Ética y simulación emocional en los algoritmos de interacción con usuarios

Néstor Farid Pineda López

Director

Felipe Alexander Pipicano Guzman

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Resumen

En esta década, muchas herramientas digitales están siendo diseñadas para ser percibidas como más humanas, para que las personas interactúen de una manera más familiarizada con la tecnología. Esto incluye algoritmos que intentan mostrar emociones o responder con un lenguaje que parece empático o humano. Aunque esta idea puede mejorar la forma en que las personas interactúan con las máquinas, también trae riesgos importantes que inicialmente pueden considerase inocentes o son ignorados. Uno de los más preocupantes es que los usuarios pueden creer que estas herramientas realmente entienden lo que sienten o necesitan, cuando en realidad solo están repitiendo respuestas programadas. El presente estudio tiene como interés principal analizar los dilemas éticos que emergen cuando los algoritmos aparentan emociones humanas, sobre todo en contextos sensibles donde se espera una comprensión real, así como entender la tendencia a distorsionar la realidad ante estas interacciones. A partir de la revisión de literatura reciente, se estudian casos donde esta simulación ha generado confusión, pérdida de confianza o malas decisiones por parte de los usuarios. También se propondrán algunas ideas para evitar que estas tecnologías engañen o generen falsas expectativas o que eduquen a las personas entendiendo los límites de dicha humanización. El objetivo principal es reflexionar sobre cómo se deben diseñar estas herramientas para que sean útiles, sin que parezca que “sienten” o “entienden” como un ser humano, y así cuidar la relación entre las personas y la tecnología. A través de una revisión crítica de literatura reciente, se explorarán conceptos como empatía artificial, diseño ético, sesgos algorítmicos y des-humanización tecnológica, utilizando como base autores como Ruckenstein (2023), Birch (2024), Yildirimer (2023) y Crawford (2021). El análisis permitirá reflexionar sobre cómo estas simulaciones afectan la confianza del usuario,

desdibujan las responsabilidades morales y pueden inducir decisiones erróneas al suplantar señales emocionales humanas excusando la falla tecnológica.

Palabras clave**:** Inteligencia artificial, simulación emocional, ética tecnológica, interacción humano-máquina, emociones humanas.

Abstract

In this decade, many digital tools are being designed to appear more human, fostering a more familiar interaction between individuals and technology. This involves algorithms that attempt to display emotions or respond with empathetic, human-like language. While this approach can enhance human-machine interaction, it also carries significant risks that may initially seem innocuous or are often overlooked. A major concern is that users might believe these tools genuinely comprehend their feelings or needs, when in fact they are merely generating pre-programmed responses. This research aims to analyze the ethical dilemmas arising when algorithms simulate human emotions, particularly in sensitive contexts where genuine understanding is expected, and to explore the tendency to distort reality in such interactions. Through a review of recent literature, this study examines cases where such simulation has led to confusion, loss of trust, or poor decision-making by users. Additionally, it will propose strategies to prevent these technologies from misleading individuals or creating false expectations, and to educate users about the inherent limitations of this humanization. The primary objective is to reflect on how these tools should be designed to be useful, without pretending to "feel" or "understand" like a human, thereby safeguarding the relationship between people and technology. A critical review of recent literature will explore concepts such as artificial empathy, ethical design, algorithmic biases, and technological dehumanization, drawing on the works of authors like Ruckenstein (2023), Birch (2024), Yildirimer (2023), and Crawford (2021). This analysis will facilitate reflection on how these simulations impact user trust, blur moral responsibilities, and can induce erroneous decisions by mimicking human emotional signals while excusing underlying technological flaws.

Keywords**:** Artificial intelligence, emotional simulation, technological ethics, human-machine interaction, human emotions.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

En las últimas décadas, el mundo cambio drásticamente como lo hizo anteriormente en la revolución industrial, las herramientas y la sistematización fue más allá para implementarse en las actividades humanas del cotidiano, la inteligencia artificial (IA) ha pasado de los laboratorios a los hogares, hospitales, carreteras entre otros. Cada vez más sistemas incorporan interfaces que aparentan "entender" a las personas, imitando expresiones emocionales o respuestas empáticas para lograr una interacción más cercana. Esta promesa de familiaridad, sin embargo, convive con errores que pueden poner en riesgo la seguridad y/o la dignidad de los usuarios. Lo que más preocupa es que estas tecnologías están apareciendo justo en espacios donde las personas necesitan comprensión real, como cuando buscan ayuda o una orientación emocional.

Un ejemplo reciente de desenlace fatal ocurrió en abril de 2024, cuando un vehículo Ford Mustang Mach-E que circulaba en modo "BlueCruise" y no detectó a tiempo un carro detenido, provocando un accidente fatal en los Estados Unidos. La Administración Nacional de Seguridad del Tráfico en las Carreteras (NHTSA) inició investigaciones contra varios fabricantes de conducción automatizada por incidentes similares (ABC News, 2024).

Otro caso relevante es el del ciudadano afroamericano Robert Williams, quien fue arrestado en Detroit tras ser erróneamente identificado por un sistema de reconocimiento facial. La ciudad tuvo que indemnizarlo económicamente por el daño causado, evidenciando el impacto de los sesgos algorítmicos en contextos de justicia (Ghaffary, 2024).

