Human-in-the-loop (HITL) as a Verification and Validation Strategy for Knowledge Generated by Generative artificial intelligence

Mauricio Rojas-Contreras[1]; Ailin Orjuela Duarte[1]; Luz Marina Santos Jaimes[1]
[1]Universidad de Pamplona, Pamplona, Norte de Santander,Colombia, {*mrojas, aorjuela,lsantos*}*@unipamplona.edu.co*

Abstract– The Human-in-the-Loop (HITL) approach describes human participation in various stages of artificial intelligence system development. This research identifies the methods employed by end users for verifying and validating knowledge generated by generative artificial intelligence (GAI). A systematic literature review was conducted following the PRISMA protocol to analyze the methods used for knowledge verification and validation in the context of the HITL approach. The search equation, developed using a generative AI tool, was applied to the Scopus database and the AI-powered search engine Undermind, retrieving a total of 95 documents. After applying inclusion and exclusion criteria, 19 articles were selected for analysis. The findings allowed for the categorization of the identified methods into two groups: those used in the design and implementation stages of GAI systems and those employed by end users. However, persistent challenges remain, particularly the lack of detailed specification and formalization of knowledge verification and validation methods at the end-user level, which impacts the accuracy of responses and the control of generated knowledge creativity. Future research should focus on specifying, testing, and formalizing these methods to optimize their application within the HITL framework. This study contributes to the field by providing a set of methods for verifying and validating knowledge generated by GAI, thereby improving response accuracy and control over creativity.

Keywords-- Artificial intelligence, generative artificial intelligence, human-in-the-loop, Verification and validation, Verification and validation methods.

Human-in-the-loop (HITL) como estrategia de verificación y validación del conocimiento generado por la IAG

Resumen– El enfoque Human-in-the-Loop (HITL) describe la participación humana en las distintas etapas del desarrollo de sistemas de inteligencia artificial. En esta investigación, se identifican los métodos empleados por los usuarios finales para la verificación y validación del conocimiento generado por la inteligencia artificial generativa (IAG). Se llevó a cabo una revisión sistemática de la literatura, basada en el protocolo PRISMA, con el propósito de analizar los métodos utilizados en la verificación y validación del conocimiento generado por la IAG en el contexto del enfoque HITL. La ecuación de búsqueda, desarrollada con una herramienta de IA generativa, se aplicó a la base de datos Scopus y al motor de búsqueda potenciado por IA Undermind, obteniendo un total de 95 documentos. Tras la aplicación de criterios de inclusión y exclusión, se seleccionaron 19 artículos para el análisis. Los hallazgos permitieron categorizar los métodos identificados en dos grupos: aquellos empleados en las etapas de diseño e implementación de los sistemas de IAG y los utilizados por los usuarios finales. No obstante, se identifican desafíos persistentes, en particular, la falta de especificación detallada y formalización de métodos de verificación y validación del conocimiento a nivel de usuario final, lo que impacta en la precisión de las respuestas y el control de la creatividad del conocimiento generado. La investigación futura debería centrarse en la especificación, prueba y formalización de estos métodos para optimizar su aplicación en el contexto HITL. Este estudio contribuye al campo al proporcionar un conjunto de métodos para la verificación y validación del conocimiento generado por la IAG, mejorando así la precisión de las respuestas y el control sobre su creatividad.

Palabras clave—Inteligencia artificial, inteligencia artificial generativa, HITL, verificación y validación, métodos de verificación y validación.

I. Introducción

La verificación y validación (V&V) de la inteligencia artificial (IA) y los sistemas de IA generativa es un desafío esencial debido a la naturaleza no determinista y compleja de estos modelos. A diferencia de los sistemas deterministas tradicionales, la IA generativa, incluidos los modelos de lenguaje grande (LLM) y los generadores de imágenes, produce resultados que pueden carecer de criterios claros de corrección, dando lugar a riesgos como alucinaciones, sesgos, comportamiento inseguro y desalineaciones éticas. Las metodologías Human-in-the-loop (HITL) están surgiendo como estrategias críticas para abordar estos desafíos mediante la integración de la retroalimentación humana en las etapas clave del proceso de V&V. Este cuerpo de investigación explora los métodos HITL específicamente adaptados a la validación de los sistemas de IA, centrándose en mejorar la seguridad, la confiabilidad y la adhesión a las preferencias del usuario.

