22 Diciembre 2025 Revista de la Carrera de Ingeniería de Sistemas
FONDO EDITORIAL ULIMA

INTERFASES

INTERFASES

Interfases
Revista de la Carrera de Ingeniería de Sistemas de la Facultad de Ingeniería
de la Universidad de Lima
N.° 22, dicicembre, 2025
https://doi.org/10.26439/interfases2025.n022
Lima, Perú

Fondo Editorial
Av. Javier Prado Este 4600
Urb. Fundo Monterrico Chico
Santiago de Surco, Lima, Perú
Código postal 15023
Teléfono (511) 437-6767, anexo 30131
fondoeditorial@ulima.edu.pe
www.ulima.edu.pe

Edición, diseño y diagramación: Fondo Editorial de la Universidad de Lima.

Correspondencia:
interfases@ulima.edu.pe

Periodicidad: semestral
Arbitraje editorial: revisión por pares doble ciego
Directorios y catálogos: Redalyc, CrossRef, Dialnet, Latindex y DOAJ

ISSN (en línea): 1993-4912
Hecho el depósito legal en la Biblioteca Nacional del Perú n.º 2020-09967

ÍNDICE

PRESENTACIÓN

https://doi.org/10.26439/interfases2025.n022.8493

La revista *Interfases*, editada por la carrera de Ingeniería de Sistemas de la Universidad de Lima, presenta en su edición 22 un total de siete artículos que abordan problemáticas y desarrollos relevantes dentro del alcance temático de la revista. Estas contribuciones abarcan áreas como las ciencias de la computación, ingeniería de *software*, sistemas de información, tecnologías de la información, ciberseguridad, ciencia de datos, entre otras. Los artículos publicados fueron seleccionados mediante un riguroso proceso de evaluación por pares bajo la modalidad de revisión por doble ciego.

Agradecemos a todos los autores que enviaron sus manuscritos como contribución a la revista. Gracias a estas propuestas, y al trabajo articulado del Comité Editorial y de nuestro equipo de revisores científicos, fue posible realizar una cuidadosa selección de los trabajos que conforman esta edición. En este número se cuenta con la participación de autores provenientes de Brasil, Costa Rica, Reino Unido y Perú, afiliados a diversas instituciones académicas, entre ellas la Universidad de Lima (Perú), la Universidad Nacional Mayor de San Marcos (Perú), la Universidad Internacional de las Américas (Costa Rica), la Universidade Federal de Goiás (Brasil) y la University of Manchester (Reino Unido). Cabe destacar que dos de los artículos fueron presentados en idioma inglés.

Las contribuciones de esta edición se enmarcan en campos temáticos como *smart cities and public safety*, arquitectura *blockchain* aplicada al registro de historiales médicos, arquitecturas agénticas de inteligencia artificial, *serious games* basados en realidad virtual, visión por computadora, implementación de inteligencia artificial en el Estado peruano, así como ética y gobernanza algorítmica en el sector financiero.

Como se puede apreciar, esta edición reúne aportes significativos que fortalecen el contenido científico de la revista. Agradecemos profundamente a los autores por confiar en *Interfases* como medio de difusión de sus trabajos de investigación científica. Asimismo, extendemos nuestro reconocimiento a los revisores que participaron en el proceso de evaluación, quienes analizaron un total de catorce artículos, de los cuales siete fueron aprobados para su publicación. Este proceso selectivo garantizó la calidad académica de los trabajos publicados.

Cabe resaltar que los revisores pertenecen a diversas instituciones académicas nacionales e internacionales, entre ellas la Fundación Universitaria Konrad Lorenz (Colombia), la Universidad Nacional del Chaco Austral (Argentina), la Universidad de Lima (Perú), la Federal University of São Paulo (Brasil), la Universidad Veracruzana (México), la Universidad Nacional Agraria de la Selva (Perú), la Pontificia Universidad Católica del Perú, la Universidad Nacional Federico Villarreal (Perú), la University of Bern (Suiza), así como profesionales del sector financiero especializados en *blockchain*.

Finalmente, expresamos nuestro más sincero agradecimiento al Fondo Editorial y al Área de Colecciones de la Biblioteca de la Universidad de Lima por su invaluable apoyo técnico en los procesos de revisión de estilo, diagramación y publicación de la edición 22 de *Interfases*.

Dra. Nadia Katherine Rodríguez Rodríguez

Directora

Interfases

ARTÍCULOS DE INVESTIGACIÓN

SMART CITIES AND PUBLIC SAFETY: EMPIRICAL EVIDENCE ON VIDEO SURVEILLANCE AND DEVELOPMENT IN APARECIDA DE GOIÂNIA CITY

MARCOS DIAS DE PAULA
https://orcid.org/0009-0003-1254-9834
marcos.paula@ufg.br
Universidade Federal de Goiás, Brazil

PEDRO H. GONÇALVES
https://orcid.org/0000-0001-9919-6557
pedrogoncalves@ufg.br
Universidade Federal de Goiás, Brazil

Received: September 5, 2025/ Accepted: October 23, 2025
doi: https://doi.org/10.26439/interfases2025.n022.8258

ABSTRACT. This article analyzes the relationship between the implementation of urban video surveillance, as part of the smart cities strategy, and social and public safety indicators in Aparecida de Goiânia (GO) in Brazil. Using a quantitative approach with ordinary least squares (OLS) regression, this study examines the influence of technology on the reduction of homicides and on the Federation of Industries of the State of Rio de Janeiro (FIRJAN) education, health, employment, and income indexes from 2014 to 2024. The theoretical framework addresses the concepts of smart cities, human development, and urban security, supported by international standards and references, including the works of Amartya Sen and the ISO 37122 guidelines. The results reveal negative correlations between the presence of cameras and violence rates; however, the robustness tests indicate model limitations, highlighting the need for additional variables and further methodological refinement. Overall, the study contributes to the broader debate on the role of technology in sustainable urban development, particularly in contexts characterized by medium socioeconomic complexity.

KEYWORDS: smart cities / video surveillance / public safety / FIRJAN indicators / human development

CIUDADES INTELIGENTES Y SEGURIDAD PÚBLICA: EVIDENCIA EMPÍRICA SOBRE VIDEOVIGILANCIA Y DESARROLLO EN LA CIUDAD DE APARECIDA DE GOIÂNIA

RESUMEN. Este artículo analiza la relación entre la implementación de la videovigilancia urbana, como parte de la estrategia de ciudades inteligentes, y los indicadores sociales y de seguridad pública en Aparecida de Goiânia (GO), Brasil. Mediante un enfoque cuantitativo con regresión de mínimos cuadrados ordinarios (MCO), el estudio examina la influencia de la tecnología en la reducción de homicidios y en los índices de educación, salud, empleo e ingresos de la Federación de Industrias del Estado de Río de Janeiro (FIRJAN) durante el período 2014–2024. El marco teórico aborda los conceptos de ciudades inteligentes, desarrollo humano y seguridad urbana, apoyándose en normas y referencias internacionales, incluyendo los aportes de Amartya Sen y las directrices ISO 37122. Los resultados revelan correlaciones negativas entre la presencia de cámaras y las tasas de violencia; sin embargo, las pruebas de robustez indican limitaciones del modelo, lo que resalta la necesidad de incorporar variables adicionales y de un mayor refinamiento metodológico. En conjunto, el estudio contribuye al debate más amplio sobre el papel de la tecnología en el desarrollo urbano sostenible, particularmente en contextos caracterizados por una complejidad socioeconómica media.

PALABRAS CLAVE: ciudades inteligentes / videovigilancia / seguridad pública / indicadores FIRJAN / desarrollo humano

INTRODUCTION

The phenomenon of accelerated urbanization, observed on a global scale in recent decades, presents a complex set of challenges for public management and the quality of life of its citizens. The growth of cities has driven the search for innovative solutions to chronic urban problems, ranging from infrastructure and mobility to public safety and social well-being. In this context, the concept of a Smart City emerges as a promising approach, proposing the integration of technology, especially Information and Communication Technologies (ICT), with urban management to optimize services, improve sustainability, and promote citizen participation (Abreu & Marchiori, 2019; Lazzaretti et al., 2019).

Although the use of technology to enhance urban life is not a novel concept, the conceptualization of the smart city as a social issue and a subject of scholarly inquiry is relatively new and continues to evolve. The term smart began to be used in the 1990s, initially referring to the infrastructures required to support information technology. Over time, however, the concept has evolved to encompass a broader perspective, in which digitalization is not an end in itself but rather a tool to support multiple dimensions of urban life, including governance, economy, environment, and quality of life (Abreu & Marchiori, 2019). Within the context of smart cities, there is an ongoing debate regarding the centrality of technology. Some approaches position technology as a key driver in a top-down framework, whereas others argue that technology should be adapted to the needs of citizens in a bottom-up approach (Dameri, 2013, as cited in Alves, 2019, p. 7).

In Brazil, the discussion on Smart Cities has been gaining ground, in both academia and on government agendas. Initiatives such as the Brazilian Charter for Smart Cities, prepared by the Ministry of Regional Development in collaboration with several societal sectors, represent an effort to establish a national public agenda for digital transformation in Brazilian cities. This Charter seeks to integrate the agendas of urban development and digital transformation, guided by the perspectives of environmental, urban, social, cultural, economic, financial, and digital sustainability. It proposes an expanded conception of the smart city, acknowledging the complexity and distinct characteristics of Brazil's various territories to reduce socio-spatial inequalities (Brasil. Ministério do Desenvolvimento Regional, 2021).

Despite increasing interest, the concept of smart cities still lacks consensus and a unifying definition in Brazil. National research on the topic is multidomain, encompassing fields such as administration, information systems, engineering, and architecture. There is a conceptual predominance that links ICTs with quality of life, as noted by Lazzaretti et al. (2019). However, the mere application of technology does not guarantee that a city becomes smart. The transformation depends on coordinated action between federal spheres, clear strategies, and overcoming barriers such as lack of technical knowledge, investment difficulties, and resistance to change (Roland, 2019).

The evaluation of the performance and degree of "intelligence" of cities is critical to directing planning and public policies. International tools and standards, such as ISO 37122:2019 and its Brazilian version, ABNT NBR ISO 37122:2020, proposed by the Associação Brasileira de Normas Técnicas (ABNT, 2020), offer sets of indicators to assess progress towards a smart city. However, it is emphasized that the evaluation should not focus only on the technological aspect but consider multiple perspectives, including sustainability, resilience, quality of life, and social aspects (Abreu & Marchiori, 2019).

Within this framework, the relationship among Smart Cities, Regional Development, and the application of social and security indicators assumes particular relevance. The socioeconomic development of Brazilian cities varies considerably, and these disparities can influence their capacity to implement smart city projects (Jordão, 2016). Synthetic indicators such as the Human Development Index (HDI) have historically been used to measure social conditions, despite criticisms about their universal application and disregard for regional particularities (Guimarães & Jannuzzi, 2005).

Similarly, violence is a social problem that directly impacts quality of life and development, with different forms of manifestation (Minayo, 2006). Indexes, such as the Criminalized Violence Index (CVI), seek to facilitate the understanding of the factors that influence criminal dynamics and support public policies focused on violence prevention and control. Social cohesion, defined as the level of coexistence between different groups in a city, is also a crucial factor for the urban context (Takiya et al., 2019).

Considering the complexity of the topic and the need to adapt the Smart City concept to the Brazilian context—with its regional particularities and social challenges—this research project aims to examine the interrelation between the implementation of technological initiatives and the performance of social and security indicators. Specifically, this study focuses on the city of Aparecida de Goiânia, which is currently implementing a Smart City project. The initiative encompasses several sectors, including public security, and the municipality has invested in technologies such as video surveillance systems. The relevance of security and human development indicators for the consolidation of Smart Cities in the Brazilian regional context still requires deeper exploration.

Within this context, this dissertation project aims to analyze the relationship between the implementation of information and communication technologies—such as video surveillance—within a Smart City initiative and the potential changes in security indicators (homicides) and quality-of-life indicators (education, health, employment, and income) in a medium-sized Brazilian city. To this end, it seeks to answer the following research questions: How did the implementation of video surveillance in Aparecida de Goiânia alter the Homicide rates and the Federation of Industries of the State of Rio de Janeiro (FIRJAN) indexes for education, health, employment, and income? The study is justified by the need to deepen the understanding of how Smart City initiatives directly contribute to citizens' well-being and to urban development, taking into account the

specificities of the Brazilian context and the importance of multidimensional indicators that extend beyond mere technological application.

RELATED WORK

The study of the relationship between the implementation of video surveillance technologies in the context of smart cities and performance in public safety and social development has been gaining relevance in the Brazilian academic and governmental scenario. This section aims to position the current article—which investigates this relationship in Aparecida de Goiânia (GO) using a quantitative approach—in comparison with other prominent works focused on Brazilian cities, such as Foz do Iguaçu (Chichoski & Marquardt, 2023), Recife (Ferreira et al., 2023), and Manaus (Filho et al., 2024).

Similarities and Conceptual Alignment

The analyzed articles significantly show strong convergence in their emphasis on video surveillance as a central instrument for public safety and contemporary urban management in Brazil.

Video Surveillance as a Smart City Tool

All studies characterize video monitoring as a technological innovation applied to public safety and aligned with smart-city principles. Across the examined cities, the primary objective is to employ this technology for crime prevention and intervention. In Manaus, for instance, the Smart Siege (Cerco Inteligente) system was implemented to enhance police response efficiency and effectiveness, thereby strengthening crime-fighting efforts.

Emphasis on Integration and Governance

There is a consensus on the need for integration and cooperation between public security agencies and municipal administration for the system's effectiveness.

- In the main article, the Aparecida de Goiânia initiative is part of a broader Smart City project focused on data-driven governance.
- In Foz do Iguaçu, the system's effectiveness is linked to the structure of the Integrated Municipal Management Cabinet (GGIM), which brings together 21 institutions for dialogue and deliberation.
- In Recife, the system is managed by the Integrated Social Defense Operations Center (CIODS), where the coordination between agencies such as the police and the fire department is considered favorable to advancing the Smart City concept.

- In Manaus, the system is under the responsibility of the Integrated Security Operations Center (CIOPS) and the Integrated Command and Control Center (CICC), which coordinate and integrate state security forces (military police, civil police, and military fire department).

Social Benefit and Sense of Security

The studies agree that video monitoring provides socioeconomic gains and, crucially, increases the sense of security in the population and public road users. In addition to combating crime, video monitoring functions as a social support mechanism, assisting in social defense activities such as managing traffic accidents, monitoring areas of drug use, and even enabling the rescue of individuals during suicide attempts, as reported in the Manaus case.

Institutional and Resource Challenges

All Brazilian contexts face barriers to the full implementation of the system. Works from Foz do Iguaçu (Chichoski & Marquardt, 2023) and Recife (Ferreira et al., 2023) report challenges related to insufficient staffing to operate the platforms—such as the overload of approximately 50 cameras per operator in Foz do Iguaçu—and limited government support for sustained investments and the acquisition of high-value equipment, respectively.

Methodological and Focal Differences

The article on Aparecida de Goiânia differs markedly from the others in its methodological approach, as it is one of the few that attempts to quantify the impact of technology on multidimensional development indicators (Table 1).

Table 1

Comparative Synthesis of Approaches

Characteristic	Article (Aparecida de Goiânia)	Chichoski y Marquardt (Foz do Iguaçu)	Ferreira, Novaes y Macedo (Recfe)	Filho, Santos y Lima (Manaus)
Main Methodology	Quantitative (OLS Regression)	Qualitative/Descriptive/Exploratory (Empirical/bibliographical research)	Qualitative (Case study, interviews)	Qualitative/Exploratory (SSP-AM Data, bibliographical)

(*continúa*)

(*continuación*)

Impact Focus	Statistical correlation between cameras and multidimensional Human Development indicators (IFDM: Education, Health, Employment/ Income) and Homicides	Operational contribution and border surveillance	Innovation in the public sector and analysis of innovation classification (incremental/technological)	Advantages and disadvantages of the "Smart Siege" system and use of images in investigations
Critical Result	Found negative correlations (cameras vs. violence), but the OLS model showed statistical limitations (heteroscedasticity, specification problems)	Highlighted the need to improve integration and investment in human resources/ maintenance costs	Insufficient government support and technological deficit (lack of AI/facial recognition)	Concerns about privacy and the risk of falling into the "myths of the video image" (Silbey, 2004)
Technological Scale	Large scale infrastructure (3,275 cameras installed, 720 km of fiber optics)	Smaller scale (288 cameras in total, migrating from radio to 240 km of fiber optics)	Existing system with technological deficit, but with expansion plan for 2,000+ cameras and addition of AI analytics	Medium scale (more than 500 cameras)

THEORETICAL FOUNDATION

The concept of Smart Cities emerged in response to contemporary urban challenges, proposing the strategic use of Information and Communication Technologies (ICT) to improve public services, promote sustainability, social inclusion, and foster regional development. The Brazilian Charter for Smart Cities reinforces this vision by highlighting that a smart city is not limited to the adoption of technologies, but rather to their use as a means of improving quality of life, promoting social justice, and reducing inequalities.

In this sense, the approach adopted by Covas and Covas (2020) shows that the smartification of territories is directly linked to the ability to mobilize collective intelligence, integrating economic, social, environmental, and cultural dimensions. Technology, therefore, acts as a catalyst; however, its effectiveness depends on the articulation between public and private actors and the civil society, which positions technology as a means to achieve objectives and not as a central element, reinforcing the more humanistic and integrative aspect of the proposal for implementing smart cities.

Sancino and Lorraine (2020) complement this perspective by arguing that leadership is a crucial element in smart city governance, fundamental for aligning technological objectives with social interests and local specificities. The absence of collaborative governance and solid public policies can constrain the benefits of digital transformation, particularly in developing countries.

The measurement of a city's degree of intelligence is a methodological challenge widely discussed in the literature. The international standard ABNT NBR ISO 37122:2020 establishes a set of indicators that enables the evaluation of the smart city performance, covering aspects such as economy, education, environment, security, mobility, and governance. However, Abreu and Marchiori (2023) point out that, although the standard is an important reference, it has limitations, such as the absence of specific indicators for data privacy and digital security, as well as a fragmentation among the topics of sustainability, intelligence, and resilience.

In the Latin American context, Gomes et al. (2016) propose a specific index that incorporates the region's socioeconomic and structural particularities and integrates the Sustainable Development Goals (SDGs) across the dimensions of economy, environment, governance, infrastructure, and social aspects. This approach aims to address the limitations of applying global models to the realities of Latin American and Caribbean countries.

The relationship of smart cities, public safety, and human development is central to the debate on sustainable urban development. Jordão (2016) argues that, beyond digitalization, smart city initiatives must be aligned with local social needs, prioritizing inclusion, social cohesion, and the reduction of inequalities. This perspective expands the traditional understanding of smart cities by emphasizing their social and human-centered dimensions.

The Human Development Index (HDI) has been widely used as a metric for assessing social well-being in cities, although it presents methodological limitations, as pointed out by Guimarães and Jannuzzi (2005). These authors emphasize that, although the HDI is a relevant indicator, it can mask internal inequalities, making it necessary to complement it with other specific metrics, such as the CVI developed by Lira (2013), which provides a territorialized analysis of crime and serves as a robust tool for supporting public security policies.

From an economic perspective, the concept of development has historically been linked to the notion of social well-being and, for a long period, was interpreted as synonymous with economic growth, especially in its regulatory aspects. However, with the advancement of theoretical reflections and the emergence of multidimensional approaches, this conception was gradually reformulated, incorporating variables that extend beyond the limits of wealth generation and focus, above all, on the elements that directly shape the population's quality of life.

At the end of the 20th century, this discussion was significantly enriched by the studies of economist Amartya Sen, whose theoretical contribution repositioned development as a process centered on expanding the real freedoms that individuals are able to exercise. For Sen (2000), development must be understood beyond an exclusively economic perspective, incorporating elements related to human capabilities and quality of life. This implies considering, in an integrated manner, both the individual and collective dimensions of well-being.

Sen's central critique lies in the observation that the mere generation of wealth is not, in itself, a sufficient to promote social well-being. This outcome depends, above all, on people's agency and their substantive freedoms—that is, the effective capacity to make choices and pursue valued life goals within contexts where the necessary means are available.

In this regard, Sen's conception of development laid the foundation for the creation of new instruments to evaluate quality of life, among which the Human Development Index (HDI) stands out. This indicator incorporated fundamental dimensions such as health, education, and income, providing an alternative and more comprehensive metric for the comparative analysis of development across countries and regions.

In the Brazilian context, Sen's theoretical proposal inspired the formulation of the Firjan Municipal Development Index (IFDM), created by FIRJAN. Although it is based on categories similar to those of the HDI—income, health, and education—the IFDM differs methodologically in the variables considered and the weights assigned in constructing the index, reflecting a contextualized application of the human development approach at the local level.

The city of Manaus, for example, adopted video surveillance systems as part of its smart city strategy, aiming to reduce crime rates and increase the perception of security. However, studies such as that of Filho et al. (2024) indicate that, while video surveillance can positively impact crime prevention and the investigation of offenses, it also raises concerns regarding privacy, the ethical use of data, and the potential to reinforce inequalities, particularly in the absence of well-structured digital governance.

The use of data is a fundamental component in the management of smart cities. However, Rezende et al. (2019) warn that there is a fine line between making data accessible to enhance public services and infringing on citizens' privacy. The General Data Protection Law (LGPD) in Brazil and the Access to Information Law (LAI) are legal frameworks that must be adhered to in the implementation of smart city projects. Additionally, Paes et al. (2021) argue that data governance should be centered on sustainable urban metabolism models, in which the collection, analysis, and use of data serve not only to enhance operational efficiency but also to promote social and environmental sustainability.

Brazilian experiences in developing smart cities reveal a range of challenges, including the lack of integration among public policies, insufficient technological infrastructure in many regions, and low maturity in the data culture. Furthermore, Martinelli et al. (2020) indicate that, although the benefits of smart cities are positively perceived, many initiatives still lack strategic planning, particularly with regard to social inclusion and meaningful citizen participation.

In the specific case of the city of Aparecida de Goiânia, the focus of this study, initiatives such as the implementation of the video surveillance system are part of a broader digital transformation effort known as the Smart City Project. The municipality of Aparecida de Goiânia has gained national prominence through the implementation of a comprehensive smart city project, which integrates advanced technological infrastructure, data-driven governance, and public security strategies grounded in urban intelligence. The project represents a convergence of technological innovation, regional development, and the promotion of security, aligning with the guidelines of the Brazilian Charter for Smart Cities and the recommendations of ABNT NBR ISO 37122.

The city of Aparecida de Goiânia took a major step towards consolidating its smart city model with the inauguration of the Center for Technological Intelligence (CIT). According to the City Hall, the CIT will serve as the "brain" of the city, receiving real-time information from approximately 700 cameras initially installed at strategic locations, including streets, squares, schools, health units, and public offices. The objective is to enhance the city's urban monitoring capacity by integrating data on security, mobility, infrastructure, and public services, thereby optimizing decision-making and incident prevention in accordance with the principles of digital governance and data-driven management.

To support this operation, an extensive infrastructure of 720 km of fiber-optic cables was installed, covering the city's entire urban territory. This high-capacity network ensures the interconnectivity of public systems, enabling high-speed data transmission with low latency—an essential requirement for the operation of critical services such as security, mobility, and integrated urban management. Additionally, the municipality established its own data center, with a storage capacity of 4,600 TB. This infrastructure supports video surveillance operations and the analysis of large volumes of data from multiple areas of public administration, enhancing the analytical and predictive capabilities of municipal management.

The system became operational in 2024 with 3,275 cameras to assist security forces in combating crime and supporting other public services. Of these, 1,632 cameras were installed in schools, 1,140 in health units, and 503 dedicated to urban monitoring. The installed cameras are high-resolution, with a range of up to 2,000 meters, and are monitored in real time by the Center for Technological Intelligence. The project resulted from research conducted at the Laboratory for Research, Development, and Innovation in Interactive Media at the Federal University of Goiás.

Based on the information presented, it is evident that the implementation of the Smart City Project of Aparecida de Goiânia (GO) has enabled significant improvements in the municipality's urban infrastructure. However, it is necessary to assess whether these initiatives have effectively influenced indicators such as the HDI and CVI, and whether such changes have made a significant contribution to consolidating the concept of a smart city within the regional context.

Understanding regional development, especially in peripheral or medium-sized urban contexts such as Aparecida de Goiânia, requires an approach that goes beyond explanations based solely on geographical factors or the accumulation of physical capital. Contemporary theories of regional development incorporate elements such as innovation, technology diffusion, institutional capabilities, and territorial cooperation networks. Within this framework, the neo-Schumpeterian approach assumes a central role, considering that regional development is intrinsically linked to the ability to generate, adapt, and diffuse innovations at a local scale, emphasizing the evolutionary dynamics of territories.

By shifting the focus from physical capital to technological, cognitive, and institutional capital, neo-Schumpeterian theorists assign particular importance to the role of the State and public innovation policies as drivers of regional transformation. The smart city, as an urban management strategy based on digital technologies and integrated information systems, represents a materialization of these ideas, embedding innovation as a fundamental component. As examined in this work, the implementation of urban video surveillance systems not only expands the State's capacity to respond to crime but also generates positive externalities related to security, public trust, investment attraction, and improvements in quality-of-life indicators.

The articulation between the smart city project and social indicators underscores the relevance of the neo-Schumpeterian approach, emphasizing that regional development results from specific trajectories of learning, institutional capacity, and social and technological innovation. From this perspective, video surveillance should not be regarded merely as a technical artifact but as part of a broader process of institutional modernization and the development of endogenous capabilities. This aligns with the concept of regional innovation systems, in which the territory functions as an active space for knowledge production and for the coordination of public, private, and social actors, capable of fostering development based on context-specific and sustainable solutions.

The reviewed literature shows that the concept of Smart Cities goes beyond the simple adoption of technologies. It is a complex process that requires collaborative governance, the development of indicators that reflect local realities, data protection, and a focus on human development and public safety. The intersection between technology, social development, and urban security is, therefore, central to the consolidation of smart cities in Brazil, especially in regional contexts like that of Aparecida de Goiânia.

METHODOLOGY

This research project adopts a quantitative, qualitative, and exploratory methodological approach, based on a case study of the implementation of video surveillance in the city of Aparecida de Goiânia.

The research considers crime hotspots from 2014 to 2024, as well as quantitative data, with a temporal division between the periods before (2014–2017) and after (2018–2024) the implementation of video surveillance in the city of Aparecida de Goiânia, GO. The crimes comprising the CVI proposed by Lira (2013), presented at the 2nd Consad Congress on Public Management, will be considered in this study. From a quantitative point of view, the data analysis will proceed with the use of descriptive statistics, followed by multivariate statistical analysis using the OLS method, with the estimation of models, analysis of results, and analysis of residuals.

In seeking to establish the methodological relevance of this research, it was found that studies examining potential relationships between socioeconomic indicators and crime are relatively common. Authors such as Silva et al. (2017) and Oliveira (2018) serve as examples. However, studies that intend to represent the phenomenon from a multivariate perspective are still rare. On the other hand, multivariate statistical methods have proven efficient in many criminological explanations.

The composition of CVI, as shown in Figure 1, comprises nine categories of crimes, with the combination of indicators formed by a set of criminal variables. According to Lira (2013), by correlating these variables with socioeconomic information, the CVI aims to facilitate the understanding of the structural factors that likely influence criminal dynamics, as well as to support the development of public policies and strategies for the prevention, control, and mitigation of violence.

Figure 1

CVI

Note. Adapted from *Índice de Violência Criminalizada (IVC) [Criminalized Violence Index (CVI)]*, by P. Lira, 2013, in *II Congresso Consad de Gestão Pública – Painel 62: Gestão em segurança pública*, Vitória (https://consad.org.br/wp-content/uploads/2013/02/%C3%8DNDICE-DE-VIOL%C3%8ANCIA-CRIMINALIZADA-IVC3.pdf).

Based on the characteristics of the indicators presented above, it can be argued that they represent typologies directly reflecting the essential conditions related to societal security and can be considered for evaluation within the scope of this research. For a better understanding, the variables included in the CVI dataset are presented in Table 1, grouped by specific indicators and accompanied by acronyms to facilitate data processing and analysis (Table 2).

Table 2

Variables of CVI

ID	Indicator	Variables
LCP	Lethal crimes against the person	Homicides, robberies, and attempted homicides
NLCP	Non-lethal crimes against the person	Bodily injury, fights, violence, and threats
SCC	Serious crimes against custom	Rape and indecent assault
RC	Robbery crimes	Total asset thefts
TC	Theft crimes	Total asset thefts
WAC	Weapons and ammunition crimes	Illegal possession of weapons, illegal manufacturing of weapons and ammunition, seizure of firearms, and discharge of weapons
DTC	Drug trafficking crimes	Trafficking of marijuana, cocaine, and other narcotics
CPUD	Crimes of possession and use of drugs	Possession and use of marijuana, cocaine, and other narcotics
CD	Crime of drunkenness	Drunkenness

Note. Adapted from *Índice de Violência Criminalizada (IVC) [Criminalized Violence Index (CVI)]*, by P. Lira, 2013, in *II Congresso Consad de Gestão Pública – Painel 62: Gestão em segurança pública*, Vitória (https://consad.org.br/wp-content/uploads/2013/02/%C3%8DNDICE-DE-VIOL%C3%8ANCIA-CRIMINALIZADA-IVC3.pdf).