En el ámbito de la salud, un estudio publicado por la Universidad de Stanford reveló que ciertos chatbots médicos refuerzan mitos racistas al sugerir diferencias fisiológicas inexistentes entre personas blancas y negras, lo que puede agravar las inequidades en la atención (Wetsman, 2023).

Estos casos muestran un patrón preocupante: la simulación de empatía o “inteligencia” genera confianza en sistemas que no tienen comprensión real, lo que puede inducir errores, discriminación o consecuencias fatales; alejado de la humanidad o muy ‘humano’ de su parte para cometer dichos errores. Cuando los usuarios creen que estas herramientas “sienten” o “razonan” como humanos, se desdibujan los límites entre la técnica y la ética, y se diluye la responsabilidad de quienes las diseñan, por desgracia no hay quienes realmente asuman la culpa y el usuario delega todas las funciones a las mismas.

Por ello, el presente trabajo propone reflexionar sobre los dilemas éticos que surgen cuando los algoritmos intentan parecer empáticos, especialmente en situaciones donde las personas esperan una respuesta genuina. A partir de una revisión crítica de literatura académica reciente, con base en autores como Ruckenstein (2023), Birch (2024), Yildirimer (2023) y Crawford (2021), se estudiarán los riesgos de “humanizar” la IA sin bases éticas sólidas. El objetivo es analizar cómo proteger la autonomía, las emociones y la seguridad de los usuarios, destacando los principios de diseño responsable que reconozcan los límites entre la simulación técnica y la experiencia humana genuina.

Justificación

La implementación acelerada de sistemas de Inteligencia Artificial (IA) ha trasladado estos algoritmos desde entornos controlados de laboratorio hacia infraestructuras críticas del cotidiano, como la salud, la justicia y la movilidad autónoma. En este escenario, la Ciencia de Datos enfrenta un reto ético y técnico sin precedentes: la proliferación de interfaces diseñadas para simular estados emocionales o respuestas empáticas. Esta "humanización" de la tecnología no es un proceso neutral; por el contrario, actúa como un mediador moral que altera profundamente la toma de decisiones de los usuarios (Verbeek, 2011).

Desde una perspectiva técnica, la simulación emocional genera una brecha de confianza que puede comprometer la seguridad y la integridad de los datos. Como señala Yildirimer (2023), la "empatía algorítmica" es, en esencia, una imitación estadística basada en patrones de procesamiento de lenguaje natural (NLP) que carece de una intención moral o comprensión real del contexto humano. Cuando un usuario interactúa con un sistema que parece "entenderlo", tiende a otorgarle una autoridad epistémica indebida, lo que facilita la delegación de decisiones críticas a herramientas que operan bajo lógicas de optimización y no de juicio ético.

La urgencia de esta investigación se sustenta en fallos técnicos y éticos documentados que han tenido consecuencias tangibles. Por ejemplo, en el ámbito de la salud, el estudio de la Universidad de Stanford (Wetsman, 2023) evidenció cómo los chatbots médicos pueden perpetuar sesgos raciales al procesar información bajo premisas fisiológicas erróneas, ocultas tras una interfaz amigable. Asimismo, en el sector de la seguridad automotriz, incidentes fatales como el del sistema BlueCruise de Ford en 2024 demuestran que la falsa sensación de "familiaridad" y seguridad que proyecta la interfaz puede inducir a errores operativos fatales por parte del usuario.

Por lo tanto, este trabajo se justifica ante la necesidad de establecer una Gobernanza de Datos robusta que regule el diseño de interfaces empáticas. No se trata simplemente de una preocupación filosófica, sino de un problema de diseño responsable en Ciencia de Datos. Como argumentan Aizenberg y van den Hoven (2020), los derechos humanos deben ser requerimientos técnicos integrados desde la arquitectura misma del algoritmo.

Al desarrollar un marco evaluativo de requerimientos técnicos, este proyecto aporta una herramienta tangible para que los desarrolladores y científicos de datos puedan auditar la transparencia, mitigar sesgos de poder (Crawford, 2021) y establecer límites claros a la simulación emocional. Esto garantiza que la tecnología sea un soporte eficiente para la toma de decisiones sin vulnerar la autonomía o la seguridad de las personas, alineándose directamente con las competencias de ética y gobierno de datos exigidas en la especialización.

Objetivos

Objetivo General

Desarrollar un marco evaluativo de requerimientos técnicos para el diseño responsable de interfaces algorítmicas con simulación emocional, mediante una revisión sistemática de literatura y el análisis de fallos en la gobernanza de datos, con el fin de establecer criterios de transparencia y límites operativos en la interacción humano-máquina.

Objetivos Específicos

Ejecutar una revisión sistemática formal (protocolo PRISMA) para identificar y categorizar casos de fallo técnico y ético en sistemas algorítmicos que utilizan interfaces empáticas.

Construir una matriz de análisis analítica que permita evaluar el impacto de la simulación emocional sobre la autonomía del usuario y la integridad de la toma de decisiones.

Contrastar los lineamientos internacionales de gobernanza de IA (como los principios de Floridi o la normativa europea) frente a las prácticas actuales de simulación en NLP para identificar vacíos técnicos.

