. Derio. Edif. 700. C.P. 48160.
raul.tabares@tecnalia.com.
ORCID:https://orcid.org/0000-0002-8149-3534.

Recibido/Received: 08-01-2025. Aceptado/Accepted: 10-05-2025
Cómo citar/How to cite: Tabarés Gutiérrez, Raúl (2025). Acerca de la idea de una Inteligencia artificial responsable. *Sociología y Tecnociencia*, 15 (2), 129-150. DOI: https://doi.org/10.24197/st.2.2025.129-150

Resumen: La idea de una Inteligencia Artificial Responsable (IAR) se ha abierto paso en los últimos años como una reacción ante los diferentes retos que plantea el desarrollo de la Inteligencia Artificial (IA) en la sociedad. A través de un estudio comparativo de dos empresas que diseñan y desarrollan sistemas de IA en nuestro país, se argumenta que el paradigma de la IAR está todavía lejos de ser popular fuera de estas plataformas digitales y cómo el desarrollo de sistemas de IA no atiende de manera satisfactoria las implicaciones éticas y sociales que tienen asociadas este tipo de sistemas.
Palabras clave: Inteligencia artificial; estudios de innovación; innovación responsable, transformación digital; culturas digitales.

Abstract: The idea of Responsible Artificial Intelligence (RAI) has emerged in recent years as a reaction to the different challenges posed by the development of Artificial Intelligence (AI) in society. Through a comparative case study, including two companies focused on the design and the development of AI systems in our country, the text argues that the RAI paradigm is still far from being popular outside these digital platforms and how the development of AI systems does

* Este trabajo se ha realizado en el marco del Proyecto de Investigación "SIRSE" (Sistemas industriales inteligentes, robustos, seguros y éticos para la Industria 5.0: paradigmas avanzados de especificación, diseño, evaluación y monitorización), financiado por el Gobierno Vasco a través del programa ELKARTEK (KK-2022/00007). También nos gustaría agradecer a todos los participantes en la investigación por compartir desinteresadamente sus puntos de vista, ideas y experiencias con el equipo investigador.

not satisfactorily address the ethical and social implications associated with the deployment of AI systems.
Keywords: Artificial intelligence, innovation studies, responsible innovation, digital transformation, digital cultures.

1. INTRODUCCIÓN

Una de las tecnologías que destaca sobremanera dentro de la transformación digital que se produce en la sociedad es la Inteligencia Artificial (IA). El desarrollo de dicha tecnología está propiciando grandes avances en diferentes campos, facilitando la automatización de un mayor número de tareas y posibilitando diferentes transformaciones en sectores como la educación, la energía o la salud (Dauvergne 2020; Nowotny 2021). Sin embargo, el desarrollo y popularización de esta tecnología en la sociedad conlleva un buen número de recelos y preocupaciones derivadas de sus numerosas implicaciones éticas, legales y sociales tales como la exacerbación de sesgos ya existentes, la falta de transparencia y confianza en su funcionamiento o las crecientes preocupaciones por sus impactos medioambientales (Bender et al. 2021; Broussard 2018; Coeckelbergh 2022; Crawford 2021; Eubanks 2018); Tabarés Gutiérrez, 2025).

Estos recelos han impulsado recientes regulaciones, en las que la Comisión Europea parece haber tomado la delantera con el desarrollo de la "AI Act" (Tambama 2022). Esta ley trata de responder a los posibles impactos, implicaciones y riesgos asociados al desarrollo de esta tecnología. Estos recelos han sido también objeto de discusiones académicas en torno a la idea de una Inteligencia Artificial Responsable (IAR) (Dignum 2017; Stahl 2023; Stahl and Wright 2018). Es decir, un modo de desarrollo tecnológico en el que los actores alrededor de los sistemas de IA puedan rendir cuentas de sus impactos y puedan responsabilizarse de los mismos. Sin embargo, existe un conocimiento limitado de cuál es la difusión de esta idea fuera de los círculos académicos y de cómo es percibida por diferentes tipos de empresas que trabajan en la transformación digital en diferentes sectores y desarrollan actualmente sistemas de IA.

Esta contribución pretende rellenar este vacío, a través de un estudio comparativo de caso, que explora las nociones de responsabilidad en dos contextos empresariales diferenciados. En particular se trata de analizar el grado de familiaridad de dos empresas con el enfoque de IAR, qué tipo de implicaciones éticas y sociales surgen en el desarrollo de sistemas de IA y qué nociones de responsabilidad situadas existen en dichos contextos empresariales. Es decir, qué significa la idea de IAR para cada una de estas organizaciones y profesionales y cómo se operacionaliza o se puede operacionalizar en sus contextos empresariales. Para ello, se adopta un enfoque cualitativo, informado a través de una revisión narrativa de la literatura de la IAR, un trabajo de campo que engloba nueve entrevistas con informantes clave, además de un taller de contraste a modo de grupo focal con las dos organizaciones, para tratar

de arrojar luz sobre cómo estas dos organizaciones entienden la responsabilidad en el desarrollo de sistemas de IA y que implicaciones éticas se revelan en sus tareas y proyectos rutinarios.

En el texto se argumenta que el desarrollo de sistemas de IA lleva asociado unas implicaciones éticas y sociales que no son atendidas debidamente por los desarrolladores de estos sistemas. Si bien hay en marcha diversas iniciativas regulatorias como la "AI Act", esta primera regulación a nivel europeo debería ser acompañada por otra serie de medidas que traten de abordar los diferentes impactos e implicaciones que tienen estos sistemas más allá del área regulatoria. Sobre todo en materia de alfabetización digital de la sociedad, al igual que en el establecimiento de mecanismos que favorezcan la rendición de cuentas por parte de las organizaciones que desarrollan estos sistemas de IA. Con esta contribución se quiere contribuir a la ingente literatura sobre los aspectos éticos, legales y sociales de la IA, pero también a la incipiente literatura que trata de operacionalizar estos principios de alto nivel a través de la idea de la IAR.

El artículo se estructura de la siguiente forma: La siguiente sección provee de una revisión de la literatura de las implicaciones éticas y sociales del desarrollo de la IA y del enfoque de IAR. La sección tercera presenta el contexto de la investigación y la metodología empleada. La cuarta sección presenta los hallazgos, antes de concluir el texto con una discusión de las implicaciones de dichos hallazgos y unas breves conclusiones finales.

2. SOBRE LA IDEA DE UNA INTELIGENCIA ARTIFICIAL RESPONSABLE

2.1. Implicaciones éticas y sociales de los sistemas de inteligencia artificial

Los sistemas de IA adolecen de una opacidad considerable debido a que las tecnologías que se utilizan en su desarrollo como Machine learning, Deep learning o el procesamiento natural del lenguaje son muchas veces inescrutables en términos humanos. La transparencia es un valor clave a la hora de que los usuarios desarrollen confianza en el sistema con el que interaccionan (Shin and Park 2019), y esta se puede fomentar de diferentes maneras, ya sea a través de favorecer un mejor entendimiento de los resultados del sistema o tratando de abrir la opacidad de los sistemas IA a los usuarios (Jammalamadaka and Itapu 2022).