Un enfoque central en esta área es la aplicación del Aprendizaje por Refuerzo de la Retroalimentación Humana (RLHF), un método en el que se utiliza la retroalimentación humana iterativa para alinear el comportamiento de la IA con los valores y objetivos humanos. RLHF se ha adoptado ampliamente para ajustar los sistemas de IA generativa, con estudios que demuestran su eficacia para mejorar el rendimiento del modelo en tareas como el resumen, la clasificación y la alineación de preferencias [1,2]. Se han propuesto varias mejoras a RLHF, incluidos mecanismos de retroalimentación de múltiples turnos para la planificación de objetivos a largo plazo [3], técnicas de aprendizaje activo para optimizar el muestreo de datos y reducir el esfuerzo humano [4, 5], y el modelado de recompensas contrastiva para mejorar la solidez del manejo de la retroalimentación [6]. Para abordar los riesgos posteriores, como la desalineación y el sobreajuste de la retroalimentación, los nuevos métodos como la simulación retrospectiva basada en RL (RLHS) se centran en evaluar el impacto a largo plazo de las interacciones [7]. Juntos, estos avances aprovechan la naturaleza iterativa de RLHF no solo para validar los modelos, sino también para refinarlos para una mayor alineación social.

Paralelamente, los enfoques de IA explicables (XAI) se están convirtiendo en una parte integral de los marcos de HITL V&V, ya que ofrecen ideas interpretables para guiar a los evaluadores humanos en la evaluación de los resultados de IA. Por ejemplo, los marcos basados en explicaciones permiten a los sistemas justificar sus decisiones, permitiendo flujos de trabajo de validación sistemáticos que incorporan retroalimentación humana para mejorar la alineación del modelo o identificar modos de falla [8]. Las herramientas pragmáticas como los Árboles de Preferencias proporcionan estructuras interpretables para verificar la adhesión a la guía humana [9]. Los enfoques basados en modelos conceptuales, como el Marco CMAG, amplían la explicabilidad a los resultados generativos al mapear los resultados en estructuras

comprensibles para el ser humano, como los gráficos de conocimiento, facilitando así la supervisión humana de alta fidelidad [10].

La escalabilidad es un desafío recurrente en la integración de los humanos en los procesos de V&V, particularmente para los modelos generativos a gran escala. Para abordar este problema, los métodos automatizados tienen como objetivo complementar los juicios humanos mientras se mantiene la calidad de la supervisión. La IA Generativa Diferencial (D-GAI) utiliza pruebas de múltiples versiones para evaluar la fiabilidad analizando la diversidad de versiones, reduciendo así la dependencia de las evaluaciones de un solo punto [11]. Métodos de exploración basados en el recuento en la exploración del equilibrio RLHF en línea y la optimización de preferencias priorizando las consultas que ofrecen un mayor valor informativo [12], mientras que los marcos de prueba basados en datos sintéticos como VerifAI validan las salidas generativas contra los conjuntos de datos multimodales para detectar inconsistencias o sesgos a escala [13].

En aplicaciones específicas de dominio, los métodos HITL V&V están diseñados para cumplir con los diferentes requisitos de seguridad y ética. Por ejemplo, en el cuidado de la salud y la conducción autónoma, la detección de anomalías y las revisiones de expertos de múltiples capas garantizan la seguridad [14, 15]. En contextos creativos, los marcos HITL se centran en la calidad y originalidad de los resultados, integrando estándares flexibles para tareas inherentemente subjetivas [16]. Las revisiones entre dominios y los análisis sistemáticos también han identificado desafíos centrales, como la gestión de riesgos éticos, la detección de la deriva conductual y el abordaje de problemas de calidad de los datos, que son críticos para el sólido V&V de los sistemas de IA en evolución [17, 18].

Por lo tanto, la integración de los enfoques HITL en AI V&V representa una evolución crítica hacia sistemas generativos más confiables. Al cerrar la brecha entre los enfoques de validación deterministas tradicionales y el comportamiento probabilístico, a menudo opaco, de la IA moderna, estos métodos proporcionan mecanismos escalables y efectivos para mejorar la colaboración humano-IA y garantizar que los resultados generativos se alineen con los objetivos sociales y específicos del dominio.

En cuanto a la pregunta de investigación que orienta esta investigación, se pretende dar respuesta a la pregunta ¿Cuáles son los métodos utilizados en el enfoque HITL para hacer verificación y validación del conocimiento generado por las inteligencias artificiales generativas?