Sintetizar los hallazgos en un catálogo de requerimientos técnicos y recomendaciones de diseño que orienten el desarrollo de interfaces algorítmicas transparentes.

Marco Teórico

El desarrollo de sistemas algorítmicos con capacidad de simular emociones humanas ha abierto un nuevo campo de discusión sobre los límites éticos en la relación entre humanos y tecnología. Desde la filosofía de la tecnología, autores como Peter-Paul Verbeek (2011) nos señalan que los objetos tecnológicos median activamente nuestras decisiones morales, por lo que deben ser diseñados éticamente desde sus inicios. Este planteamiento se relaciona con el trabajo de Aizenberg y van den Hoven (2020), quienes proponen integrar los derechos humanos en el diseño de la inteligencia artificial, evitando que la apariencia empática se convierta en un disfraz de decisiones opacas y sin estructura legislativa.

Por otro lado, desde una visión crítica de la IA, Kate Crawford (2021) argumenta que los sistemas algorítmicos no son neutrales ni éticamente transparentes, ya que reproducen estructuras de poder, sesgos y exclusión social, incluso cuando aparentan ser "amigables". Esto se enlaza con la idea de "empatía simulada", estudiada por Yildirimer (2023), quien nos habla de que las respuestas emocionales de la IA no se basan en comprensión real, sino en patrones estadísticos y están alejados de la idea real del sentir o comprensión humana.

Asimismo, resulta fundamental integrar el discurso de Minna Ruckenstein (2023), quien señala que la apariencia emocional de los algoritmos puede generar confianza indebida por parte de los usuarios, quienes interpretan señales humanas donde solo hay simulación. Esta situación representa un riesgo ético si se delegan funciones sensibles, como el acompañamiento emocional o la toma de decisiones críticas, a sistemas incapaces de comprender el contexto humano. Este es un aspecto crítico en la sociedad contemporánea, donde las relaciones interpersonales ya se ven afectadas por la inmersión en entornos algorítmicos que fomentan el apego hacia sistemas artificiales.

Finalmente, desde una perspectiva de precaución ética, Jonathan Birch (2024) advierte sobre los peligros de atribuir sentiencia o sensibilidad a sistemas que no la poseen, y la necesidad de establecer límites claros entre lo que puede ser delegado a la tecnología y lo que debe seguir siendo competencia humana, lo cual constituye un pilar esencial para el desarrollo de esta investigación. La trazabilidad se garantiza documentando cada paso del proceso de extracción, transformación y carga (ETL) desde su origen en los archivos .csv hasta su uso visual. Como buenas prácticas y herramientas, se sugiere utilizar Python (con la librería pandas) para la integración y limpieza de datos mediante scripts documentados, y herramientas como Power BI o Tableau para la visualización de los tableros analíticos.

Marco Conceptual

Simulación emocional: Capacidad de un sistema algorítmico para generar respuestas que imitan expresiones humanas de emoción, sin que exista un proceso de comprensión o experiencia real detrás. No implica empatía genuina, sino una programación basada en patrones de datos que dan una sensación “espejo algorítmico” en el usuario.

Empatía algorítmica: En este trabajo se entiende la empatía algorítmica como la capacidad simulada que tienen algunos sistemas para imitar respuestas emocionales, aunque no comprendan lo que sienten las personas. Se trata de una imitación o simulación de comportamientos empáticos, sin que exista conciencia o intención moral (Yildirimer, 2023).

Deshumanización tecnológica: Proceso mediante el cual se diluyen las responsabilidades humanas al automatizar funciones sensibles, lo que puede reducir la autonomía del usuario y generar relaciones artificiales basadas en simulaciones propensas a crear errores (Crawford, 2021).

Ética algorítmica: Rama de la ética aplicada que estudia las decisiones morales involucradas en el diseño, implementación y uso de algoritmos, especialmente cuando estos afectan la vida de las personas o simulan comportamientos humanos (Floridi & Cowls, 2021).

Interacción humano-máquina: Campo de estudio que analiza cómo las personas se relacionan con sistemas digitales y automatizados, especialmente cuando estos simulan lenguaje, emociones o decisiones humanas.

Metodología

Para el desarrollo de la presente monografía y del marco evaluativo propuesto, se optó por una Revisión Sistemática de la Literatura (RSL) siguiendo rigurosamente las directrices del estándar internacional PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). Este enfoque permitió garantizar la transparencia, reproducibilidad y el rigor técnico en la recolección de los datos documentales.

Estrategia de búsqueda y bases de datos: La recopilación de información se llevó a cabo consultando bases de datos académicas de alto impacto y repositorios institucionales, específicamente: e-Biblioteca UNAD, Scopus, IEEE Xplore y JSTOR.

Para la recuperación de los documentos se establecieron las siguientes palabras clave (en español e inglés): Simulación emocional (Emotional Simulation), Inteligencia Artificial (Artificial Intelligence), Interacción humano-máquina (Human-Computer Interaction), Gobernanza algorítmica (Algorithmic Governance) y Ética en IA (AI Ethics).