Es importante aclarar que los usuarios que interactúan con los sistemas IA no tienen por qué ser expertos en esta tecnología, ni tienen porque tener conocimientos avanzados en estadística y/o computación, lo cual hace que se tengan que relajar las barreras de entrada para el uso de esta tecnología si se quiere popularizar su uso. En este sentido la transparencia es un valor que contribuye innegablemente a dicho propósito (Csiszar et al. 2020) y que comúnmente se enfatiza en estrategias y guías a nivel internacional (Jobin, Ienca, and Vayena 2019).

Otra de las problemáticas comunes alrededor de los sistemas de IA es el de la interpretabilidad, alrededor del cual la industria ha realizado diversos esfuerzos, desarrollando componentes (en su mayoría interfaces gráficas) para ayudar a los usuarios (normalmente no expertos en IA) a interactuar con estos sistemas y favorecer la comprensión de sus resultados en términos humanos (Jammalamadaka and Itapu 2022; Kordzadeh and Ghasemaghaei 2022; Zhu et al. 2021). La falta de interpretabilidad que presentan estos sistemas es común y varias tecnologías de IA como las redes neuronales (deep learning) presentan una problemática muy difícil de abordar, ya que en muchos casos es casi imposible ofrecer una comprensibilidad en términos humanos para las decisiones que toman este tipo de sistemas (Anagnostou et al. 2022; Nakao et al. 2022).

Por ello los sistemas de IA suelen adolecer de una falta de confianza generalizada por parte de sus usuarios debido a las características de la tecnología. La necesidad de promover una IA de confianza ha estado muy presente en el debate académico (Buruk, Ekmekci, and Arda 2020; Jobin et al. 2019; Nowotny 2021; Ryan 2020) e instituciones como la Comisión Europea (CE) han promovido grupos de expertos, guías y recomendaciones con el objetivo de fortalecer la confianza en esta tecnología a través del paradigma del "Trustworthy AI" (High-Level Expert Group on Artificial Intelligence 2019).

Sin embargo, la operacionalización de este tipo de valores sigue siendo una asignatura pendiente, ya que la confianza en los sistemas IA no sólo depende de un valor determinado, sino de varias familias de valores a la vez (Zhu et al. 2021). Al mismo tiempo, tampoco se puede delimitar el desarrollo de esta confianza en los sistemas en base al desarrollo de ciertos componentes de software o interfaces que traten de promover la confianza en los sistemas IA, sino que hay que acompañar la implementación de estos componentes con una serie de buenas prácticas a nivel organizacional y/o institucional que prioricen estos valores en el diseño, desarrollo y operacionalización de estos sistemas. Este hecho es crítico si se quieren combatir las visiones tecno-solucionistas y tecno-optimistas que se han promovido activamente desde Sillicon Valley en los últimos años (Broussard 2018; Morozov 2013). Estas aproximaciones se caracterizan por reducir la multidimensionalidad de las problemáticas sociales y de dotar a la tecnología de una excesiva confianza a la hora de abordar diferentes problemáticas sociales.

Por último, el desarrollo paralelo a estos sistemas de IA de mecanismos de evaluación, auditoría y responsabilidad externas se torna como imprescindible (Shneiderman 2021). En este sentido, es importante considerar al desarrollo de sistemas IA no como una tarea restringida a perfiles asociados a la IA, sino también a otro tipo de perfiles que tengan conocimientos de usabilidad, desarrollo de software, requerimientos legales e implicaciones éticas, sociales y medioambientales, entre otros (Csiszar et al. 2020; Tabarés Gutiérrez, 2025). Todo ello con el objetivo de empoderar y capacitar a los usuarios que tienen que lidiar con estos sistemas para que puedan evaluar su uso de una forma informada y crítica (Nakao et al. 2022).

2.2. Inteligencia Artificial Responsable

Debido a estos condicionantes, en los últimos años hemos asistido a la emergencia de la idea de una IAR (Dignum 2017). Esta idea se entronca a su vez en la popularización mundial del enfoque de innovación responsable, el cual ha atraído la atención de la comunidad académica de manera notable durante la última década (Owen and Pansera 2019; Stilgoe, Owen, and Macnaghten 2013). Sin embargo, la idea de una innovación responsable pese a estar muy presente en la literatura sigue adoleciendo de una limitada implementación en programas de financiación de la investigación a escala nacional e internacional (Tabarés et al 2022), y de manera particular en la industria (van de Poel et al. 2020).

Esto último es de gran importancia, ya que la mayoría de los sistemas más importantes de IA se encuentran en manos privadas. La idea de una IAR hereda en buena medida las proposiciones de la innovación responsable, aunque también bebe de otras propuestas académicas como la responsabilidad social corporativa, el enfoque ELSA (Rip 2014), la evaluación constructiva de tecnologías (Burget, Bardone, and Pedaste 2017) o los criterios FAT (en inglés "Fairness, Accountability and Transparency") (Shin and Park 2019). Y es que el desarrollo de la IA conlleva una serie de implicaciones éticas, legales y sociales inherentes al desarrollo de la tecnología (Crawford 2021; Mittelstadt et al. 2016; Nowotny 2021). La idea de una IAR más comúnmente interpelada es la de Virginia Dignum, la cual enuncia como:

> "la IA Responsable es más que unos tics en unos apartados sociales o que unos añadidos a los sistemas de IA". (página 4, Dignum, 2017)

Para Dignum, esta visión de una IAR se apoya en tres pilares de igual importancia. En primer lugar, la sociedad debe estar preparada para responder a los impactos de la IA, lo que significa que investigadores y desarrolladores deben estar equipados para ser responsabilizados de los impactos de sus desarrollos. En segundo lugar, la IAR implica el desarrollo de mecanismos que permitan que los sistemas de IA puedan razonar y actuar acorde a los valores humanos, lo que requiere de modelos y algoritmos que puedan representar decisiones basadas en base a valores humanos, ya que técnicas como el "Deep learning" manifiestan una falta de interpretabilidad que lo imposibilita. Por último, la participación es necesaria para entender como diferentes investigadores y desarrolladores trabajan en torno a la IA para desarrollar marcos de IAR (Dignum 2017).

La idea de una IAR se caracteriza por la inclusión, protección y salvaguarda de varios valores presentes en la sociedad, mediante un diálogo activo entre industria y sociedad. Y es que el desarrollo de diversas soluciones de IA puede poner en riesgo ciertos valores o no ser tenidos en cuenta por el desarrollo de la tecnología (Buruk et al. 2020). Debido a estos riesgos durante los últimos

años se ha incidido en la regulación de estos sistemas, especialmente en Europa, con el desarrollo de la "AI Act" (Tambama 2022).