II. Materiales y métodos

La estrategia de investigación empleada en este trabajo fue una revisión sistemática para sintetizar el conocimiento existente sobre los métodos utilizados para la verificación y validación del conocimiento generado por IA generativa utilizando estrategias Human-in-the-loop (HITL). Se desarrolló una estrategia de búsqueda integral para identificar literatura relevante en la base de datos SCOPUS y el motor de búsqueda impulsado por IA, Undermind. Se utilizó la ecuación de búsqueda ***("Human-in-the-loop" OR HITL) AND ("verification" OR "validation") AND ("generative AI" OR "AI-generated knowledge" OR "artificial intelligence")*** para extraer estudios pertinentes**.**

Criterios de Inclusión

Se incluyeron estudios que abordaran estrategias HITL en el contexto de la verificación o validación del conocimiento generado por IA. En forma específica, se planearon los siguientes criterios de inclusión:

I1: Artículos que hayan sido publicados en el dominio de tiempo 2022 - 2024.

I2: Se eligieron artículos los cuales su campo de estudio fueran las Ciencias Sociales, las Ciencias de la computación y la ingeniería.

I3: Documentos que fuesen del tipo artículo de revisión o artículo de investigación.

Criterios de Exclusión

Se excluyeron artículos que no abordaran la verificación o validación del conocimiento generado por IA o que no involucraran estrategias HITL. En forma particular, se especificaron los siguientes criterios de exclusión:

E1: Se excluyen aquellos artículos en los que al analizar el título y el resumen no tenían un aporte importante al objeto de de estudio.

E2: Se excluyen artículos en los que al analizar de forma completa el trabajo se identifica que no está completamente alineado al objeto de estudio.

Recolección de Datos

Los resultados de la búsqueda se exportaron desde la base de datos SCOPUS y el motor de búsqueda Undermind en formato RIS e importaron en la herramienta Rayyan para el proceso de selección.

Análisis de Datos

Se analizaron los estudios seleccionados para identificar temas y metodologías comunes relacionados con las estrategias HITL para la verificación y validación.

Declaración de Aprobación/Ética

Dado que este estudio implicó el análisis de literatura existente, no se requirió aprobación ética.

Declaración de Consentimiento Informado

El consentimiento informado no fue aplicable, ya que el estudio no involucró a participantes humanos.

TABLA I
ECUACIÓN DE BÚSQUEDA

Prompt	Herramienta de inteligencia artificial generativa	Versión	Ecuación de búsqueda
Actue como investigador en el enfoque HITL para la verificación y vlidación del conocimiento generado por la IAG. Genere una ecuación de búsqueda optimizada acerca de el enfoque Human in the Loop (HITL) para la verificación y validación del conocimiento generado por la IAG. Configuración: Temperatura: 0.01.	Chatgpt	4.o	("Human-in-the-loop" OR HITL) AND ("verification" OR "validation") AND ("generative AI" OR "AI-generated knowledge" OR "artificial intelligence")

Posteriormente, se identificaron las bases de datos y motores de búsqueda relevantes para el campo de las Ciencias sociales, Ciencias de la Computación y la Ingeniería; se tuvieron en cuenta aquellas que cubren las conferencias y revistas más importantes en el campo de la tecnología educativa, siendo esta, Scopus y el motor de búsqueda impulsado con IA Undermind.

Teniendo en cuenta la ecuación de búsqueda se realizó una búsqueda en la base de datos electrónica y en el motor de búsqueda; se aplicaron filtros para los años 2022-2024 y solo artículos de investigación o revisión, las cuales arrojaron la siguiente cantidad de documentos representados en la Tabla II.

TABLA II
CANTIDAD DE DOCUMENTOS EN FUENTES

Ecuación de búsqueda	SCOPUS	UNDERMIND
("Human-in-the-loop" OR HITL) AND ("verification" OR "validation") AND ("generative AI" OR "AI-generated knowledge" OR "artificial intelligence")	63	32

III. Resultados

La figura 1 ilustra el proceso de identificación realizado en la base de datos SCOPUS y el motor de búsqueda Undermind, utilizando la ecuación de búsqueda especificada en la tabla I. Como resultado de aplicar la ecuación de búsqueda a las fuentes de datos se obtuvieron 95 documentos. A continuación, se aplicaron los criterios de inclusión (I1, I2, I3) y de exclusión (E1, E2), obteniendo como resultado 19 artículos para análisis, revisión y dar respuesta a la pregunta objeto de estudio de esta investigación.