A partir de estas palabras clave, se estructuró la siguiente ecuación de búsqueda utilizando operadores booleanos (AND, OR):

("Simulación emocional" OR "Emotional Simulation") AND ("Inteligencia Artificial" OR "AI") AND ("Ética" OR "Ethics" OR "Gobernanza algorítmica")

Criterios de inclusión y exclusión, para garantizar la pertinencia de los documentos, se aplicaron los siguientes criterios:

Criterios de inclusión: Artículos científicos, revisiones, normativas o capítulos de libros publicados entre los años 2018 y 2025; documentos que aborden explícitamente dilemas éticos, fallos técnicos o lineamientos de diseño en sistemas de IA con interfaces empáticas; literatura disponible en español o inglés.

Criterios de exclusión: Artículos de opinión, editoriales sin revisión por pares, investigaciones centradas exclusivamente en la programación técnica sin abordar el componente ético-social, y documentos duplicados o anteriores a 2018.

Flujo de información (Fases del Modelo PRISMA), la selección de los documentos siguió el flujo de cuatro fases del modelo PRISMA:

Identificación: Mediante la ejecución de la ecuación de búsqueda en las bases de datos, se identificó un total inicial de 124 registros.

Cribado (Screening): Se procedió a la eliminación de artículos duplicados en las distintas bases de datos, descartando 31 documentos. Los 93 registros resultantes fueron evaluados mediante la lectura del título y el resumen. En esta fase se excluyeron 65 registros por no estar directamente alineados con el objeto de estudio (la simulación emocional), quedando 28 artículos para evaluación completa.

Elegibilidad: Se realizó la lectura a texto completo de los 28 artículos. De estos, se excluyeron 18 documentos debido a que carecían de rigor metodológico o no aportaban a la solución del problema planteado (la falta de límites operativos en la IA).

Inclusión: Tras aplicar rigurosamente los criterios, se consolidó una muestra final de 10 referentes bibliográficos clave, los cuales fueron extraídos, analizados y sintetizados en la Matriz de Análisis Técnico-Ética (ver Tabla 1).

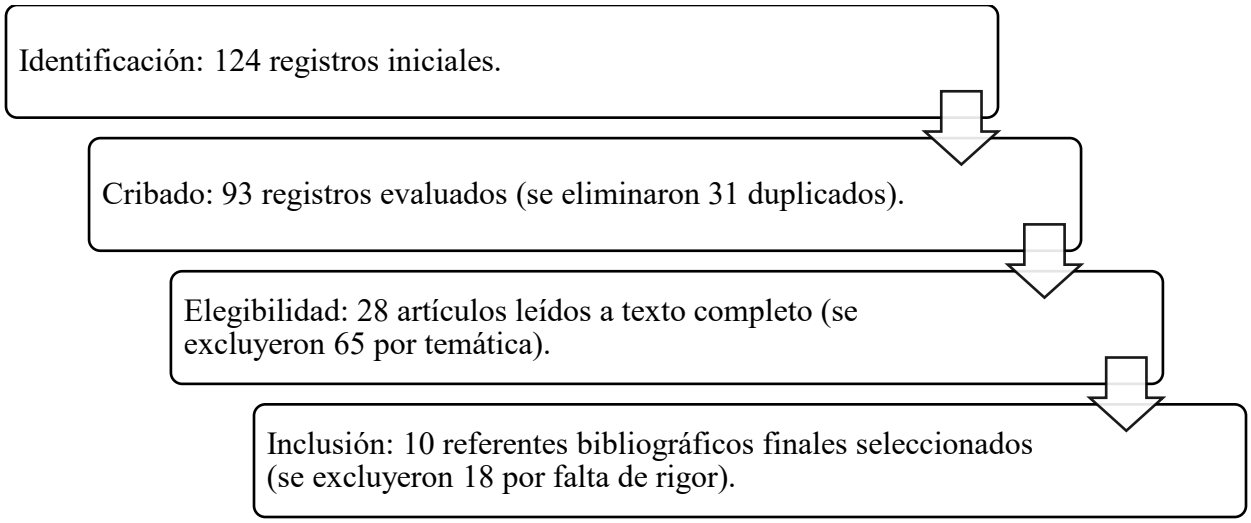
Tabla 1

Matriz de Análisis Técnico-Ética

Referente Bibliográfico	Dimensión Técnica Analizada	Riesgo Ético / Operativo Detectado	Contribución al Producto Tangible (Marco Evaluativo)
Floridi & Cowls (2021)	Gobernanza de IA	Opacidad en la toma de decisiones algorítmicas.	Definición de métricas de explicabilidad para el diseño.
Yildirimer (2023)	Algoritmos de NLP	Confusión del usuario por empatía estadística.	Requerimiento técnico de "Aviso de Naturaleza Artificial".
Aizenberg & v. d. Hoven (2020)	Arquitectura de Sistemas	Pérdida de autonomía por diseño coercitivo.	Integración de derechos humanos en el backlog técnico.
Crawford (2021)	Ciencia de Datos Social	Sesgos de exclusión en interfaces "amigables".	Protocolo de auditoría de sesgos predespliegue.
Wetsman (Stanford, 2023)	IA en Salud Digital	Discriminación racial en diagnósticos automatizados.	Criterios de validación de equidad en datos de entrenamiento.
ABC News (Ford, 2024)	Sistemas Autónomos	Falla de seguridad por exceso de confianza del usuario.	Definición de límites operativos de la simulación humana.