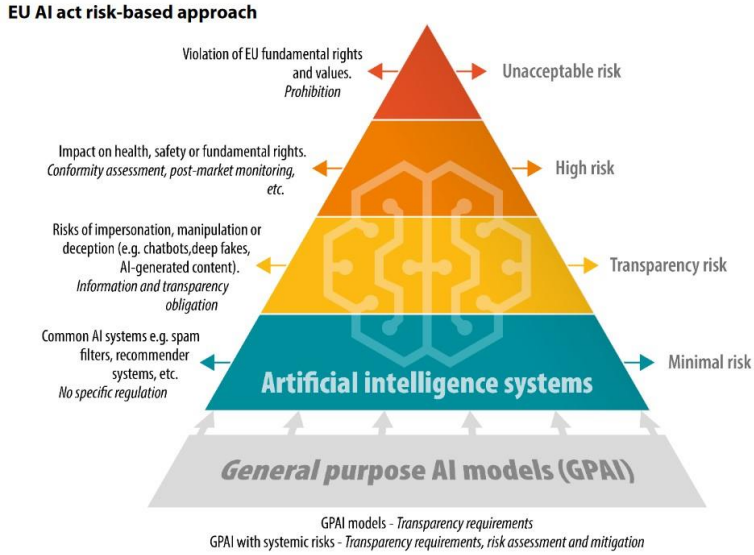

Figura 1 Pirámide de riesgos en el "AI Act". Fuente: CE

Esta ley es la primera regulación específica para este tipo de tecnologías que además exige una autoevaluación de impacto de los sistemas IA a las empresas desarrolladoras, de cara a su introducción en el mercado y con una categorización del riesgo (ver figura 1) de estos sistemas en base a cuatro niveles (Floridi et al. 2022). Esta ley sigue la ambición regulatoria de la Unión al igual que en su día se promovió el Reglamento General de Protección de Datos[1] o la más reciente Ley de Servicios Digitales[2], que trata de poner el foco sobre las grandes plataformas digitales (Prieto Vertel y Tabarés Gutiérrez, 2025).

La idea de una IAR también ha sido promovida por parte de algunas plataformas digitales como Microsoft o Google, las cuales ofrecen diferentes herramientas tecnológicas, principalmente, y otro tipo de recursos en sus respectivos portales dedicados a tal efecto[3]. Estas herramientas se enfocan en el tratamiento y análisis de los datos, pruebas para analizar sesgos, plantillas para proveer de

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?toc=OJ%3AL%3A2022%3A277%3ATOC&uri=uriserv%3AOJ.L_.2022.277.01.0001.01.ENG

[3] Ver https://www.microsoft.com/es-mx/ai/responsible-ai; https://www.tensorflow.org/responsible_ai?hl=es-419

transparencia, etc. Dichas herramientas tratan de facilitar la operacionalización de una IAR, aunque desde un punto de vista tecno-solucionista y/o tecno-optimista (Morozov 2013; Broussard 2018) que muchas veces subestima el impacto social de estos sistemas de IA una vez son desplegados sobre un contexto determinado. Es decir, desligando el funcionamiento de dichos sistemas respecto a sus posibles impactos e implicaciones sociales sobre un contexto determinado.

Y es que aparte de un creciente énfasis regulatorio, que a su vez es clave para el desarrollo y aceptación de la tecnología, la idea de la IAR también se ha extendido en los últimos años hacia el concepto de ecosistema de innovación, enfatizando el carácter relacional de los actores que desarrollan sistemas de IA y sus impactos e implicaciones para el resto de los actores con los que interactúan en dicho ecosistema (Minkkinen, Zimmer, and Mäntymäki 2023; Stahl 2023). Este carácter relacional pone de manifiesto la necesidad de desarrollar también enfoques multidisciplinares e interdisciplinares a lo largo del desarrollo tecnológico, que puedan prever, anticipar y mitigar los efectos no deseados de los sistemas de IA en diferentes contextos.

3. METODOLOGÍA Y CONTEXTO

El contexto en el que se desarrolla esta investigación se produce dentro del proyecto SIRSE, en el que se pudo hacer trabajo de campo compuesto por 9 entrevistas y un taller de contraste a modo de grupo focal durante la primavera de 2023, con diferentes profesionales de dos empresas que desarrollan sistemas de IA; una PYME y una gran empresa. Se seleccionaron dichas empresas por su diferente tamaño, pero por su papel activo en torno a la transformación digital de sectores como la salud, la industria o la energía, a través del desarrollo de sistemas de IA en el ámbito industrial. Más concretamente una lo hacía en el desarrollo de un "gemelo digital" para la simulación del comportamiento de diferentes generadores en un parque eólico. Es decir, un modelo virtual de un objeto físico, diseñado para simular su funcionamiento con precisión a través de la adquisición de datos fidedignos del entorno físico y poder anticiparse a problemas futuros en su funcionamiento. El objetivo de este gemelo digital consistía en mejorar la eficiencia de dichos generadores y su rendimiento, funcionamiento y mantenimiento a largo plazo. Por otro lado, la otra empresa abordaba la automatización de un puente grúa introduciendo diferentes componentes predictivos con el objetivo de dotar al sistema de "inteligencia" en la gestión del almacenaje de bobinas de metal en diferentes naves industriales. El objetivo de esta segunda actuación era mejorar la eficiencia de los procesos de almacenaje a través de la automatización.

Gracias a este trabajo de campo se mantuvieron entrevistas con nueve informantes clave que fueron seleccionados entre las dos empresas participantes y de acuerdo a los perfiles que ocupaban en sus respectivas organizaciones, tales como ingeniero de datos, gestor de proyecto, experto en usabilidad y/o responsable de responsabilidad social corporativa, entre otros. Posteriormente, cuatro de los

entrevistados participaron también en un taller de contraste de tipo grupo focal que se detalla más adelante en este apartado metodológico. De esas nueve entrevistas, cinco fueran mantenidas con una organización y cuatro con la otra. La selección de dichos informantes clave siguió un proceso acumulativo, de tipo "bola de nieve", comenzando por las dos personas de las dos organizaciones que habían sido designadas como representantes en el proyecto.

Posteriormente y tras realizar estas dos entrevistas se prosiguió con compañeros suyos que fueron identificados como relevantes de cara a su posición y conocimiento en la organización. Para ello se contactó con perfiles que diseñan y desarrollan sistemas de IA en las dos organizaciones, pero también con perfiles que tienen gran capacidad de decisión e influencia en la cultura corporativa de las organizaciones y los valores que se promueven organizacionalmente, hasta llegar a las nueve entrevistas. Este número de entrevistas permitió reflejar de manera representativa los diferentes perfiles asociados a la IA en las dos organizaciones, a pesar de que en las dos organizaciones los puestos y perfiles recibían diferentes nomenclaturas (ver tabla 1) y sus responsabilidades eran diferentes respecto a las diferentes demandas organizacionales. Al mismo tiempo se intentó también mantener un balance de género en la realización de las entrevistas, aunque la realidad de determinados perfiles asociados a la analítica de datos en las dos empresas mostró la imposibilidad de llegar a este balance, ya que la mayoría de los perfiles relacionados con la IA en las dos organizaciones eran ocupados por hombres (ver tabla 1). En cuanto a la edad de los entrevistados se intentó contactar con diversos segmentos demográficos que pudieran aportar reprsentatividad generacional y por ello los perfiles entrevistados disponen de una experiencia laboral diferenciada.

Para el desarrollo de las entrevistas se siguió un guion que contó con una batería de preguntas orientada a entender las nociones de responsabilidad existentes entre los informantes clave y las organizaciones que representan. Esta guía contenía preguntas alusivas a las diversas nociones de responsabilidad existentes (social, ambiental, legal, etc.), así como al uso de herramientas existentes para su operacionalización (ver anexos). Dicho guion se desarrolló en base a una revisión narrativa del estado de la técnica en torno a la IAR y que se ha presentado en la anterior sección. Esta revisión narrativa se operacionalizó a través de diferentes búsquedas en las bases de datos Scopus y Google Scholar del término "IA responsable" y otros términos relacionados con el concepto como ingeniería, software, diseño y/o ética en IA. Dichas búsquedas se realizaron en inglés, ya que es la lengua en la que se han escrito la mayoría de las contribuciones más importantes a nivel mundial. Dichas búsquedas se enfocaron en los últimos cinco años antes de la investigación (2018-2022), debido a lo emergente del tema. Las búsquedas se centraron en la identificación de literatura que pudiera aportar información sobre el paradigma de "IA Responsable" en el desarrollo de la ingeniería de software de IA, así como sus particularidades, necesidades y factores organizacionales asociados a este tipo de desarrollos.