The variables Theft Crimes (TC) and Robbery Crimes (RC) are categories of property crimes included in the CVI, which aims to facilitate the understanding of criminal dynamics. Within the composition of the CVI, the article states that both categories, RC and TC, are represented by the descriptive variable "Total Asset Thefts". However, the core distinction between TC and RC lies in the manner in which the asset is taken: RC implies the use of violence or serious threat against a person, whereas TC involves the taking of the asset without such means. Thus, although both variables in the CVI share the same description, namely "total asset thefts", the categorical differentiation is crucial for the analysis of criminalized violence, with RC being a crime of violence against the person and TC being a strictly property crime.

Once the data had been accessed, the OLS method was applied to estimate the relationship between the video surveillance implemented in the city of Aparecida de Goiânia and the selected explanatory variables. OLS is an estimation method widely

used in econometrics, aiming to minimize the sum of the squares of the residuals, which are the differences between the observed values of the dependent variable and the values predicted by the model. OLS relies on a set of classical assumptions that guarantee its estimators are unbiased, consistent, and efficient. One of these assumptions is homoscedasticity, which requires the error term to exhibit constant variance across all observations. As per the analysis performed, this assumption was violated, indicating the presence of heteroscedasticity in the model's error term.

To assess the robustness of the results and verify the adequacy of the model estimated via OLS, several diagnostic tests were performed on the residuals and explanatory variables. The characteristics of the tests used are:

- **Breusch-Pagan test**: This test is used to assess the presence of heteroscedasticity in the residuals. The null hypothesis assumes homoscedasticity, meaning that the error term exhibits constant variance. Rejecting the null hypothesis provides evidence of heteroscedasticity.
- **Multicollinearity test (variance inflation factor, VIF)**: This test examines the existence of multicollinearity among the explanatory variables.
- **RESET test**: This is a residual-based test used to evaluate model specification. A high *p* value indicates that the model is adequately specified.
- **Durbin-Watson test**: This test evaluates the presence of autocorrelation in the residuals. The alternative hypothesis states that the autocorrelation coefficient (ρ) differs from zero. In addition to the original linear model, a log–log specification will also be estimated by transforming all variables into their natural logarithms. The same diagnostic tests will be applied to this transformed model to determine whether the logarithmic functional form improves the robustness of the results.

RESULTS AND DISCUSSION

The data used in this analysis corresponds to a time series covering the period from 2014 to 2024, enabling the examination of trends in key social, economic, and public security indicators over the past decade. The dataset was structured around the variable *lethal crimes against the person* (LCP), which is part of the CVI, and was labeled "homicides" for clarity.

The variables used are detailed in the Table 3 and include indicators related to the number of homicides, education, health, employment, and income (based on the IFDM), as well as the variable representing the coverage of urban video surveillance. This database was statistically processed and organized with the support of the RStudio software, enabling the execution of econometric tests with methodological robustness and alignment with the best practices of quantitative research applied to regional analysis.

Table 3

Variables Used in the Analysis

Variable	Description	Behavior in analysis
Homicides	Number of homicides in the period	Explanatory variable
IFDM	FIRJAN Municipal Development Index	Explanatory variable
CAM	Presence or absence of cameras installed in the period	Explanatory variable
EDU	FIRJAN Education Index	Explanatory variable
SAL	FIRJAN Salary Index	Explanatory variable
ER	FIRJAN Employment and Income Index	Explanatory variable

The descriptive statistics of the variables were used to gain an initial understanding of their behavior, as presented in Table 2, which summarizes the following characteristics (Table 4):

Table 4

Descriptive Statistics

Variables	Homicides	CAM	EDU	SAL	ER
Mean	267.5	1.00	0.4719	0.5843	0.7919
Minimum	139.0	0.00	0.3726	0.5291	0.7395
Maximum	342.0	1.00	0.6133	0.6694	0.8823
Variance	5280.16	0.24	0.005858829	0.001683414	0.002853324
Standard deviation	72.66470945	0.489897949	0.076542987	0.04102943	0.053416511
Coefficient of variation	183.6963855	0.002939388	0.000358367	0.00024065	0.00043012

To estimate the relationship between the *Homicides* variable and the explanatory variables, the OLS method was used. OLS is a widely used estimation method in econometrics, that aims to minimize the sum of squared residuals, which represent the differences between the observed values of the dependent variable and those predicted by the model. The classical assumptions of OLS ensure that its estimators are unbiased, consistent, and efficient. One of the assumptions of OLS is homoscedasticity, which requires the error term to have constant variance.

The data analysis and econometric model estimation were conducted with the help of RStudio software, an open-source statistical tool widely used in the field of applied econometrics. RStudio was used to perform data pre-processing, which included organizing variables, conducting descriptive analyses, and checking for inconsistencies, as well as estimating regressions using the OLS method. The programming environment also

allowed for the application of essential statistical tests to evaluate the model's assumptions, such as homoscedasticity, normality of residuals, multicollinearity, and functional specification. The use of RStudio facilitated the transparency, reproducibility, and robustness of the empirical analysis, in line with the best practices of quantitative research.

A model was estimated with *Homicides* as the dependent variable and the remaining variables as explanatory variables, and all necessary diagnostic tests were conducted to assess the robustness of the results, as presented below.

Estimated Model

Homicides = 817.47 – 44.55 CAM – 516.85 EDU – 101.89 SAL - 293.32 ER + error

The estimated model was implemented using RStudio software, and the results are presented in the Table 5.

Table 5

Results of the Model Execution in RStudio

	Estimate standard	Error	Value	Pr (>t)
(Intercept)	817.47	258.22	3.166	0.0249 *
CAM	- 44.55	41.01	-1.086	0.3270
EDU	- 516.85	335.97	-1.538	0.8470
SAL	- 101.89	501.46	-0.203	0.8470
ER	- 293.32	227.73	227.73	0.2541
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

The results of the estimated model, including the behavior of the variables, are presented below, with the statistical significance levels indicated at the bottom of the table. Residuals are presented in the Table 6.

Table 6

Residuals

Residual standard error	Multiple R^2	Adjusted R^2	*F*-statistic	*p* value
33.13 on 5 degrees of freedom (df)	0.8961	0.813	10.78 on 4 and 5 df	0.01128

To evaluate the robustness of the results and verify the adequacy of the model estimated by OLS, several diagnostic tests were performed on the residuals and explanatory variables. The characteristics of the tests used, as described in the sources, are:

- **Breusch-Pagan test**: This test is used to assess the presence of heteroscedasticity in the residuals. The null hypothesis assumes homoscedasticity, meaning that the error term exhibits constant variance. Rejecting the null hypothesis provides evidence of heteroscedasticity. An analysis where the value points to a "very low (<2)" *p* value leads to the rejection of the null hypothesis.
- **Shapiro-Wilk test**: This test assesses whether the residuals follow a normal distribution. In this analysis, a *p* value greater than 0.5 indicates that the residuals deviate from normality, thereby violating the normality assumption.
- **Multicollinearity test (VIF)**: This test examines the existence of multicollinearity among the explanatory variables. If the analysis yields a VIF value greater than 10, it is considered indicative of severe multicollinearity. Conversely, a VIF value below 10 suggests the absence of severe multicollinearity, which is favorable for the stability of the estimated coefficients.
- **RESET test**: This is a residual-based test used to evaluate model specification. A high *p* value indicates that the model is adequately specified. If the analysis indicates a low *p* value ("ideally >= 0.05"), this suggests a potential specification problem, which may result from omitted relevant variables or from the relationship between variables not being strictly linear.
- **Durbin-Watson test**: This test checks for the existence of autocorrelation in the residuals. The alternative hypothesis of the test states that the autocorrelation coefficient (ρ) differs from zero. According to the cited sources, a *p* value greater than 0.05 in the normality test indicates that the residuals deviate from normality, while for the multicollinearity and autocorrelation tests, results are considered problematic if the *p* value is below 0.02. Applying a logarithmic functional form may improve the robustness of the results.

A linear model and a log-log model were estimated, as presented in Table 7:

Table 7

Comparison of Results Between the Linear and Log-Log Models

Test / Metric	Linear model	Log-Log model
Breusch-Pagan test		
BP statistic	4.0906	3.2673
df	4	4
p value	0.3939	0.5141
Shapiro-Wilk test		
W statistic	0.94076	0.89527
p value	0.5615	0.1943

(*continúa*)

(*continuación*)

Test / Metric	Linear model	Log-Log model
Multicollinearity test (VIF)		
VIF (CAM / lCAM)	3.67902	4.219989
VIF (EDU / lEDU)	6.027090	5.489638
VIF (SAL / lSAL)	3.857900	3.365490
VIF (ER / lER)	1.348632	1.251529
RESET test		
RESET statistic	3.6872	3.0856
df1	2	2
df2	3	3
p value	0.1555	0.1871
Durbin-Watson test		
DW statistic	1.663671	1.857104
p value	0.044	0.112

Analysis of the Linear Model Results

The estimated equation for the linear model, using the OLS method, can be represented as follows:

Homicides = $\beta_0 + \beta_1$ VideoSurveillance + β_2 Education + β_2 Health + β_3 Employment/ Income + error

Based on the data incorporated into the model, the equation can be represented as follows:

Homicides = 817.47 – 44.55 CAM – 516.85 EDU – 101.89 SAL - 293.32 ER + error

The OLS estimation results indicate that:

- The intercept is 817.47 and is statistically significant. The CAM variable has a coefficient of -44.55 and is also statistically significant, suggesting a negative relationship between the presence of cameras and the Homicides variable.
- The EDU variable has a coefficient of -516.85 and is highly statistically significant, suggesting a negative relationship between the FIRJAN Education Index and the *Homicides* variable.
- The SAL variable has a coefficient of -101.89 and is statistically significant, suggesting a negative relationship between the FIRJAN Health Index and the *Homicides* variable.
- The model's goodness of fit, measured by the multiple R-squared, is 0.8961, and the adjusted R-squared is 0.813. The overall F-test is highly significant

(F-statistic = 10.78, *p* = 0.01128), indicating that the explanatory variables, collectively, are relevant in explaining Homicides.

The diagnostic tests for the linear model yielded the following results and analyses:

- **Breusch-Pagan test**: BP = 4.0906, df = 4, *p* = 0.3939. **Analysis**: The test indicated that the residuals are heteroscedastic, indicating a problem of heteroscedasticity in the model. The null hypothesis of homoscedasticity is accepted, considering that the *p* value showed a considerable value (>2).
- **Shapiro-Wilk test**: W = 0.94076, *p* = 0.5615. **Analysis**: The test indicates that the residuals do not have a normal distribution due to the *p* > 0.5, thereby satisfying the normality assumption.
- **Multicollinearity test (VIF)**: VIF for CAM = 3.679022, EDU = 6.027090, SAL = 3.857900, ER = 1.348632. **Analysis**: The VIF results do not indicate the presence of multicollinearity, as all values are below 10, which is favorable for the stability of the estimated coefficients.
- **RESET test**: RESET = 3.6872, df1 = 2, df2 = 3, *p* value = 0.1555. **Analysis**: The test indicates that the model has a specification problem, suggesting it may be misspecified, possibly with important variables omitted, given that the *p* value was very low (ideal *p* value >= 0.05). The relationship between the variables may not be strictly linear.
- **Durbin-Watson test**: DW statistic = 1.663671, *p* value = 0.044. **Analysis**: The test result suggests that the error has a normal distribution due to the *p* value < 5. According to additional analysis, the multicollinearity and autocorrelation tests would not be good because the *p* value should be greater than 2 (Note: As described in the methodology section, this analysis of the DW *p* value differs from the standard interpretation).

Based on the considerations above for the linear model, the results are not entirely satisfactory, suggesting that additional relevant variables should be included.

Analysis of the Log-Log Model Results

After transforming the variables to their natural logarithm, the estimated equation for the log-log model is as follows:

Homicides = 3.96143 – 0.07529 lCAM – 1.15769 lEDU – 0.68532 lSAL - 1.40505 lER + error

The OLS estimation results for the log-log model indicate that:

- The intercept is 3.96143, and is highly statistically significant. None of the other variables showed statistical relevance.

In terms of fit, the Multiple R^2 for the log-log model is 0.8941, and the Adjusted R^2 is 0.8094. These values are slightly higher than those of the linear model, suggesting a marginally better fit. The overall F-test of the log-log model does not show high statistical significance (*F*-statistic: 10.55, *p* value: 0.01181).

The diagnostic tests for the log-log model yielded the following results and analyses:

- **Breusch-Pagan test**: BP = 3.2673, df = 4, *p* = 0.5141. **Analysis**: Unlike the linear model, the test indicated that the residuals are heteroscedastic (heteroscedasticity problem). The null hypothesis of homoscedasticity is rejected, considering that the *p* value was very low (<2). The presence of heteroscedasticity violates one of the classical assumptions of the OLS method.
- **Shapiro-Wilk test**: W = 0.89527, *p* = 0.1943. **Analysis**: Unlike the linear model, the test indicates that the residuals have a normal distribution due to the *p* value < 0.5, thereby satisfying the normality assumption.
- **Multicollinearity test (VIF)**: VIF for lCAM = 4.219989, lEDU = 5.489638, lSAL = 3.365490, and lER = 1.251529. **Analysis**: The VIF results do not indicate the presence of multicollinearity, as all values are below 10, which is favorable for the stability of the estimated coefficients.
- **RESET test**: RESET = 3.0856, df1 = 2, df2 = 3, *p* = 0.1871. **Analysis**: The test indicates that the log-log model also has a specification problem, suggesting it may be misspecified, possibly with important variables omitted, given that the *p* value was very low (ideal *p* value >= 0.05). The relationship between the variables may not be strictly linear.
- **Durbin-Watson test**: DW statistic = 1.857104, *p* = 0.112. **Analysis**: The result suggests that the error has a normal distribution due to the *p* value < 5. According to additional analysis, the multicollinearity and autocorrelation tests would not be good because the *p* value should be greater than 2 (Note: As described in the methodology section, this analysis of the DW *p* value differs from the standard interpretation).

Based on the considerations above, even after transforming the variables into the log–log form, the results of the diagnostic tests (heteroscedasticity, residual normality, and model specification) are not entirely satisfactory. The analysis suggests that the model, in both forms, still exhibits issues, indicating the need for improvements, such as the potential inclusion of additional explanatory variables to better capture the variation in *Homicides* and address the problems of heteroscedasticity and misspecification. The absence of severe multicollinearity represents a positive aspect in both models.

In summary, the results of the diagnostic tests indicate that neither OLS model (linear nor log–log) fully satisfies all of the classical assumptions, suggesting the need

for improvements, such as the inclusion of additional explanatory variables, to achieve more robust and satisfactory results.

Threats to the Validity of the Results

Using the OLS method the results of the quantitative analysis are subject to significant threats to internal validity, preventing the findings from being considered fully satisfactory or robust without further methodological refinement. Internal validity was compromised due to violations of classical OLS assumptions. In the linear model, the Shapiro-Wilk test indicated that the residuals did not present a normal distribution, since the *p* value (0.5615) was greater than 0.5, thus failing to adhere to the normality assumption.

Although the transformation to the log-log model resolved the issue of residual normality (*p* value = 0.1943), it exposed a problem of heteroscedasticity, according to the Breusch-Pagan test analysis in the log-log model, which led to the rejection of the null hypothesis of homoscedasticity. The presence of heteroscedasticity, together with the deviation from normality in the linear model, is a key violation of the classical OLS assumptions, affecting both the efficiency and consistency of the estimators.

The main threat to construct and external validity arises from the model misspecification, suggesting the omission of important explanatory variables. The RESET test, applied to both the linear model (*p* value = 0.1555) and the log-log model (*p* value = 0.1871), indicated that both models have a specification problem. However, the results suggest that the relationship between the variables may not be strictly linear and that some relevant factors could be missing. Although the Multicollinearity test (VIF) yielded positive results in both cases (values below 10), the recurrence of heteroscedasticity and specification problems limits the predictive and explanatory capacity of the results found.

Therefore, the analysis conclusions suggest the imperative need to include additional variables in the model to better capture the variation in the Homicides variable and achieve more robust and satisfactory results.

CONCLUSION

Research such as the present study is essential for empirically and locally assessing the contributions of smart city initiatives within the broader context of regional development. In medium-sized urban territories, like Aparecida de Goiânia, the strategic use of emerging technologies—such as video surveillance—can not only enhance public safety but also drive structural transformations in local social and economic indicators. The interdisciplinary analysis of technology, public policies, and human development proves to be indispensable for guiding managers, planners, and institutions in understanding the real impacts of these initiatives, providing technical and scientific support for more effective and contextualized decisions.

In the specific case of Aparecida de Goiânia, the results indicate a negative correlation between the expansion of video surveillance camera coverage and homicide rates, suggesting a possible deterrent effect of this technology on urban crime. Furthermore, the analysis showed positive interactions between the advances of the smart city project and the FIRJAN Municipal Development Index (IFDM), particularly in health and education components, although with variations across the historical series. Despite the statistical limitations identified in the estimated models—such as the presence of heteroscedasticity and the need to refine the set of explanatory variables—the study fulfills its purpose by providing initial evidence of the potential of technological solutions to strengthen urban governance and promote collective well-being.

To complement the evaluation of the results, it is imperative to consider the Employment and Income (ER) dimension, a key component of the IFDM. Although subject to statistical limitations, the econometric analysis indicated a negative correlation between the Employment and Income Index and the Homicides variable (coefficient of −293.32 in the linear model). This result suggests that advances in economic development and income generation—core elements of human development as proposed by Amartya Sen—tend to move in the opposite direction of lethal violence. However, the ER coefficient did not demonstrate statistical relevance in any of the executed models, estimated models, with the *p* value in the linear specification equal to 0.2541. The absence of statistical significance for the ER coefficient corroborates the need pointed out in the discussion of results: model misspecification (as indicated by the RESET test, *p* value = 0.1555 in the linear model) and the violation of OLS assumptions require the inclusion of additional variables to robustly capture the contribution of Employment and Income to crime reduction, thereby aligning the smart city strategy with the promotion of quality of life and the reduction of inequalities.

REFERENCES

Abreu, J. P. M., & Marchiori, F. F. (2019). Ferramentas de avaliação de desempenho de cidades inteligentes: uma análise da norma ISO 37122:2019 [Performance assessment tools for smart cities: An analysis of ISO 37122:2019]. *PARC Pesquisa Em Arquitetura E Construção*, *14*, e023002. https://doi.org/10.20396/parc.v14i00.8668171

Alves, L. A. (2019). Cidades saudáveis e cidades inteligentes: Uma abordagem comparativa [Healthy cities and smart cities: A comparative approach]. *Sociedade & Natureza*, *31*, e47004. https://doi.org/10.14393/sn-v31-2019-47004

Associação Brasileira de Normas Técnicas. (2020). *ABNT NBR 37122: Cidades e comunidades sustentáveis — Indicadores para cidades inteligentes* [ABNT NBR

37122: Sustainable cities and communities — Indicators for smart cities]. Associação Brasileira De Normas Técnicas. https://abnt.org.br/certificacao/smartcities/

Brasil. Ministério do Desenvolvimento Regional. (2021). *Carta Brasileira de Cidades Inteligentes*. Ministério do Desenvolvimento Regional. https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/desenvolvimento-urbano-e-metropolitano/projeto-andus/carta-brasileira-para-cidades-inteligentes

Chichoski, A., & Marquardt, J. F. (2023). O uso de tecnologias através do videomonitoramento do gabinete de segurança integrada municipal de Foz do Iguaçu – PR pelas agências de segurança pública e defesa social [The use of technologies through videomonitoring by the Integrated Municipal Security Office of Foz do Iguaçu – PR by public security and social defense agencies]. *Revista (RE)Definições das Fronteiras*, *1*(1), 197-217. https://doi.org/10.59731/vol1iss1pp219-239

Covas, M. M. C. M., & Covas, A. M. A. (2020). Cidades inteligentes e criativas e smartificação dos territórios: Apontamentos para reflexão [Smart and creative cities and the smartification of territories: Notes for reflection]. *DRd – Desenvolvimento Regional em Debate*, *10*(ed. esp.), 40-59. https://doi.org/10.24302/drd.v10ied.esp..2896

Ferreira, D. L. S., Novaes, S. M., & Macedo, F. G. L. (2023). Cidades inteligentes e inovação: A videovigilância na segurança pública de Recife, Brasil [Smart cities and innovation: Video surveillance in the public security system of Recife, Brazil]. *Cadernos Metrópole, 25*(58), 1095-1122. https://doi.org/10.1590/2236-9996.2023-5814

Filho, N. B., Santos, A. A. R., & Lima, H. C. P. (2024). Manaus como cidade inteligente – Vantagens e desvantagens do videomonitoramento aplicado à segurança pública [Manaus as a smart city – Advantages and disadvantages of video surveillance applied to public security]. *Revista Contemporânea*, *4*(7), 1-26. https://doi.org/10.56083/RCV4N7-010

Gomes, M., Albernaz, L. R., Nascimento, A. C., & Torres, F. R. (2016). Accountability e transparência na implementação da Agenda 2030: As contribuições do Tribunal de Contas da União [Accountability and transparency in the implementation of the 2030 Agenda: Contributions from the Federal Court of Accounts]. *Revista do Tribunal de Contas da União*, *48*(136), 76-91.

Guimarães, J. R. S., & Jannuzzi, P. M. (2005). IDH, indicadores sintéticos e suas aplicações em políticas públicas: Uma análise crítica [HDI, synthetic indicators and their applications in public policies: A critical analysis]. *Revista Brasileira de Estudos Urbanos e Regionais*, *7*(1), 73-90. https://doi.org/10.22296/2317-1529.2005v7n1p73

Jordão, K. C. P. (2016). *Cidades inteligentes: Uma proposta viabilizadora para a transformação das cidades brasileiras* [*Smart cities: A viable proposal for the transformation of Brazilian cities*] [Master's thesis, Pontifícia Universidade Católica de Campinas]. PUC Repository. https://repositorio.sis.puc-campinas.edu.br/handle/123456789/15131

Lazzaretti, K., Sehnem, S., Bencke, F. F., & Machado, H. P. V. (2019). Cidades inteligentes: Insights e contribuições das pesquisas brasileiras [Smart cities: Insights and contributions from Brazilian research]. *Urbe. Revista Brasileira de Gestão Urbana*, *11*. https://doi.org/10.1590/2175-3369.011.001.e20190118

Lira, P. (2013). *Índice de violência criminalizada (IVC) [Criminalized violence index (CVI)].* In *II Congresso Consad de Gestão Pública – Painel 62: Gestão em segurança pública*, Vitória, Brasil. https://consad.org.br/wp-content/uploads/2013/02/%C3%8DNDICE-DE-VIOL%C3%8ANCIA-CRIMINALIZADA-IVC3.pdf

Martinelli, M. A., Achcar, J. A., & Hoffmann, W. A. M. (2020). Cidades inteligentes e humanas: Percepção local e aderência ao movimento que humaniza projetos de smart cities [Smart cities and human-centered cities: Local perception and adherence to the movement that humanizes smart-city projects]. *Revista Tecnologia e Sociedade*, *16*(39), 164-181. https://doi.org/10.3895/rts.v16n39.9130

Minayo, M. C. S. (2006). *Violência e saúde* [Violence and health] (Temas em Saúde series, Vol. 1). Editora Fiocruz. https://static.scielo.org/scielobooks/y9sxc/pdf/minayo-9788575413807.pdf

Oliveira, E. F. T. (2018). *Estudos métricos da informação no Brasil: Indicadores de produção, colaboração, impacto e visibilidade* [Metric studies of information in Brazil: Indicators of production, collaboration, impact, and visibility]. Editora Unesp. https://doi.org/10.36311/2018.978-85-7983-930-6

Paes, C. F. C., Gonçalves, P. H., & Ziebell, C. S. (2021). *Smart city como ação determinante ao desenvolvimento do urbanismo sustentável* [Smart city as a determinant action for the development of sustainable urbanism]. In Anais do IX Encontro de Sustentabilidade em Projeto – UFSC (pp. 257-268). Universidade Federal de Santa Catarina. https://repositorio.ufsc.br/bitstream/handle/123456789/227788/257-268.pdf?sequence=1&isAllowed=y

Rezende, L. V. R., Cruz-Riascos, S. A., & Rodrigues, A. C. T. (2019). *Dados abertos em cidades inteligentes: Uma análise da fronteira entre acesso e privacidade* [Open data in smart cities: An analysis of the frontier between access and privacy]. e-LIS: e-prints in Library and Information Science. http://eprints.rclis.org/38639/

Roland Berger. (2019). *Smart city strategy index: Vienna and London leading in worldwide ranking*. https://www.rolandberger.com/en/Insights/Publications/

Smart-City-Strategy-Index-Vienna-and-London-leading-in-worldwide-ranking.html

Sancino, A., & Lorraine, H. (2020). Leadership in, of, and for smart cities – Case studies from Europe, America, and Australia. *Public Management Review*, *22*, 701-725. https://doi.org/10.1080/14719037.2020.1718189

Silva, A. F. da, Sousa, J. S. de, & Araujo, J. A. (2017). Evidências sobre a pobreza multidimensional na região Norte do Brasil [Evidence on Multidimensional Poverty in the Northern Region of Brazil]. *Revista de Administração Pública*, *51*, 219–239. https://doi.org/10.1590/0034-7612160773

Takiya, Carolina, A., Junior, S., & Marè, R. (2019). Cidades Inteligentes e o Índice Cities in Motion [Smart Cities and the Cities in Motion Index] – Case São Paulo. *Revista Simetria do Tribunal de Contas do Município de São Paulo*, *1*(5), 65-81. https://doi.org/10.61681/revistasimetria.v1i5.46

ARQUITECTURA *BLOCKCHAIN* HÍBRIDA PARA EL REGISTRO DE HISTORIALES MÉDICOS

ALEJANDRO HERNÁN SON ROMERO
https://orcid.org/0009-0005-5504-3803
20202016@aloe.ulima.edu.pe
Universidad de Lima, Perú

NICOLAS XAVIER HERRERA MEDINA
https://orcid.org/0009-0009-1859-8956
20201004@aloe.ulima.edu.pe
Universidad de Lima, Perú

PABLO ALBERTO ROJAS JAEN
https://orcid.org/0000-0002-9955-6740
projas@ulima.edu.pe
Universidad de Lima, Perú

Recibido: 5 de julio del 2025 / Aceptado: 15 de octubre del 2025
doi: https://doi.org/10.26439/interfases2025.n022.8092

RESUMEN. En este trabajo se presenta una arquitectura de *blockchain* híbrida para mejorar la gestión e interoperabilidad de los historiales médicos electrónicos. Se combina una *blockchain* pública permisionada basada en Polkadot, la cual gestiona roles y permisos con una *blockchain* privada implementada en Hyperledger Fabric que, a la vez, está encargada del almacenamiento de los datos médicos de los pacientes. Los registros textuales se guardan en CouchDB, mientras que las imágenes en el IPFS se representan como *tokens* no fungibles, bajo un enfoque centrado en el paciente. Las pruebas de estrés mostraron latencias promedio de 2050 ms para la creación de historiales médicos electrónicos y 2000 ms para su intercambio, con un uso de CPU del 65 % y memoria de 170 MB, lo que evidencia eficiencia y estabilidad. La arquitectura propuesta demuestra ser una solución escalable y segura para entornos hospitalarios, ya que optimiza recursos y fortalece la confidencialidad de la información médica del paciente.

PALABRAS CLAVE: *blockchain* híbrida / historial médico electrónico / *blockchain* privada / *blockchain* pública / interoperabilidad / seguridad

HYBRID BLOCKCHAIN ARCHITECTURE FOR MEDICAL RECORD MANAGEMENT

ABSTRACT. This article introduces a hybrid blockchain architecture to enhance Electronic Medical Record (EMR) management and interoperability. It integrates a permissioned public blockchain on Polkadot—managing roles and permissions—with a private blockchain on Hyperledger Fabric responsible for EMR storage. Text data are stored in CouchDB and medical images in IPFS as Non-Fungible Tokens (NFTs), following a patient-centric model. Stress tests yielded average latencies of 2050 ms for EMR creation and 2000 ms for sharing, with 65 % CPU and 170 MB memory usage, indicating system stability and efficiency. The proposed architecture provides a scalable, secure, and interoperable solution suitable for healthcare environments that demand data confidentiality and controlled access.