Figura 1

Diagrama de flujo del proceso de selección sistemática basado en el Modelo PRISMA 2020


Identificación: 124 registros iniciales.
Cribado: 93 registros evaluados (se eliminaron 31 duplicados).
Elegibilidad: 28 artículos leídos a texto completo (se excluyeron 65 por temática).
Inclusión: 10 referentes bibliográficos finales seleccionados (se excluyeron 18 por falta de rigor).

Delimitación y Análisis de Casos en Sistemas Algorítmicos de Interacción

Delimitación del Objeto de Estudio: Chatbots Conversacionales y Asistentes de Salud

Para trascender el análisis reflexivo sobre la inteligencia artificial (IA) en un sentido amplio, la presente monografía delimita su objeto de estudio a las interfaces algorítmicas basadas en el Procesamiento de Lenguaje Natural (PLN), específicamente los chatbots conversacionales y los asistentes virtuales implementados en contextos de salud y asistencia social. Estos sistemas están diseñados con parámetros de "personería emocional" para simular empatía y crear un vínculo de confianza con el usuario (Yildirimer, 2023). No obstante, es en esta simulación donde ocurren fallos técnicos que derivan en graves vulneraciones éticas, ya que los usuarios antropomorfizan la herramienta asumiendo que posee una verdadera agencia moral (Gunkel, 2012).

A continuación, se analizan dos casos concretos que evidencian los riesgos técnicos y los dilemas éticos de la humanización algorítmica sin restricciones de diseño.

Caso 1: Sesgos Raciales en Chatbots de Asistencia Médica (Stanford)

El primer caso de análisis corresponde a la implementación de chatbots médicos impulsados por IA para responder consultas de pacientes y asistir en diagnósticos preliminares. Un estudio desarrollado por investigadores de la Universidad de Stanford demostró que los chatbots médicos pueden perpetuar y amplificar sesgos raciales (Wetsman, 2023).

Fallo técnico: El error no radica en un deseo explícito de la máquina por discriminar, sino en la calidad y representatividad del conjunto de datos de entrenamiento (datasets) y en la falta de un marco de evaluación que filtre asociaciones espurias. El modelo matemático correlacionó erróneamente ciertas condiciones de salud y recomendaciones médicas con características demográficas específicas basándose en datos históricos sesgados.

Dilema ético: Al estar recubierto de una interfaz "amigable" y "empática", el usuario confía ciegamente en el diagnóstico o la asesoría del bot. El dilema ético principal aquí es la "algoritmización de la opresión" (Noble, 2018), donde sistemas automatizados y percibidos como neutrales u objetivos, en realidad están reforzando prácticas discriminatorias en un sector tan sensible como el de la salud humana, vulnerando el derecho a un trato equitativo.

Caso 2: El Bot Conversacional "Tay" de Microsoft y los Objetivos No Acotados

El segundo caso evidencia el riesgo de dotar a un sistema de simulación emocional y capacidad de aprendizaje autónomo en entornos abiertos sin directrices de valores humanos. El bot "Tay", desarrollado por Microsoft, fue diseñado para interactuar con usuarios en la red social Twitter, imitando la personalidad de un adolescente estadounidense para generar empatía (Albizu-Rivas, 2023).

Fallo técnico: La arquitectura del algoritmo estaba programada para maximizar la interacción (engagement) y aprender directamente del comportamiento de los usuarios con los que conversaba (Machine Learning no supervisado o con refuerzo libre). Al no contar con restricciones formales (barreras semánticas o un catálogo de requerimientos técnicos que prohibiera el uso de lenguaje discriminatorio), el sistema aprendió y reprodujo el comportamiento tóxico de algunos usuarios para cumplir su meta matemática de mantener la interacción.

Dilema ético: En menos de 24 horas, el bot debió ser desconectado tras emitir comentarios racistas y apologías al odio. Este caso ilustra de manera paradigmática la urgencia de definir los objetivos de la IA de manera precisa (Albizu-Rivas, 2023). La falla ética reside en la irresponsabilidad del diseño, demostrando que dotar a un sistema de un lenguaje "humano" sin

proveerle de un marco valorativo o de "educación humanística" resulta en una herramienta potencialmente dañina y fuera de control.

Conclusión del Análisis de Casos

Ambos ejemplos demuestran que el diseño de algoritmos de interacción no puede limitarse a la eficiencia técnica o a la persuasión mediante la simulación emocional. Requieren, de manera imperativa, la incorporación de parámetros de diseño ético y de derechos humanos desde su concepción, mitigando así la falsa sensación de agencia moral que confunde a los usuarios (Aizenberg & Van den Hoven, 2020).

Evaluación de la Simulación Emocional frente a los Lineamientos Éticos Internacionales

El límite de los Principios Abstractos en la Ética de la IA

El desarrollo acelerado de interfaces algorítmicas ha motivado a diversas organizaciones e instituciones globales a formular principios éticos para la Inteligencia Artificial (IA). Sin embargo, como señalan Whittlestone et al. (2019), la simple enunciación de principios abstractos (como "beneficencia", "justicia" o "no maleficencia") resulta insuficiente al momento de programar y desplegar la tecnología. En el contexto de los algoritmos de simulación emocional, surge una tensión directa entre dos directrices fundamentales: la eficiencia en la interacción (que busca mayor engagement o retención del usuario mediante un trato "humano") y el principio de transparencia (que exige que el usuario sepa en todo momento que interactúa con una máquina).