Tabla 1 Caracterización de los informantes clave seleccionados. Elaboración propia

Número de informante clave	Sexo	Perfil	Tipo de empresa	Experiencia laboral
IC1	Masculino	Responsable de Transformación Digital	Gran empresa	15-20 años
IC2	Masculino	Gestor/a de Proyecto	Gran empresa	5-10 años
IC3	Masculino	Ingeniero/a de Automatización	Gran empresa	5-10 años
IC4	Masculino	Ingeniero/a de Automatización	Gran empresa	5-10 años
IC5	Masculino	Ingeniero/a de Software	PYME	15-20 años
IC6	Masculino	Ingeniero/a de Software	PYME	10-15 años
IC7	Masculino	Ingeniero/a de Software	PYME	50-10 años
IC8	Masculino	Responsable de Diseño	PYME	10-15 años
IC9	Femenino	Director/a de Responsabilidad Social Corporativa	PYME	15-20 años

También se incluyeron preguntas orientadas a conocer qué tipo de relaciones disponían los entrevistados con otras organizaciones y agentes del ecosistema de IA, adoptando un enfoque relacional, y con el propósito de entender su posición y actividades dentro de él. Por último, también se introdujeron preguntas encaminadas a explorar los diferentes tipos de impactos asociados a los sistemas IA que los entrevistados esperan en la sociedad y en la economía en un futuro próximo, con el objetivo de explorar sus percepciones alrededor de este tema (Knott et al. 2022).

Las entrevistas fueron objeto de un análisis temático realizado por el equipo de investigación, identificando los temas más recurrentes a través de las ideas, testimonios o conceptualizaciones que los entrevistados realizaban a partir de las diferentes preguntas (Ryan & Bernard, 2003). Este análisis se llevó a cabo una vez se transcribieron las entrevistas y siguió un proceso de codificado en dos fases. Es decir, revisando, agrupando y refinando con posterioridad los códigos que se habían identificado en la primera revisión de las transcripciones. También se generó un archivo maestro con el fin de visualizar todos los códigos asignados. Para ilustrar las ideas principales de dicho análisis temático se han utilizado fragmentos de dichas transcripciones, los cuales han sido editados cuando ha sido necesario, para facilitar su lectura por el lector y evitar formas y expresiones que dificultasen su comprensión (Laureau 2021). A partir de este análisis temático se han estructurado los cuatro bloques que se presentan en la siguiente sección y que se repetían de manera continuada en las transcripciones de las entrevistas realizadas.

Las entrevistas se realizaron de manera telemática a través de la aplicación Microsoft Teams para favorecer la involucración de un número diverso de participantes, evitar costes y desplazamientos. La duración de dichas entrevistas osciló entre treinta minutos y una hora, dependiendo de la profundidad de las conversaciones y de la apertura de los entrevistados. Hay que recordar que una entrevista es un proceso social y como tal está expuesto a diferentes dinámicas de interacción entre el entrevistador y el entrevistado, asimetrías de información y otras variables que hacen que su desarrollo pueda ser irregular. Las variaciones en la duración de las entrevistas respondieron a estas y otras cuestions que afectaron la profundidad y apertura de las conversaciones con los informantes clave. Así mismo, los participantes recibieron un documento de consentimiento informado que explicaba los objetivos del proyecto y el equipo investigador solventó las diferentes dudas que se plantearon por parte de los entrevistados de cara a la investigación en sí misma. Se siguió además un protocolo ético en base a la recopilación y el tratamiento de los datos, en concordancia con el Reglamento Europeo de Protección de Datos y el responsable del tratamiento de datos de la institución (ver anexos). Además de anonimizarse los datos, las menciones directas a personas, organizaciones o instituciones durante las entrevistas se han cambiado por seudónimos o se han anonimizado para proteger las identidades de las personas, organizaciones e instituciones ajenas a la investigación. Este proceso se ha realizado tanto para las entrevistas como para el taller de contraste de tipo grupo focal que se hizo con las dos empresas.

Dicho taller se realizó durante el mes de noviembre de 2023 en la sede del equipo de investigación y tuvo una duración de unas dos horas y media, con la presencia de dos personas de cada organización involucrada (cuatro en total). Tres de las cuatro personas ya habían participado en las entrevistas y respondían a perfiles con capacidad de decisión anteriormente enunciados (ver tabla 1). Así como en las entrevistas tampoco se pudo obtener una representación balanceada de género, ya

que asistieron tres hombres y una mujer. En dicho taller se ofreció una pequeña presentación de los resultados de las entrevistas, así como otros resultados del proyecto, para poder contrastar con los participantes los hallazgos de dicha investigación. A los asistentes se les facilitó una agenda del taller previamente, además de unas breves instrucciones sobre cómo se realizaría y los documentos de consentimiento informado asociados.

Aparte de este contraste también se habilitó una pequeña dinámica a modo de grupo focal, con el propósito de conocer sus intereses a medio plazo en torno a las implicaciones éticas y sociales de los sistemas de IA y la emergente regulación europea y sus posibles implicaciones para el sector empresarial. De esta manera se intentaba identificar qué necesidades perciben las empresas en torno a estas cuestiones y de qué manera se pueden ir preparando y trabajando para su adaptación a dicha futura normativa, además de contrastar y validar los testimonios resultantes de las entrevistas. Con el objetivo de coordinar dicha sesión se articularon varias preguntas que pudieran guiar el debate, principalmente:

- Desde el punto de vista socio-ético… ¿Cuáles creéis que deberían ser los aspectos que debierais reforzar? ¿Por qué?
- ¿Cuáles son las principales incertidumbres que las crecientes regulaciones a nivel Europeo/nacional os plantean? ¿Por qué? ¿Os sentís capacitados para hacer una autoevaluación del riesgo de vuestros sistemas? ¿En qué manera?
- ¿Qué tipo de herramientas os podrían ser de utilidad para abordar estos nuevos requisitos legales y normativos? ¿Financiación para poner en marcha procesos estandarizados? ¿Formación específica en estos temas? ¿Información detallada como guías, metodologías, etc.?

4. RESULTADOS DEL TRABAJO DE CAMPO

4.1. Carencia de prácticas normalizadas y dependencia contextual

La mayoría de los entrevistados que participaron en el trabajo de campo coincidían en el carácter emergente de los proyectos de IA a los que se enfrentaban en sus organizaciones y la falta de prácticas establecidas o metodologías consolidadas para el desarrollo de este tipo de proyectos. Los entrevistados reflejaban el carácter casi "artesanal" y particular de cada proyecto de IA que se abordaba en sus respectivas organizaciones y cómo se afrontaba con la involucración activa y ayuda de los expertos del dominio de conocimiento, campo y/o sector en el que operaban para su desarrollo satisfactorio.