KEYWORDS: hybrid blockchain / EHR / private blockchain / public blockchain / interoperability / security

INTRODUCCIÓN

A pesar de los avances tecnológicos en infraestructura para telecomunicaciones y el desarrollo de *software*, todavía persisten deficiencias para garantizar un servicio de salud de calidad a los pacientes. En este sector, el manejo de historiales médicos ha evolucionado de ser registros físicos que contienen información delicada —sobre la salud y diagnósticos— almacenados localmente, a ser historiales médicos electrónicos (HME) que se almacenan en la nube (International Organization for Standardization, 2019). Si bien esta transformación digital ha mejorado la gestión y el almacenamiento de información médica, los HME presentan problemas de interoperabilidad entre instituciones, así como un alto riesgo de ataques cibernéticos, fallos humanos y ausencia de estándares (Jayabalan & Jeyanthi, 2022; Mani et al., 2021; Miyachi & Mackey, 2021; Samala & Rawas, 2024). Estas vulnerabilidades exponen la información privada de los HME a terceros no autorizados, lo que, según Mulligan y Braunack-Mayer (2004), podría presentar un riesgo de discriminación por condiciones médicas.

Adicionalmente, la falta de acceso a un historial médico unificado puede dar lugar a la creación de múltiples versiones de este, lo que afectaría significativamente la calidad del servicio (Mulligan & Braunack-Mayer, 2004). Por ejemplo, en el 2010, en Lima (Perú), se identificó que, de 450 historiales médicos, 147 estaban incompletos y 140 se habían perdido (Montañez-Valverde et al., 2015). Asimismo, entre 2016 y 2017, se reportó un incremento del 89 % en ataques cibernéticos relacionados al sector salud en Estados Unidos (Awad Abdellatif et al., 2021), en los que 3 620 000 registros de pacientes fueron comprometidos en lo que se conoció como el mayor incidente del 2016 (Abouelmehdi et al., 2018).

Dada esta problemática, diversos autores han propuesto soluciones innovadoras basadas en tecnología de *blockchain*. Haas et al. (2011) sugirieron dar más control al paciente sobre su información, mientras que Jin et al. (2019) recomendaron aprovechar las ventajas de sistemas autorizados y sin permisos en arquitecturas de *blockchain*, y Liu et al. (2023) propusieron el uso de almacenamiento dentro y fuera de la cadena para reducir la carga del sistema. Por su parte, Mani et al. (2021) plantearon una arquitectura segura para la gestión de datos sanitarios centrada en el paciente, basada en contratos inteligentes (Cintel). En esta línea, Jayabalan y Jeyanthi (2022) propusieron un *framework* que integra la *blockchain* con el sistema de archivos interplanetario (*interplanetary file system*, IPFS), a fin de crear una arquitectura descentralizada y distribuida, y preservar la privacidad mediante un enfoque centrado en el paciente. Finalmente, Guo et al. (2019) sugirieron un sistema en el que los doctores consultan, modifican y agregan información mediante enlaces de un solo uso para almacenar los registros en nodos en el borde de la red.

Para abordar los desafíos de seguridad y confidencialidad en los registros médicos, en este trabajo se propone una arquitectura de *blockchain* híbrida (BHIB), compuesta por una *blockchain* privada (BPRIV) para la gestión de los HME y una *blockchain* pública

(BPUB) permisionada para la gestión de permisos de los usuarios. Para el almacenamiento fuera de la cadena, se utilizó el IPFS y para las imágenes los *tokens* no fungibles (*non-fungible token*, NFT). Este enfoque brinda tres ventajas fundamentales: alta privacidad de los datos, puesto que el enfoque centrado en el paciente le otorga el total dominio de sus datos a estos; alta seguridad, ya que, en adición a la seguridad de la propia *blockchain*, se emplea un almacenamiento dentro y fuera de la cadena; y mayor flexibilidad de datos, puesto que se proponen NFT para el almacenamiento de imágenes, como placas de rayos X fuera de la cadena y los datos textuales de los HME dentro de la cadena.

El presente artículo se organiza de la siguiente manera: en la sección 2, se discuten las diferentes soluciones tradicionales existentes y otras propuestas basadas en la *blockchain*; en la sección 3, se profundiza en la propuesta planteada y en la metodología empleada; en la sección 4, se detalla el proceso de implementación; en la sección 5, se presentan los resultados; en la sección 6, se hace una discusión de estos; finalmente, en la sección 7, se comparten las conclusiones.

ESTADO DEL ARTE

Tipos de almacenamiento usados en la gestión de historiales médicos electrónicos

La gestión de HME en sistemas de *blockchain* emplea tres modelos de almacenamiento —dentro de la cadena, fuera de la cadena y el modelo híbrido—, con el fin de resolver problemas de seguridad y escalabilidad. El almacenamiento dentro de la cadena, que guarda los datos directamente en ella, es cada vez menos utilizado debido a su baja disponibilidad (Jayabalan & Jeyanthi, 2022). En contraste, el almacenamiento fuera de la cadena se ha convertido en una estrategia común que mejora la escalabilidad, pues registra solo referencias criptográficas en la *blockchain*. Investigadores como Dagher et al. (2018), Hussien et al. (2021), Nhan et al. (2024) y Kang et al. (2022) utilizan IPFS para este fin, almacenando los HME o sus *hashes*; sin embargo, este enfoque introduce nuevas preocupaciones de seguridad y dependencia de plataformas externas.

Para solucionar estas preocupaciones, el almacenamiento híbrido —que combina el almacenamiento dentro y fuera de la cadena— equilibra seguridad, escalabilidad y eficiencia (Amanat et al., 2022; Chelladurai et al., 2021; Kaur et al., 2023; Lee et al., 2022; Samala & Rawas, 2024; Tanwar & Thakur, 2023). En este contexto, estudios como los de Jayabalan & Jeyanthi (2022) y Mani et al. (2021), que se resumen en la Tabla 2.1, muestran cómo el uso del IPFS contribuye a abordar los problemas de escalabilidad y almacenamiento de grandes volúmenes de datos médicos. En ambos casos, los investigadores almacenan los HME fuera de la cadena y preservan dentro de la *blockchain* los *hashes* o metadatos cifrados, lo que fortalece la integridad de los registros y permite implementar múltiples capas de seguridad mediante cifrado.

Del mismo modo, Rajput et al. (2021), Haddad et al. (2024) y Uddin et al. (2021) emplearon Cintel para gestionar los accesos y las transacciones, a fin de guardar los registros de auditoría dentro de la cadena, mientras que los HME permanecen fuera de ella. Por su parte, Shen et al. (2019) (Tabla 2.1) han propuesto una red *peer-to-peer* externa en lugar de un IPFS y así conservar dentro de la cadena únicamente los metadatos y *hashes* de los historiales. Estos ejemplos ilustran la diversidad de estrategias *off-chain* que buscan mantener la trazabilidad sin comprometer la eficiencia del sistema.

Como se detalla en la Tabla 1, el enfoque híbrido logra combinar seguridad y escalabilidad, pero introduce nuevos retos relacionados con la gestión de claves criptográficas y la integridad de los datos fuera de la cadena. Sobre la base de lo visto en el estado del arte sobre el almacenamiento, este trabajo propone una arquitectura de *blockchain* híbrida para la gestión de HME debido a su capacidad de equilibrar las demandas de seguridad, escalabilidad y eficiencia. Este enfoque híbrido permite almacenar los metadatos y las referencias a los datos médicos en la *blockchain*, lo que asegura su trazabilidad e integridad mediante la BPRIV, mientras que los HME de mayor tamaño —como las imágenes y documentos médicos— se almacenan fuera de la cadena en IPFS. Esto garantiza que la *blockchain* no se sobrecargue, mejora la escalabilidad del sistema y asegura que los pacientes y médicos puedan acceder a los datos de manera eficiente y segura a través de los Cintel que gestionan los accesos.

Tipos de *blockchain* utilizados en la gestión de historiales médicos electrónicos

Los tres tipos principales de *blockchain* que se utilizan en la gestión de HME son la BPUB, la BPRIV y la BHIB. Cada una de estas opciones ha sido explorada en distintos estudios para abordar desafíos relacionados con interoperabilidad, seguridad y escalabilidad.

El uso de la BPUB ha sido propuesto en diversos trabajos, tales como los de Fatokun et al. (2021), Hossain Faruk et al. (2021), Kumari et al. (2024), Mauricio et al. (2024), Puneeth y Parthasarathy (2024) y Wang et al. (2021). Asimismo, Mandarino et al. (2024) y Omar et al. (2021) —mencionados en la Tabla 1— emplearon Ethereum como BPUB para garantizar la descentralización y la transparencia. En particular, Mandarino et al. (2024) integraron computación en el borde (*edge computing*) para almacenar datos de manera local en los dispositivos, lo que redujo la dependencia de servidores centralizados. Por su parte, Omar et al. (2021) priorizaron la seguridad mediante el cifrado de HME en la nube, empleando Cintel para regular el acceso. De modo similar, Zhang et al. (2022) plantearon un modelo en el que la BPUB facilita la trazabilidad de los datos médicos, utilizando protocolos de pago justo que incentivan la participación de los pacientes. Aunque la BPUB ofrece ventajas en trazabilidad y transparencia, enfrenta limitaciones notables de escalabilidad y eficiencia (Tabla 1).

En contraste, el uso de la BPRIV ha sido explorado para reforzar la seguridad, privacidad y control de acceso en la gestión de HME (Abunadi & Kumar, 2021; Ali, Rahim,

Pasha et al., 2021; Ali, Rahim, Ali, et al. 2021; Antwi et al., 2021; Guo et al., 2019; Hashim et al., 2022; Huang et al., 2021; Selvarajan & Mouratidis, 2023; Singh et al., 2021; Wu et al., 2022). En particular, Awad Abdellatif et al. (2021) —mencionados en la Tabla 1— presentaron MEdge-Chain, una BPRIV combinada con computación en el borde para gestionar de manera eficiente grandes volúmenes de HME distribuidos en múltiples hospitales. Este sistema permite la monitorización automatizada y la detección temprana de eventos médicos críticos, ejecutando transacciones a través de Cintel. Además, existen BPRIV de consorcio (Exceline & Nagarajan, 2024; Li et al., 2024; Liu et al., 2023; Mohey Eldin et al., 2023; Xiao et al., 2021), las cuales son *blockchains* permisionadas administradas por un grupo de organizaciones o entidades, lo que optimiza el control de acceso y la privacidad, aunque con limitaciones de interoperabilidad, como se aprecia también en la Tabla 1.

Finalmente, la BHIB combina características de *blockchains* públicas y privadas, ofreciendo una solución equilibrada frente a los desafíos de seguridad, escalabilidad e interoperabilidad en la gestión de los HME. Algunos ejemplos representativos se encuentran en Uppal et al. (2023) y Chelladurai y Pandian (2022), quienes, como se detalla en la Tabla 1, emplearon Hyperledger Fabric (HF) para gestionar permisos de acceso, compartir datos entre proveedores y permitir a los pacientes controlar sus propios historiales. Asimismo, Kaur et al. (2023) integraron BPUB y BPRIV en el contexto de la cirugía médica, empleando la primera para garantizar la inmutabilidad de los datos y la segunda para proporcionar acceso controlado a los pacientes autorizados.

Como se observa en la Tabla 1, el enfoque híbrido representa una síntesis de las ventajas de ambos tipos de *blockchain*, lo que resuelve los problemas de escalabilidad de las BPRIV y los de seguridad de las BPUB. Entonces, sobre la base de lo discutido en el estado del arte, y tal como ya se ha mencionado, este trabajo propone una arquitectura BHIB para la gestión de los HME. Esta solución permite superar las limitaciones de escalabilidad y eficiencia de las BPUB, a la vez que aborda los desafíos de interoperabilidad y seguridad propios de las BPRIV. En la arquitectura propuesta, la BPUB se encarga de la verificación de accesos y la interoperabilidad interhospitalaria, mientras que la BPRIV gestiona los permisos de acceso a los HME, lo que garantiza que solo los actores autorizados puedan modificar los datos. Este enfoque asegura un manejo eficiente de los HME, por lo que logra ofrecer privacidad, seguridad y control de acceso, sin comprometer la interoperabilidad entre instituciones.

Tabla 1

Comparativa de los trabajos más resaltantes para sistemas de historiales médicos electrónicos

Autor	Metodología	Tecnología de almacenamiento	Seguridad y privacidad	Resultados	Limitaciones
Shen et al. (2019)	Diseño de un sistema de BPUB para el intercambio de datos médicos	Dentro de la cadena y fuera de la cadena (BPUB)	Control de acceso mediante *blockchain*, pero con limitaciones en escalabilidad y latencia	Implementación de BPUB para compartir datos médicos y lograr mayor trazabilidad y seguridad	Problemas de escalabilidad y latencia en la red pública
Mandarino et al. (2024)	Integración de *blockchain* con computación en el borde (*edge computing*) para manejar grandes volúmenes de datos médicos	Fuera de la cadena (*edge computing* para dispositivos IoT)	Mejora de la privacidad al procesar datos cerca de su origen y garantiza seguridad mediante Cintel	Eficiencia en el manejo de grandes volúmenes de datos en tiempo real y reducción de la latencia	Abordaje inadecuado de la interoperabilidad con otros sistemas de salud
Awad Abdellatif et al. (2021)	Implementación de un sistema híbrido que integra *blockchain* con *edge computing* para la gestión eficiente de datos médicos	Fuera de la cadena (*edge computing* e IPFS)	Garantía de privacidad al descentralizar el procesamiento de datos y manejar accesos mediante Cintel	Mejora de la escalabilidad y eficiencia al integrar *blockchain* con *edge computing* y IPFS	Complejidad en la sincronización de datos entre nodos y dispositivos IoT
Uddin et al. (2021)	Uso de HF para gestionar de manera segura y eficiente los accesos a los HME	Dentro de la cadena (HF) y fuera de la cadena (base de datos externa)	Gestión granular de accesos y transacciones mediante Cintel	Seguridad robusta en la gestión de HME mediante HF (eficiente para instituciones de salud privadas)	Limitada interoperabilidad con otras instituciones de salud que no utilicen HF
Omar et al. (2021)	Propuesta de un sistema híbrido que combina *blockchain* y almacenamiento en la nube para gestionar los HME	Fuera de la cadena (IPFS y almacenamiento en la nube)	Encriptado de datos fuera de la cadena y manejo de claves dentro de la cadena para la protección contra accesos no autorizados	Alta transparencia en el acceso a los datos médicos y escalabilidad mediante almacenamiento fuera de la cadena	Riesgos de seguridad inherentes a la dependencia en la nube para almacenamiento fuera de la cadena

(*continúa*)

(*continuación*)

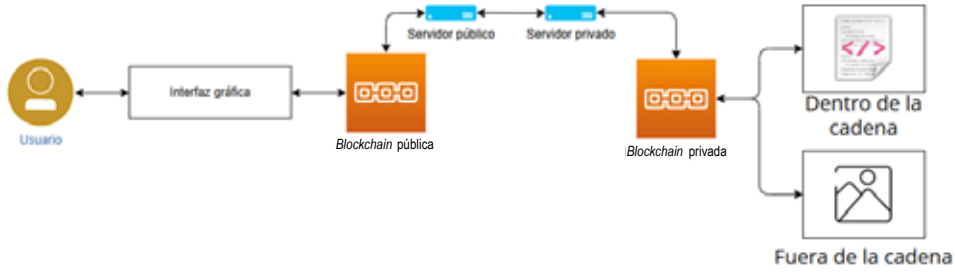
Autor	Metodología	Tecnología de almacenamiento	Seguridad y privacidad	Resultados	Limitaciones
Mani et al. (2021)	Desarrollo de un sistema basado en BPRIV para gestionar HME mediante almacenamiento en IPFS	Fuera de la cadena (IPFS)	Gestión de permisos de acceso mediante Cintel en HF	Optimización del uso de IPFS para el almacenamiento de grandes volúmenes de datos médicos, con control de acceso a través de BPRIV	Interoperabilidad limitada con otros sistemas de salud al depender de una BPRIV
Jayabalan y Jeyanthi (2022)	Modelo de BHIB para gestionar HME mediante almacenamiento distribuido en IPFS	Fuera de la cadena (IPFS)	Gestión de claves y permisos mediante Cintel para asegurar el acceso a los datos médicos	Escalabilidad mejorada mediante el uso de IPFS para almacenamiento distribuido de datos grandes, como imágenes médicas	Seguridad de los datos fuera de la cadena que depende de la solidez de IPFS y el manejo de claves criptográficas
Rajput et al. (2021)	Propuesta de un marco de BPRIV para el intercambio seguro de datos médicos encriptados	Dentro de la cadena (BPRIV)	Enfoque en la privacidad mediante el uso de encriptación avanzada y control de acceso descentralizado	Seguridad mejorada mediante encriptación avanzada y control granular sobre el acceso a los HME	Baja escalabilidad y limitada interoperabilidad debido al uso exclusivo de BPRIV
Chelladurai y Pandian (2022)	Sistema automatizado de gestión de HME basado en Cintel y BPRIV	Dentro de la cadena (BPRIV)	Control de acceso mediante contratos inteligentes para garantizar la privacidad de los datos médicos	Automatización del control de acceso a HME, lo que mejora la seguridad y la eficiencia del sistema	Limitada interoperabilidad al depender de una BPRIV
Zhang et al. (2022)	Sistema híbrido de *blockchain* para preservar la privacidad de los HME mediante el almacenamiento en IPFS	Fuera de la cadena (IPFS)	Encriptado de datos médicos antes de ser almacenados fuera de la cadena, mientras los permisos y claves se gestionan dentro de la cadena	Mejora de la privacidad mediante el almacenamiento en IPFS y la gestión de claves en la *blockchain*, con alta escalabilidad	Riesgo en la dependencia de IPFS para el almacenamiento fuera de la cadena y posible complejidad en la gestión de claves criptográficas

METODOLOGÍA

Como se observa en la Figura 1, por medio de la interfaz gráfica, los usuarios interactúan con la arquitectura. En la BPUB se encuentra el Cintel que asigna roles a los usuarios, los permisos de visualización y los datos básicos de las cuentas de los usuarios. En la BPRIV se encuentra la información de los HME de los pacientes, la cual verifica la acción que el usuario solicita mediante el rol y permiso correspondiente. Una vez realizada la validación, la información solicitada retorna a la BPUB o guarda los cambios realizados. De esta manera, la lógica de las cuentas y los HME asociados a cada cuenta están separados, lo que permite a las *blockchains* estar centralizadas en una función en particular.

Figura 1

Arquitectura de la propuesta

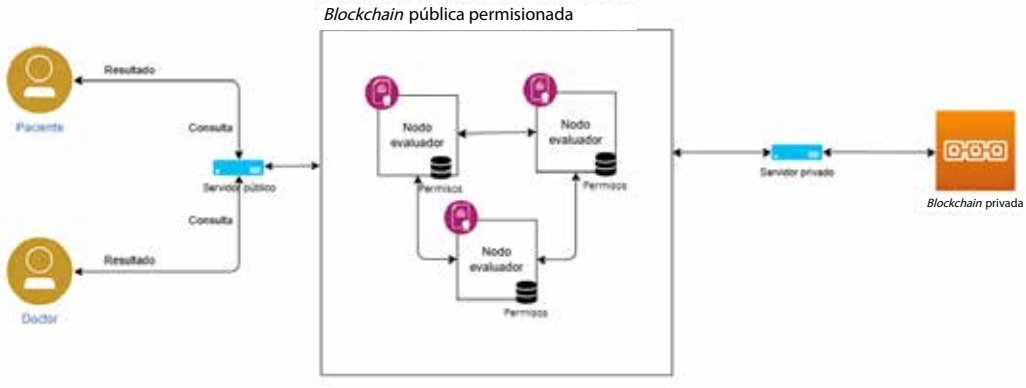


Blockchain pública permisionada

En las *blockchain* públicas permisionadas (BPUBPER) se permite que cualquier usuario que se una a la red tenga conocimiento de la identidad de los demás usuarios; sin embargo, su capacidad de acción está restringida a los privilegios otorgados según el rol designado. En la Figura 2, se observa que los usuarios pueden interactuar con la arquitectura mediante el servidor de la parte pública y este servidor. A través del Cintel "Control de acceso", se le asigna a cada usuario un rol, el cual puede ser de doctor o de paciente; de esta manera, al ser paciente, se puede garantizar o revocar permisos a doctores. Asimismo, el servidor de la parte pública se puede comunicar con el servidor de la parte privada para consultar la información de los HME.

Figura 2

Arquitectura de la blockchain *pública permisionada*

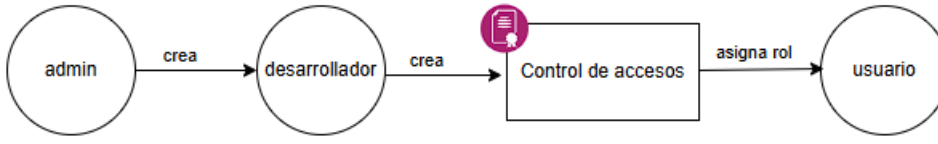


Cuentas de desarrollador y administrador

En la Figura 3, se observa cómo se genera una cuenta de desarrollador con la capacidad de crear Cintel, utilizando la cuenta de administrador de la *blockchain* que cuenta con los permisos más avanzados. A partir de esta cuenta, se establece el Cintel "Control de acceso". Este contrato se encarga de asignar roles a los usuarios, en caso de que su clave secreta esté asociada a un centro médico. Esto determina si el usuario es un doctor, en el caso de que su clave privada está registrada en la red, o si es un paciente, en el caso de que no contar con información disponible. Una vez que se tenga el rol asociado a la cuenta, el usuario podrá hacer uso de las funciones para visualizar el HME y otorgar o revocar permisos de visualización. Además, se usará el Cintel "Control de acceso" para que la BPRIV verifique los roles de quien hace la consulta, así como los permisos de visualización.

Figura 3

Diagrama de cuentas de desarrollador



Control de acceso y permisos

De acuerdo con lo mencionado, el control de acceso tiene como principal funcionalidad asignar un rol específico a cada usuario que se conecte a la red. Con este, se le asignan

los permisos respectivos. En la Tabla 2, se detallan las especificaciones de cada rol. En el caso de que se decida crear una cuenta, el usuario tendrá que usar la clave privada para crearla en la BPUBPER. Luego, ingresa su información personal y la asocia al rol de doctor o paciente, respectivamente. En caso de que el usuario posea una cuenta creada como doctor y desee crear una cuenta como paciente o viceversa, tendrá que asociar su cuenta con la información.

Tabla 2

Asignación de rol

Rol	Motivo	Consecuencia	Funcionalidades
Doctor	La clave privada se encuentra presente en una organización de la BPRIV	Se crea su cuenta y se asocia con su información personal	Modificar el HME, solicitar acceso de visualizar el HME y visualizar el HME de un paciente
Paciente	No existe una cuenta como paciente previamente	Se crea su cuenta y se asocia con su información personal	Visualizar el HME, otorgar y revocar permiso de visualizar el HME

En la Tabla 3, se observan los casos de asignación y revocación de permisos. En el primer caso, un doctor solicita acceso de visualización a un paciente, quien puede aceptar o rechazar dicha solicitud. En el segundo caso, el paciente quiere revocar el permiso a un doctor con acceso de visualización, se realiza la solicitud mediante el servidor de la parte pública y se revoca el permiso. El último caso, implica que un paciente se atendió en un centro médico que no pertenece a la arquitectura híbrida propuesta, por lo que el doctor no tiene acceso al HME de su paciente; entonces, el paciente le muestra su HME. Se utiliza el enfoque centrado al paciente, debido a que el paciente tiene el control de acceso de su HME. Esto incrementa la confiabilidad y seguridad de la información sensible.

Tabla 3

Asignación de permisos

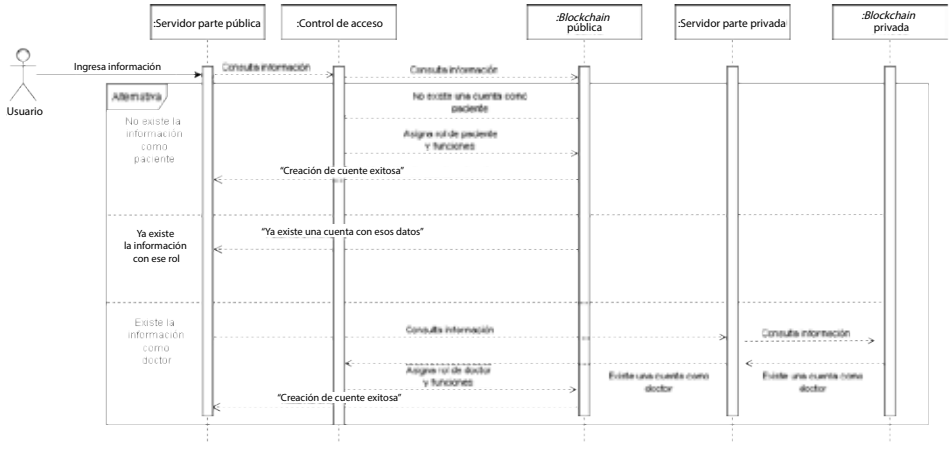
Caso	Acción del paciente
Un doctor perteneciente a una organización quiere visualizar el HME de un paciente	El paciente puede otorgar el permiso o rechazar la solicitud
El paciente quiere revocar el permiso de visualización del HME	El paciente dueño del HME puede revocar el permiso
Un doctor no perteneciente a una organización quiere visualizar el HME de un paciente	El paciente dueño del HME puede mostrarle desde su cuenta su HME

Funcionamiento de control de acceso

En la Figura 4, el usuario ingresa su información por medio del servidor público para crear una cuenta. Al registrarse, se crea de forma automática una cuenta en la red BPUBPER que le sirve para conectarse posteriormente. Cuando se suben los datos, el Cintel verifica en la *blockchain* si estos están asociados a un rol (doctor o paciente). En caso de que no lo estén, significa que es un usuario que recién está ingresando al sistema, por lo que se le otorga el rol de paciente y se le otorgan las funciones correspondientes. Si se trata de un doctor, el servidor de la parte pública se comunica con el servidor de la parte privada para verificar si su documento de identidad es de un doctor. A continuación, si se verifica su estatus, se le asigna el rol de doctor y las funcionalidades correspondientes. En el caso de que la información se encuentre en la *blockchain* y se intente crear nuevamente una cuenta con la misma información como doctor, se rechaza; como paciente, se permite guardar. Una persona solo puede tener dos cuentas: una de doctor y otra de paciente. Esta restricción existe porque un usuario podría ser doctor en un hospital y luego paciente en otro, o ser primero paciente y luego doctor.

Figura 4

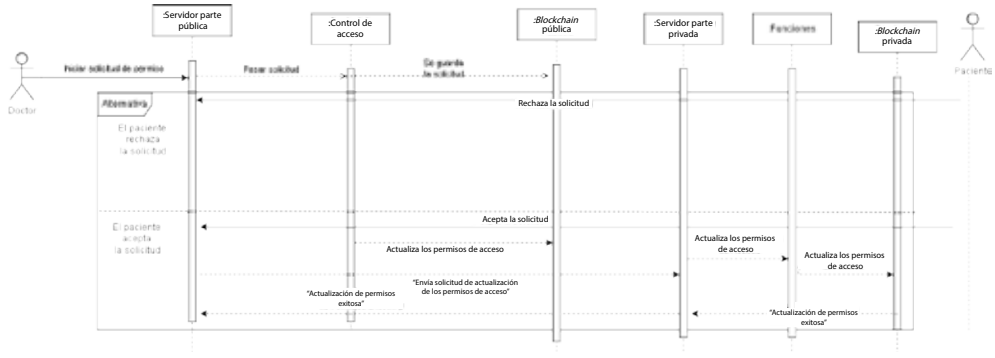
Diagrama de secuencia del control de acceso



Para realizar el control de acceso, se emplea el Cintel "Control de acceso". Este asigna el rol al usuario y las funcionalidades que aparecen en la Tabla 2. De igual manera, como se muestra en la Figura 5, la solicitud del doctor pasa por el servidor de la parte pública. Por medio del Cintel "Control de acceso", se registra la solicitud en la BPUBPER. Luego, el paciente puede decidir si rechaza la solicitud o la acepta. Si la rechaza, los permisos no se cambian. En caso contrario, se actualizan los permisos de visualización de la BPUBPER, mientras que los permisos en la BPRIV se realizan por medio del Cintel "Funciones". Finalmente, se le informa al doctor que se aceptó su solicitud.

Figura 5

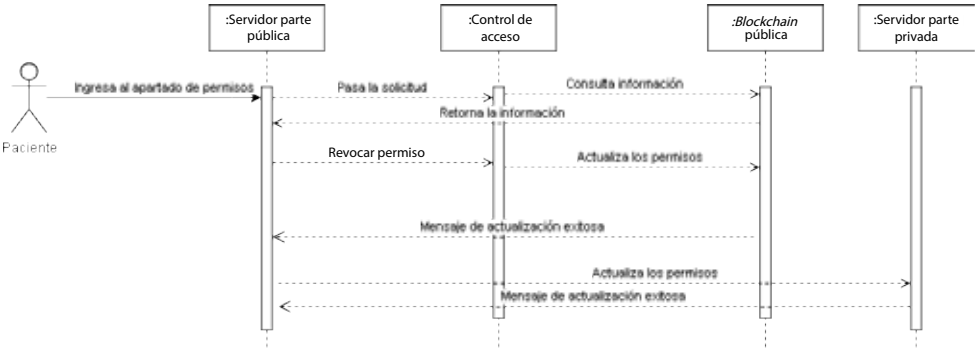
Otorgar permisos de acceso al HME



La Figura 6 detalla que el paciente puede revocar los permisos asignados a los doctores, gracias al Cintel "Control de acceso". El paciente, por medio del servidor de la parte pública, revoca los permisos del doctor y los actualiza en la BPRIV.