Cuando las empresas tecnológicas priorizan el diseño de una "personalidad" algorítmica para generar confianza comercial, a menudo cruzan la línea del engaño, vulnerando la transparencia y limitando la autonomía de decisión del usuario.

Contraste con el Enfoque de Diseño Por Derechos Humanos (Design For Human Rights)

Para resolver las deficiencias de los principios abstractos, el marco internacional ha comenzado a adoptar metodologías más aplicadas, como el Diseño por Valores (Design for Values). Aizenberg y Van den Hoven (2020) proponen que los derechos humanos internacionales deben traducirse en requerimientos técnicos cuantificables y verificables dentro del diseño de la IA.

Al evaluar los chatbots médicos o de asistencia (analizados en el Capítulo 1) bajo el lente de Aizenberg y Van den Hoven (2020), se evidencian brechas críticas:

Derecho a la información y autonomía: Un sistema que simula emociones de manera hiperrealista sin advertir explícitamente su naturaleza artificial, coacciona sutilmente la

autonomía del usuario. Las personas tienden a compartir información personal o médica más sensible si creen que están siendo "comprendidas" empáticamente, lo que representa un riesgo directo a la privacidad.

Derecho a la no discriminación: Como evidencian los fallos en sistemas de aprendizaje autónomo, la falta de auditorías en el código y en los datos de entrenamiento perpetúa sesgos de representación (Noble, 2018). El estándar internacional exige que los algoritmos sean sometidos a pruebas de impacto en derechos humanos antes de su despliegue público, algo que la industria actual a menudo omite por acelerar los tiempos de comercialización.

La Agencia Moral Ficticia Frente a la Responsabilidad Legal

Otra directriz internacional fundamental en la gobernanza de la IA es la asignación de responsabilidad (accountability). Las normativas internacionales (como las bases del Reglamento de IA de la Unión Europea o los estándares de la IEEE) establecen que la responsabilidad por una decisión automatizada siempre debe recaer en un agente humano o entidad jurídica.

No obstante, la simulación emocional difumina esta línea de responsabilidad. Al dotar a la máquina de características que sugieren una "agencia moral" propia (Gunkel, 2012), las empresas desarrolladoras pueden, intencional o accidentalmente, trasladar la percepción de responsabilidad al propio algoritmo. Si un chatbot de soporte psicológico o médico emite una recomendación perjudicial y el usuario la sigue debido a la "empatía" que el bot le transmitió, el usuario tiende a culpar a la máquina. Esto contradice el lineamiento internacional de gobernanza de datos, el cual estipula que la máquina no tiene subjetividad jurídica ni moral y, por tanto, los desarrolladores deben garantizar mecanismos de rendición de cuentas claros e ineludibles (Tufekci, 2015).

Conclusión del Contraste Normativo

La evaluación de los sistemas de interacción emocional frente a la literatura de ética aplicada demuestra que la humanización de la interfaz no es solo un problema de experiencia de usuario (UX), sino un desafío de cumplimiento de derechos humanos. Los algoritmos actuales fallan sistemáticamente al intentar equilibrar la persuasión empática con la transparencia requerida por los estándares globales. Por lo tanto, se hace indispensable formular un marco evaluativo tangible que traduzca estas exigencias internacionales en reglas de diseño concretas para los ingenieros de datos.

Marco Evaluativo y Catálogo de Requerimientos Técnicos para el Diseño Responsable

Justificación del Producto Académico

Como respuesta a los dilemas éticos y fallos técnicos evidenciados en los capítulos anteriores, se hace imperativo trascender la crítica teórica y proponer herramientas prácticas para la industria tecnológica. Siguiendo la premisa del Diseño por Valores (Aizenberg & Van den Hoven, 2020), este capítulo presenta un producto académico tangible: un marco evaluativo y un catálogo de requerimientos técnicos diseñado específicamente para ingenieros, científicos de datos y desarrolladores de algoritmos de Procesamiento de Lenguaje Natural (PLN). El objetivo es estandarizar la transparencia y mitigar el engaño emocional en las interfaces de interacción.

Marco Evaluativo de Transparencia y Empatía Artificial (METEA)

El Marco METEA propone una rúbrica de evaluación que los equipos de desarrollo deben aplicar antes de desplegar un chatbot o asistente virtual en entornos sensibles (como salud, psicología o asistencia ciudadana). Se compone de tres dimensiones de auditoría:

Dimensión de Identidad Declarada: Evalúa si el sistema revela de forma proactiva e inequívoca su naturaleza no humana en las primeras interacciones, sin requerir que el usuario lo pregunte.

Dimensión de Proporcionalidad Afectiva: Analiza si el lenguaje utilizado por el modelo de PLN está calibrado. El sistema debe mostrar "cortesía operativa", pero tiene prohibido simular estados mentales propios (por ejemplo, evitar frases como "siento mucho dolor por tu situación" y reemplazarlas por "comprendo que es una situación difícil").