Muchos de los entrevistados respondían a perfiles técnicos como científico de datos, ingeniero informático, y/o similares. Dichos perfiles enfatizaban la necesidad de contar con la involucración del usuario, técnicos y actores implicados en el dominio y/o contexto donde el sistema IA se iba a implementar. Este conocimiento

del dominio era muchas veces subrayado como crítico, sobre todo a la hora de elegir qué tipo de datos eran más representativos y/o útiles de cara al desarrollo del sistema IA. Determinar dónde y bajo qué situaciones se pueden recoger datos, elegir qué datos son pertinentes y cuales no, etc., de cara al desarrollo de estos proyectos era un factor recurrente en los testimonios de los participantes. La complejidad del dominio donde se desarrollaban los proyectos de IA era algo que los participantes destacaban y el acceso a este tipo de entornos se recalcaba que debía de ser de la mano de los verdaderos conocedores del entorno en cuestión.

De igual manera era común observar cómo los participantes en el estudio también tenían diversas relaciones de colaboración con otros centros de investigación y universidades de cara al desarrollo de capacidades de manera conjunta que les permitieran afrontar los diversos retos tecnológicos en el desarrollo de sistemas de IA. Así, estas colaboraciones respondían a diferentes necesidades ya fueran de carácter puntual o enmarcadas en alianzas más estables.

Así mismo, también se aludían a las similitudes que este tipo de proyectos tienen con proyectos de desarrollo de software más tradicionales. Varios de los entrevistados indicaban como metodologías ágiles y marcos de trabajo en equipo y productividad consolidados en la industria del software tales como SCRUM o Design Sprint pueden ser de ayuda en su desarrollo. Sin embargo, y al mismo tiempo, también se enfatizaban comúnmente las diferencias existentes entre modelos de desarrollo de software clásico y modelos de desarrollo de sistemas IA, ya que estos últimos presentan mayores incertidumbres y producen resultados no deterministas o estocásticos que afectan comúnmente a su desarrollo, difusión y aceptación.

> "Un proyecto de IA tiene muchos puntos en común con un proyecto de software, sobre todo la parte inicial de la edición, tratamiento y demás de datos y el lanzamiento digital del proyecto... Sí que es verdad que la IA es un poco más probabilística y hay varias soluciones que pueden llegar a los mismos resultados, mientras que en el software al ser una industria un poco más madura tienes ya como muchos procesos ya bien definidos. Pero en la IA, en cuanto empiezas a meter cosas...empieza ya a crecer en tamaño el proyecto, el caso de uso y su complejidad. Vas a tener que tener una recepción de los datos, la captación, el tratamiento de los datos, la etapa de preprocesamiento de los datos. Luego también todos estos modelos grandes incluyen procesos de infraestructura de servicios, temas de seguridad, etc. Los modelos de IA en cuanto empiezan a crecer, empiezan a tener más puntos en común con proyectos de software. Sí que tienen particularidades como todas las tecnologías, pero sí que comparten muchos procesos con el software." (IC6)

4.2. Conflictos de valores, dilemas morales e impactos sociales

Cuando se preguntaba a los entrevistados por los diferentes tipos de retos técnicos, sociales, éticos y ambientales que se afrontaban en los diferentes proyectos de IA que abordaban, se aludía comúnmente a retos relacionados con las necesidades

de sus clientes. Muchas veces relativos a la necesidad de ofrecer una mayor transparencia e interpretabilidad en sus sistemas IA, debido al componente de "caja negra" y la opacidad inherente a este tipo de tecnologías. En esta línea los entrevistados confesaban que dichas características de la tecnología suponen una barrera a la hora de vender este tipo de tecnologías y que los usuarios confíen en dichas tecnologías. Así mismo, la indeterminación de los resultados que ofrecen este tipo de sistemas también añade complicaciones a la hora de que el usuario desarrolle confianza en ellos. Estos elementos, en general, eran motivo de preocupación por parte de los entrevistados de cara a la difusión y aceptación de la tecnología.

> "Estamos cobrando por la mejora. Nos interesa que el modelo sea lo más explicable y lo más transparente posible, es algo que nos preocupa, si es una caja negra...o sea, nosotros podemos explicarlo. De hecho, hemos hecho informes en los que explicamos cómo es el modelo con total transparencia, pero sí que va variando, va cambiando. Entonces esa explicabilidad, esa parte de fiabilidad, de transparencia... esa parte sí que es algo que desde el punto de vista en este proyecto y en el caso de uso concreto, pues es lo que digamos habría que trabajar...sobre todo la explicabilidad." (IC5)

Por otro lado, los participantes en el estudio también hacían hincapié en otros dilemas éticos que el uso y popularización de la IA conlleva en sus proyectos. Entre ellos, principalmente, la privacidad de los datos, además de sus diferentes implicaciones y derivadas tales como la seguridad de los mismos, su tratamiento y reutilización. También se hacía hincapié en el uso y aceptación de la tecnología en diversos entornos, ya que en algún caso particular se planteaba la necesidad de realizar formaciones específicas dentro de las plantas industriales de sus clientes, debido a que los trabajadores no confiaban en este tipo de tecnologías y mostraban su rechazo de manera activa.

> "De hecho, nosotros nos hemos encontrado con el sabotaje de equipos, pero no externo, sino interno. Que un sistema empiece a tomar decisiones, que empiece a transformar, a generar valor, sustituir capacidad humana…digamos que es necesario capacitar a los operadores, que ellos entiendan que es un tema de supervisión, que el sistema, por más que sea autónomo, tiene que estar supervisado, tiene que haber alguien que verifique y confirme." (Grupo focal de contraste)

Por último y de manera más residual diversos participantes también interpelaban al creciente coste energético del entrenamiento de los sistemas de IA y de cómo se han desarrollado diversas iniciativas en sus organizaciones para optimizar este tipo de tareas. Así se aludía a la programación de este tipo de tareas durante los fines de semana, cuando el coste de la energía es menor o por las noches, para no contribuir a engrosar su coste energético. Alguno de los participantes también indicaba, de manera más anecdótica, otras acciones alineadas con la promoción de la

sostenibilidad asociadas al aprovechamiento del calor disipado por estas infraestructuras.

4.3. Diligencia, seguridad, privacidad y fiabilidad como responsabilidad "de facto"

La mayoría de los participantes desconocían por completo dichos enfoques y herramientas asociadas, pero sin embargo mostraban interés por ello. Las nociones de responsabilidad "de facto" con las que los participantes concebían su trabajo y actividades asociadas al desarrollo de sistemas IA estaban relacionadas en su mayoría a un desempeño responsable de sus tareas y actividades, mantener la seguridad de los datos que se utilizaban en ellas, gestionar la privacidad de los mismos y asegurar la fiabilidad de los sistemas que se desarrollaban, principalmente.

> "Bueno, pues quizás hacer un uso responsable de los datos. Es que no sabría muy bien decírtelo, la verdad. Mantener la seguridad de los datos. El no compartirlo con cualquiera para poder realizar modelos. Y poco más. Un poco la seguridad de los datos de cada..., bueno en este caso no sé si incluiría personas, pero...herramientas, personas o cualquier tipo. Siempre y cuando el modelo responda, sea coherente, que no sea algo fuera de lo común, pero por el momento no hemos tenido ningún acercamiento a esos enfoques." (IC3)

Por otro lado, y a pesar del desconocimiento generalizado de los participantes de este tipo de enfoques éticos y responsables en el desarrollo de sistemas IA, algunos de los participantes en el estudio indicaban como el contexto europeo está ejerciendo una presión en este sentido que probablemente se traslade a las organizaciones a través de nuevas herramientas, instrumentos y enfoques. Varios entrevistados enfatizaban como los modelos de IA generativa como ChatGPT han supuesto un punto de inflexión y han provocado una gran atención mediática que ha provocado a su vez posteriores movimientos regulatorios asociados al impacto social de estos sistemas en la sociedad.