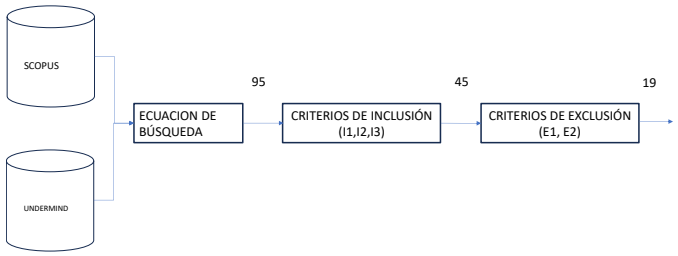

Fig. 1 Proceso de filtrado de artículos

En forma específica, el proceso de screening (aplicación de los criterios de exclusión E1, E2) se llevó a cabo con el apoyo de la herramienta impulsada con IA Rayyan, la cual permite analizar los artículos en dos niveles, un primer nivel solo utilizando el título y el abstract, un segundo nivel analizando el texto completo del artículo como se describe en la figura 2.

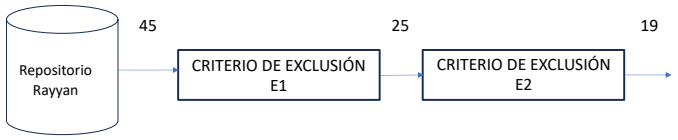

Fig. 2 Proceso de screening

En la Tabla III se presentan los resultados del proceso de selección (Screening) y el análisis completo de los artículos seleccionados, los cuales están relacionados con los métodos utilizados en el enfoque HITL para la verificación y validación del conocimiento generado por las inteligencias artificiales

generativas. De forma complementaria, se incluye la respuesta a la pregunta de investigación Q1, basada en el conocimiento generado por la herramienta de inteligencia artificial generativa Chatpdf. Con el fin de verificar y validar dicho conocimiento, se detalla en la columna "Triangulación hermeneútica" (análisis comparativo del conocimiento generado por la herramienta de IA generativa con los referentes teóricos y la posición del investigador) el resultado de aplicar este método cualitativo a cada uno de los artículos identificados durante el proceso de revisión sistemática.