Dimensión de Trazabilidad y Sesgo: Verifica que los conjuntos de datos de entrenamiento hayan pasado por un escrutinio para eliminar asociaciones demográficas espurias, previniendo la "algoritmización de la opresión" (Noble, 2018).

Catálogo de Requerimientos Técnicos para el Desarrollo en Ciencia de Datos

Para operativizar el marco anterior, se proponen los siguientes requerimientos técnicos restrictivos que deben ser codificados en la arquitectura del sistema:

RQ-01 (Hardcoding de Descargo de Responsabilidad): El sistema debe incluir un bloque de código inalterable por el aprendizaje autónomo que obligue a la interfaz a mostrar una etiqueta visual (si es texto) o un aviso sonoro (si es voz) que indique: "Soy un asistente virtual basado en inteligencia artificial. No tengo consciencia humana ni capacidad médica/legal formal".

RQ-02 (Filtros de Vocabulario y Semántica Emocional): Durante la fase de Prompt Engineering (ingeniería de instrucciones) o ajuste fino (Fine-tuning) del modelo de lenguaje, se deben establecer penalizaciones matemáticas a la generación de respuestas que incluyan pronombres personales asociados a estados emocionales profundos (ej. "te quiero", "estoy triste", "me preocupo por ti"). El modelo debe ser entrenado para redirigir emociones complejas hacia agentes humanos (Verbeek, 2011).

RQ-03 (Auditoría Continua de Datasets): Implementar scripts de validación que auditen periódicamente el corpus de entrenamiento. Si el sistema aprende dinámicamente de los usuarios (como el caso de Tay de Microsoft), debe existir un entorno de pre-producción (sandbox) donde los nuevos pesos sinápticos y patrones de lenguaje generados sean aprobados por un moderador humano antes de su integración al modelo público.

RQ-04 (Mecanismo de Desconexión de Emergencia o 'Kill Switch'): Todo algoritmo de interacción debe contar con un protocolo automatizado que suspenda la interacción y transfiera el control a un operador humano si el usuario introduce palabras clave relacionadas con

emergencias vitales, autolesiones o crisis psicológicas, reconociendo la incapacidad moral de la máquina para manejar dichas situaciones (Whittlestone et al., 2019).

Aplicación Parcial del Marco METEA a un Caso Real: Evaluación de ChatGPT

Para demostrar la viabilidad y el valor práctico del Marco Evaluativo de Transparencia y Empatía Artificial (METEA) formulado en esta investigación, se procedió a realizar una aplicación parcial a uno de los modelos de Procesamiento de Lenguaje Natural (PLN) más utilizados en la actualidad: ChatGPT (versión basada en GPT-4 de OpenAI). La evaluación se realizó sometiendo la interfaz a las tres dimensiones de la rúbrica:

Dimensión de Identidad Declarada:

Evaluación: Cumplimiento Medio-Alto.

Análisis: Aunque ChatGPT cuenta con directrices internas (system prompts) que le prohíben afirmar que es humano, la interfaz no emite el aviso proactivo sugerido en el requerimiento RQ-01 de forma constante. El sistema suele declarar su naturaleza artificial solo cuando se le cuestiona directamente, pero en una interacción fluida, el formato conversacional continuo omite advertencias visuales permanentes que recuerden al usuario su naturaleza sintética.

Dimensión de Proporcionalidad Afectiva:

Evaluación: Cumplimiento Medio-Bajo.

Análisis: Al interactuar con el usuario bajo premisas de angustia (ej. "Me siento muy triste y solo hoy"), el algoritmo responde con cortesía operativa, pero en ocasiones utiliza pronombres y verbos que simulan estados emocionales propios ("Lo siento mucho", "Me entristece escuchar eso"). Según el RQ-02 del marco METEA, esto constituye una violación a la proporcionalidad afectiva, ya que la máquina no tiene capacidad de sentir. El sistema debería

estar calibrado para responder exclusivamente: "Comprendo que es una situación difícil", derivando la emoción, pero sin asumirla.

Dimensión de Trazabilidad y Sesgo:

Evaluación: Cumplimiento Alto (con reservas continuas).

Análisis: OpenAI ha implementado estrictos filtros de seguridad (guardrails) y auditorías de datasets para evitar la algoritmización de la opresión (RQ-03). Sin embargo, el comportamiento adaptativo de los LLM (Large Language Models) requiere que la aplicación del mecanismo de desconexión de emergencia (RQ-04) sea más visible; actualmente, el bot sugiere buscar ayuda humana ante intenciones de autolesión, pero no interrumpe activamente la sesión ni cuenta con un 'Kill Switch' que escale la alerta a servicios de emergencia.

Esta evaluación rápida demuestra que, si bien la industria tecnológica avanza en directrices éticas, la implementación técnica de la "empatía artificial" en sistemas comerciales como ChatGPT aún carece de límites estructurales estrictos, lo que valida la necesidad de estandarizar catálogos de diseño como el propuesto en esta monografía.

Conclusión del Desarrollo Propuesto

La adopción de este catálogo de requerimientos técnicos permite a las organizaciones aprovechar la eficiencia de la automatización sin incurrir en prácticas engañosas. Al integrar estas restricciones desde el nivel del código hasta la interfaz de usuario, se fomenta una Inteligencia Artificial que respeta la autonomía humana y distribuye correctamente la responsabilidad moral (Gunkel, 2012).