4.4. Impactos en el trabajo y la creatividad

Por último, la mayoría de los entrevistados reflexionaban comúnmente sobre los posibles impactos sociales y económicos que pudieran tener la popularización de los sistemas IA. Así, muchos de los entrevistados remarcaban como la IA puede tener un impacto directo en el empleo en muchos sectores, provocando grandes transformaciones laborales y resaltando los riesgos asociados al desempleo masivo que la popularización de la IA puede tener a gran escala. Muchos de los entrevistados reflexionaban sobre la dualidad de la tecnología y como un mal uso de este tipo de tecnologías podría desencadenar consecuencias negativas en la sociedad. Así mismo se enfatizaba el auge de la IA generativa y como este tipo de tecnologías disruptivas

pueden suponer un gran cambio de paradigma. En particular, algunos entrevistados resaltaban como este tipo de IAs pueden acarrear grandes cambios en torno a la propiedad intelectual, ya que el modo de que cómo se alimentan estos sistemas IA es cuestionable (derechos de autor de las imágenes y textos con los que se producen los entrenamientos) y sus resultados también (remezcla de contenidos y/o patrones ya establecidos), debido a la gran opacidad existente en torno a cómo funcionan estos sistemas y si sus productos vulneran o no la legislación existente.

> "Está claro que la inteligencia artificial va a sustituir el trabajo de mucha gente y yo creo que es uno de los retos más grandes. O sea, quizá nosotros que estamos haciendo trabajos más especializados y que incluso pensamos que no nos van a afectar, también nos podríamos ver afectados. Entonces yo creo que es uno de los retos, no sé cómo plantearlo, pero bueno, de la resistencia de millones de personas, básicamente. Y aparte con el tema de los riesgos está, pues obviamente las infracciones de derechos de autor y demás. Con la aparición de todos estas IAs que replican comportamientos u obras de arte...ya sea propiedad intelectual en general, entonces bueno, ese es otro de los riesgos que también van a acarrear. O sea, el tema de quién se va a hacer cargo de todos esos desarrollos que hagamos con la IA y que infrinjan la ley de propiedad intelectual." (IC4)

Por otro lado, otros entrevistados también argumentaban que el rol de la creatividad en la sociedad en su conjunto también podría verse influenciado por este tipo de tecnologías de IA generativa, ya que son capaces de remezclar contenidos y crear nuevas obras a partir de diferentes instrucciones suministradas por los usuarios. Se resaltaba que la IA supone también una oportunidad para realizar trabajos que disponga de mayor valor y abandonar tareas que pudieran ser redundantes o de poco valor añadido. En este sentido se argumentaba que una vez la sociedad aprenda a trabajar con este tipo de herramientas y supere los temores que puede producir, podrá aprovechar todo su potencial.

> "Lo que sí veo yo en la gente, sobre todo en la gente que no está dentro de este sector es el tema del miedo con Chat GPT, que todo el mundo nos vamos a quedar sin trabajo y demás. Esto lo decían también con la invención de Internet. Cuando hay tecnologías disruptivas y como ahora mismo está pasando con los grandes modelos de lenguaje, siempre hay ese miedo al cambio, que al final pasará y al final aprenderemos a convivir con ello. Aprenderemos a usarlo en nuestro favor y en ese momento ya será la adopción total de IA. Y ahora mismo está en un estado en el que hay mucho miedo. Los reguladores están intentando sacar planes de regularización para las inteligencias artificiales y vigilar lo que se hace con los datos en los entrenamientos. Pero una vez ya que superemos esto, yo creo que la IA ha venido para quedarse, para ayudar a la gente y a la sociedad a seguir avanzando." (IC8)

5. EL PROBLEMA DE UNA VISIÓN TECNO-OPTIMISTA Y TECNO-SOLUCIONISTA DE LA TECNOLOGÍA

Como se ha evidenciado en esta investigación, los entrevistados presentan un desconocimiento general respecto al concepto de IAR, así como de las herramientas desarrolladas en torno a esta idea, ya sean de carácter tecnológico o no (van de Poel, 2020; Tabarés et al, 2022). Pero esto no se puede achacar a una falta de interés en el concepto, ya que la mayoría de los informantes mostraron interés por ello, además de por cómo podrían beneficiarse de su aplicación para mejorar sus sistemas de IA y mejorar la venta de estos sistemas a sus clientes. Este desconocimiento también parece guardar una estrecha relación con otro de los hallazgos de la investigación, que no es otro que la ausencia de prácticas normalizadas y dependencia contextual para el desarrollo de sistemas de IA. Como muchos participantes enfatizaron, este tipo de proyectos son emergentes y no están tan formalizados y codificados como otros proyectos de desarrollo de software clásico. Algo que se asociaba a la incertidumbre que acompaña a estos proyectos de IA y sus resultados (Buruk, Ekmekci, and Arda 2020; Jobin et al. 2019; Ryan 2020), a pesar de que también resaltaban otras similitudes existentes.

Las implicaciones éticas y sociales que surgen en el desarrollo de los sistemas IA es otra de las cuestiones que se ponen de relevancia en esta investigación. El desarrollo de sistemas IA está íntimamente asociado a dilemas éticos, legales y sociales, ya que desplazar la capacidad de acción, agencia y autonomía de humanos a máquinas en procesos sociales como el trabajo, implica encontrarse con diversos conflictos de valores y dilemas morales que producen diferentes impactos (Coeckelbergh 2022; Nowotny 2021). Este "trabajo moral" parece quedar fuera de las tareas de los desarrolladores de estos sistemas de IA, ya que normalmente se refieren a este tipo de impactos y/o implicaciones que producen estos sistemas, como una responsabilidad que deben adquirir los clientes o receptores de los sistemas de IA (Rip 2014). Si bien se enfatizó en numerosas ocasiones durante el trabajo de campo que la mayoría de los sistemas de IA necesitan supervisión humana, curiosamente la mayoría de los argumentos de venta que manejaban estas empresas para con sus clientes enfatizaban la lógica de la sustitución humano-máquina como mejora de la competitividad.

De igual manera, las implicaciones medioambientales del desarrollo de la IA, como son su gran consumo energético, recursos hídricos y promoción de la basura electrónica, eran obviadas o tratadas de manera anecdótica por los entrevistados. Un debate que, sin embargo, cobra cada vez más relevancia por el desarrollo de los centros de datos a nivel global, que proveen de la infraestructura necesaria para alojar buena parte de los sistemas de IA más avanzados en la actualidad, tales como ChatGPT o Gemini. El diseño, entrenamiento y desarrollo de estos sistemas de IA, al igual que el desarrollo de estos centros de datos por buena parte de la geografía internacional, plantea numerosas preguntas acerca de la sostenibilidad de estas

tecnologías y de sus numerosos impactos sobre el medio ambiente en forma de consumo energético, basura electrónica, consumo de agua y/o aumento de las emisiones de CO2 a la atmosfera (Bender et al 2021, Crawford 2021).