Tabla III
Resultados revisión sistemática

Título artículo y referencia	RQ1	Triangulación hermenéutica
Advice Conformance Verification by Reinforcement Learning agents for Human-in-the-Loop. Verma et al. (2022) [9]	-Preference trees. -Extracción de árboles de preferencia. -Experimentos en entornos controlados. -Estudios con usuarios humanos.	El análisis permitió identificar que los métodos descritos por los autores son utilizados en la fase de pruebas del sistema de IAG.
Verification and Validation of AI Systems Using Explanations. Mahmuud et al. (2024) [8]	-Evaluacion de explicaciones. -Retroalimentación crítica. -Evaluación cuantitativa y cualitativa. -Explotación de atribución de características. -Iteración continua. -Pruebas de verificación basadas en consulta. -Reforzamiento a través de feedback. -Estudio de caso REVEALE.	El enfoque HITL es usado en el diseño e implementación del sistema a través de métodos o técnicas como: Desarrollo de modelos explicativos, prototipos de evaluación, integración de feedback humano, pruebas de verificación predictiva. HITL uso por el usuario del sistema: Evaluación directa de resultados, feedback crítico, pruebas informales de verificación, iteración y refinamiento.
How Can I Trust AI? : Extending a UXer-AI Collaboration Process in the Early Stages. Yoon et al. (2024) [19]	-Revisión por expertos. -Talleres colaborativos. -Etapas de verificación y toma de decisiones. -Métodos de verificación diversos.	Los métodos descritos en este artículo están orientados al diseño UX y a la interacción con la tecnología de los sistemas de IA.
Understanding the Interplay Between Trust, Reliability, and Human Factors in the Age of Generative AI. Thorne (2024) [16]	-Validación del usuario: Triangulación de hechos, Conocimiento del dominio. - Verificación del código. -Estrategias de mitigación de alucinaciones.	Validación del usuario: Triangulación de hechos, Conocimiento del dominio. Estrategias de mitigación de alucinaciones: Redacción de prompts, Evaluación de respuestas.
	- Calibración de la confianza. - Pruebas de estrés y escenarios de incertidumbre.	
Certifiable Trust in Autonomous Systems: Making the Intractable Tangible. Lyons et al. (2017) [20]	-Transparencia del Sistema: Modelos de transparencia, comunicación de factores analíticos. - Entrenamiento basado en escenarios: Pruebas en diversos escenarios, escenarios moralmente contenciosos. -Calibración de la confianza: evaluación de la confiabilidad, modelos de trabajo en equipo. -verificación continua. -Modelos de intención y entorno.	Métodos y técnicas para usuarios del sistema: -Calibración de la confianza: transparencia y comunicación, entrenamiento en diversos escenarios. -Verificación continua: verificación en tiempo de ejecución.
Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities. Retzlaff et al. (2024) [1]	-Desarrollo inicial del agente. -Aprendizaje del agente. -Evaluación del agente. -Despliegue del agente.	Fase de uso por los usuarios del sistema: -Evaluación del agente: resumen de políticas, explicaciones basadas en grafos, evaluación de seguridad. -Despliegue del agente: Indicadores visuales y auditivos, explicaciones tetuales.
Dual Active Learning for Reinforcement Learning from Human Feedback. Liu et al. (2024) [4]	-Aprendizaje active dual. -Aprendizaje por refuerzo refuerzo con retroalimentación humana. -Política pesimista para aprendizaje offline. -Uso de modelos de selección de expertos.	Los métodos descritos están diseñados para ser utilizados en las fases de diseño e implementación del sistema de IA y no se explicita su uso por el usuario final del sistema.
Combining Theory of Mind and Kindness for Self-Supervised Human-AI Alignment. Hewson (2024) [3]	-Aprendizaje por refuerzo desde la retroalimentación humana. -Teoría de la mente (ToM). -Neuronas espejo. -Simulación e imitación. -Empatia y teoría de la mente avanzada.	Uso por parte del usuario para verificación y validación: -Evaluación de salidas del modelo. -Retroalimentación continua. -Pruebas de escenarios. -Monitoreo y auditoría.
Human-in-the-loop Learning for Safe Exploration through Anomaly Prediction and Intervention.	-Aprendizaje por demostración. -Aprendizaje por intervención.	Uso por parte del usuario final: -Exploración segura. -Ejecución conjunta.