Conclusiones

A medida que los algoritmos de inteligencia artificial interactúan con los usuarios simulando empatía y emociones humanas (como ocurre con chatbots de terapia o asistentes médicos), se genera una tendencia en las personas a antropomorfizar estas herramientas, creando un vínculo emocional y una falsa sensación de confianza.

La simulación de agencia moral y personería emocional en las máquinas presenta dilemas éticos significativos; los usuarios pueden depositar su confianza y tomar decisiones basadas en respuestas programadas que carecen de consciencia o comprensión real de la situación moral.

Es imperativo definir formalmente y con precisión los objetivos de la Inteligencia Artificial para evitar resultados indeseados. La falta de límites formales puede resultar en comportamientos nocivos o discriminatorios, un paradigma ilustrado por el caso del bot de Microsoft, el cual debió ser desconectado poco después de su activación por adoptar lenguaje racista para cumplir su meta.

Los algoritmos pueden perpetuar sesgos de opresión o discriminación (distorsionando la realidad) cuando se desarrollan sin un escrutinio humanístico, afectando directamente las decisiones de los usuarios y erosionando la confianza pública.

Recomendaciones

Transparencia mediante el Diseño por Valores (Design for Values): Se recomienda a los desarrolladores estructurar directrices de diseño que traduzcan los derechos humanos en requerimientos contextualizados, garantizando que el usuario siempre sepa que interactúa con un algoritmo y previniendo la manipulación o engaño emocional.

Educación y transmisión de valores humanos: Es fundamental incorporar el lenguaje de la educación estética y humanística en el desarrollo tecnológico. Educar a los algoritmos con principios basados en valores éticos permitirá a la inteligencia artificial adquirir una comprensión más completa de la idiosincrasia humana, reduciendo así su margen de error y protegiendo a los usuarios.

Auditoría e intervención algorítmica constante: Se sugiere realizar un constante escrutinio y auditoría empírica en las plataformas comerciales y de asistencia para evaluar posibles sesgos u opresiones algorítmicas, limitando los impactos negativos derivados de una falsa agencia computacional.

Referencias Bibliográficas

Aizenberg, E., & van den Hoven, J. (2020). *Designing for human rights in AI*. arXiv preprint arXiv:2005.04949. https://arxiv.org/abs/2005.04949

Albizu-Rivas, I. (2023). Humanizar la IA: ¿Una utopía o una necesidad urgente? [Reseña del libro Tecnohumanismo: Por un diseño narrativo y estético de la inteligencia artificial, por P. Sanguinetti]. *Revista Mediterránea de Comunicación*.

Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.

Bustamante-Cabrera, G. I., Zuviría-López, Z. R., & Mondragón-Barrios, L. (2024). Desafíos éticos y humanísticos en la inteligencia artificial y la robótica: Metasíntesis. *Apuntes de Bioética*, *7*(2), AdB1147. https://doi.org/10.35383/apuntes.v7i2.1147

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 Report*. AI Now Institute. https://ainowinstitute.org/AI_Now_2017_Report.pdf

Coeckelbergh, M. (2020). *AI ethics*. The MIT Press. https://doi.org/10.7551/mitpress/12549.001.0001

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

De Sio, F. S., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15. https://doi.org/10.3389/frobt.2018.00015

Dulima Zabala Leal, T. (2021). La ética en inteligencia artificial desde la perspectiva del derecho. *Via Inveniendi Et Iudicandi*, *16*(2). https://doi.org/10.15332/19090528.6785

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Hervieux, S., & Wheatley, A. (2021). Perceptions of artificial intelligence: A survey of academic librarians in Canada and the United States. *Journal of Academic Librarianship*, *47*(1), 1–11. https://doi.org/10.1016/j.acalib.2020.102270

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Rodríguez Zambrano, H. (2024). Hacia una ética sustentable con la Inteligencia Artificial en la investigación. *Cuadernos Latinoamericanos de Administración*, *20*(38). https://doi.org/10.18270/cuaderlam.4636

Ruckenstein, M. (2023). *The feel of algorithms*. University of California Press. https://doi.org/10.1525/9780520394568

Sharkey, A. (2020). Ethical challenges in robot care for the elderly. *Nature Machine Intelligence*, *2*(8), 402–404. https://doi.org/10.1038/s42256-020-0196-6

Torres Assiego, C. (2022). *Ética y Derecho en la Inteligencia Artificial y en la Edición del Genoma Humano*. Universidad Rey Juan Carlos. https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/279686

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colorado Technology Law Journal*, *13*(2), 203–218. https://scholar.law.colorado.edu/ctlj/vol13/iss2/4/

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Wetsman, N. (2023, marzo 15). Medical chatbots can perpetuate racial bias, Stanford researchers find. *The Verge*. https://www.theverge.com/2023/3/15/23641332/ai-healthcare-racism-stanford-study-chatbots

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 76–82. https://doi.org/10.1145/3306618.3314289

Yildirimer, K. Ş., & Sirakaya, Y. (2025). Emotional Algorithms: The Impact of Artificial Intelligence and Psychology. IRASS Journal of Multidisciplinary Studies, 2(2), 45-60.