Figura 6

Revocar permiso de visualización a terceros



Blockchain privada

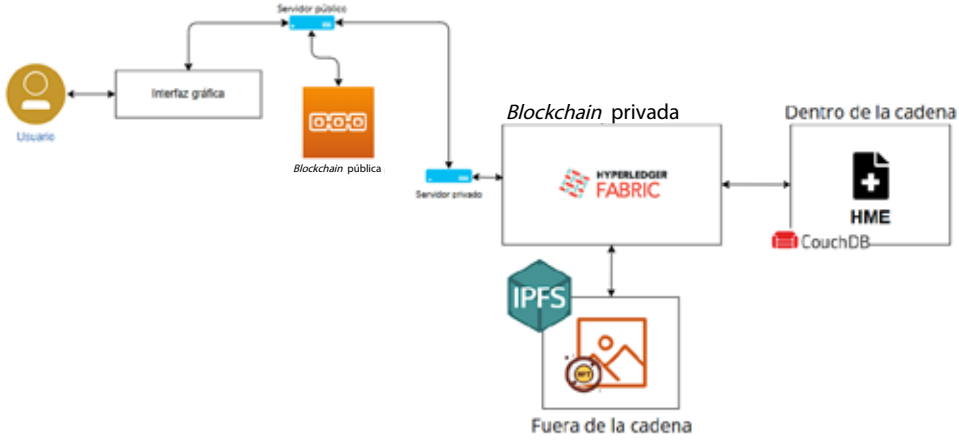
La creciente necesidad de gestionar HME de manera segura, eficiente y centrada en el paciente ha impulsado la adopción de tecnologías avanzadas como la *blockchain*. En este contexto, HF se presenta como una solución robusta para implementar una BPRIV que permita almacenar, acceder y modificar HME con altos estándares de seguridad y privacidad.

Arquitectura de la blockchain privada

Como se visualiza en la Figura 7, la BPRIV está diseñada para soportar la gestión descentralizada de HME entre dos organizaciones, pues representa a diferentes hospitales dentro de una cadena hospitalaria. Cada organización actúa como un nodo evaluador, responsable de validar transacciones y mantener una copia actualizada del libro mayor distribuido. Esta configuración permite una colaboración fluida y segura entre los diferentes hospitales, lo que garantiza la disponibilidad y consistencia de los HME. En este caso, dichos hospitales se encuentran dentro de la red de HF, mientras que los HME se encuentran divididos en el almacenamiento dentro y fuera de la cadena. Dentro de la cadena, se encuentra la información textual de los HME y el *hash* respectivo de los NFT relacionados a cada HME. Por otro lado, fuera de la cadena, se encuentran las imágenes convertidas en NFT para garantizar su trazabilidad.

Figura 7

Visualización del almacenamiento de un HME



Organizaciones y nodos

Cada hospital perteneciente a la BPRIV se configura como una organización independiente. Los nodos evaluadores de cada organización son responsables de validar transacciones, ejecutar el Cintel y mantener la integridad del libro mayor. Esta estructura permite una gestión distribuida y segura de los HME, en el que cada nodo tiene la capacidad de verificar y registrar transacciones, lo que asegura la transparencia y confiabilidad del sistema.

Contratos inteligentes

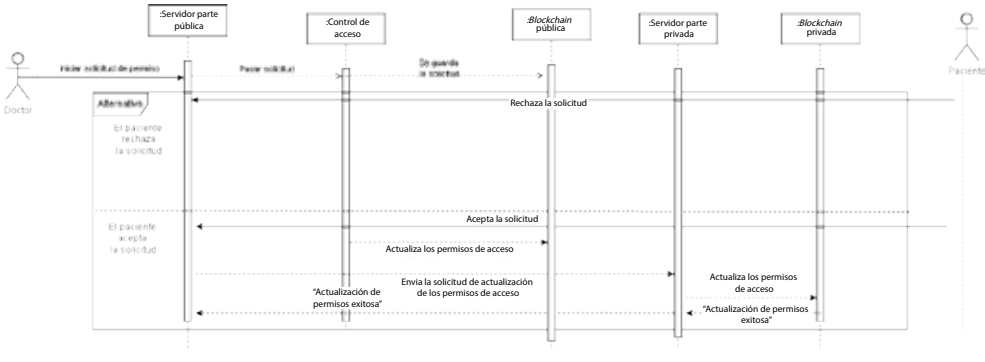
Los Cintel —conocidos en HF como códigos de cadena— son fundamentales para gestionar las operaciones sobre los HME. Estos Cintel definen las reglas y procedimientos para la creación, almacenamiento, actualización y acceso a los registros médicos. A continuación, se describen algunas de sus funciones clave.

- Creación del HME: permite registrar un nuevo HME en la BPRIV. Esta función toma los datos del paciente y crea un registro único en CouchDB.
- Modificación del HME: facilita la modificación de HME existentes. Solo los usuarios con permisos adecuados pueden actualizar la información, lo que garantiza la integridad y exactitud de los datos.
- Visualización del HME: permite que los usuarios visualicen los HME que tienen acceso.
- Verificación de permisos: antes de permitir cualquier operación sobre un HME, esta función verifica los permisos del usuario solicitante y asegura que solo los usuarios autorizados puedan acceder o modificar los datos.
- Agregación y eliminación de hashes de imágenes: permite añadir y eliminar los hashes de las imágenes generadas por IPFS, lo que modifica el HME.

Como se muestra en la Figura 8, el doctor, tras enviar su solicitud al servidor público, puede visualizar el HME de un paciente. Esto se comprueba utilizando el Cintel "Control de acceso", si el usuario tiene permiso de visualización. En el caso de que lo tenga, se retorna el historial médico; de lo contrario, no podrá visualizarlo y deberá solicitar el acceso al paciente.

Figura 8

Visualización de historial médico electrónico de un tercero



Almacenamiento de datos

Como se puede apreciar en la Figura 7, se plantea una solución de almacenamiento híbrida que combina las ventajas del almacenamiento dentro y fuera de la cadena:

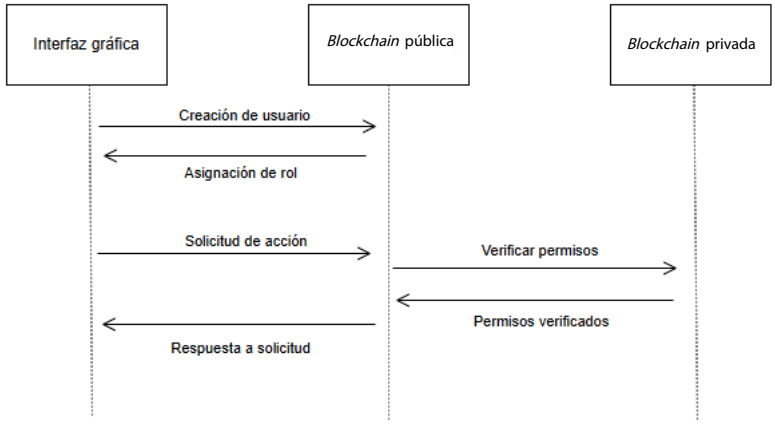
- Dentro de la cadena: utilizada para almacenar la parte textual de los HME directamente en la blockchain. Asimismo, se almacenan los hashes que vinculan a los archivos fuera de la cadena, lo que garantiza la integridad de los datos y su recuperación eficiente.
- Fuera de la cadena: la información visual crítica, como radiografías y tomografías, se almacena fuera de la cadena como NFT.

Verificación de permisos

Como se visualiza en la Figura 9, cuando un usuario intenta acceder o modificar un HME, la BPRIV verifica sus permisos consultando los roles y autorizaciones almacenados en la BPUBPER. Este proceso de verificación incluye la autenticación del usuario y la confirmación de sus permisos específicos. Solo los usuarios con los permisos adecuados pueden proceder con la visualización o modificación de los HME, para asegurar que los datos se mantengan protegidos y privados.

Figura 9

Verificación de permisos



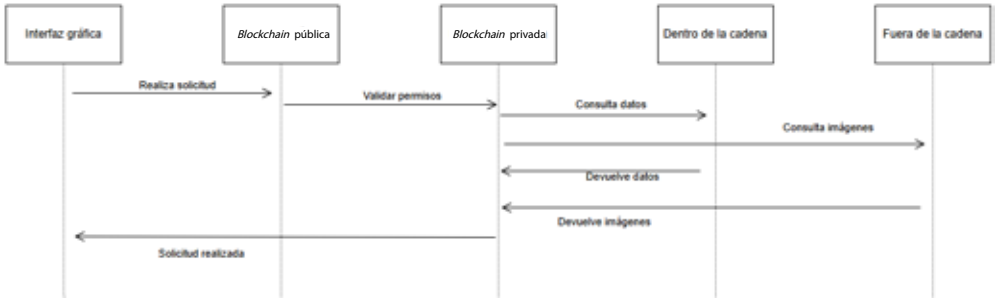
Flujo de trabajo de consulta de HME

Como se aprecia en la Figura 3.7, para consultar un HME, el usuario envía una solicitud a través de la interfaz gráfica. La BPUBPER la recibe y valida los permisos respectivos con la BPRIV, y esta verifica los permisos del usuario consultando a la BPUBPER. Si los permisos son válidos, la BPRIV recupera el HME desde el almacenamiento dentro y

fuera de la cadena y retorna la información solicitada a la interfaz gráfica. Este proceso asegura que solo los usuarios autorizados puedan acceder a los datos médicos, manteniendo la privacidad y seguridad de la información.

Figura 10

Flujo de trabajo de consulta de HME

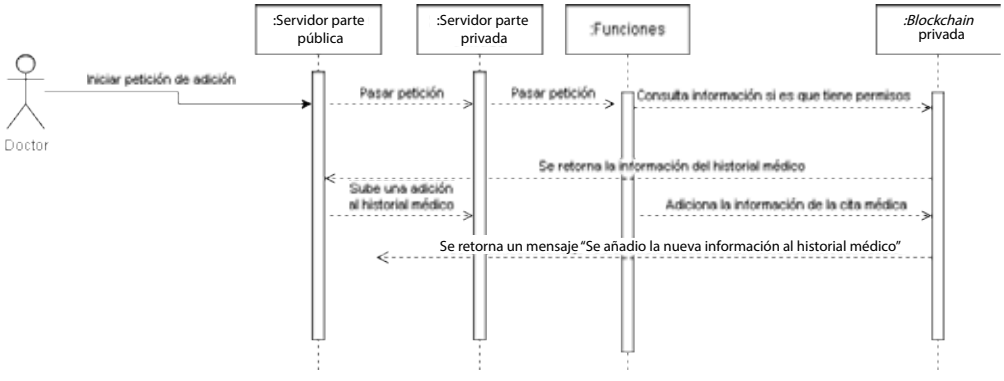


Flujo de trabajo de modificación de HME

Como se visualiza en la Figura 11, para modificar un HME, el usuario envía una solicitud de modificación a través del servidor de la parte pública. La BPRIV verifica los permisos del usuario consultando a la BPUBPER. Si los permisos son válidos, los cambios se aplican y el HME se actualiza en la base de datos dentro de la cadena. Los cambios también pueden involucrar actualizaciones de *hashes* que apuntan a nuevas versiones de archivos en la base de datos fuera de la cadena. Este flujo de trabajo garantiza que solo los usuarios autorizados puedan modificar los datos médicos, lo que preserva la integridad y exactitud del HME.

Figura 11

Flujo de trabajo de modificación de HME



Mecanismos y seguridad

La BPRIV implementa varios mecanismos de seguridad para proteger los datos médicos, incluyendo:

- Cifrado de datos: los datos se cifran tanto en tránsito como en reposo para protegerlos contra accesos no autorizados.
- Autenticación de usuarios: se utiliza autenticación basada en certificados digitales para asegurar que solo los usuarios autorizados puedan acceder al sistema.
- Políticas de control de acceso: las políticas de control de acceso detalladas aseguran que solo los usuarios con los permisos adecuados puedan realizar operaciones específicas sobre los HME.

Privacidad de los datos

La privacidad de los HME se garantiza mediante un enfoque centrado en el paciente. Los datos médicos solo son accesibles para los usuarios que fueron autorizados explícitamente por el paciente. Los *hashes* en CouchDB, que apuntan a datos en IPFS, aseguran que los datos almacenados fuera de la cadena sean recuperables de manera segura y eficiente. Este enfoque asegura que la privacidad del paciente se mantenga en todo momento.

Interconexión con la *blockchain* pública permisionada

La arquitectura se fundamenta en una estricta separación de responsabilidades: la BPUBPER gestiona exclusivamente los roles y permisos de los usuarios, mientras que la BPRIV se dedica al almacenamiento y manejo de los datos médicos (HME). Esta especialización optimiza la seguridad y eficiencia del sistema.

La interconexión se materializa a través de dos servidores dedicados. El servidor público gestiona las peticiones del usuario, comunicándose con la BPUBPER a través del Cintel "Control de acceso". Para validar cualquier solicitud, el servidor privado, que interactúa con la BPRIV mediante el Cintel "Funciones", consulta al servidor público para verificar los permisos correspondientes en la BPUBPER. Este flujo garantiza que solo usuarios autorizados interactúen con los HME.

IMPLEMENTACIÓN

Se realizaron las pruebas en computadoras con el sistema operativo Windows 11 Enterprise. Con respecto al *hardware*, se utilizó un procesador 12th Gen Intel (R) Core (TM) i7-12700 2,10 GHz, con 64 GB (63,7 GB usable) con una GPU Nvidia GeForce RTX 3080 de 10 GB.

Configuración del entorno de desarrollo

El entorno de desarrollo se configuró integrando una BPRIV sobre HF y una BPUBPER sobre Polkadot. La implementación de HF requirió Docker, Docker Compose y sus herramientas CLI para configurar una red de dos nodos y un canal. Para el almacenamiento de datos, esta red se conectó con CouchDB (textuales) e IPFS (no textuales, vinculados por *hash*). Paralelamente, se desplegó el nodo de Polkadot. La comunicación entre ambas *blockchains* se gestionó mediante servidores desarrollados en Node.js, mientras que Postman se utilizó como interfaz para ejecutar y validar las solicitudes.

Desarrollo de contratos inteligentes

El desarrollo de Cintel en HF implica la creación de código de cadena en Go para la gestión de HME, control de acceso y permisos. Estos contratos se despliegan en la red de HF y se someten a pruebas unitarias para asegurar su funcionalidad.

En el caso de Polkadot, se utiliza el *framework* Substrate para desarrollar Cintel que gestionan la asignación y verificación de roles y permisos de usuarios. Los contratos se despliegan en la red Polkadot y se realizan pruebas de integración con HF para garantizar una comunicación fluida.

El Cintel "Control de acceso" presenta once funciones: add_user, assign_role, user_exists, user_role, get_role, grant_permission, revoke_permission, has_permission, get_grantees, approve_access y request_access. De todas ellas, las únicas con las que interactúa el usuario directamente son add_user, assign_role, grant_permission y revoke_permission, ya que permiten registrar la cuenta, elegir el rol y otorgar o revocar permisos.

Integración de componentes

La arquitectura integra HF con CouchDB para gestionar los datos textuales de los HME. Paralelamente, IPFS almacena los datos no textuales, mientras sus *hashes* se registran en CouchDB para que las transacciones de HF puedan recuperar y modificar el registro completo.

Blockchain pública permisionada

Como se muestra en la Figura 2, la BPUBPER utiliza Polkadot en un despliegue local, basado en la plantilla de una red local de Substrate (Ink, s. f.), modificada para ser permisionada y compatible con Cintel desarrollados en ink! (Ink, s. f.). Se eligió Polkadot por cumplir con los requerimientos mínimos: rápido crecimiento, amplia documentación y facilidad para crear Cintel. El servidor público está implementado con Polkadot JS API, Express JS y Node JS.

Interconexión

La comunicación se hace entre el servidor de la parte pública con el de la privada. Cada uno de estos servidores se comunica con sus *blockchains* correspondientes.

DISEÑO EXPERIMENTAL

El conjunto de datos (*dataset*) utilizado para las pruebas fue "EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-Ray Images" (Bae et al., 2023). Este está compuesto por la combinación de dos *dataset*: MIMIC-CXR-VQA y EHRSQL (MIMIC-IV). El primero de ellos es un recurso integral diseñado para tareas de respuesta a preguntas visuales (*visual question answering*, VQA) en el ámbito médico, centrado principalmente en radiografías de tórax. Además, es derivado principalmente de los *datasets* MIMIC-CXR-JPG y Chest ImaGenome, obtenidos de Physionet. El segundo es una base de datos que contiene registros de salud electrónicos de pacientes, así como la información demográfica, los resultados de pruebas médicas, los diagnósticos, los procedimientos médicos, las medicaciones administradas, entre otros.

Asimismo, el *dataset* está estructurado en archivos JSON. Esto facilita su manipulación, análisis y transferencia a la BPRIV de HF, donde su integridad y seguridad se garantizan mediante Cintel y políticas de control de acceso.

Se realizaron pruebas de las funciones de los Cintel utilizando un fragmento del *dataset* previamente mencionado. Estas pruebas evaluaron métricas de uso de recursos, latencia, velocidad de consulta *off-chain* y costo por función (Mani et al., 2021; Singh et al., 2021). Para las mediciones se empleó un código presente en el servidor, que facilita la comunicación con la red pública.

Es importante señalar que, en BPUB, la autosostenibilidad se logra mediante recompensas a los nodos que realizan tareas específicas. Por ejemplo, en Bitcoin, los mineros reciben recompensas por validar bloques, mientras que en Ethereum y EOSIO, los validadores obtienen compensaciones por la creación de bloques de transacciones. Cada función ejecutada mediante Cintel genera una comisión que varía según la *blockchain* utilizada, además de poder dar propinas para incrementar la prioridad de la transacción enviada (Polkadot, 2024a).

Pruebas de estrés

Para evaluar el rendimiento de la BPUBPER y realizar pruebas de estrés, se enviaron mil peticiones en tres modalidades: concurrente, secuencial y por lotes (Tabla 4). Cabe destacar que tanto la BPRIV como la BPUBPER, junto con los servidores respectivos, se ejecutan en la misma máquina.

Tabla 4

Pruebas de estrés y descripción

Tipo de prueba	Descripción
Secuencial	Consiste en enviar una petición a la vez de manera secuencial al servidor sin esperar una respuesta exitosa o de rechazo
Concurrente	Consiste en enviar todas las peticiones al mismo tiempo al servidor sin esperar una respuesta exitosa o de rechazo
Lotes (batches)	Consiste en enviar un grupo de peticiones (diez peticiones) cada seis segundos, debido a que el tiempo de bloques de Polkadot es de seis segundos (Polkadot, 2024b), sin esperar una respuesta exitosa o de rechazo

Para crear las cuentas, los nombres y apellidos tienen de longitud entre cinco y seis caracteres, respectivamente, pues es la longitud media de los nombres y apellidos en el Perú (Sociedad LR, 2022), mientras que los *e-mails* tienen alrededor de veintidós caracteres (AtData, 2021).

Por otro lado, se realizaron pruebas de estrés con cien HME distribuidos aleatoriamente en los dos hospitales pertenecientes a HF, lo que garantizó la interoperabilidad entre ambas entidades. Esta prueba de estrés consistió en la generación de los cien HME, incluyendo los datos textuales en CouchDB y las imágenes convertidas a NFT en IPFS.

RESULTADOS

En la Tabla 5, se presentan las métricas obtenidas para las funcionalidades add_user, get_role y has_permissions, tras las pruebas de estrés con mil peticiones. En GET / has_permission/, la latencia media fue de 3,65 ms y la mediana de 3,51 ms, prácticamente con un consumo de CPU y RAM nulo. Esto confirma que las consultas de solo lectura al contrato son muy ligeras y pueden ejecutarse con alta frecuencia sin afectar los recursos del sistema.

La ruta GET /role/ mostró resultados similares, con una latencia media de 3,88 ms y mediana de 3,76 ms, manteniendo un uso de CPU mínimo y un incremento de memoria insignificante (0,0001 %) (Tabla 5). Ambas operaciones reflejan la eficiencia de las consultas de lectura en la red.

En cuanto a las operaciones de escritura, POST /create_user_with_dynamic_gas presentó una latencia media de 309,41 ms y mediana de 318,89 ms en modo secuencial, con un aumento casi nulo en el uso de CPU y solo un 0,0007 % de incremento en RAM (Tabla 5). El retardo provino principalmente del envío de la transacción a la red de Polkadot y del cálculo del tip dinámico previo a la firma, una sobrecarga razonable que mejora la fiabilidad sin penalizar el rendimiento local.

En modo *batch* (envío agrupado en lotes), la latencia media se elevó ligeramente a 365,66 ms (mediana de 364,55 ms) y el consumo de RAM aumentó hasta 0,009 %. Incluso así mantuvo un rendimiento eficiente (Tabla 5). Esto indica que el envío agrupado introduce una mínima sobrecarga, pero sigue siendo una opción viable para procesar múltiples transacciones de forma controlada y estable.

Tabla 5

Métricas estadísticas con mil peticiones

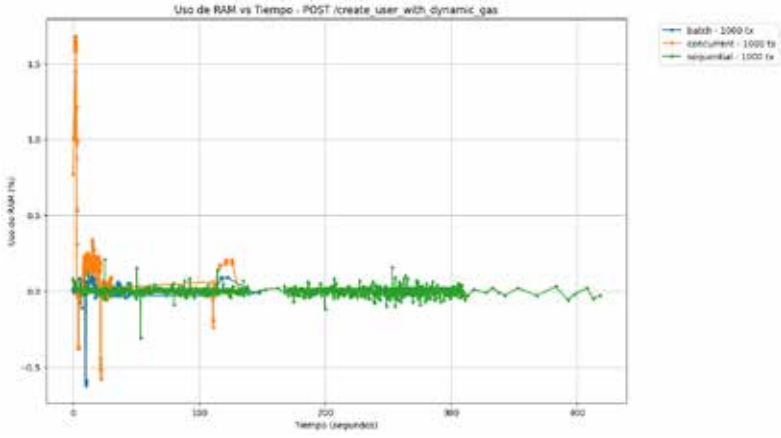
Ruta	Latencia media (ms)	Latencia mediana (ms)	Promedio de incremento del CPU (%)	Promedio de incremento de RAM (%)
GET /has_permission (secuencial)	3,65	3,51	0	0,0000
GET /role/ (secuencial)	3,88	3,76	0	0,0001
POST /create_user_with_dynamic_gas (secuencial)	309,41	318,89	0	0,0007
POST /create_user_with_dynamic_gas (batch)	365,66	364,55	0	0,0090

Como se observa en la Figura 12, la curva de uso de RAM muestra cómo, bajo carga secuencial (mil transacciones), disminuye y baja el uso del recurso, se mantiene uniforme a lo largo del tiempo con valores que oscilan entre 0,0 % a 0,3 % de uso. Con respecto a la modalidad concurrente de mil peticiones simultáneas, se observa que llega hasta 1,7 % para luego bajar rápidamente a 0 y tener una ligera subida posterior. Este comportamiento ocurre porque se colocan en cola las solicitudes y se van procesando lo más rápido posible.

Finalmente, con respecto al modo *batch*, este ofrece un comportamiento similar al modo secuencial en el uso de RAM, pero finaliza las peticiones más de cuatrocientos segundos antes. Los valores negativos en la gráfica aparecen porque, independientemente del uso de la arquitectura, el sistema operativo ejecuta procesos en segundo plano que se desarrollan en paralelo.

Figura 12

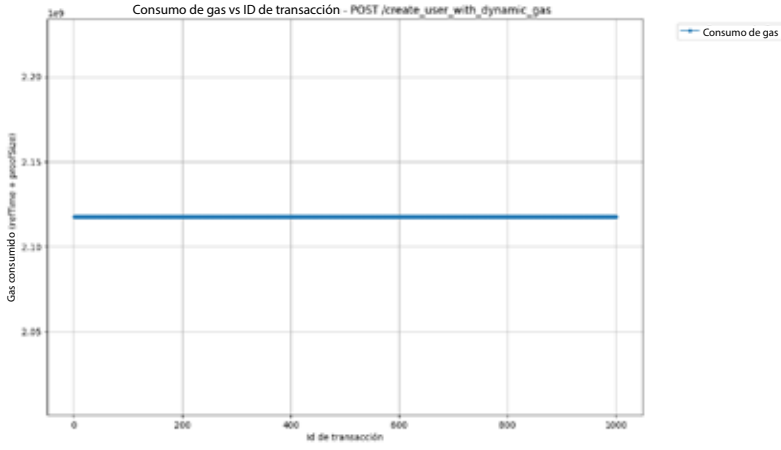
Uso de RAM en mil peticiones en el tiempo



En relación con la Figura 13, al graficar el *gas used* contra el identificador de cada transacción, vemos una línea prácticamente horizontal en torno a 2 117 705 426 unidades de gas. En cada transacción de las mil, la variabilidad es prácticamente nula, lo que implica que el mecanismo de gas dinámico consume una cantidad de gas similar en cada llamada. Esto se debe a que el peso de la transacción es el mismo para todas (misma longitud de nombre, apellido y *e-mail*) y que el tip aleatorio añadido sirve para priorizar las transacciones que llegan, sin afectar el costo base de gas. Cabe resaltar que el costo de transacción en la BPUBPER es fundamental para incentivar a los nodos de mantener la red funcionando.

Figura 13

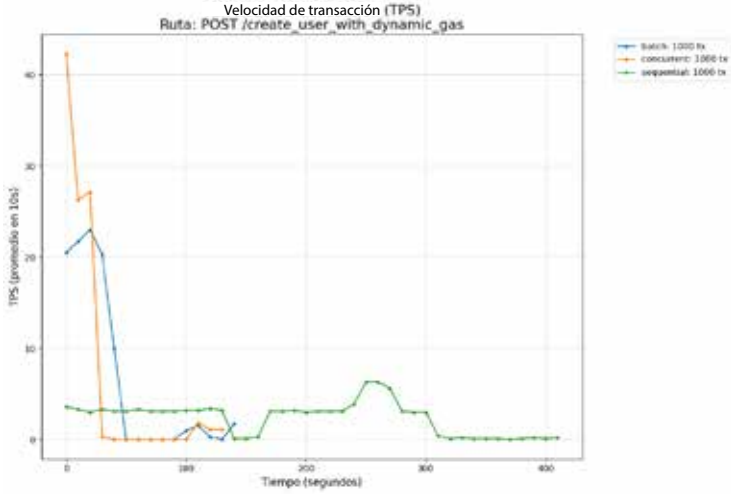
Uso de recursos consumo de gas vs ID de transacción



Con respecto a la Figura 14, el histograma de velocidad de transacción sobre tiempo muestra una acumulación muy marcada en el primer intervalo de cuarenta segundos. El pico inicial, fruto de inyectar las consultas en modo *batch* y el concurrente, significa que la *blockchain* ha recibido una gran cantidad de transacciones que van a ser procesadas por la *blockchain*. Asimismo, a pesar del modo concurrente de brindar todas las transacciones al inicio, tuvo una conclusión de envío de sus transacciones —al terminar de enviar las mil transacciones— casi igual al modo *batch*. Esto implica que no necesariamente hay una mejora significativa en la velocidad en envío de manera concurrente frente al envío en lotes. Con respecto al modo secuencial, este tuvo transacciones por segundo más estables que en los modos previos y una duración de más del doble de envío de transacciones.

Figura 14

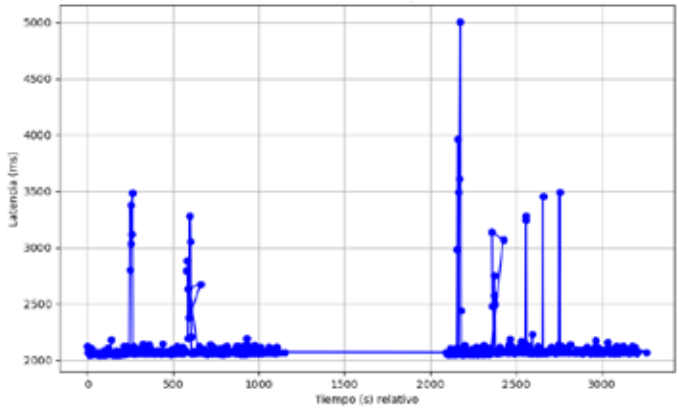
Velocidad de transacción sobre el tiempo



Por otro lado, se realizó la prueba de estrés en la BPRIV para la generación de cien HME, considerando las mismas métricas (latencia, uso de CPU y uso de RAM).