Rajendran et al. (2022) [14]	-Aprendizaje por evaluación. -Exploración Segura. -Ejecución conjunta.	
Facilitating Human Feedback for GenAI Prompt Optimization. Sherson & Vinchon (2024) [21]	-Bucle de entrenamiento humano-AI: Evaluación de salidas generadas, retroalimentación comparativa. -Evaluación sistemática: escala de evaluación, comparación de salidas. -Estudios piloto: experimentos con estudiantes. -Análisis descriptivo: Número de palabras utilizadas.	Uso por parte de los usuarios del sistema: -Retroalimentación continua: evaluación de salidas, comparación de salidas. -Interacción humano-AI: Ajuste de prompts, descripciones y justificaciones.
Online Preference Alignment for Language Models via Count-based Exploration. Bai et al. (2025) [5]	-Optimización de preferencias directas (DPO): modelado de recompensas, alineación de preferencias. -Optimización de preferencias en línea (RLHF): generación de nuevas respuestas, exploración y optimización. -Exploración basada en conteo (DOPO): módulo de conteo, balance entre exploración y optimización.	Uso por parte de usuarios del sistema: proporcionar retroalimentación, evaluación de respuestas, uso de herramientas de verificación.
Enhancing Large Language Model Performance with Reinforcement Learning from Human Feedback: A Comprehensive Study on Q&A, Summarization, and Classification. Rawal et al. (2024) [2]	-Generación de datos. -Entrenamiento del modelo de recompensa: Señales de recompensa derivadas de la retroalimentación humana. -Aprendizaje por refuerzo: Proximal Policy optimización (PPO).	Utilización por parte de los usuarios: -Recolección de retroalimentación humana. -Verificación y validación continua. -Ajuste fino basado en retroalimentación.
Improving Discriminative Capability of Reward Models in RLHF Using Contrastive Learning. Chen et al. (2024) [6]	-Aprendizaje por refuerzo con retroalimentación humana (RLHF). -Aprendizaje contrastivo para modelado de recompensa: SimCSE(Simple Contrastive Learning of Sentence Embeddings), SmAV (Swapping Assignments between Views).	A bivel de usuario final: -Verificación manual de respuestas. -Evaluación con herramientas externas. -Retroalimentación humana en plataformas con HITL. -Uso de sistemas de IA con verificación incorporada.
	-Modelado de preferencias humanas. -Ajuste fino con optimización de políticas proximales (PPO). -Evaluación con benchmarking de modelos de lenguaje. -Muestreo y aumento de datos.	
CMAG: A Framework for Conceptual Model Augmented Generative Artificial Intelligence. Fill et al. (2024) [10]	-Uso de modelos conceptuales. -Estrategias de prompting estructurado. -Validación basado en modelos semánticos. -Interacción en múltiples pases; generación inicial, transformación y representación, validación humana, ajustes y refinamientos. -Uso de modelos de dominio específico.	Fase de usuario final: -Uso de modelos conceptuales para inspección. -Validación semántica por expertos. -Refinamiento mediante interacción humano-IA. -Aplicación de métodos de representación visual.
RLHS: Mitigating Misalignment in RLHF with Hindsight Simulation. Liang et al. (2025) [7]	-Evaluación directa de la salida. -Revisión de pares. -Feedback iterativo. -Simulaciones de escenarios. -Uso de hindsight simulation. -Análisis de regreso (Post-hoc análysis).	Por usuarios del sistema: -Evaluación directa de la salida. -Revisión de pares. -Feedback iterativo. -Simulaciones de escenarios. -Uso de hindsight simulation. -Análisis de regreso (post-hoc análisis).
VerifAI: Verified Generative AI. Tang et al. (2023) [13]	-Revisión humana. -Feedback de usuarios. -Verificación basada en datos. -Corroboración de múltiples fuentes. -Ajuste de parámetros del modelo.	Por usuarios del sistema: -Revisión manual. -Uso de herramientas de comparación. -Proporcionar retroalimentación. -Consultas a expertos.
N-Version Assessment and Enhancement of Generative AI: Differential GAI. Kessel & Atkinson (2024) [11]	-Generación de multiples versions. -Análisis comparativo. -Plataforma Lasso. -Evaluación de la calidad del código. -Pruebas diferenciales. -Recomendación de códigos y pruebas.	Uso por parte de los usuarios del sistema: -Análisis comparativo. -Pruebas diferenciales. -Recomendación de código y pruebas. -Evaluación y mejora de pruebas.
Toward a Methodology for the Verification and Validation of AI-Based Systems.	-Modelo probabilístico del dominio de diseño operacional.	-Modelo probabilístico del dominio de diseño operacional. -Análisis de modos y efectos de fallos (FMEA) de la IA.

Paardekooper & Borth (2024). [15]	-Análisis de modos y efectos de fallos (FMEA) de la IA. -Modelo de aptitud para la IA.	-Modelo de aptitud para la IA.
Verification & Validation Methods for Complex AI-enabled Cyber-Physical Learning-Based Systems: A Systematic Literature Review. Meyer & Oosthuizen (2023). [17]	-Inspección. -Demostración. -Pruebas. -Análisis. -Gestión de datos. -Lista de complicaciones.	Utilizados por usuarios del sistema. -Inspección. -Demostración. -Pruebas. -Análisis. -Gestión de datos.

IV. DISCUSIÓN

Los enfoques HITL, como FairCaipi y D-BIAS, demuestran que los bucles de retroalimentación iterativos con intervención humana pueden mitigar eficazmente los sesgos refinando conjuntos de datos y ajustando relaciones causales. Además, las intervenciones en tiempo real mediante explicaciones contrafactuales permiten a los humanos corregir decisiones sesgadas dinámicamente. Las herramientas de IA explicables, como FAIRVIS y D-BIAS, mejoran la transparencia y la equidad al permitir auditorías interactivas. Los sistemas HITL también equilibran la equidad y la utilidad mediante marcos de optimización multiobjetivo y estrategias de aplazamiento de decisiones inciertas al juicio humano.