Por otro lado, la mayoría de participantes en la investigación valoraba la necesidad de regulación de estos sistemas como una fase rutinaria de cara a la aceptación y popularización de la tecnología en la sociedad. Además, varios participantes mostraban síntomas de una clara confianza en la tecnología que se podrían enmarcar como visiones tecno-optimistas y tecno-solucionistas (Broussard 2018; Morozov 2013), al enfatizar que la regulación es algo que sucede al desarrollo de la tecnología, pero no al revés, y que el "miedo" de la sociedad a este tipo de tecnologías como ocurrió con Internet es algo injustificado, ya que una vez pasados estos temores, la sociedad sabrá utilizarlo a su favor. Sin embargo, hay que recordar que a pesar de que Internet y la IA han tenido y pueden tener efectos positivos en la sociedad, sus efectos en materia de desigualdad han sido notables y pueden ser devastadores a futuro.

Un ejemplo de esto es el desarrollo de las plataformas digitales estadounidenses, las cuales han sido capaces de acumular recursos y poder de manera alarmante desde el inicio del S.XXI, redefiniendo sectores productivos enteros, actuando como intermediarios culturales y excluyendo del plano online buena parte de la diversidad cultural, lingüística y social que existe en el planeta. Un fenómeno que corre el riesgo de verse intensificado si no se actúa en contra (Tabarés Gutiérrez, 2025). Algo con lo que los entrevistados se mostraban de acuerdo, ya que asumían que el desarrollo de la IA conllevaría un aumento de la desigualdad, a través de diferentes impactos en el empleo, facilitando la automatización de diferentes tareas, no sólo de carácter más manual y físico, sino también otras de carácter más intelectual, y en línea con diferentes referencias de la literatura anteriormente presentada (Bender et al. 2021; Coeckelbergh 2022; Eubanks 2018)..

Así mismo, los entrevistados también disponían de opiniones diferenciadas en torno a como favorecer el desarrollo y puesta en marcha de una regulación específica para este tipo de tecnologías. Mientras que algunos informantes clave con mayor experiencia entrevían diferentes retos sociales asociados a la regulación, en materia de empleo, derechos de propiedad intelectual o procesos de descualificacación y recualificación asociados, otros perfiles con menos experiencia poseían una opinión más indiferente y destacando el "poder transformador" de la tecnología. Comúnmente, se aludía a como la tecnología también puede crear nuevos empleos y altamente remunerados que puedan suplir la pérdida de otros. En este sentido, la irrupción de ChatGPT por parte de Open AI a finales de 2022, ha supuesto un punto de inflexión para las plataformas digitales, las cuales se han lanzado a movilizar un ingente número de recursos con el objetivo de desarrollar IAs que puedan convertirse en asistentes inteligentes y capaces de generar textos, código, vídeos o imágenes.

El creciente uso y adopción de estas herramientas nos predispone como sociedad a una mayor "necesidad" de dichas tecnologías relacionadas con la creación

de contenidos, pese a que todavía está por ver los efectos que tienen para el aumento de la productividad en diferentes sectores y cuando ya se han observado sus numerosas implicaciones éticas, legales, medioambientales y sociales (Crawford, 2021; Bender et al, 2021; Eubanks, 2018). Una necesidad creada y alimentada por las grandes tecnológicas norteamericanas que se han beneficiado de la gran cantidad y disponibilidad de datos de usuarios de la Web, y por los avances en procesamiento y computación desarrollados gracias a la computación en la nube (Tabarés, 2021).

La IA, quizás, sea la forma más visible de una nueva forma de capitalismo que se abre paso a través de la transformación digital de la sociedad, la economía y el trabajo, donde la recolección, análisis y reutilización de datos en diversos entornos impulsa diferentes innovaciones y ventajas competitivas en un capitalismo digital cada vez más explotador, extractivo e insostenible (Tabarés Gutiérrez, 2024). La sociedad, sin embargo, necesita y reclama mecanismos e instrumentos que la ayuden a a lidiar con las diversas implicaciones que se asocian a la IA. En este sentido la "AI Act" puede ser un primer paso, pero seguramente se necesitarán muchos más instrumentos y no solamente de carácter legal, para poder hacer frente a estas transformaciones de manera satisfactoria. Fomentar enfoques multi e interdisciplinares en estos proyectos de desarrollo tecnológico se antoja como necesario, además del desarrollo de códigos deontológicos, la formación de colegios profesionales, o la necesidad de formar a estos profesionales en cuestiones relativas a los valores democráticos se antojan como medidas complementarias a la regulación.

6. CONCLUSIONES

Los avances que se han desarrollado en torno a la IA en las últimas décadas han despertado grandes incertidumbres alrededor de sus implicaciones éticas y sociales. Por ello, la idea de una IAR ha surgido con fuerza en la comunidad académica y se ha extendido de manera desigual por la industria, la administración pública y la sociedad (Dignum 2017; Stahl and Wright 2018). Este artículo ha tratado de conocer cuál es la difusión y operacionalización de esa idea de IAR, a través de un estudio comparativo que ha involucrado a dos empresas que desarrollan sistemas de IA, observando cómo interpretan la idea de una IAR y la responsabilidad en general, y analizando cuáles son los valores que se ponen en cuestión con el desarrollo de sistemas de IA.

En esta contribución se argumenta que el desarrollo de sistemas de IA lleva asociado unas implicaciones éticas, legales y sociales que no son atendidas debidamente por los desarrolladores de estos sistemas. Si bien hay en marcha diversas iniciativas regulatorias como la "AI Act", esta primera regulación a nivel europeo debería ser acompañada por otra serie de medidas que traten de abordar los diferentes impactos e implicaciones que tienen estos sistemas más allá del área regulatoria. De esta manera, en este texto se ha contribuido a la ingente literatura

sobre los aspectos éticos, legales y sociales de la IA (Broussard 2018; Coeckelbergh 2022; Nowotny 2021), pero también a la incipiente literatura que trata de operacionalizar estos principios de alto nivel a través de la idea de una IAR en el desarrollo tecnológico (Dignum 2017; Stahl 2023).

En particular es muy importante que la sociedad en su conjunto entienda las diferentes implicaciones éticas, legales y sociales asociadas al desarrollo de estas tecnologías por parte de diferentes actores. Por ello, acciones de alfabetización digital de la población en torno a la IA son importantes, pero también mecanismos de rendición de cuentas por parte de las empresas y organizaciones que están detrás del desarrollo de estos sistemas. En este sentido, iniciativas legislativas que puedan combatir las diferentes brechas digitales asociadas al desarrollo de la IA (acceso, conocimiento, implicaciones, etc.), al igual que instrumentos que permitan a las organizaciones que desarrollan estas tecnologías poder ser auditadas e inspeccionadas, adquieren una importancia estratégica desde el punto de vista del paradigma de una IAR (Prieto Vertel y Tabarés Gutiérrez, 2025).

Por último, también es importante señalar las limitaciones de este estudio de caso comparativo, ya que se ha restringido a dos empresas de diferente tipo y que desarrollan sus actividades en contextos delimitados y diversos. El enfoque cualitativo que se ha adoptado es profundo y rico en detalles para comprender el tipo de implicaciones éticas, legales y sociales asociadas al desarrollo de la IA en estas dos organizaciones, pero también presenta limitaciones para su generalización por las características de la muestra.

REFERENCIAS BIBLIOGRÁFICAS

Anagnostou, Marianna, Olga Karvounidou, Chrysovalantou Katritzidaki, Christina Kechagia, Kyriaki Melidou, Eleni Mpeza, Ioannis Konstantinidis, Eleni Kapantai, Christos Berberidis, Ioannis Magnisalis, and Vassilios Peristeras. 2022. "Characteristics and Challenges in the Industries towards Responsible AI: A Systematic Literature Review." *Ethics and Information Technology* 24(3).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Pp. 610–23 in *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc.