Figura 15

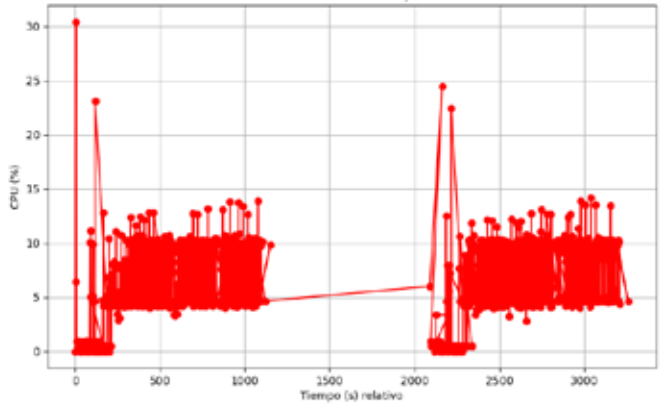
Latencia de consultas en milisegundos vs tiempo para cien creaciones de historiales médicos electrónicos



Como se puede visualizar en la Figura 16, con respecto a la latencia media de cada transacción, esta se sitúa muy cerca de los 2000 ms, lo que refleja el tiempo requerido para la propuesta, validación y *commit* en HF. Sin embargo, se aprecian brotes puntuales que llegan hasta los 3000 a 5000 ms, concentrados al inicio de cada gran bloque de transacciones (tanto en la primera fase como en el *minting*). Estos *outliers* coinciden con los momentos de alta concurrencia de llamadas (envío masivo de IPFS o de *minting* de NFT) y muestran cómo, ante picos de carga, el *ordering service* y los *peers* tardan más en procesar y confirmar los bloques. Aun así, la dispersión de la nube principal de puntos alrededor de los 2000 ms confirma un rendimiento bastante estable bajo condiciones normales.

Figura 16

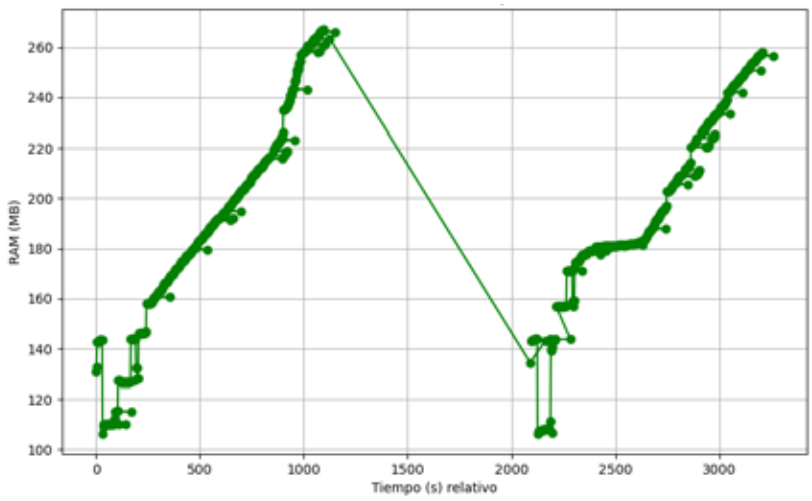
Porcentaje de uso de CPU vs tiempo para cien creaciones de historiales médicos electrónicos



Como podemos observar en la Figura 17, el gráfico de CPU revela dos fases bien diferenciadas a lo largo del *script*. En el arranque (primeros 200 segundos), el consumo oscila muy bajo, entre 0 % y 5 %, con picos aislados que alcanzan el 30 % al matricular al administrador y al procesar las primeras pocas transacciones. A continuación, durante la fase intensiva de creación de HME y subida de imágenes a IPFS (aproximadamente de 200 a 1000 segundos), el consumo se estabiliza entre un 6 % y un 14 %, lo que marca la mayor parte del trabajo de contrato más operaciones IPFS. Tras una pausa en la que el proceso se reinicia para la parte del *minting*, vuelve a aparecer un patrón muy similar (6 % a 12 %) con nuevos picos al iniciar esta segunda fase. En conjunto, se muestra que la mayor carga de CPU viene de los *loops* de invocación de Cintel y de *input/output* con IPFS, pero que la infraestructura de HF mantiene un uso moderado, incluso con cientos de transacciones.

Figura 17

Uso de RAM en megabytes vs tiempo para cien creaciones de historiales médicos electrónicos



Como podemos deducir de la Figura 17, el consumo de memoria arranca en unos 105 MB y crece de manera casi lineal hasta rondar los 265 MB conforme avanza la fase de HME más subida a IPFS (entre 0 y 1100 s). Después del corte (cuando se pasa al *minting*), observamos una caída brusca hasta ~135 MB, seguida de un nuevo crecimiento (hasta 260 MB en la segunda fase). Este patrón de subida sostenida indica que muchos objetos (*gateways*, contratos, *buffers* de IPFS, estructuras de datos de métricas) permanecen en memoria mientras el bucle avanza. Para evitar este *memory bloat*, convendría reciclar o liberar explícitamente las conexiones y datos intermedios entre lotes, o bien volcar métricas y desconectar *gateways* más a menudo.

En la Tabla 6, se presentan los resultados estimados para las métricas clave obtenidas a través de la simulación del despliegue de la propuesta de tesis en una red de HF configurada con dos organizaciones. Cabe resaltar que las métricas incluyen latencia y utilización de recursos.

Tabla 6

Resultados promedios de la simulación de la blockchain *privada*

Métricas	Resultados promedio
Latencia (creación)	2100 ms
Utilización de CPU	7 %
Utilización de memoria	190 MB

Se usaron librerías como *matplotlib* y *pidusage* para poder extraer la latencia y utilización de recursos respectivamente. Como se puede apreciar en la Tabla 6, la latencia promedio de las funciones "Creación de HME" y "Compartir los HME" con un doctor de la organización demoraron aproximadamente 2100 ms en generar cada caso particular. Por otro lado, en la sección de utilización de recursos, al realizar un descarte de los procesos no relacionados a la arquitectura, se puede evidenciar una utilización de CPU del 7 %. Esta es una métrica adecuada para el despliegue de ambas *blockchains* y sus respectivos servidores. Además, la utilización de la memoria ha sido baja, cuyo promedio fue de 190 MB para todo el proceso y mostró un cese de crecimiento conforme el proceso duraba más.

DISCUSIÓN

Los experimentos evidencian que las consultas puramente lectoras (GET /has_permission, GET /role) en modo secuencial se resuelven entre 3,65 y 3,88 ms, con un uso de CPU e incremento de RAM imperceptible para ambos casos (< 0,01 %). Estas cifras sitúan nuestro enfoque por encima de soluciones basadas únicamente en BPUB, las cuales suelen registrar latencias superiores y un mayor consumo de recursos. En el caso de operaciones de escritura (POST /create_user_with_dynamic_gas), la introducción de lógica de gas dinámico añade una sobrecarga de 100 a 200 ms en promedio, pero mantiene la estabilidad de uso de CPU (< 0,5 %) y RAM (< 0,3 %), lo que indica que el cálculo de tip no penaliza la ejecución local. Bajo concurrencia intensa, se generan picos de latencia de hasta 5,8 s, atribuibles principalmente a colas en la red de nodos y al proceso de consenso distribuido en Polkadot, y no al procesamiento en el servidor, lo que sugiere que futuros trabajos podrían investigar técnicas de escalado horizontal o *batching* más sofisticado. En la BPRIV, la creación masiva de cien HME arrojó una latencia media de ≈2 s, con CPU estabilizada en torno al 7 % y un crecimiento de RAM hasta

265 MB en fase intensiva. La aparición de *memory bloat* en picos sucesivos indica la conveniencia de liberar o reciclar conexiones IPFS y *gateways* entre lotes.

En conjunto, los resultados confirman que la separación de roles (pública para permisos y privada para datos clínicos) ofrece un equilibrio óptimo entre seguridad, rendimiento y escalabilidad, si bien requiere ajustes de optimización en la capa de red y en la gestión de memoria.

CONCLUSIONES

En conclusión, nuestra propuesta de BHIB demuestra que es posible integrar de manera efectiva la trazabilidad e inmutabilidad de una red pública permisionada (Polkadot) con la privacidad y el control de una red privada (HF) para la gestión de HME. Los resultados experimentales confirman lecturas ultrarrápidas (3,65-3,88 ms) con consumo local prácticamente nulo y escrituras estables (200-500 ms para usuarios y ≈2 s para HME), incluso bajo cargas de hasta 150 TPS, lo que mantiene el uso de CPU por debajo del 7 % y de RAM por debajo del 0,3 %. Aunque se observaron picos de latencia en escenarios altamente concurrentes y un crecimiento de memoria durante la fase de HME, estas limitaciones apuntan a la necesidad de incorporar mecanismos de *batching* adaptativo, autoescalado de nodos y *pooling* de conexiones IPFS para optimizar aún más el rendimiento.

En conjunto, este trabajo valida no solo la viabilidad técnica y la eficiencia operativa de un sistema de HME basado en BHIB, sino también su potencial para el despliegue en entornos sanitarios distribuidos, proporcionando una base sólida para futuras investigaciones en *sharding*, contratos más ligeros y auditoría clínica.

Finalmente, a pesar de los resultados positivos, la implementación presenta limitaciones, como la necesidad de una infraestructura robusta para mantener el rendimiento y la seguridad. Además, estudios futuros podrían explorar la optimización de los Cintel. Se recomienda la creación de sesiones temporales que permitan eliminar los permisos de visualización luego de un periodo de tiempo.

REFERENCIAS

Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018). Big healthcare data: Preserving security and privacy. *Journal of Big Data*, *5*(1), 1. https://doi.org/10.1186/s40537-017-0110-7

Abunadi, I., & Kumar, R. (2021). BSF-EHR: Blockchain security framework for electronic health records of patients. *Sensors*, *21*(8), 2865. https://doi.org/10.3390/s21082865

Ali, A., Rahim, H. A., Ali, J., Pasha, M. F., Masud, M., Rehman, A. U., Chen, C., & Baz, M. (2021). A novel secure blockchain framework for accessing electronic health records using multiple certificate authority. *Applied Sciences*, *11*(21), 9999. https://doi.org/10.3390/app11219999

Ali, A., Rahim, H. A., Pasha, M. F., Dowsley, R., Masud, M., Ali, J., & Baz, M. (2021). Security, privacy, and reliability in digital healthcare systems using blockchain. *Electronics*, *10*(16), 2034. https://doi.org/10.3390/electronics10162034

Amanat, A., Rizwan, M., Maple, C., Zikria, Y. B., Almadhor, A., & Kim, S. W. (2022). Blockchain and cloud computing-based secure electronic healthcare records storage and sharing. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.938707

Antwi, M., Adnane, A., Ahmad, F., Hussain, R., Habib Ur Rehman, M., & Kerrache, C. A. (2021). The case of HyperLedger Fabric as a blockchain solution for healthcare applications. *Blockchain: Research and Applications*, *2*(1), 100012. https://doi.org/10.1016/j.bcra.2021.100012

AtData. (2021, 22 de junio). *How long is the average email address?* https://atdata.com/blog/long-email-addresses/

Awad Abdellatif, A., Samara, L., Mohamed, A., Erbad, A., Chiasserini, C. F., Guizani, M., O'Connor, M. D., & Laughton, J. (2021). MEdge-Chain: Leveraging edge computing and blockchain for efficient medical data exchange. *IEEE Internet of Things Journal*, *8*(21), 15762-15775. https://doi.org/10.1109/JIOT.2021.3052910

Bae, S., Kyung, D., Ryu, J., Cho, E., Lee, G., Kweon, S., Oh, J., Ji, L., Chang, E. I., Kim, T., & Choi, E. (2023). EHRXQA: A multi-modal question answering dataset for electronic health records with chest X-ray images. En A. Oh, T. Naumann & A. Globerson (Eds.), *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 3867-3880). Curran Associates. https://dl.acm.org/doi/10.5555/3666122.3666292

Chelladurai, U., & Pandian, S. (2022). A novel blockchain based electronic health record automation system for healthcare. *Journal of Ambient Intelligence and Humanized Computing*, *13*(1), 693-703. https://doi.org/10.1007/s12652-021-03163-3

Chelladurai, U., Pandian, S., & Ramasamy, K. (2021). A blockchain based patient centric electronic health record storage and integrity management for e-Health systems. *Health Policy and Technology*, *10*(4), 100513. https://doi.org/10.1016/j.hlpt.2021.100513

Dagher, G. G., Mohler, J., Milojkovic, M., & Marella, P. B. (2018). Ancile: Privacy-preserving framework for access control and interoperability of electronic health records

using blockchain technology. *Sustainable Cities and Society*, *39*, 283-297. https://doi.org/10.1016/j.scs.2018.02.014

Exceline, E., & Nagarajan, S. (2024). Flexible access control mechanism for cloud stored EHR using consortium blockchain. *International Journal of System Assurance Engineering and Management*, *15*, 503-518. https://doi.org/10.21203/rs.3.rs-397642/v1

Fatokun, T., Nag, A., & Sharma, S. (2021). Towards a blockchain assisted patient owned system for electronic health records. *Electronics*, *10*(5), 580. https://doi.org/10.3390/electronics10050580

Guo, H., Li, W., Nejad, M., & Shen, C.-C. (2019). Access control for electronic health records with hybrid blockchain-edge architecture. En *2019 IEEE International Conference on Blockchain (Blockchain)* (pp. 44-51). IEEE. https://doi.org/10.1109/Blockchain.2019.00015

Haas, S. Wohlgemuth, I. Echizen, S., Sonehara, N., & Müller, G. (2011). Aspects of privacy for electronic health records. *International Journal of Medical Informatics*, *80*(2), e26-e31. https://doi.org/10.1016/j.ijmedinf.2010.10.001

Haddad, A., Habaebi, M. H., Elsheikh, E. A. A., Islam, Md. R., Zabidi, S. A., & Suliman, F. E. M. (2024). E2EE enhanced patient-centric blockchain-based system for EHR management. *Plos One*, *19*(4), e0301371. https://doi.org/10.1371/journal.pone.0301371

Hashim, F., Shuaib, K., & Sallabi, F. (2022). Connected blockchain federations for sharing electronic health records. *Cryptography*, *6*(3), 47. https://doi.org/10.3390/cryptography6030047

Hossain Faruk, M. J., Shahriar, H., Valero, M., Sneha, S., Ahamed, S. I., & Rahman, M. (2021). Towards blockchain-based secure data management for remote patient monitoring. En *2021 IEEE International Conference on Digital Health* (pp. 299-308). IEEE. https://doi.org/10.1109/ICDH52753.2021.00054

Huang, H., Sun, X., Xiao, F., Zhu, P., & Wang, W. (2021). Blockchain-based eHealth system for auditable EHRs manipulation in cloud environments. *Journal of Parallel and Distributed Computing*, *148*, 46-57. https://doi.org/10.1016/j.jpdc.2020.10.002

Hussien, H., Yasin, S., Udzir, N., & Ninggal, M. (2021). Blockchain-based access control scheme for secure shared personal health records over decentralised storage. *Sensors*, *21*(7), 2462. https://doi.org/10.3390/s21072462

Ink. (s. f.). *Ink vs. CosmWasm*. https://use.ink/docs/v5/ink-vs-cosmwasm/

International Organization for Standardization. (2019). *ISO 13606-1:2019(en) Health informatics — Electronic health record communication — Part 1: Reference model*. https://www.iso.org/obp/ui/en/#iso:std:iso:13606:-1:ed-2:v1:en

Jayabalan, J., & Jeyanthi, N. (2022). Scalable blockchain model using off-chain IPFS storage for healthcare data security and privacy. *Journal of Parallel and Distributed Computing*, *164*, 152-167. https://doi.org/10.1016/j.jpdc.2022.03.009

Jin, H., Luo, Y., Li, P., & Mathew, J. (2019). A review of secure and privacy preserving medical data sharing. *IEEE Access*, *7*, 61656-61669. https://doi.org/10.1109/ACCESS.2019.2916503

Kang, P., Yang, W., & Zheng, J. (2022). Blockchain private file storage-sharing method based on IPFS. *Sensors*, *22*(14), 5100. https://doi.org/10.3390/s22145100

Kaur, J., Rani, R., & Kalra, N. (2023). Attribute-based access control scheme for secure storage and sharing of EHRs using blockchain and IPFS. *Cluster Computing*, *27*, 1047-1061. https://doi.org/10.1007/s10586-023-04038-2

Kumari, D., Singh, A., Sunil, G., Mishra, K., & Panda, S. (2024). HealthRec-chain: Patient-centric blockchain enabled IPFS for privacy preserving scalable health data. *Computer Networks*, *241*, 110223. https://doi.org/10.1016/j.comnet.2024.110223

Lee, J.-S., Chew, C.-J., Liu, J.-Y., Chen, Y.-C., & Tsai, K.-Y. (2022). Medical blockchain: Data sharing and privacy preserving of EHR based on smart contract. *Journal of Information Security and Applications*, *65*, 103117. https://doi.org/10.1016/j.jisa.2022.103117

Li, P., Zhou, D., Ma, H., & Lai, J. (2024). Flexible and secure access control for EHR sharing based on blockchain. *Journal of Systems Architecture*, *146*, 103033. https://doi.org/10.1016/j.sysarc.2023.103033

Liu, J., Jiang, W., Sun, R., Kashif, A., Dahman, M.i, Hua, Q., & Yu, K. (2023). Conditional anonymous remote healthcare data sharing over blockchain. *IEEE Journal of Biomedical and Health Informatics*, *27*(5), 2231-2242. https://doi.org/10.1109/JBHI.2022.3183397

Mandarino, V., Pappalardo, G., & Tramontana, E. (2024). A blockchain-based electronic health record (EHR) system for edge computing enhancing security and cost efficiency. *Computers*, *13*(6), 132. https://doi.org/10.3390/computers13060132

Mani, V., Manickam, P., Alotaibi, Y., Alghamdi, S., & Khalaf, O. I. (2021). Hyperledger healthchain: Patient-centric IPFS-based storage of health records. *Electronics*, *10*(23), 3003. https://doi.org/10.3390/electronics10233003

Mauricio, D., Llanos-Colchado, P. C., Cutipa-Salazar, L. S., Castañeda, P., Chuquimbalqui-Maslucán, R., Rojas-Mezarina, L., & Castillo-Sequera, J. L. (2024). Electronic health record interoperability system in Peru using blockchain. *International Journal of Online and Biomedical Engineering*, *20*(3), 136-153. https://doi.org/10.3991/ijoe.v20i03.44507

Miyachi, K., & Mackey, T. K. (2021). hOCBS: A privacy-preserving blockchain framework for healthcare data leveraging an on-chain and off-chain system design. *Information Processing & Management*, *58*(3), 102535. https://doi.org/10.1016/j.ipm.2021.102535

Mohey Eldin, A., Hossny, E., Wassif, K., & Omara, F. A. (2023). Federated blockchain system (FBS) for the healthcare industry. *Scientific Reports*, *13*(1), 2569. https://doi.org/10.1038/s41598-023-29813-4

Montañez-Valverde, R. A., Montenegro-Idrogo, J. J., & Vásquez-Alva, R. (2015). Pérdida de información en historias clínicas: más allá de la calidad en el registro. *Revista Médica de Chile*, *143*(6), 812. https://doi.org/10.4067/S0034-98872015000600017

Mulligan, E., & Braunack-Mayer, A. (2004). Why protect confidentiality in health records? A review of research evidence. *Australian Health Review*, *28*(1), 48-55. https://doi.org/10.1071/AH040048

Nhan, T., Upadhyay, K., & Poudel, K. (2024). Towards patient-centric healthcare: Leveraging blockchain for electronic health records. En S. Zaza & C. Riemenscheinder (Eds.), *SIGMIS-CPR '24: Proceedings of the 2024 Computers and People Research Conference* (pp. 1-8). https://doi.org/10.1145/3632634.3655883

Omar, A. A., Jamil, A. K., Khandakar, A., Uzzal, A. R., Bosri, R., Mansoor, N., & Rahman, M. S. (2021). A transparent and privacy-preserving healthcare platform with novel smart contract for smart cities. *IEEE Access*, *9*, 90738-90749. https://doi.org/10.1109/ACCESS.2021.3089601

Polkadot. (2024a). *Transaction fees*. https://wiki.polkadot.network/docs/learn-transaction-fees

Polkadot. (2024b). *Consensus*. https://wiki.polkadot.com/learn/learn-consensus/

Puneeth, R. P., & Parthasarathy, G. (2024). Blockchain-based framework for privacy preservation and securing EHR with patient-centric access control. *Acta Informatica Pragensia*, *13*(1), 1-23. https://doi.org/10.18267/j.aip.225

Rajput, A. R., Li, Q., & Ahvanooey, M. T. (2021). A blockchain-based secret-data sharing framework for personal health records in emergency condition. *Healthcare*, *9*(2), 206. https://doi.org/10.3390/healthcare9020206

Samala, A. D., & Rawas, S. (2024). Transforming healthcare data management: A blockchain-based cloud EHR system for enhanced security and interoperability. *International Journal of Online and Biomedical Engineering*, *20*(2), 46-60. https://doi.org/10.3991/ijoe.v20i02.45693

Selvarajan, S., & Mouratidis, H. (2023). Author correction: A quantum trust and consultative transaction-based blockchain cybersecurity model for healthcare systems. *Scientific Reports*, *13*, 9409. https://doi.org/10.1038/s41598-023-36573-8

Sociedad LR. (2022). Top de los nombres más populares de niños y niñas en Perú, según Reniec. *La República*. https://larepublica.pe/sociedad/2022/09/12/reniec-top-de-los-nombres-mas-populares-de-ninos-y-ninas-en-peru-dni-nombres-comunes-peru-2022

Shen, B., Guo, J., & Yang, Y. (2019). MedChain: Efficient healthcare data sharing via blockchain. *Applied Sciences*, *9*(6), 1207. https://doi.org/10.3390/app9061207

Singh, A. P., Pradhan, N. R., Luhach, A. K., Agnihotri, S., Jhanjhi, N. Z., Verma, S., Kavita, Ghosh, U., & Roy, D. S. (2021). A novel patient-centric architectural framework for blockchain-enabled healthcare applications. *IEEE Transactions on Industrial Informatics*, *17*(8), 5779-5789. https://doi.org/10.1109/TII.2020.3037889

Tanwar, N., & Thakur, J. (2023). Patient-centric soulbound NFT framework for electronic health record (EHR). *Journal of Engineering and Applied Science*, *70*(33). https://doi.org/10.1186/s44147-023-00205-9

Uddin, M., S. Memon, M., Memon, I., Ali, I., Memon, J., Abdelhaq, M., & Alsaqour, R. (2021). Hyperledger fabric blockchain: Secure and efficient solution for electronic health records. *Computers, Materials & Continua*, *68*(2), 2377-2397. https://doi.org/10.32604/cmc.2021.015354

Uppal, S., Kansekar, B., Mini, S., & Tosh, D. (2023). HealthDote: A blockchain-based model for continuous health monitoring using interplanetary file system. *Healthcare Analytics*, *3*, 100175. https://doi.org/10.1016/j.health.2023.100175

Wang, M., Guo, Y., Zhang, C., Wang, C., Huang, H., & Jia, X. (2021). MedShare: A privacy-preserving medical data sharing system by using blockchain. *IEEE Transactions on Services Computing*, *16*(1). https://doi.org/10.1109/TSC.2021.3114719

Wu, G., Wang, S., Ning, Z., & Zhu, B. (2022). Privacy-preserved electronic medical record exchanging and sharing: A blockchain-based smart healthcare system. *IEEE Journal of Biomedical and Health Informatics*, *26*(5), 1917-1927. https://doi.org/10.1109/JBHI.2021.3123643

Xiao, Y., Xu, B., Jiang, W., & Wu, Y. (2021). The health chain blockchain for electronic health records: Development study. *Journal of Medical Internet Research*, *23*(1), e13556. https://doi.org/10.2196/13556

Zhang, G., Yang, Z., & Liu, W. (2022). Blockchain-based privacy preserving e-health system for healthcare data in cloud. *Computer Networks*, *203*, 108586. https://doi.org/10.1016/j.comnet.2021.108586

ARQUITECTURAS AGÉNTICAS DE IA: UN ESTUDIO COMPARATIVO DE *WORKFLOWS* (FLUJOS DE TRABAJO) VERSUS A2A (*AGENT TO AGENT*) Y SU APLICACIÓN EN LOS SECTORES DE EDUCACIÓN, INDUSTRIA Y SERVICIOS

OLDA BUSTILLOS ORTEGA
https://orcid.org/0000-0003-2822-3428
obustillos@uia.ac.cr
Escuela de Ingeniería Informática, Universidad Internacional de las Américas, Costa Rica

JORGE MURILLO GAMBOA
https://orcid.org/0000-0001-5548-8283
jmurillo@uia.ac.cr
Escuela de Ingeniería Informática, Universidad Internacional de las Américas, Costa Rica

DANIEL MENA BOCKER
https://orcid.org/0009-0002-1062-6675
dfmenab@edu.uia.ac.cr
Escuela de Ingeniería Informática, Universidad Internacional de las Américas, Costa Rica

CARLOS DE LA O FONSECA
https://orcid.org/0009-0002-3413-4990
cdelaof@uia.ac.cr
Escuela de Ingeniería Informática, Universidad Internacional de las Américas, Costa Rica

CARLOS AGUILAR MORA
https://orcid.org/0000-0003-0946-6053
chaguilarm@uia.ac.cr
Escuela de Ingeniería Informática, Universidad Internacional de las Américas, Costa Rica

Recibido: 26 de setiembre del 2025 / Aceptado: 23 de octubre del 2025
doi: https://doi.org/10.26439/interfases2025.n022.8430

RESUMEN. El desarrollo de arquitecturas agénticas de inteligencia artificial (IA) ha dado lugar a nuevos modelos de interacción y automatización, entre los que destacan los flujos

de trabajo basados en agentes de IA (*agentic workflows*) y la comunicación entre agentes A2A (*agent-to-agent*). Se presenta un estudio comparativo entre ambas arquitecturas, analizando sus principios de diseño, capacidades de coordinación humano-IA, escalabilidad y adaptabilidad en distintos entornos, así como sus obstáculos y desafíos. El análisis evidencia que, si bien los *workflows* agénticos proporcionan eficiencia y control en entornos estables con procesos delimitados, las arquitecturas A2A sobresalen en contextos distribuidos y heterogéneos, al ofrecer mayor flexibilidad y autonomía. Todo ello bajo la premisa de que la tecnología debe fortalecer, y no sustituir, las capacidades humanas, generando un impacto significativo en los ámbitos en los que más se requiere. Se examinan ejemplos representativos en los sectores de educación, salud e industria y se analiza cómo estas arquitecturas transforman procesos clave. Se presentan figuras y tablas comparativas que integran diversos enfoques entre ambas arquitecturas agénticas de IA. Finalmente, se discuten los desafíos técnicos y éticos asociados con su implementación y se plantean líneas futuras de investigación para una adopción responsable y eficaz de estas tecnologías.

PALABRAS CLAVE: agentes / educación / flujos de trabajo / inteligencia artificial / tecnología

AI AGENT ARCHITECTURES: A COMPARATIVE STUDY OF WORKFLOWS VERSUS A2A (AGENT TO AGENT) AND THEIR APPLICATION IN THE EDUCATION, INDUSTRY AND SERVICES SECTORS

ABSTRACT. The development of agentic architectures for artificial intelligence (AI) has led to new models of interaction and automation, most notably AI agent-based workflows and agent-to-agent (A2A) communication. This paper presents a comparative study of both architectures, analyzing their design principles, human-AI coordination capabilities, scalability and adaptability in different environments, as well as obstacles and challenges. The analysis shows that while agentic workflows provide efficiency and control in stable environments with well-defined processes, A2A architectures excel in distributed and heterogeneous contexts, offering greater flexibility and autonomy. Based on the premise that technology should strengthen, not replace, human capabilities, this study aims to generate a significant impact in the areas where it is most needed. Representative examples from the education, healthcare, and industrial sectors are examined, demonstrating how these architectures transform key processes. Comparative figures and tables are presented, integrating various approaches between the two agentic AI architectures. Finally, the technical and ethical challenges associated with their implementation are discussed, and future lines of research are proposed for a responsible and effective adoption of these technologies.

KEYWORDS: agents / artificial intelligence / education / technology / workflows

INTRODUCCIÓN

En el ámbito de las ciencias de la computación, la inteligencia artificial (IA) se puede definir como una disciplina orientada al diseño de sistemas informáticos y modelos algorítmicos capaces de reproducir o emular procesos cognitivos propios de la inteligencia humana. Los sistemas de automatización inteligente, que combinan procesos automatizados e inteligencia artificial para gestionar flujos de trabajo estructurados, constituyen la base de los agentes de IA modernos.

Un agente de IA generativa se puede definir como una aplicación autónoma diseñada para alcanzar un objetivo específico, que consigue al percibir su entorno y contar con herramientas disponibles para tomar decisiones y ejecutar acciones. Puede razonar y actuar sin supervisión humana directa ni instrucciones explícitas (Viswanathan et al., 2025).