Los marcos iterativos HITL son efectivos para reducir sesgos en datos, modelos y resultados mediante la retroalimentación humana continua. Las explicaciones contrafactuales y las herramientas visuales interactivas permiten a los usuarios identificar y rectificar problemas de equidad en tiempo real. La combinación de expertos en dominios y usuarios finales en el proceso de mitigación de sesgos asegura una evaluación más completa y precisa. Además, los sistemas que equilibran equidad y utilidad mediante optimización multiobjetivo destacan la importancia de la toma de decisiones éticas en contextos críticos.

Los métodos HITL de verificación y validación pueden implementarse tanto en la fase de diseño de sistemas de IA como a nivel de usuario final. Esto permite una mitigación de sesgos más efectiva y una mejora en la transparencia y equidad de los sistemas de IA. La integración de la retroalimentación humana en tiempo real y las herramientas explicables facilita la auditoría continua y la corrección dinámica de sesgos, lo que es crucial para aplicaciones en contextos de alto riesgo.

Las fortalezas de los métodos HITL incluyen la mejora de la precisión y la minimización de alucinaciones en IA generativa. Sin embargo, aún se deben desarrollar métodos adicionales para controlar las alucinaciones y abordar limitaciones relacionadas con la escalabilidad, sesgos y la inclusión humana. La dependencia de la intervención humana puede limitar la eficiencia y la capacidad de los sistemas para operar a gran escala sin supervisión constante.

Es esencial continuar investigando en métodos de verificación y validación del conocimiento generado por IA, así como en estrategias para controlar las alucinaciones. Además, se deben explorar enfoques para mejorar la escalabilidad y reducir la dependencia de la intervención humana. La investigación futura también debería centrarse en la integración de HITL en diversos dominios críticos para asegurar decisiones éticas y equitativas en aplicaciones de IA.

V. CONCLUSIONES

Los procesos de verificación y validación del conocimiento generado por las inteligencias artificiales generativas se llevan a cabo en las diferentes etapas del proceso de desarrollo de los sistemas de inteligencia artificial y en particular a nivel de usuario final con la intervención de humanos materializando el concepto de Human-in-the-loop (HITL), el cual especifica la participación de humanos en todo el ciclo de desarrollo de sistemas de inteligencia artificial y a nivel de usuario final de este tipo de tecnologías.

En cuanto a los patrones, se puede enunciar que existen diferentes métodos para hacer verificación y validación del conocimiento generado por las inteligencias artificiales generativas con intervención humana; particularmente, se pueden identificar dos categorías de métodos: aquellos que se utilizan en las etapas de diseño e implementación de los sistemas de inteligencia artificial generativa y la segunda categoría corresponde a métodos utilizados por los usuarios finales de este tipo de tecnologías.

En cuanto a las brechas identificadas, en la literatura existente, se evidencia la falta de especificación detallada de cómo llevar a cabo los procesos de verificación y validación del conocimiento por los usuarios finales de las inteligencias artificiales generativas.

En lo referente a la contribución teórica, los hallazgos relacionados con los métodos de verificación y validación del conocimiento generado por las inteligencias artificiales generativas en el contexto del enfoque HITL, permiten categorizarlos en métodos utilizados en el proceso de diseño e implementación y métodos utilizados a nivel de usuario final. En forma complementaria, la contribución práctica de esta investigación se centra en la identificación de métodos para hacer verificación y validación del conocimiento generado por las IAG, lo cual redunda en la mejora de la precisión de las respuestas y el control de la creatividad del conocimiento generado por la IAG.

Se identifica un campo de investigación futura en la especificación detallada y formalización de métodos para hacer verificación y validación del conocimiento generado por IAG a nivel de usuario final; se podrían especificar técnicas para mejorar la precisión de las respuestas desde el prompt a través del ajuste de variables de configuración y especificación detallada de métodos de análisis para verificar y validar las respuestas generadas por las IAG, principalmente en los contextos académicos y de investigación.

Finalmente, ante el reconocimiento de la generación de respuestas con altos indicadores de creatividad por las IAG y particularmente con la gestión de las alucinaciones por parte de los usuarios finales de las IAG, se hace necesario que los usuarios finales conozcan métodos para hacer verificación y validación del conocimiento generado por las IAG, con el fin de minimizar los riesgos relacionados con la desinformación en contextos críticos donde la precisión es un punto clave para la toma de decisiones.

REFERENCIAS

[1] Retzlaff, C., Das, S., Wayllace, C., Mousavi, P., Afshari, M., Yang, T., Saranti, A., Angerschmid, A., Taylor, M.E., & Holzinger, A. (2024). Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities. J. Artif. Intell. Res., 79, 359-415.