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: The MIT Press.

Burget, Mirjam, Emanuele Bardone, and Margus Pedaste. 2017. "Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review." *Science and Engineering Ethics* 23(1):1–19. doi: 10.1007/s11948-016-9782-1.

Buruk, Banu, Perihan Elif Ekmekci, and Berna Arda. 2020. "A Critical Perspective on Guidelines for Responsible and Trustworthy Artificial Intelligence." *Medicine, Health Care and Philosophy* 23(3):387–99. doi: 10.1007/s11019-020-09948-1.

Coeckelbergh, Mark. 2022. *The Political Philosophy of AI. An Introduction*. Cambridge: Polity Press.

Crawford, Kate. 2021. *Atlas of AI. Power, Politics and the Planetary Costs of Artificial Intelligence*. New Haven and London: Yale University Press.

Csiszar, Akos, Philipp Hein, Michael Wachter, Alexander Verl, and Angelika C. Bullinger. 2020. "Towards a User-Centered Development Process of Machine Learning Applications for Manufacturing Domain Experts." Pp. 36–39 in *Proceedings - 2020 3rd International Conference on Artificial Intelligence for Industries, AI4I 2020*. Institute of Electrical and Electronics Engineers Inc.

Dauvergne, Peter. 2020. *AI in the Wild. Sustainability in the Age of Artificial Intelligence.* Cambridge, MA: The MIT Press.

Dignum, Virginia. 2017. "Responsible Artificial Intelligence: Designing AI for Human Values." *ICT Discoveries* 25(1):1–8.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Floridi, Luciano, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. 2022. *CapAI A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act*.

High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworhty AI*. Brussels.

Jammalamadaka, Krishna Ravali, and Srikanth Itapu. 2022. "Responsible AI in Automated Credit Scoring Systems." *AI and Ethics*. doi: 10.1007/s43681-022-00175-3.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "Artificial Intelligence: The Global Landscape of Ethics Guidelines." *Nature Machine Intelligence* 389–99. doi: https://doi.org/10.1038/s42256-019-0088-2.

Knott, Eleanor, Aliya Hamid Rao, Kate Summers, and Chana Teeger. 2022. "Interviews in the Social Sciences." *Nature Reviews Methods Primers* 2(1). doi: 10.1038/s43586-022-00150-6.

Kordzadeh, Nima, and Maryam Ghasemaghaei. 2022. "Algorithmic Bias: Review, Synthesis, and Future Research Directions." *European Journal of Information Systems* 31(3):388–409.

Laureau, Annette. 2021. *Listening to People. A Practical Guide to Interviewing, Participant Observation, Data Analysis, and Writing It All Up*. Chicago: The University of Chicago Press.

Minkkinen, Matti, Markus Philipp Zimmer, and Matti Mäntymäki. 2023. "Co-Shaping an Ecosystem for Responsible AI: Five Types of Expectation Work in Response to a Technological Frame." *Information Systems Frontiers* 25(1):103–21. doi: 10.1007/s10796-022-10269-2.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data and Society* 3(2). doi: 10.1177/2053951716679679.

Morozov, Eugeny. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.

Nakao, Yuri, Lorenzo Strappelli, Simone Stumpf, Aisha Naseer, Daniele Regoli, and Giulia Del Gamba. 2022. "Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness." *International Journal of Human–Computer Interaction* 1–27. doi: https://doi.org/10.1080/10447318.2022.2067936.

Nowotny, Helga. 2021. *In AI We Trust. Power, Illusion and Control of Predictive Algorithms*. Cambridge: Polity Press.

Owen, Richard, and Mario Pansera. 2019. "Responsible Innovation and Responsible Research and Innovation." Pp. 26–48 in *Handbook on Science and Public Policy*, edited by D. Simon, S. Kuhlmann, J. Stamm, and W. Canzle. Edward Elgar publishing.

Prieto Viertel G. y Tabarés Gutiérrez R. (2025). El reto de la moderación de contenidos en inteligencia artificial generativa: ChatGPT bajo el marco regulatorio de la Unión Europea. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales, Avance en línea*, 1-14. https://doi.org/10.5209/tekn.97799

van de Poel, Ibo, Lotte Asveld, Steven Flipse, Pim Klaassen, Zenlin Kwee, Maria Maia, Elvio Mantovani, Christopher Nathan, Andrea Porcari, and Emad Yaghmaei. 2020. "Learning to Do Responsible Innovation in Industry: Six Lessons." *Journal of Responsible Innovation* 1–11. doi: 10.1080/23299460.2020.1791506.

Rip, Arie. 2014. "The Past and Future of RRI." *Life Sciences, Society and Policy* 10(17):1–15. doi: 10.1902/jop.1939.10.1.31.

Ryan, Gery W., and H. Russell Bernard. 2003. "Techniques to Identify Themes." *Field Methods* 15(1):85–109. doi: 10.1177/1525822X02239569.

Ryan, Mark. 2020. “In AI We Trust: Ethics, Artificial Intelligence, and Reliability.” *Science and Engineering Ethics* 26(5):2749–67. doi: 10.1007/s11948-020-00228-y.

Shin, Donghee, and Yong Jin Park. 2019. “Role of Fairness, Accountability, and Transparency in Algorithmic Affordance.” *Computers in Human Behavior* 98:277–84. doi: 10.1016/j.chb.2019.04.019.

Shneiderman, Ben. 2021. “Responsible AI: Bridging from Ethics to Practice.” *Communications of the ACM* 64(8):32–35.

Stahl, Bernd Carsten. 2023. “Embedding Responsibility in Intelligent Systems: From AI Ethics to Responsible AI Ecosystems.” *Scientific Reports* 13(1). doi: 10.1038/s41598-023-34622-w.

Stahl, Bernd Carsten, and David Wright. 2018. “Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation.” *IEEE Security and Privacy* 16(3):26–33. doi: 10.1109/MSP.2018.2701164.

Stilgoe, Jack, Richard Owen, and Phil Macnaghten. 2013. “Developing a Framework for Responsible Innovation.” *Research Policy* 42(9):1568–80. doi: http://dx.doi.org/10.1016/j.respol.2013.05.008.

Tabarés Gutiérrez, R. (2025). Interfaces conversacionales, tecnolenguajes y tecnodesigualdades. *RECERCA. Revista De Pensament I Anàlisi*, *30*(1). https://doi.org/10.6035/recerca.8402

Tabarés Gutiérrez, R. (2024). Plataformización, automatización y aceleración en los medios sociales. *Daimon Revista Internacional de Filosofia*, (93), 137–152. https://doi.org/10.6018/daimon.612051

Tabarés, R., Loeber, A., Nieminen, M., Bernstein, M. J., Griessler, E., Blok, V., … Frankus, E. (2022). Challenges in the implementation of responsible research and innovation across Horizon 2020. *Journal of Responsible Innovation*, *9*(3), 291–314. https://doi.org/10.1080/23299460.2022.2101211

Tambama, Madiega. 2022. *Artificial Intelligence Act*. European Parliament.

Zhu, Liming, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. 2021. “AI and Ethics - Operationalising Responsible AI.” in *Humanity Driven AI*, edited by F. Chen and J. Zhou. Cham: Springer.