Estos agentes de IA requieren autonomía y flexibilidad para operar en entornos dinámicos, impredecibles y sociales, siendo su objetivo tomar decisiones proactivas y apoyar al usuario en tareas, enseñanza y supervisión. La capacidad de aprendizaje, esencial para el comportamiento inteligente, les permite adaptarse a entornos desconocidos y cambiantes mediante técnicas sofisticadas. Se vislumbra un futuro en el que estos sistemas mejoran significativamente sectores como la industria, la salud, la educación y la resolución de problemas complejos, generando un impacto real más allá de la mera automatización (Shoham & Leyton-Brown, 2008).

De acuerdo con Xu y King (2025), la IA se encuentra en un punto de inflexión, pues sus rápidos avances están transformando industrias, economías y sociedades, al tiempo que generan una serie de desafíos complejos que requieren atención tanto por parte de la comunidad investigadora como de los profesionales del sector. Desde una perspectiva técnica, la IA presenta limitaciones significativas en cuanto a generalización, interpretabilidad, robustez y eficiencia energética.

Si bien los sistemas actuales muestran un alto rendimiento en dominios específicos, aún enfrentan dificultades en el aprendizaje por transferencia, el razonamiento bajo incertidumbre y la integración coherente de información multimodal. Esta brecha evidencia la necesidad de innovar en algoritmos, arquitecturas y paradigmas de aprendizaje.

Ramalingam (2025), basado en su experiencia implementando agentes de IA, ha enfrentado diversos obstáculos, entre los que destaca los siguientes:

- La integración con sistemas heredados y el cumplimiento de regulaciones estrictas.
- La resistencia a la reingeniería de procesos dentro de los flujos de trabajo establecidos complica la adopción de soluciones impulsadas por IA.
- Los beneficios potenciales incluyen niveles sin precedentes de eficiencia y escalabilidad.

Bajo la premisa de que la tecnología de IA debe fortalecer, y no sustituir, las capacidades humanas, a fin de generar un impacto significativo en los ámbitos en los que más se requiere, es fundamental que se desarrollen máquinas inteligentes que aprendan de la experiencia, se adapten a nuevas condiciones y ejecuten tareas tradicionalmente asociadas al pensamiento humano. Por ejemplo, el reconocimiento del lenguaje hablado, la toma de decisiones, la traducción automática o el procesamiento visual mediante visión por computadora, entre otros usos.

Se evidencia la necesidad de avanzar hacia una IA agéntica que integre de manera eficaz el procesamiento y razonamiento con la ejecución autónoma, lo cual permitiría transformar la interacción entre humanos y máquinas, optimizando los flujos de procesos y la orquestación colaborativa entre agentes autónomos, además de potenciar la automatización inteligente.

El foco principal del trabajo descrito en este artículo son las arquitecturas agénticas de IA, particularmente los modelos de flujos de trabajo (*AI workflows)* y los protocolos entre agentes A2A: no solo de manera comparativa, sino como herramientas conceptuales y prácticas que sirvan para avanzar teórica y prácticamente en el campo de la IA.

Los términos *AI workflow* y flujo de trabajo de IA se emplean de manera intercambiable en este artículo, aludiendo al mismo significado.

Este artículo se subdivide en los siguientes temas.

- Automatizando con IA: despliegue y limitaciones, agentes de IA, colaboración humano-IA, marco de referencia SPAR y la metáfora del árbol de IA.
- AI workflows y protocolos A2A: dimensiones de comparación, operando con flujos de trabajo de IA y agentes A2A, relevancia de los enfoques y tabla comparativa de características clave.
- Ejemplos de automatización en educación, salud, industria y servicios.
- Discusión y resultados: implicaciones para el diseño de sistemas inteligentes (workflow y A2A), consideraciones éticas y sociales de la adopción de estas arquitecturas.

Como productos de esta investigación, se ofrecen también tablas, figuras y glosarios.

OBJETIVOS DE LA INVESTIGACIÓN

Examinar las arquitecturas agénticas de inteligencia artificial, los flujos de trabajo automatizados (*AI workflows*) y la interacción de agentes autónomos (A2A), con el fin de comprender su estructura funcional, capacidades operativas, coordinación, escalabilidad y potencial de aplicación en los sectores de salud, educación, industria y servicios.

Entre los objetivos específicos de la investigación, se consideran los siguientes:

1. Caracterizar conceptualmente las arquitecturas *AI workflows* y A2A, sus fundamentos teóricos, modelos de diseño y principios operativos en el marco de la inteligencia artificial distribuida.
2. Analizar las diferencias estructurales y funcionales entre ambas arquitecturas, las dimensiones de autonomía, escalabilidad, adaptabilidad, coordinación y eficiencia.
3. Explorar aplicaciones representativas de cada arquitectura en los sectores educativo, industrial y de servicios, estudios de caso e impactos en los procesos organizacionales.
4. Proponer orientaciones teóricas y metodológicas que contribuyan al desarrollo, implementación y evaluación de arquitecturas agénticas de IA desde una perspectiva crítica, ética y contextualizada.

Se busca responder a la pregunta de investigación: ¿cómo se comparan las arquitecturas agénticas de *workflow* y de agente-a-agente (A2A) en términos de autonomía, escalabilidad, adaptabilidad y eficacia en su aplicación a los sectores de educación, salud, industria y servicios?

METODOLOGÍA

Se aplicará un enfoque metodológico comparativo, exploratorio y documental, utilizando criterios de comparación entre modelos de *workflows* agénticos y protocolos de agente-a-agente (A2A), aplicado a los sectores educación, salud, industria y servicios.

La revisión de literatura (análisis de documentos) se realizó utilizando los criterios de búsqueda en temas sobre *workflows* agénticos y protocolos A2A y sus impactos en diversos sectores.

Se accedió a diferentes bases de datos, tales como Google Académica, ProQuest Digital Dissertation and Theses, IEEE Xplore y Academia.edu.

Los criterios de selección aplicados se definieron utilizando búsquedas con palabras clave sobre IA en *workflows* y A2A, principalmente.

Se excluyen de esta investigación las búsquedas sobre ChatGPT, lenguajes de programación o bases de datos asociadas con el tema de la inteligencia artificial.

AUTOMATIZANDO CON IA

Despliegue y limitaciones de la IA

De acuerdo con Sedat Arslan (2025), la IA está transformando la industria alimentaria al incorporar técnicas de aprendizaje automático, procesamiento de lenguaje natural y redes neuronales para fortalecer tanto la seguridad como la nutrición. Ofrece recomendaciones individualizadas mediante la detección de contaminantes, la optimización del almacenamiento y la trazabilidad con *blockchain*[1], basadas en datos genéticos, metabólicos, del microbioma y del estilo de vida. Tenemos nutricionistas digitales impulsados por IA ofreciendo apoyo dietético en ambientes en los que existe un acceso limitado a profesionales de la salud.

Surge acá una paradoja: muchas organizaciones que buscan implementar automatizaciones se ven restringidas por el hecho de que, pese a contar con sistemas de IA generativa altamente sofisticados en su capacidad de razonamiento, estos carecen en la práctica de la habilidad para ejecutar acciones efectivas. Pueden analizar datos complejos en segundos, crear presentaciones convincentes y ofrecer perspectivas sobre cualquier tema. Sin embargo, aunque la IA ha alcanzado un alto nivel de razonamiento, sigue sin poder actuar de manera autónoma, lo que obliga a que los trabajadores del conocimiento dediquen gran parte de su tiempo (hasta 60 %) a tareas mecánicas, como supervisar y ejecutar manualmente las recomendaciones generadas por estos sistemas. "Tratamos a los humanos como robots y a la IA como creativos y es hora de invertir la ecuación" (Bornet et al., 2025, p. 2).

A manera de ejemplos ilustrativos, se presentan tres casos sobre automatizaciones con IA: empresa de servicios, industria alimentaria y robots de inspección inteligentes.

- Una empresa del sector servicios enfrentaba dificultades para optimizar su atención al cliente. Frente a ello, implementó un chatbot de alta precisión que interpreta y responde consultas, junto con bots de automatización de procesos (RPA)[2] que ejecuta secuencias dentro de flujos de trabajo predefinidos. No obstante, existía una limitación clave: la falta de un enlace efectivo entre la comprensión proporcionada por el chatbot y la ejecución automática de acciones. El personal de atención al cliente seguía actuando como intermediario, trasladando manualmente las recomendaciones del chatbot y

1 La cadena de bloques (*blockchain*) constituye un registro digital en continua expansión, integrado por múltiples bloques de datos organizados de forma cronológica y enlazados entre sí mediante mecanismos criptográficos que garantizan su integridad y seguridad (Binance Academy, s. f.).

2 La RPA (*robotic process automation*). es un método de automatización de los procesos de negocio. Emplea robots de *software* (bots) para automatizar tareas digitales que suelen realizar los humanos. Trabaja junto con herramientas de IA, lo que incluye la IA generativa (Automation Anywhere, 2025).

activando los flujos de trabajo. La próxima evolución de la IA pretende mejorar la capacidad de razonamiento, con mayor autonomía y capacidad para actuar de forma independiente.

- En la industria alimentaria, la IA estaba revolucionando los procesos mediante la aplicación de aprendizaje automático, procesamiento de lenguaje natural y redes neuronales para generar recomendaciones nutricionales personalizadas y mejorar la trazabilidad a través de tecnología blockchain. A pesar de que ya poseían una alta capacidad de razonamiento, los sistemas de IA generativa carecían de autonomía para ejecutar acciones, lo que obligaba a los trabajadores a realizar tareas repetitivas y supervisar manualmente los resultados.
- Los robots de inspección inteligentes han demostrado sus ventajas en diversos escenarios, como galerías de tuberías, túneles y subestaciones eléctricas, pero su aplicación generalizada aún es insuficiente y su desarrollo diversificado requiere mejoras urgentes. Se han logrado avances significativos en el control y monitoreo remoto, pero persisten ciertas limitaciones, tales como problemas de seguridad, procesamiento, análisis de datos y respuesta en tiempo real, entre otros. Debido a la complejidad e inestabilidad en los sistemas de monitoreo remoto multirobot, se han propuesto arquitecturas multiagente adaptadas a la teleoperación y con mayor autonomía realizando tareas (Li et al., 2025).

Estos tres ejemplos ponen de manifiesto un desfase entre comprensión y acción, junto con la necesidad de desarrollar una inteligencia artificial agéntica que integre de manera coherente el análisis de datos, el razonamiento y la ejecución autónoma. Un avance de este tipo permitiría transformar la interacción entre humanos y máquinas (sectores de servicios e industria alimentaria) o bien optimizar los flujos de procesos y orquestar a múltiples agentes autónomos, como sucede en el caso de los robots inteligentes para inspección remota.

Se evidencia un despliegue a gran escala en soluciones de IA y, de acuerdo con Frank X. Shaw (2025), unos quince millones de desarrolladores utilizan GitHub Copilot para programar, revisar y desplegar código. Cientos de miles de usuarios se sirven de él para hacer investigación y desarrollar soluciones y más de 230 000 organizaciones emplean Copilot Studio para crear agentes de IA y automatizaciones. En suma, no solo están transformando los procesos de desarrollo de *software* de IA, sino también la forma en que los individuos, equipos y organizaciones realizan su trabajo.

En resumen, se expande una visión emergente de la web agentiva abierta, en la que agentes de IA operan de manera autónoma en contextos individuales, organizacionales y empresariales de extremo a extremo, tomando decisiones y ejecutando tareas en representación de los usuarios e instituciones. Sin embargo, aparecen limitaciones con respecto a la habilidad para ejecutar acciones efectivas de manera completamente autónoma.

A pesar de sus limitaciones, la IA se consolida rápidamente como una disciplina central de las ciencias de la computación, orientada a reproducir funciones cognitivas humanas mediante algoritmos capaces de aprender, adaptarse y actuar en entornos complejos.

Agentes de IA

Los agentes de IA son entidades de *software* que emplean técnicas de inteligencia artificial para ejecutar tareas y alcanzar objetivos sin necesidad de entradas de datos explícitos ni resultados predefinidos. Pueden recibir instrucciones, diseñar planes de acción, utilizar herramientas para completar tareas y generar resultados dinámicos.

Las arquitecturas de IA apoyada con agentes están reconfigurando procesos fundamentales: desde la gestión del conocimiento y la personalización del aprendizaje, hasta la automatización industrial y la optimización de los servicios de atención al cliente.

¿Cuándo aparecieron los agentes de IA?

El año 2022 se considera como el nacimiento de los agentes de IA, cuyo campo sigue evolucionando rápidamente y casi a diario surgen nuevas posibilidades, con el objetivo de explicar estas nuevas tecnologías, y también de ofrecer herramientas para las personas y empresas; todo con el propósito de construir un mundo mejor, en el que la computación basada en agentes constituye un dominio científico con amplia posibilidad de difusión. Se busca superar la brecha entre el razonamiento y la acción, por lo que se requiere avanzar hacia una inteligencia artificial de carácter agéntico, capaz de integrar de forma autónoma los procesos de comprensión y ejecución, como una nueva etapa en la relación entre tecnología, conocimiento y práctica humana.

¿Cómo funcionan los agentes de IA?

Existen dos tipos de agentes de IA: la generativa, cuyo enfoque principal es la producción de contenido, y la IAA (inteligencia artificial agéntica), concebida para actuar de forma proactiva en representación de los usuarios, minimizando (o eliminando) la necesidad de contar con una supervisión humana y adquiriendo la capacidad de tomar decisiones y ejecutar acciones de manera autónoma (Empresa Actual, 2025).

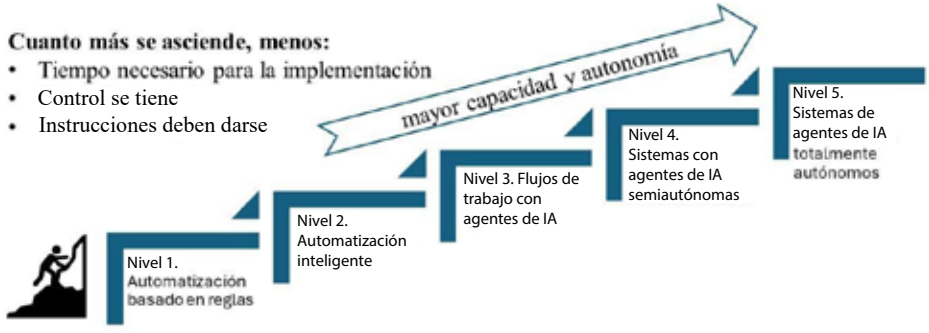
De acuerdo con Coshow (2024), la IAA emplea sensores o diversas fuentes de datos para recopilar información en tiempo real del entorno en el que opera. Posteriormente, el sistema utiliza algoritmos de aprendizaje automático y redes neuronales avanzadas para procesar la información, reconocer patrones y formular predicciones sobre escenarios futuros. La IAA mejora su rendimiento a través de un aprendizaje continuo, incorporando mecanismos de acción que le permiten ejecutar tareas específicas de acuerdo con los objetivos definidos y mantener una interacción efectiva con su entorno.

¿Hay diferentes niveles de automatización con agentes de IA?

Para comprender los niveles de automatización y de autonomía de los agentes de IA, Bornet et al. (2025) proponen cinco niveles para medir la automatización (más que madurez) del uso de la inteligencia artificial dentro de una organización (ver Figura 1).

Figura 1

Niveles de progresión de automatización con la IA



Nota. Adaptado de *Agentic Artificial Intelligence: Harnessing AI Agents to reinvent Business, Work and Life* (p. 78, Figura 1.3. The Agentic AI Progression Framework), por P. Bornet, J. Wirtz, T. H. Davenport, D. De Cremer, B. Evergreen, P. Fersht, R. Gohel y S. Khiyara, 2025, World Scientific Publishing Company.

De acuerdo con este marco de ascenso o progresión de la automatización con IA, el mejor nivel no necesariamente es el último (nivel 5), sino que depende sobre todo de las necesidades y circunstancias específicas de la organización y del nivel de impacto deseado. Esto es similar a contar con un catálogo de diferentes tipos de agentes, cada uno entrenado y adecuado para las necesidades y contextos específicos en los que cada uno se ubicaría en diferente nivel.

Por ejemplo, los automóviles modernos que vienen equipados con asistentes para la conducción automatizada: si bien es posible técnicamente activar la conducción totalmente autónoma en una autopista, muchos conductores prefieren ser más precavidos y activan solo el control de crucero básico (por ejemplo, los niveles 1 o 2).

Es posible diseñar catálogos complejos con múltiples funciones para los automóviles, el hogar y la industria, pero el humano es quien decide, finalmente, el nivel de automatización de IA requerido para cada caso específico. El marco de progresión anterior viene a servir de guía para ubicar el nivel esperado al realizar una propuesta de automatización.

COLABORACIÓN HUMANO-IA

En las últimas décadas, los avances en investigación sobre interacción humano-computadora (HCI) e inteligencia artificial (IA) han transformado las formas de colaboración

entre las personas y los agentes de *software* inteligentes. El desarrollo de un marco teórico sólido que explique la relación entre el comportamiento humano y el computacional sigue en investigación. A medida que las tecnologías de IA se materializan en sistemas y establecen interacciones directas con los humanos, estos conceptos se vuelven fundamentales. Comprender la autonomía de la IA es crucial para el control y la influencia colaborativa, las relaciones de confianza entre humanos e IA, y la gestión de los flujos de trabajo (Samdani et al., 2025).

La interacción humano-IA puede ubicarse en un espectro de niveles de autonomía, que va desde la actuación independiente hasta la supervisada, en la que la intervención humana plantea diversos desafíos éticos y legales (entre los cuales se encuentran la falta de confianza, la descoordinación y la disparidad en el control de las decisiones).

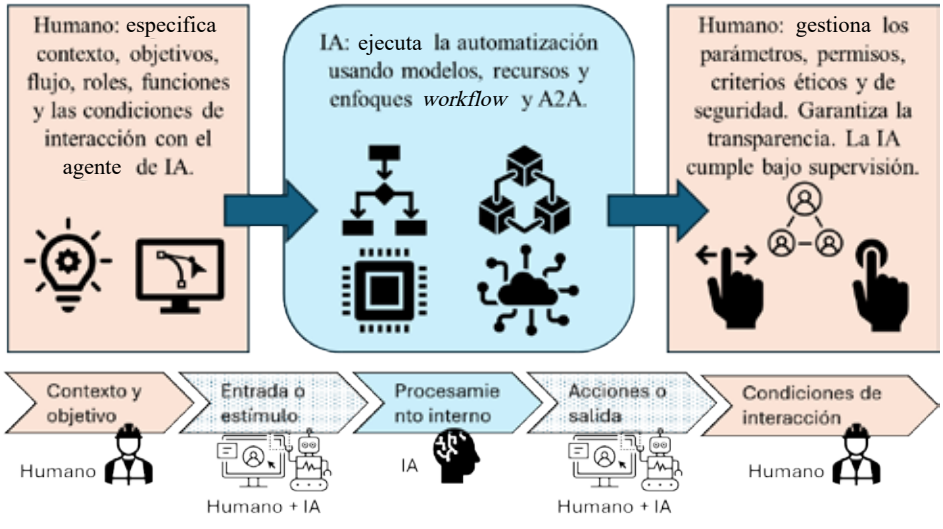
La colaboración humano-IA, por otro lado, se fundamenta en la interacción entre usuarios y sistemas de IA, que abarcan una amplia variedad de contextos de actividad.

Por ello, resulta esencial garantizar la transparencia en el diseño de las soluciones de IA y definir con claridad las responsabilidades en los procesos de toma de decisiones.

Al respecto, se propone un marco de referencia para la colaboración humano-IA, el cual se estructura en cinco etapas que describen la interacción entre una persona humana y un entorno compuesto por agentes de IA. La Figura 2 muestra esta secuencia, conformada por los siguientes pasos: contexto y objetivo, entrada o estímulo, procesamiento interno, acciones o salidas, y condiciones de interacción.

Figura 2

Colaboración humano-IA



A continuación, se describen en detalle los pasos y sus elementos.

- Paso 1 (contexto y objetivo). El humano delimita con claridad el propósito de la interacción con el agente, el contexto, objetivos, alcance, flujo, roles, funciones y las condiciones del dominio de aplicación y las tareas que se espera que ejecute el agente o el workflow.
- Paso 2 (entrada y estímulo). El humano provee y estructura los datos que la IA interpreta. Debe especificar la naturaleza y origen de los datos que el agente de IA recibirá, así como el formato y la estructura de la información (texto, voz, imágenes, sensores), además de incluir las condiciones iniciales que orientan su procesamiento.
- Paso 3 (procesamiento interno). Enfoques de A2A o workflow, a través de los cuales la IA ejecuta de acuerdo con el modelo de razonamiento preestablecido (basado en reglas, aprendizaje automático o generación), con recursos de conocimiento, métodos de adaptación o aprendizaje continuo.
- Paso 4 (acciones o salida). La IA, bajo supervisión humana, incluye el modo en que el agente generará respuestas (textuales, visuales o a través de la ejecución de acciones) e indicará el nivel de autonomía con el que actúa y los mecanismos de retroalimentación previstos. La IA genera las salidas correspondientes en respuestas textuales, visuales o acciones ejecutables.
- Paso 5 (condiciones de interacción). El humano preestablece lo que la IA cumple dentro de los parámetros definidos. El humano gestiona los roles y permisos que delimitan la acción del agente, los criterios éticos y de seguridad que regulan la protección de datos, la mitigación de sesgos y el cumplimiento normativo para garantizar la transparencia en la interacción.

Vemos entonces que la interacción con un agente de IA requiere una distribución clara de responsabilidades entre el usuario humano y el propio sistema. En esta colaboración, la IA asume el procesamiento interno mediante modelos de razonamiento, recursos de conocimiento y mecanismos de adaptación. Las acciones se realizan bajo la supervisión del usuario humano, quien conserva el rol de garante ético y regulador, al definir los criterios de seguridad, privacidad, transparencia y explicabilidad que guían la operación del agente.

A continuación, se analizan dos casos de colaboración humano-IA:

- Un estudio de Casini et al (2023) con modelos de segmentación semántica como medio de colaboración entre humanos con la IA, permitieron detectar sitios arqueológicos en las llanuras de Mesopotamia. Al usar modelos de aprendizaje profundo preentrenados y apoyados en imágenes satelitales, fue posible detectar sitios arqueológicos con una precisión cercana al 80 %. La

colaboración entre expertos en arqueología y científicos de datos fue crucial para la creación y evaluación del conjunto de datos con el que se contaba. Se mostró que la integración de la experiencia, conocida como humanos en el bucle (human in the loop), fue esencial para mejorar la precisión de un modelo de IA y refinar el conjunto de datos con nuevas anotaciones basadas en las predicciones del modelo. Para optimizar la detección de futuros sitios arqueológicos, se propone modelar un flujo de trabajo de IA agéntico que combine las predicciones del modelo junto con la experiencia humana.

- Otra experiencia consistió en integrar agentes de IA a través de enfoques empresariales (Enterprise AI - EAI), en combinación con sistemas de intercambio electrónico de datos (Electronic Data Interchange - EDI). Se buscó potenciar la eficiencia operativa y optimizar los procesos de toma de decisiones, estimando mejoras de hasta un 30 % en el desempeño global. Para la interacción con los clientes, se incorporó un chatbot de servicio como interfaz, respaldado por un agente orquestador de IA que asignaba tareas a agentes especializados. Finalmente, la capa de gobernanza cierra el ciclo, asegurando la integridad del sistema y el cumplimiento de las normativas vigentes (Ramalingam, 2025).

De acuerdo con Viswanathan et al. (2025), se espera que el futuro de los agentes de inteligencia artificial experimente avances significativos en cuanto a autonomía, adaptabilidad y colaboración humano-IA. Estos agentes autónomos y automejorables emplearán técnicas como el aprendizaje por refuerzo, el aprendizaje continuo y el metaaprendizaje para optimizar la toma de decisiones en entornos dinámicos.

De acuerdo con Balic (2005), los sistemas autónomos de IA multiagente están llamados a transformar diversos sectores, en particular la industria del desarrollo de *software* y las actividades basadas en el conocimiento. A pesar del rápido avance en sus capacidades técnicas, existe una brecha significativa en la comprensión de cómo los profesionales perciben estos sistemas, incluyendo sus capacidades, limitaciones, implicaciones éticas y el posible impacto en los empleos actuales. Estas percepciones trascienden el ámbito teórico, ya que influyen de manera directa en las tasas de adopción, las decisiones de inversión, y las estrategias de formación y preparación de la fuerza laboral.

En resumen, la interacción humana–IA se concibe como un proceso colaborativo en el que las capacidades cognitivas y normativas del humano enmarcan y orientan la autonomía de la IA.

De la Figura 1 se desprende que el ser humano es quien establece los objetivos, los datos de entrada y los criterios éticos, mientras que la IA (*workflow* o agentes A2A) se encarga del procesamiento y la generación de respuestas. De esta manera,

las capacidades normativas y de supervisión humanas delimitan y condicionan la autonomía técnica del agente, asegurando que su desempeño se enfoque en la eficiencia y la responsabilidad. Para más detalle, ver el Glosario (sección Colaboración humano-IA).

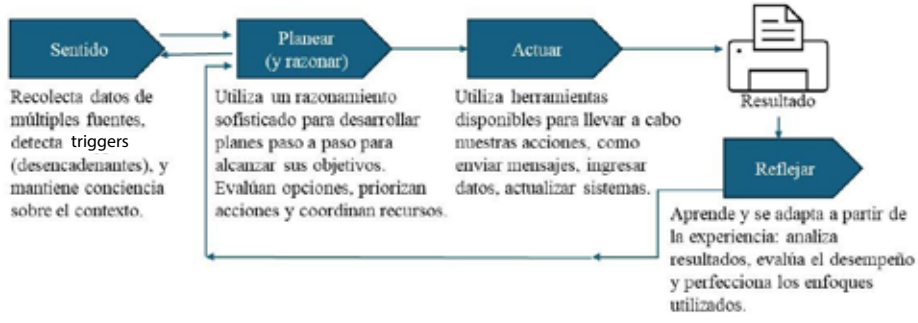
Marco de referencia SPAR

¿Qué sucede al darle a un agente de IA esta instrucción?: "genere el resultado que necesito".

Para responder a esta pregunta, Pascal Bornet et al. (2025) proponen utilizar el marco integral SPAR (acrónimo de los términos *sense*, *plan*, *act* y *reflect*), que ayudar a comprender las capacidades de los agentes de IA. Para analizar el ciclo de comportamiento de un agente de IA, podemos utilizar el marco de referencia SPAR mostrando la secuencia de pasos a seguir.

Figura 2

El marco de referencia SPAR -cómo actúa un agente de IA



Nota. De *Agentic Artificial Intelligence: Harnessing AI Agents to reinvent Business, Work and Life* (p. 23, Figura 1.2. How an AI agent takes action: The SPAR Framework), por P. Bornet, J. Wirtz, T. H. Davenport, D. De Cremer, B. Evergreen, P. Fersht, R. Gohel y S. Khiyara, 2025, World Scientific Publishing Company.

El modelo SPAR inicia con el seguimiento básico de reglas hasta lograr resultados, luego iterar y regresa a planear y actuar de nuevo. Este ciclo de retroalimentación se realiza con suficiente autonomía, lo cual permite al agente desenvolverse ante un panorama complejo, apoyado en decisiones informadas sobre las soluciones más adecuadas para sus necesidades.

A partir de la Figura 2, se describen paso a paso cada una de las acciones de los agentes de IA:

a) Sentido (*sensing*): son los ojos y los oídos de los agentes de IA. Integra datos de diversas fuentes, detecta condiciones desencadenantes y sostiene una

conciencia contextual dinámica del entorno operativo. A manera de ejemplo, un automóvil autónomo necesita comprender su entorno de manera integral. De igual manera, los agentes de IA deben ser capaces de percibir su espacio de trabajo digital. Cuando se introduce una ruta destino en la pantalla del auto, se le está estableciendo el objetivo o destino hacia el cual debe dirigirse.

b) Planear y razonar (*plan and reason*): emplea procesos de razonamiento avanzado para formular planes estructurados orientados al cumplimiento de objetivos específicos. El agente de IA analiza alternativas, establece prioridades y gestiona los recursos disponibles.

c) Actuar (*act*): emplea herramientas tecnológicas para ejecutar acciones operativas, como la transmisión de mensajes, el ingreso de datos y la actualización de sistemas. Facilita la implementación de decisiones previamente planificadas y la interacción del entorno digital con otros componentes del sistema.

d) Reflejar (*reflect*): adquiere conocimiento con cada iteración y mejora su desempeño mediante un proceso de aprendizaje basado en la experiencia. Se incluyen el análisis de resultados, la evaluación del rendimiento y el perfeccionamiento continuo de los enfoques aplicados.