[2] Rawal, N., Tavva, P., & Selvakumar, P. (2024). Enhancing Large Language Model Performance with Reinforcement Learning from Human Feedback: A Comprehensive Study on Q&A, Summarization, and Classification. 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET), 1-6.

[3] Hewson, J.T. (2024). Combining Theory of Mind and Kindness for Self-Supervised Human-AI Alignment. ArXiv, abs/2411.04127.

[4] Liu, P., Shi, C., & Sun, W.W. (2024). Dual Active Learning for Reinforcement Learning from Human Feedback. ArXiv, abs/2410.02504.

[5] Bai, C., Zhang, Y., Qiu, S., Zhang, Q., Xu, K., & Li, X. (2025). Online Preference Alignment for Language Models via Count-based Exploration.

[6] Chen, L., Zheng, R., Wang, B., Jin, S., Huang, C., Ye, J., Zhang, Z., Zhou, Y., Xi, Z., Gui, T., Zhang, Q., & Huang, X. (2024). Improving Discriminative Capability of Reward Models in RLHF Using Contrastive Learning. Conference on Empirical Methods in Natural Language Processing.

[7] Liang, K., Hu, H., Liu, R., Griffiths, T.L., & Fisac, J.F. (2025). RLHS: Mitigating Misalignment in RLHF with Hindsight Simulation.

[8] Mahmud, S., Saisubramanian, S., & Zilberstein, S. (2024). Verification and Validation of AI Systems Using Explanations. Proceedings of the AAAI Symposium Series.

[9] Verma, M., Kharkwal, A., & Kambhampati, S. (2022). Advice Conformance Verification by Reinforcement Learning agents for Human-in-the-Loop. ArXiv, abs/2210.03455.

[10] Fill, H., Härer, F., Vasic, I., Borcard, D., Reitemeyer, B., Muff, F., Curty, S., & Bühlmann, M. (2024). CMAG: A Framework for Conceptual Model Augmented Generative Artificial Intelligence. International Conference on Conceptual Modeling.

[11] Kessel, M., & Atkinson, C. (2024). N-Version Assessment and Enhancement of Generative AI. ArXiv, abs/2409.14071.Hazari, S. (2024). Justification and Roadmap for Artificial Intelligence (AI) Literacy Courses in Higher Education. Journal of Educational Research and Practice.

[12] Bai, C., Zhang, Y., Qiu, S., Zhang, Q., Xu, K., & Li, X. (2025). Online Preference Alignment for Language Models via Count-based Exploration..

[13] Tang, N., Yang, C., Fan, J., & Cao, L. (2023). VerifAI: Verified Generative AI. ArXiv, abs/2307.02796.

[14] Rajendran, P.T., Espinoza, H., Delaborde, A., & Mraidha, C. (2022). Human-in-the-loop Learning for Safe Exploration through Anomaly Prediction and Intervention. SafeAI@AAAI..

[15] Paardekooper, J., & Borth, M. (2024). Toward a Methodology for the Verification and Validation of AI-Based Systems. SAE International Journal of Connected and Automated Vehicles.

[16] Thorne, S. (2024). Understanding the Interplay Between Trust, Reliability, and Human Factors in the Age of Generative AI. International journal of simulation: systems, science & technology.

[17] Meyer, W., & Oosthuizen, R. (2023). Verification & Validation Methods for Complex AI-enabled Cyber-Physical Learning-Based Systems: A Systematic Literature Review. 2023 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), 1-7.

[18] Angelis, E.D., Angelis, G.D., & Proietti, M. (2023). A Classification Study on Testing and Verification of AI-based Systems. 2023 IEEE International Conference On Artificial Intelligence Testing (AITest), 1-8.

[19] Yoon, H., Oh, C., & Jun, S. (2024). How Can I Trust AI? : Extending a UXer-AI Collaboration Process in the Early Stages. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems.

[20] Lyons, J.B., Clark, M.A., Wagner, A.R., & Schuelke, M.J. (2017). Certifiable Trust in Autonomous Systems: Making the Intractable Tangible. AI Mag., 38, 37-49.

[21] Sherson, J., & Vinchon, F. (2024). Facilitating Human Feedback for GenAI Prompt Optimization. HHAI.