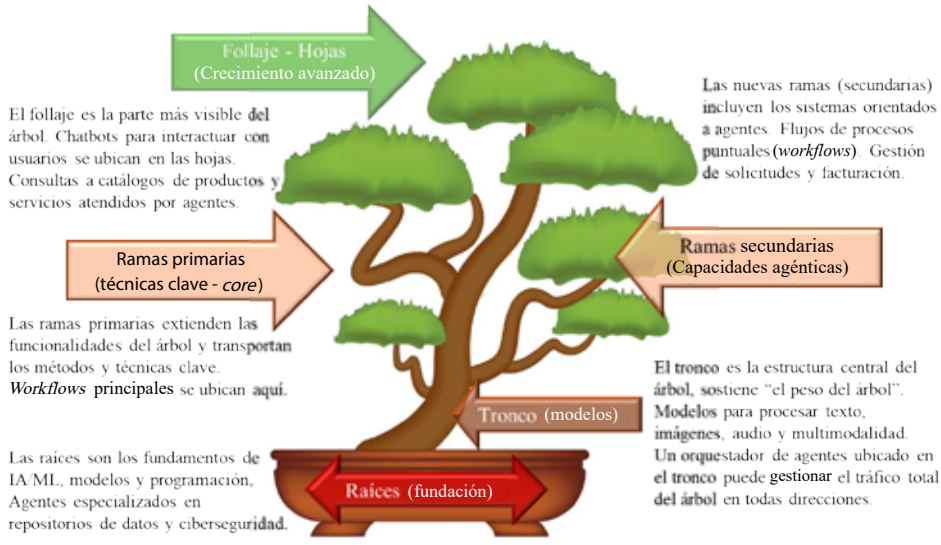
En resumen, el modelo SPAR facilita que los agentes de IA operen mediante un ciclo iterativo compuesto por cuatro fases fundamentales: En la fase de percepción (*sensing*) se recolectan los datos y objetivos. En la fase de planificar y razonar (*plan and reason*), se aplican técnicas de razonamiento para priorizar los pasos a seguir. En la fase de actuación (*act*) se llevan a cabo las acciones. Finalmente, en la fase de reflexión (*reflect*) se logra el aprendizaje con base en la evaluación del desempeño y el análisis de resultados, para luego iterar hacia la fase de planeación. Este ciclo integral SPAR se convierte en una guía de referencia útil para el desarrollo de agentes autónomos de IA capaces de operar eficazmente en entornos dinámicos y complejos.

Metáfora del árbol de IA

Una manera de conceptualizar los componentes, modelos de agentes y flujos de trabajo con IA es por medio de una analogía que denominamos la metáfora del árbol de IA, en la que cada elemento tiene su funcionalidad y ubicación estratégica, simulando una organización.

Figura 4

Metáfora del árbol de IA



¿En cuáles áreas se están desarrollando proyectos de IA? Ante esta pregunta, una manera rápida y simple de responder consiste en apoyarse en la metáfora del árbol de IA. En el árbol podemos ubicar las diferentes iniciativas de automatización con IA:

- Raíces: modelos y herramientas de IA a seguir. Agentes de IA especializados en el acceso a los repositorios de datos y esquemas de ciberseguridad.
- Tronco: "soporte" del árbol. Modelos para procesar todo tipo y formato de datos. Orquestador de agentes de IA pueden gestionar todo el tráfico dentro del árbol.
- Ramas primarias: métodos y técnicas clave. Workflows de uso general permiten trasladar información de la raíz hasta las hojas. Por ejemplo, la cadena de abastecimiento (supply chain), gestión de bodegas y almacenes e inventarios.
- Ramas secundarias: sistemas y workflows puntuales usando agentes de IA. Aplicaciones para gestionar funcionalidades de la organización. Por ejemplo, facturación y solicitudes.
- Follaje – hojas: es la parte "visible" del árbol. Los usuarios interactúan por este medio con agentes, gestionando catálogos de bienes y servicios en línea. Si usan "carro de compras", este puede ser otro agente de IA especializado el cual es llamado para gestionar el pago.

La elección final de la ubicación dependerá de una decisión estratégica y de las particularidades de cada caso. Alternativamente, se puede considerar la ubicación de

un agente orquestador en una rama primaria, organizando el flujo específico de *workflows* de IA al resto del árbol.

Un *workflow* para la cadena de abasto puede interactuar con varios agentes y flujos de trabajo, siempre orquestados por un agente central.

En cualquier caso, es fundamental incorporar principios de autonomía y escalabilidad en todos los componentes del sistema, para de esa manera maximizar su eficacia y adaptabilidad.

Esta metáfora del árbol de IA ofrece un marco conceptual, simple y práctico, para gestionar proyectos de IA en empresas, instituciones académicas o sectores industriales. Para más detalle, véase el Glosario (Sección Metáfora del árbol de IA).

Desafíos y ética con la IA

La adopción de la IA enfrenta diversas dificultades, entre las que destacan la privacidad y la seguridad de los datos, especialmente en sectores sensibles. La complejidad inherente a la integración de la IA con sistemas existentes puede extender los plazos de los proyectos. La necesidad de una capacitación adecuada de la fuerza laboral, esencial para garantizar una adopción exitosa, es otro desafío. Las restricciones en costos y recursos, por su parte, pueden ser mitigadas mediante el uso de plataformas en la nube y modelos preentrenados.

De acuerdo con Xu & King (2025), existen otros desafíos con el desarrollo impulsado por la IA:

- como frontera emergente
- alineada con valores humanos
- aprendizaje en contextos de recursos limitados,
- en la investigación científica

Estos desafíos deben ir acompañados de una reflexión ética constante respecto de los sesgos, la transparencia y la rendición de cuentas, a la vez que evaluar, a nivel mundial, la demanda creciente de marcos regulatorios adecuados. En este contexto, lograr un equilibrio entre innovación y gobernanza será crucial para mantener la confianza en general.

Por otro lado, se hace imprescindible instaurar marcos normativos y auditables, como el Reglamento General de Protección de Datos de la Unión Europea (General Data Protection Regulation, GDPR) y la Ley de Portabilidad y Responsabilidad del Seguro Médico (Health Insurance Portability and Accountability Act, HIPAA), ley que regula la confidencialidad y seguridad de la información médica y de salud, además de establecer mecanismos de gobernanza que incluyan la detección de sesgos y una evaluación ética previa a su implementación.

En este contexto, la convergencia entre tecnología y gestión pública transforma constantemente los paradigmas tradicionales e impulsa la apertura de nuevos escenarios de oportunidad, lo cual permite optimizar procesos y aumentar la eficiencia, pero también habilitar espacios para la innovación, el análisis y el aprovechamiento de grandes volúmenes de datos (Ramalingam, 2025).

Dentro del ámbito académico, Corvalán & Sánchez Caparrós (2025) señalan que uno de los desafíos centrales consiste en la formación de profesionales con la capacidad de adaptarse y gestionar, con base en criterios éticos y de eficiencia, el diseño e implementación de sistemas emergentes basados en componentes de inteligencia artificial.

Según Hauser (2019), la reflexión ética sobre la tecnología, y en particular sobre la IA, ha tendido a aplicarse directamente desde teorías en principio concebidas para los seres humanos hacia los sistemas de IA. Un ejemplo de esta tendencia es el enfoque que considera a la IA como un instrumento peligroso: se enfatizan las diferencias entre los sistemas de IA y los humanos bajo el argumento de que dichos sistemas no deben ser considerados agentes morales, pues las teorías de agencia ética fueron desarrolladas únicamente para sujetos humanos.

Desde la academia, se propone abordar la ética desde una perspectiva más amplia, centrada en los sistemas de información, considerando a la IA como un subconjunto dentro de este marco. Esta aproximación facilita la incorporación de los estudios de la información en el análisis de los desafíos éticos planteados por las tecnologías de IA.

Asimismo, resulta fundamental comprender las percepciones de los profesionales en estos ámbitos para anticipar los retos relacionados con la adopción de una IA segura, las implicaciones éticas y la evolución del mercado laboral.

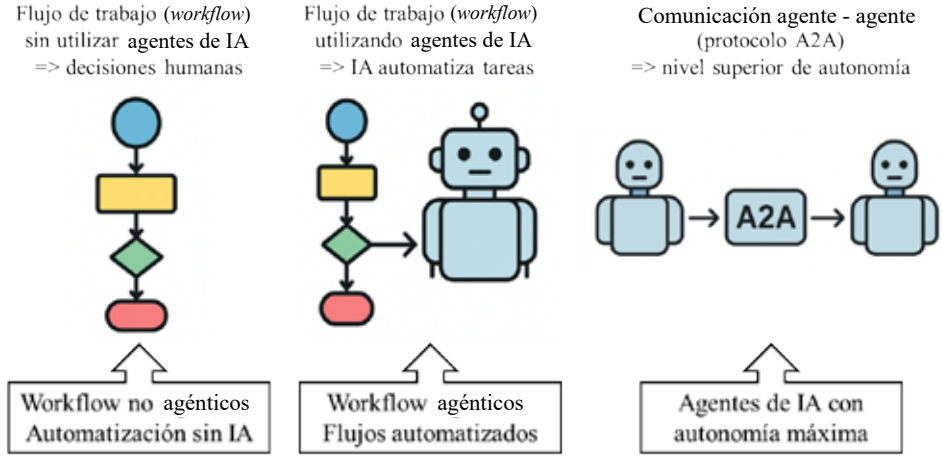
AI WORKFLOWS Y PROTOCOLOS A2A

Categorizando *workflows* de IA y agentes A2A

Michael Lieben (2025) hace una propuesta para categorizar *workflows* y agentes de IA, utilizando tres tipos de agentes de IA. Estos se muestran en la Figura 5.

Figura 5

Tipos de workflows *y agentes de IA*



Nota. Adaptado del flujograma *3 Types of AI Workflows*, por M. Lieben, 2025. LinkedIn. https://www.linkedin.com/posts/the-best-of-ai_ai-agents-are-replacing-human-employees-activity-7364786657318838273-SH--/

Las principales diferencias se analizan a continuación.

a) *Workflows* no agénticos: se distinguen porque la responsabilidad de las decisiones recae enteramente en el ser humano, mientras que la inteligencia artificial se limita a actuar como herramienta de apoyo. En este esquema, el usuario dirige cada acción. Por ejemplo, cuando se consulta un sistema de gestión y luego se solicita a un modelo de lenguaje la redacción de un correo, que finalmente es enviado por el humano de manera manual. La IA, en este caso, cumple una función instrumental, subordinada al control humano.

b) *Workflows* agénticos: la inteligencia artificial se integra dentro de un sistema regido por reglas y disparadores que permiten automatizar determinadas tareas. Así, por ejemplo, el propio sistema puede detectar la necesidad de programar una reunión, verificar la disponibilidad en la agenda y generar automáticamente un correo de invitación. Este tipo de flujos incrementa la eficiencia operativa, aunque su capacidad de respuesta es limitada ante situaciones no previstas.

c) Agentes de IA: nivel superior de autonomía. Actúan en función de metas establecidas, como la satisfacción del cliente. Estos agentes son capaces de interpretar datos contextuales, monitorear interacciones previas, generar comunicaciones personalizadas, dar seguimiento a la respuesta del interlocutor y coordinar actividades posteriores, como la elaboración de una agenda. A diferencia de los flujos anteriores, los agentes no solo ejecutan tareas, sino que articulan acciones estratégicas orientadas a objetivos.

De acuerdo con Corvalán & Sánchez Caparrós (2025), los flujos de trabajo automatizados que integran modelos de lenguaje combinan herramientas y modelos de IA bajo rutas de código establecidas. Los agentes de IA, en contraste, poseen mayor grado de autonomía al no basarse en rutas preestablecidas y siendo más adaptables a situaciones de decisión.

La progresión de los sistemas de IA puede comprenderse en tres niveles:

- En primer lugar, los flujos de trabajo sin agentes, en los que la IA cumple un papel estrictamente instrumental bajo control humano.
- En segundo término, los flujos agénticos (workflows), caracterizados por la automatización de tareas mediante reglas y disparadores, lo que incrementa la eficiencia, pero con limitada adaptabilidad frente a contingencias.
- Finalmente, los agentes A2A introducen un grado superior de autonomía al operar con metas definidas, interpretar información contextual y coordinar acciones orientadas a objetivos estratégicos. De este modo, se consolidan como un avance sustantivo en la transición de la automatización hacia la toma de decisiones inteligente (The Best of AI, 2025).

Comparando *workflows* de IA y protocolos A2A

Para llevar a cabo el análisis comparativo, utilizaremos los criterios de autonomía, escalabilidad, flexibilidad y coordinación, bajo los cuales ubicaremos las características de *workflows* y A2A.

Autonomía y adaptabilidad

Workflows

- Los AI workflows generalmente siguen una secuencia de pasos predefinidos, aunque pueden incorporar componentes de IA (como modelos predictivos o generativos) que permitan cierta adaptabilidad.
- Su autonomía depende del diseño, que suele requerir supervisión humana.
- La adaptabilidad está limitada por las reglas y modelos integrados, y responde a cambios dentro de lo previsto, pero no a contextos totalmente nuevos.

A2A

- Los protocolos A2A poseen autonomía inherente, toman decisiones en función de su entorno y objetivos.
- Son altamente adaptativos y pueden modificar su comportamiento al interactuar con otros agentes y aprender de la dinámica del sistema.
- Permiten mayor descentralización y autogestión que los flujos tradicionales.

- En cuanto a autonomía y adaptabilidad, la ventaja la tiene el protocolo A2A, por su naturaleza autoorganizativa y adaptativa.

Escalabilidad y eficiencia

Workflow

- Los AI workflows escalan de manera lineal. Añadir más procesos o nodos implica extender el flujo, lo que a su vez puede aumentar la complejidad.
- La eficiencia depende de la optimización del pipeline y del poder de cómputo.
- Funcionan bien en entornos centralizados (por ejemplo, en el procesamiento de grandes volúmenes de datos).
- Los AI workflows son mejores en sistemas centralizados de alto volumen.

A2A

- Los protocolos A2A escalan de manera distribuida: cada agente gestiona sus tareas y la carga se reparte entre múltiples entidades.
- La eficiencia se puede ver afectada por la sobrecarga de comunicación entre agentes, especialmente en sistemas muy grandes.
- Se desempeñan mejor en escenarios descentralizados o dinámicos (IoT, redes inteligentes).
- Los protocolos A2A son mejores en entornos distribuidos dinámicos.

Flexibilidad y robustez

Workflow

- Los AI workflows son flexibles en cuanto a la integración de distintos módulos o servicios (por ejemplo, IA para visión, procesamiento del lenguaje natural - NLP, análisis de datos).
- Poco robustos ante fallos inesperados fuera del diseño del flujo. Por tanto, una interrupción puede afectar todo el proceso.
- La resiliencia depende de redundancias planificadas.

A2A

- Los protocolos A2A son altamente flexibles; los agentes pueden reorganizarse, asumir roles distintos y redistribuir tareas.
- Robustos frente a fallos locales: si un agente cae, otros pueden compensar o reorganizarse.
- Adecuados para entornos inciertos o con cambios frecuentes.
- A2A es más ventajosa, por su descentralización y resiliencia.

Coordinación y comunicación

Workflows

- Los AI workflows poseen una coordinación jerárquica y secuencial, en la que cada etapa depende de la anterior.
- Comunicación limitada, generalmente de tipo unidireccional (paso de datos de un módulo a otro).
- Adecuados para procesos determinísticos o en cadena.

A2A

- Los protocolos A2A coordinan de manera colaborativa y negociada; los agentes intercambian información, establecen acuerdos y se adaptan en tiempo real.
- Comunicación rica, bidireccional y en ocasiones con protocolos complejos (cooperación, competencia y negociación).
- Ideales para entornos heterogéneos y con múltiples actores.
- A2A es más ventajosa, dada su mayor comunicación y cooperación dinámica.

Resumen de los criterios anteriores

- En los criterios de autonomía y adaptabilidad, la A2A es más ventajosa.
- En escalabilidad y eficiencia, los workflows de IA son mejores en sistemas centralizados de alto volumen y los A2A lo son en entornos distribuidos dinámicos.
- En cuanto a flexibilidad y robustez y a coordinación y comunicación, A2A es más ventajosa.

Usos prácticos de *workflows* y agentes de IA

Utilizando los criterios anteriores (autonomía, escalabilidad, flexibilidad y coordinación) se analizarán algunos ejemplos de la operación con IA *workflows*.

- Autonomía y adaptabilidad. Los flujos de trabajo de IA en el ámbito bancario, como la evaluación de solicitudes de crédito, siguen pasos predefinidos (verificación documental, análisis de scoring y decisión final). Su capacidad de adaptación es limitada y depende de reconfiguraciones humanas cuando se enfrentan a nuevos formatos o condiciones no previstas. Si aparece un documento en un formato nuevo, el flujo podría fallar o requerir intervención humana para actualizar la lógica.
- Escalabilidad y eficiencia. En plataformas de comercio electrónico, los workflows permiten procesar transacciones masivas de manera centralizada,

aunque escalar requiere incrementar infraestructura (por ejemplo, más servidores). La eficiencia se sostiene en entornos estables y controlados y escalar implica añadir más servidores que repliquen el flujo.

- Flexibilidad y robustez. En el ámbito hospitalario, un workflow que integra IA para diagnóstico puede verse interrumpido si uno de sus módulos falla (por ejemplo, el de procesamiento de lenguaje NLP para historiales clínicos); se compromete la totalidad del proceso y el flujo entero puede interrumpirse.
- Coordinación y comunicación. En procesos de reclutamiento automatizado, la coordinación se estructura de forma jerárquica y secuencial (filtrado de currículos, ranking, entrevistas automatizadas), con comunicación secuencial y del tipo unidireccional entre las etapas.

Un caso específico es el de la industria 4.0, en la que la convergencia entre tecnologías digitales y físicas está transformando el *workflow* de los procesos de manufactura. Entre estos, la integración de inteligencia artificial (IA) con la manufactura aditiva (AM) destacan como ejes fundamentales de innovación, permitiendo un control preciso de cada capa en la fabricación de productos: desde la recopilación de datos en tiempo real mediante sensores hasta la optimización de parámetros de diseño, geometría y desempeño estructural. Esta relación sinérgica entre IA y AM no solo resulta esencial para alcanzar la visión de la industria 4.0, sino que también se proyecta hacia la industria 5.0, caracterizada por la sostenibilidad, la colaboración humano-máquina y la hiper personalización de los procesos productivos (Bassey et al., 2025).

Algunos ejemplos de la operación utilizando protocolos A2A son los siguientes:

- Autonomía y adaptabilidad: en entornos de negociación financiera, los agentes A2A muestran alta autonomía al ajustar sus decisiones de compra y venta en tiempo real, coordinando con otros agentes y adaptándose a la aparición de nuevos activos sin necesidad de rediseñar todo el sistema.
- Escalabilidad y eficiencia: en sistemas de redes eléctricas inteligentes (smart grids), cada agente gestiona su consumo y producción local, logrando escalabilidad distribuida sin depender de un nodo central. La eficiencia se mantiene en contextos dinámicos mediante la optimización local y, al aumentar el número de dispositivos, la coordinación sigue siendo posible sin necesidad de un controlador central.
- Flexibilidad y robustez: en escenarios en los que se utilicen robots para rescate, los agentes cooperan y redistribuyen las tareas cuando alguno falla, garantizando así la continuidad de la misión y demostrando resiliencia ante fallos locales. Ante un robot que falla, son capaces de reorganizar la búsqueda y de compensar la pérdida, manteniendo la misión activa.

- Coordinación y comunicación: en flotas de drones de reparto, la comunicación es bidireccional y negociada; los agentes ajustan rutas en tiempo real para evitar colisiones y optimizar entregas colectivamente. La coordinación es bidireccional y si un dron detecta tráfico aéreo congestionado, negocia con otros para replanificar la ruta en tiempo real.

A continuación, y siempre utilizando los mismos criterios, se presenta un análisis comparativo de los ejemplos analizados de flujos de trabajo y protocolos A2A.

Tabla 1

Criterios y análisis comparativo de workflow *y A2A*

Criterio	Análisis de *AI workflow*	Análisis de protocolos A2A
Autonomía y adaptabilidad	- Los *workflows* mantienen autonomía restringida de acuerdo con sus reglas predefinidas. - Automatización de procesos productivos acorde con la industria 4.0. - La aparición de un documento en un formato nuevo puede interrumpir el flujo y requerir intervención humana para su adaptación.	- Los protocolos A2A exhiben mayor capacidad adaptativa y autoorganización frente a escenarios imprevistos. - Muestran alta autonomía al coordinar con otros agentes y adaptarse a los cambios.
Escalabilidad y eficiencia	- Los *workflows* escalan en arquitecturas centralizadas y homogéneas. - La eficiencia se mantiene en entornos estables; para escalar, requiere replicar el flujo utilizando más servidores.	- Los A2A son más apropiados para sistemas distribuidos y heterogéneos. - Al aumentar el número de dispositivos, la coordinación sigue siendo posible sin necesidad de un controlador central.
Flexibilidad y robustez	- Los *workflows* ofrecen integración modular, pero son más vulnerables a fallos críticos. - El fallo de un módulo puede interrumpir todo el flujo del proceso.	- Los A2A muestran mayor robustez y flexibilidad gracias a su naturaleza descentralizada. - Si un robot falla, el sistema puede reorganizarse para compensar la pérdida y mantener la misión.
Coordinación y comunicación	- Los *workflows* se sustentan en coordinación lineal y control centralizado. - La coordinación bidireccional permite a los drones replanificar rutas en tiempo real ante congestión aérea.	- Los A2A promueven cooperación dinámica y una comunicación distribuida. - Los drones se coordinan bidireccionalmente para replanificar rutas en tiempo real.

De acuerdo con este análisis, los *AI workflows* son buenos para procesos estandarizados, repetitivos y centralizados, mientras que los protocolos A2A son ideales para entornos distribuidos, dinámicos y con alta interacción entre entidades autónomas.

DISCUSIÓN DE HALLAZGOS

La revisión de los aportes recientes en torno a la colaboración humano-IA y el desarrollo de agentes inteligentes evidencia una convergencia hacia modelos más autónomos, adaptativos y escalables. Se propone una categorización fundamental de los flujos de trabajo y agentes de IA —desde los no agénticos hasta los plenamente autónomos— que permita comprender el grado de delegación cognitiva y operativa entre humanos y sistemas inteligentes.

De manera complementaria, se sugiere la necesidad de replantear la relación entre humanos e IA a través del marco de referencia SPAR —*sensing*, *planning*, *acting*, *reflecting*— que orienta los niveles de automatización esperados y subraya la importancia de la percepción contextual, el razonamiento estructurado, la acción coordinada y el aprendizaje iterativo. Utilizar la metáfora del árbol nos ofrece una visión estructural del ecosistema de la IA generativa, pues ayuda a identificar las capas de fundamentos, modelos, técnicas y capacidades que sustentan el desarrollo y ubicación de agentes inteligentes y *workflows* (raíces, tronco, ramas y hojas), además de detallar los componentes esenciales para su implementación: desde marcos agénticos y sistemas de memoria, hasta mecanismos de gobernanza y despliegue.

Los sistemas de flujos de trabajo basados en agentes de IA (*agentic workflows*) y la comunicación entre agentes A2A (*agent-to-agent*), destacan por su autonomía y adaptabilidad para enfrentar desafíos complejos. Se recomienda su implementación apoyándose en los modelos, referencias y guías analizados en este artículo, con el objetivo de diseñar novedosas propuestas de automatización, aplicables en sectores laborales y académicos en los que se proyecta que, hacia 2027, aproximadamente el 82 % de las organizaciones habrá incorporado estas prácticas, ya sea en la modalidad de flujos de trabajo de IA o con protocolos A2A.

El análisis comparativo evidencia que los *AI workflows* son mejores para procesos estandarizados, repetitivos y centralizados, mientras que los protocolos A2A son ideales para entornos distribuidos, dinámicos y con alta interacción entre entidades autónomas. Otro hallazgo es que los *AI workflows* resultan más adecuados en contextos centralizados, estables y de alta carga transaccional, en los que la predictibilidad y la eficiencia dependen de procesos secuenciales bien definidos. Sin embargo, su autonomía y resiliencia son limitadas, pues requieren mayor supervisión humana para adaptarse a cambios imprevistos y presentan vulnerabilidad frente a fallos críticos en sus componentes.

Por el contrario, los agentes de IA trabajando en conjunto y comunicándose vía protocolos A2A destacan en entornos distribuidos, dinámicos y heterogéneos, gracias a su capacidad de autoorganización, comunicación bidireccional y resiliencia ante la pérdida de nodos individuales. Su flexibilidad los hace idóneos para escenarios en los

que la cooperación, la negociación y la adaptación en tiempo real son esenciales, así como en sistemas de energía inteligente, robótica colaborativa o logística distribuida.

CONCLUSIONES

La IA aún enfrenta retos significativos en materia de transparencia, responsabilidad y equidad, lo cual puede amplificar sesgos y socavar la confianza social. Para mitigar estos riesgos, resulta esencial el cumplimiento de marcos normativos auditables, como el GDPR y la HIPAA, junto con la implementación de mecanismos de gobernanza que contemplen la detección de sesgos y una evaluación ética previa a su despliegue.

El desafío actual de la IA trasciende la mera creación de agentes y se centra en su despliegue confiable a gran escala. El éxito en su adopción dependerá de concebir sistemas que, desde el inicio, tengan capacidad de escalabilidad, integrando de forma coherente marcos de referencia, memoria, gobernanza y mecanismos efectivos de implementación.

Las arquitecturas agénticas de *AI workflows* y A2A no deben entenderse como enfoques excluyentes, sino más bien complementarios. Mientras los primeros optimizan procesos repetitivos y de gran volumen en marcos controlados, los segundos permiten gestionar la complejidad e incertidumbre en redes autónomas. La convergencia de ambos enfoques puede abrir nuevas posibilidades para el diseño de sistemas híbridos, capaces de combinar eficiencia estructurada con adaptabilidad distribuida.

La construcción de un agente de IA constituye solo la etapa inicial; el verdadero desafío se encuentra en su escalamiento hacia entornos de producción. En este proceso, la fiabilidad se erige como el elemento diferenciador entre un prototipo y un producto consolidado, ya que sin ella los sistemas de IA permanecen limitados a la experimentación, sin generar un impacto real. En ese sentido, sería deseable, por ejemplo, enfocarse en el desarrollo de sistemas robustos de memoria y gobernanza como clave para el éxito de la implementación de *workflows* y agentes de IA.

Bajo este marco, llevar a cabo el análisis de una arquitectura por capas y de los componentes esenciales evidencia la necesidad de una planificación estratégica orientada a aplicaciones concretas y flujos de trabajo. Esto resulta especialmente relevante en un campo dinámico y en constante evolución como la inteligencia artificial.

En conclusión, la integración de tecnologías de IA ha transformado profundamente a las organizaciones, orientándolas hacia niveles crecientes de autonomía y optimización operativa. Sin embargo, este avance plantea la necesidad de desarrollar arquitecturas de agentes impulsadas por IA que sean explicables, auditables y que respalden procesos de toma de decisiones confiables y transparentes. Ello exige mantener una reflexión ética permanente sobre aspectos como los sesgos algorítmicos, la transparencia y la rendición de cuentas, en consonancia con la creciente demanda global de marcos regulatorios robustos y responsables.

En síntesis, las investigaciones revisadas consolidan una visión de la inteligencia artificial como un ecosistema en continua evolución, en el cual la colaboración humano-IA progresa hacia esquemas de autonomía compartida, aprendizaje adaptativo y gobernanza ética. Este enfoque integrador proyecta una nueva etapa en la ciencia computacional, caracterizada por la convergencia entre el diseño algorítmico, la interacción humana y la sostenibilidad tecnológica, sentando las bases para el desarrollo de sistemas inteligentes más confiables, transparentes y orientados al bien común.

Finalmente, se recomienda utilizar las figuras y tablas presentadas como recursos prácticos y concisos, a fin de facilitar la planificación, programación y ejecución de flujos de trabajo y de iniciativas de colaboración basadas en protocolos de agentes de IA.

GLOSARIOS

Colaboración humano-IA

Elemento de automatización	Descripción de aspectos a especificar	Responsable principal
Paso 1. Contexto y objetivo	- Propósito de la interacción - Dominio de aplicación - Alcance de las tareas	Humano
Paso 2. Entrada o estímulo	- Tipo de datos (texto, voz, imágenes, sensores, etcétera) - Formato y estructura- Condiciones iniciales	Humano provee y estructura datos; la IA interpreta
Paso 3. Procesamiento interno	- Modelo de razonamiento (reglas, ML, generativo) - Recursos de conocimiento - Capacidad de adaptación	IA
Paso 4. Acciones o salida	- Modo de respuesta (texto, visualización, acción en sistema) - Nivel de autonomía - Retroalimentación	IA bajo supervisión humana
Paso 5. Condiciones de interacción	- Roles y permisos - Ética y seguridad (privacidad, sesgos, normativa) - Usabilidad y explicabilidad	Humano establece roles y permisos que delimitan la acción. La IA interpreta y opera dentro de las reglas y parámetros definidos.

Metáfora del árbol de IA

Nivel (metáfora)	Componentes	Ejemplos de productos de IA
Raíces (fundación)	Fundamentos de IA/ML, programación y matemáticas aplicadas	PyTorch, TensorFlow, Keras; Python (Jupyter, VS Code); Pandas, NumPy, SciPy, Matplotlib.

(continúa)

(continuación)

Nivel (metáfora)	Componentes	Ejemplos de productos de IA
Tronco (modelos generativos)	Modelos centrales para texto, imágenes, audio y multimodalidad	GPT, Claude, LLaMA; Stable Diffusion, MidJourney, Adobe Firefly; Runway, Descript, ElevenLabs; GPT-4o, Gemini, LLaVA, Kosmos-1.
Ramas (técnicas core)	Métodos clave para optimizar y personalizar modelos	Prompt Engineering (LangChain, DSPy, FlowGPT); Fine-Tuning (LoRA, QLoRA, PEFT); RAG (Pinecone, ChromaDB, Weaviate).
Nuevas ramas (capacidades agénticas)	Sistemas orientados a agentes y orquestación de flujos	AutoGPT, CrewAI, LangGraph, Microsoft Autogen; Airflow, n8n, Zapier, Make.com.
Nuevas hojas (crecimiento avanzado)	Implementación, escalamiento y casos de uso especializados	Docker, Kubernetes, AWS Bedrock, GCP Vertex AI; Healthcare AI, FinTech AI, Creative AI, Enterprise Automation.

Evolución de la automatización con agentes de IA

Niveles de automatización	Descripción	Década	Agentes de IA
Primer nivel: agentes conversacionales iniciales	Primeros bots de IA usando lenguage natural, con capacidades limitadas	1960	Introducción de Eliza, el primer agente conversacional
Segundo nivel: la IA generativa	Mejoró la versatilidad y accesibilidad de la IA	2011 2022	Lanzamiento de agentes modernos: Siri y Alexa. El auge de la IA generativa con el lanzamiento de ChatGPT.
Tercer nivel: avance multimodal	La IA se volvió multimodal, al integrar múltiples formatos de entrada y salida.	2023	Lanzamiento de Google Gemini para texto, imágenes y audio. Microsoft Copilot para desarrollo de software, sugiriendo líneas de código y ayudando a resolver problemas de programación.
Cuarto nivel: mejora del razonamiento	La IA mejoró el razonamiento: maneja tareas complejas con menos intervención humana.	2024 inicios	Avance de la IA multimodal. Lanzamiento de OpenAI Sora, modelo multimodal para generar contenido audiovisual. Claude AI para documentos extensos, análisis de datos, seguridad, ética y personalización.
Quinto nivel: mayor autonomía	La IA con más autonomía y adaptabilidad, con herramientas, memoria y acceso a internet	2024 finales	Evolución de los sistemas de IA siendo más autónomos y adaptables. Claude AI crea chatbots y asistentes virtuales que proporcionan respuestas seguras y responsables.

(continúa)

(continuación)

Niveles de automatización	Descripción	Década	Agentes de IA
Sexta generación: agentes especializados y de carácter generalista	Agentes de IA altamente especializados, demostrando alta autonomía en tareas específicas.	2025 inicios	Introducción de: • Agentes altamente autónomos: Gemini 1.5, Deep Research y Computer Use. • Agentes de carácter generalista: Anthropic's Computer Use, Google's Project Mariner y OpenAI's Operator.

Modelos de lenguaje

Modelo	Descripción general	Fortalezas	Aplicaciones relevantes
GPT-5 / GPT-4o (OpenAI)	Modelo de propósito general, ampliamente reconocido por su versatilidad en tareas textuales y computacionales.	Equilibrio entre redacción, programación y razonamiento; capacidades multimodales en GPT-4o.	Redacción académica y creativa, generación de código, asistentes conversacionales, análisis de datos.
Claude 4 (Anthropic)	Modelo orientado a la seguridad y el razonamiento avanzado, diseñado para contextos extensos.	Manejo de largos contextos, razonamiento profundo, énfasis en alineación ética.	Investigación académica, análisis de documentos extensos, apoyo en tareas de reflexión compleja.
Gemini 2.5 (Google DeepMind)	Modelo multimodal de última generación con énfasis en integración de texto, imágenes y otras modalidades.	Conocimiento actualizado, razonamiento lógico, capacidades multimodales robustas.	Sistemas interactivos, análisis multimedia, aplicaciones educativas y científicas.
Grok (xAI)	Modelo desarrollado por xAI con un enfoque en interacción directa y expresiva.	Inmediatez informativa, estilo comunicativo más dinámico y expresivo.	Plataformas de comunicación, asistentes personales, generación de contenido en tiempo real.
Mistral 7B / Mixtral	Modelo de código abierto, optimizado para eficiencia y personalización.	Ligereza, velocidad, bajo costo computacional, flexibilidad en entornos de innovación.	Laboratorios de investigación aplicada, *startups*, despliegues en entornos con recursos limitados.
LLaMA 4 (Meta)	Modelo abierto y ampliamente adoptado por la comunidad investigadora y de desarrolladores.	Transparencia, soporte comunitario, independencia de APIs propietarias.	Investigación académica, desarrollo de aplicaciones personalizadas, proyectos que requieren soberanía tecnológica.

REFERENCIAS

Arslan, S. (2025). Artificial intelligence in food safety and nutrition practices: opportunities and risks. *Academia Nutrition and Dietetics*, *2*(3). https://doi.org/10.20935/AcadNutr7904

Automation Anywhere. (2025). *¿Qué es la automatización robótica de procesos (RPA)? Una guía empresarial*. https://www.automationanywhere.com/la/rpa/robotic-process-automation

Balic, N. (2025). *Will agents replace us? Perceptions of autonomous multi-agent AI*. arXiv. https://doi.org/10.48550/arXiv.2506.02055

Bassey, M., Akpan, J., Ikpe, A., David, V., & Kehinde, T. O. (2025. 26 de agosto). The convergence of additive manufacturing and artificial intelligence with LLMs for smart product design. *Academia Materials Science*, *2*(3). https://doi.org/10.20935/AcadMatSci7868

Binance Academy. (s. f.). *Blockchain*. https://academy.binance.com/en/glossary/blockchain

Bornet, P., Wirtz, J., & Davenport, T. H., De Cremer, D., Evergreen, B. Fersht, P. Gohel, R. & Khiyara, S. (2025). *Agentic artificial intelligence. Harnessing AI agents to reinvent business, work and life*. World Scientific Publishing Company.

Casini, L., Marchetti, N., Montanucci, A., Orrù, V., & Roccetti, M. (2023). A human–AI collaboration workflow for archaeological sites detection. *Scientific Reports*, *13*(1), 8699. https://doi.org/10.1038/s41598-023-36015-5

Corvalán, J. G., & Sánchez Caparrós, M. (2025). *Agentes de inteligencia artificial y workflows agénticos: la nueva frontera de la automatización*. Laboratorio de Inteligencia Artificial de la Facultad de Derecho UBA. https://www.web.ialab.com.ar/webia/wp-content/uploads/2025/libros/Agentes-de-inteligencia-artificial-y-workflows-agenticos/Agentes-de-inteligencia-artificial-y-workflows-agenticos.pdf

Coshow, T. (2024, 1 de octubre). Los agentes inteligentes de IA pueden operar de forma autónoma. *Gartner*. https://www.gartner.es/es/articulos/agente-inteligente-de-ia

Empresa Actual. (2025, 11 de febrero). ¿Qué son los agentes de IA o la inteligencia artificial agéntica (IAA)? *EmpresaActual.com* https://www.empresaactual.com/que-son-los-agentes-de-ia-o-la-inteligencia-artificial-agentica-iaa/

Hauser, E. (2019, 9-13 de noviembre). *AI systems as ethical agents* [Presentación dentro del taller Good Systems: Ethical AI]. 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), Austin, Texas, Estados Unidos.

Li, Y. H., Chang, Z. H., Shan Ou, Y., Chen, Y., Yu, L. Y., Tan, J. Q., Liu, Y., & Chang, Y. Z. (2025). Exploration of the development and technical features of intelligent inspection robots. *International Journal of Advanced Engineering Research and Science*, *12*(10), 61-70. https://doi.org/10.22161/ijaers.1210.9

Lieben, M. (2025). *3 types of AI workflows* [flujograma]. LinkedIn. https://www.linkedin.com/posts/the-best-of-ai_ai-agents-are-replacing-human-employees-activity-7364786657318838273-SH--/

Ramalingam, B. C. (2025). Efficient implementation of AI agents in enterprise application integration (EAI) and electronic data interchange (EDI). *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *11*(2), 150-170. https://doi.org/10.32628/cseit251112397

Samdani, G., Viswanathan, G., & Dasu Jegadeesh, A. (2025). Human-AI collaboration: Balancing agentic AI and autonomy in hybrid systems. *International Journal on Cloud Computing: Services and Architecture*, *15*(1), 1-15. https://doi.org/10.5121/ijccsa.2025.15101

Shaw, F. X. (2025, 19 de mayo). Microsoft Build 2025: The age of AI agents and building the open agentic web. *Official Microsoft Blog.* https://blogs.microsoft.com/blog/2025/05/19/microsoft-build-2025-the-age-of-ai-agents-and-building-the-open-agentic-web/

Shoham, Y., & Leyton-Brown, K. (2008). Multiagent systems. Algorithmic, game-theoretic, and logical foundations. *Cambridge University Press*.

The Best of AI. (2025). *AI agents are replacing human employees. Build yours* [Publicación]. LinkedIn. https://www.linkedin.com/posts/the-best-of-ai_ai-agents-are-replacing-human-employees-activity-7364786657318838273-SH--/

Viswanathan, G., Samdani, G., & Dixit, Y. (2025). AI Agents. *International Journal of Advanced Information Technology*, *15*(1/2), 9-17. https://doi.org/10.5121/ijait.2025.15202

Xu, Z., & King, I. (2025, 2 de septiembre). Introducing *Academia AI and Applications*: A new platform for responsible and interdisciplinary AI research. *Academia AI and Applications*. https://doi.org/10.20935/AcadAI7885

UNIVQUAKE: SERIOUS VIRTUAL REALITY GAME ABOUT EARTHQUAKES IN UNIVERSITY SCENARIOS

Sebastian Herrera Solis
https://orcid.org/0009-0009-9837-4257
sebas.herrera.solis@gmail.com
Universidad de Lima, Peru

Hernan Quintana-Cruz
https://orcid.org/0000-0002-7037-4302
hquintan@ulima.edu.pe
Universidad de Lima, Peru

Received: August 31, 2025 / Accepted: November 3, 2025
doi: https://doi.org/10.26439/interfases2025.n022.8229

UNIVQUAKE: SERIOUS VIRTUAL REALITY GAME ABOUT EARTHQUAKES IN UNIVERSITY SCENARIOS

ABSTRACT. Conventional earthquake drills often fail to provide safe, realistic, and repeatable training for procedural knowledge. This limitation is particularly critical in high-density settings, such as university environments, where it is vital to be prepared in the event of such disasters. Virtual reality (VR) combined with serious games (SGs) provides an opportunity for users to experience and practice responses to such events within a controlled environment. This research describes the design, development, and validation of VR-based SG for earthquake preparedness. The primary objective was to assess the effectiveness of the SG in improving university students' knowledge of the measures to be taken in the event of earthquakes. The secondary objective was to analyze the correlation between the level of presence experienced in the SG and the knowledge gain. A pre-post study was conducted using a questionnaire validated by expert judgment to measure knowledge acquisition. The results demonstrated a statistically significant increase in participants' average knowledge scores after using the VR-based SG. Additionally, a noteworthy finding was that no direct correlation could be confirmed between the level of presence reported by the participants and the knowledge they acquired. This research concludes that VR-based SG are an effective technological tool for improving procedural knowledge in earthquake preparedness in university settings, even when subjective presence is not a significant factor in the learning process.

KEYWORDS: earthquake, virtual reality, presence, serious game, undergraduate

UNIVQUAKE: JUEGO SERIO DE REALIDAD VIRTUAL SOBRE TERREMOTOS EN ESCENARIOS UNIVERSITARIOS

RESUMEN. Los simulacros de terremotos convencionales a menudo no proporcionan una formación segura, realista y repetible sobre los procedimientos a seguir. Esta deficiencia es especialmente grave en entornos de alta densidad, como los universitarios, donde la preparación ante este tipo de desastres resulta fundamental. La realidad virtual (RV), combinada con los juegos serios (JS), ofrece la oportunidad de «vivir» y practicar las respuestas a estos eventos en un entorno controlado. Esta investigación describe el diseño, desarrollo y validación de un JS basado en RV para la preparación ante terremotos. El objetivo principal era medir la eficacia del JS para mejorar los conocimientos de los estudiantes universitarios sobre las medidas que deben tomarse en caso de terremoto. El objetivo secundario era analizar la correlación entre el nivel de presencia experimentado en el JS y la adquisición de conocimientos. Se llevó a cabo un estudio pre-post utilizando un cuestionario validado por expertos para medir la adquisición de conocimientos. Los resultados demostraron un aumento estadísticamente significativo en las puntuaciones medias de conocimientos de los participantes después de utilizar el JS de RV. Además, un hallazgo notable fue que no se pudo confirmar una correlación directa entre el nivel de presencia reportado por los participantes y los conocimientos que obtuvieron. Este estudio concluye que los JS de RV son una herramienta tecnológica eficaz para mejorar los conocimientos procedimentales en materia de preparación para terremotos en entornos universitarios, incluso cuando la presencia subjetiva no es un factor significativo en el proceso de aprendizaje.

Palabras clave: realidad virtual, terremotos, juego serio, presencia, universitarios

INTRODUCTION

Peru is part of the so-called Pacific Ring of Fire (Guardia & Tavera, 2012, p. 1), which means that the possibility of an earthquake is a constant risk in the country. This vulnerability is shared by other countries, such as Japan, which is considered an international reference in disaster risk reduction due to its effective strategies for managing such disasters (Pastrana-Huguet et al., 2022). In contrast, Peru faces challenges in fostering an effective culture of prevention and safety, as only approximately 55% of the population participated in drills during 2019 (Instituto Nacional de Defensa Civil [INDECI], 2020).

This vulnerability presents a critical risk for high-density populations. In 2023, Peruvian private universities had combined a total of over 1.5 million enrolled students (National Institute of Statistics and Informatics [INEI], 2023). Given the unpredictability of earthquakes, they can occur at any time, including during class hours. Therefore, this research focuses on undergraduate students, as preparedness on university campuses is a key factor in mitigating risks.

Currently, the primary preparedness tool used in this context is national drills. However, most earthquake drills lack sufficient effectiveness and realism. Consequently, students and administrators do not gain a truly realistic experience during these activities (Gong et al., 2015, p. 2242).

Given the need for simulated environments that provide a realistic experience of these events, a number of studies have proposed the use of new technologies, including VR, to recreate disaster-training scenarios at a lower cost and with significantly reduced risk compared to real-world exercises. This approach has proved to be useful in both educational and training contexts (Lu et al., 2020). Additionally, users must be able to learn from the simulation in an interactive and engaging way. According to Checa and Bustillo (2020), serious games (SGs) offer a student-centered educational approach with specific, well-defined tasks that enhance learning. Through this approach, students engage in an interactive and motivating process that will improve their overall, active, and critical learning. Therefore, the objective of this study is to design, develop, and validate a VR-based SG for earthquake preparedness in a university scenario.

The article is organized into three main parts. First, it introduces a review of the relevant literature from the last seven years related to the research proposal. Next, it describes the methodology proposed for this research. Finally, it addresses the experimentation conducted.

STATE OF THE ART

This section provides a compilation of articles on VR-based SG for emergencies, VR-based SG focused on earthquakes, and VR-based earthquake simulators. It also summarizes recent advances and identifies existing research gaps.

Serious Immersive and Non-Immersive Games for Emergencies

VR-based SG can support training for emergency scenarios and enhance individuals' chances of survival in highly hazardous situations (Irshad et al., 2021). Furthermore, VR technology offers an innovative approach to delivering this training in a scalable and resource-efficient manner (Rickenbacher-Frey et al., 2023). Table 1 presents a compilation of articles on the design, development, and implementation of VR-based SG for emergency situations.

Table 1

VR-based SG for Emergency Situations

Emergency	Subjects	I or N	Type of navigation	Scenario	Origin	Authors	Contribution / Gap
Rescue in confined spaces	20 regular workers and 20 specialized workers	I	Limited to decisions	Incident in Foshan, China	China	Lu et al. (2020)	Gap: Participant cannot control movement
Flooding	55 volunteers	N	Open navigation via keyboard and mouse	Urban buildings in Italy	Italy	D'Amico et al. (2023)	Contribution: Focus on safe routes and object interaction
Terrorist strike	32 volunteers	I	Teleportation	University building	New Zealand	Lovreglio et al. (2022)	Contribution: University scenario
Fire	30 people aged 60 to 80	I	Teleportation	Realistic residences	China	Fu & Li (2023)	Contribution: Dynamic events
Fire	140 junior students	I	-	High-rise building	Taiwan	Chen & Chien (2022)	Contribution: Teaching situations based on levels
Fire	78 hospital staff members	N	Open navigation via keyboard and mouse	Vincent Van Gogh Hospital	Belgium	Rahouti et al. (2021)	Contribution: Immersive experience positively affects "Presence"

Note. I = immersive, N = non-immersive.

VR-BASED SG FOR EARTHQUAKE TRAINING

Serious games in VR about earthquakes

Table 2 presents a compilation of articles that address the design, development, and implementation of immersive VR-based SG focused on earthquake preparedness training.

Table 2

Immersive VR-based SG for Earthquakes

Subjects	Type of navigation	Scenario	Origin	Authors	Contribution / Gap
91	Predefined	High school and office building	-	Feng, González, Mutch et al. (2020)	Contribution: 6-framework for customizable learning
93	Predefined	Fifth floor of Auckland City Hospital	New Zealand	Feng, González, Amor et al. (2020)	Contribution: Qualitative strategy for damages
147	Open using remote control	Shopping mall	-	Ahmadi et al. (2023)	Gap: No object interaction
42	Teleportation	Juvenile room	Greece	Maragkou et al. (2023)	Contribution: Interactive mechanics

The literature reveals a lack of studies that integrate natural locomotion systems with object interaction across different scenarios.

VR Simulators for Earthquakes

During this section, the reviewed literature uses VR to measure and observe participant behavior. For example. Zhang et al. (2021) developed a VR simulator to evaluate the safe actions people take during an earthquake in an office and in a room, aiming to understand people's behavior during an earthquake. Similarly, Mitsuhara et al. (2021) created a system to observe behaviors during the evacuation of people, highlighting the differences between a single-person and a two-person simulation. Their primary goal is to understand what people do in these types of emergencies.

On the other hand, the literature also focuses on the environmental replication of real scenarios. Xu et al. (2023) implemented a method using mixed reality where they scanned the simulation area and showed safe and dangerous zones where the participant should be located, highlighting the ability to represent an environment in VR. Also, Suzuki et al.

(2018) conceived an earthquake simulator that generates a copy of the indoor environment using artificial intelligence (AI) technologies and allows it to be projected into the VR world.

Meanwhile, other literature validates the educational impact. Rajabi et al. (2022) decided to evaluate the effect of VR education on decisions during an earthquake, using a classroom in Tehran as a case study. Liuwandy et al. (2020) explored an affordable way to conduct earthquake simulations in VR, creating an earthquake simulator that can be run on cell phones. Finally, Shu et al (2019) compared the presence and effectiveness of an earthquake simulator viewed on a computer screen with that of a VR headset. They concluded that VR simulators help people become familiar with earthquake simulations.

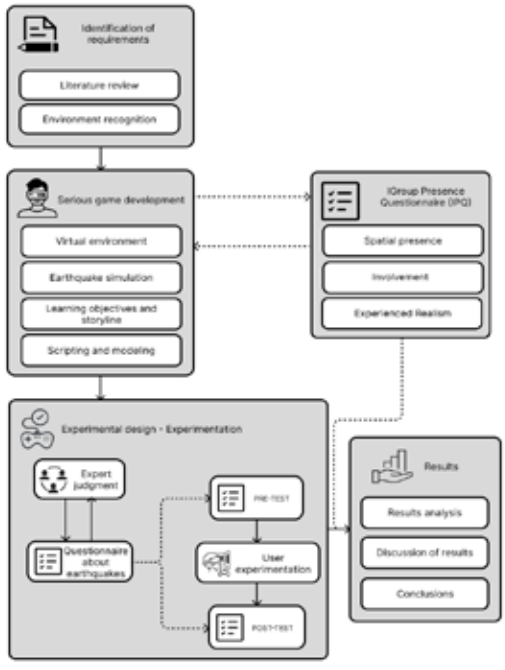
A review of the literature reveals significant deficiencies. We identified a shortcoming in locomotion and navigation, with most applications opting for restrictive navigation methods. Secondly, physical interactivity is clearly limited. Current applications focus on decision-making rather than on interactions with the environment. Finally, reviewed simulators are primarily used for behavioral observation. This suggests an opportunity to gamify these simulations, transforming them into effective training tools.

3. MATERIALS AND METHODS

The methodology proposed for this research can be seen in Figure 1.

Figure 1

Methodological Scheme



Identification of Requirements

During the requirements identification process, the material provided by INDECI was reviewed, from which the relevant recommendations and instructions for these disaster scenarios were extracted. This information is essential for defining the learning objectives and developing the game's storyline.

As shown in Table 3, the learning objectives incorporate INDECI's recommendations for actions to be taken before, during, and after an earthquake.

Table 3

Learning Objectives During the Stages of an Earthquake

Stage	Learning objectives
Before an earthquake	• Identify safe areas • Prepare an emergency backpack • Keep exits clear of obstacles
During an earthquake	• Remain calm • Avoid falling objects • Stay away from windows • Go to safe areas or perform the "drop, cover, and hold on" procedure
After an earthquake	• Check your health and others' health • Do not use elevators • Cover your head during evacuation • Avoid returning to the building • Evacuate in an orderly and safe manner • Call emergency numbers, if necessary

Note. Learning objectives based on interviews and a review of official guidelines (INDECI, 2024).

Additionally, the learning objectives address other preparedness activities, such as identifying the essential items that should be included in an emergency kit: (i) water, (ii) canned food, (iii) coats, (iv) a flashlight, (v) a battery-powered radio, (vi) a whistle, and (vii) a first-aid kit. Furthermore, the objectives include identifying safe areas inside buildings in the event of an earthquake, such as (i) columns, (ii) life-triangle zones, (iii) areas away from glass or falling objects, (iv) doorways, (v) locations under beams, and (vi) areas outside elevator shafts (INDECI, 2024).

Regarding the recognition of environments, the approach taken by Xiao et al. (2017) was followed, in which the author proposed a complete review of the infrastructure of the environments. Classrooms L3-402 and O2-202 at the Universidad de Lima were selected for analysis and research, along with their respective evacuation routes and points of interest, due to the availability and accessibility of infrastructure data.

SG Development

During the development of the virtual environment, basic 3D models were used to create the buildings aiming to replicate reality as closely as possible, along with textures designed to simulate the surrounding environment. Several objects within the environments were sourced from Sketchfab (https://sketchfab.com/) and Unity Asset Store (https://assetstore.unity.com/). As seen in Figure 2, these models in the virtual classroom must exhibit physics, dimensions, and behaviors consistent with their real-world counterparts (Rajabi et al., 2022).

According to Lovreglio (2018), this study will use earthquake simulation using a qualitative strategy. This approach involves simulating the damage caused by the disaster without incorporating information about the structures in the simulated environment. In addition, this strategy was used because it is more suitable for designing a SG, as it allows threats to be represented strategically to enhance training effectiveness (Lovreglio, 2018). The Modified Mercalli Intensity (MMI) scale was used to choose the earthquake magnitude, as it aligns with the qualitative approach, based on work of Feng, González, Mutch et al. (2020). The scale chosen for this research is MM9, as this causes partial damage to buildings, total breakage of glass, and cracks in walls (Dowrick et al., 2008).

Based on Lu et al. (2020) and Rahouti et al. (2021), a story for the game is proposed using the scenarios and learning objectives (see Table 3). The story, which lasts approximately 30 minutes, consists of the following elements:

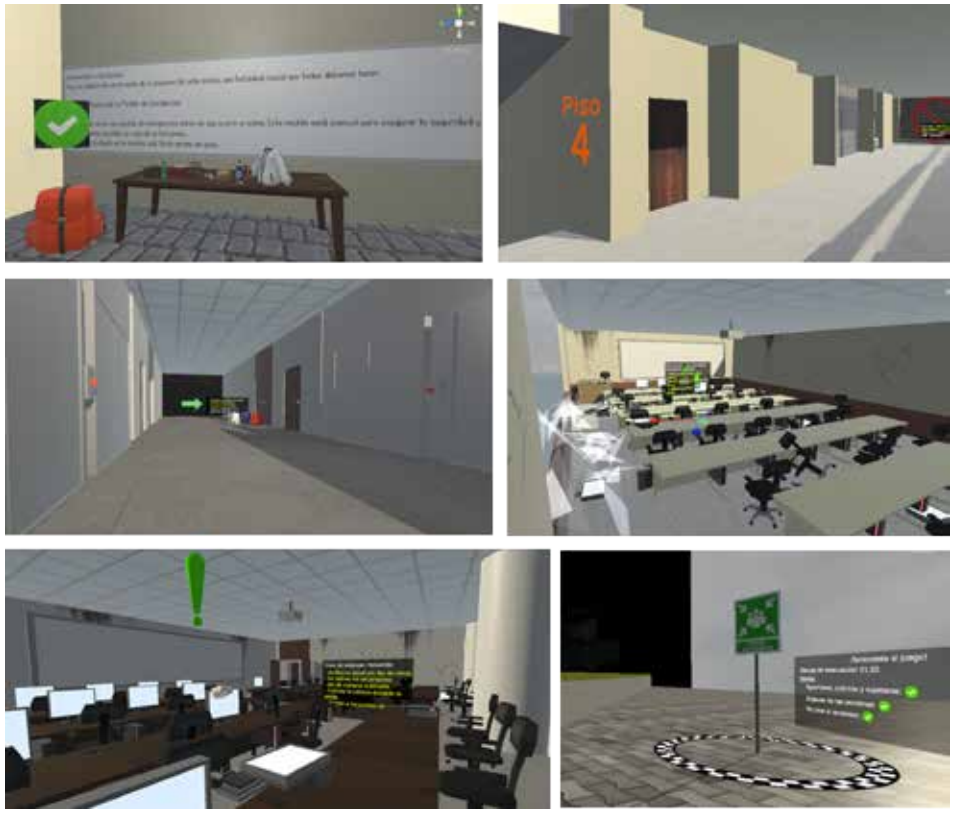
- At the beginning, participants must prepare their emergency kit before proceeding to their classroom.
- Participants begin outside their classroom building and are required to proceed toward it.
- During their journey, they must recognize safe areas and ensure that there are no obstacles along the evacuation route.
- When they arrive at their classroom and take their seats, an earthquake begins. Participants must remain calm and decide whether to take shelter under a table, move to a safe area, or evacuate as quickly as possible using the available controls. Depending on their choice, they may need to repeat this level.
- After the earthquake, participants must wait for any aftershocks or move toward the exit.
- During the evacuation, participants must avoid any objects that may fall or cause injury, while keeping their heads protected as a precaution.
- Once the evacuation is complete, participants must proceed to an emergency meeting point and call the emergency numbers, if necessary, to end the game.

They will also have the option to return to the building, in which case they will restart the level.

At the end of the game, the players will be provided with a report detailing their actions throughout the simulation, along with the total evacuation time (Figure 2).

Figure 2

Screenshots of UnivQuake



Note. UnivQuake scenes: "Emergency kit" (top left), "L3 hallway" (top right), "O2 hallway" (middle left), "L3 classroom after earthquake" (middle right), "O2 classroom after earthquake" (bottom left), and "Evacuation complete" (bottom right).

Unity Game engine was used to create the game, as this Integrated Development Environment (IDE) provides the audio, physics, and graphics tools required to create an enjoyable experience in a space in which users can freely navigate the virtual space (Ilongwe et al., 2023). Within Unity, version 2021.3.23f1 was used, along with the XR Interaction Toolkit, an interaction system for creating augmented and VR experiences through interactive objects that operate using Unity events (Unity, 2023).

Regarding the sounds implemented for the earthquake simulation, the audio files were sourced from Zapslat (https://www.zapslat.com/) and were used to illustrate the noises that occur during an earthquake.

During the development stage, a pilot test was conducted to measure the presence level of the SG using the Igroup Presence Questionnaire (IPQ), an instrument designed to assess the sense of presence experienced by users in a virtual environment. The IPQ consists of one general item and three subscales: (i) Spatial Presence, (ii) Involvement, and (iii) Experienced Realism (Igroup, 2016). This pilot test enabled adjustments to several presence-related aspects of the SG.

Questionnaire About Earthquakes

The questionnaire used in this research to measure knowledge consists of five open-ended questions related to the learning objectives, as shown in Table 4. The score for each question will be determined by the number of correct responses provided by the participants for that question. The use of this type of questionnaire is supported by De Fino et al. (2023), Feng, González, Mutch et al. (2020), and Lu et al. (2020). These studies use pre-test and post-test questionnaires, methodologically designed to measure the knowledge acquired through their respective SGs.

Table 4

Earthquake Knowledge Questionnaire

Questions	Score
What actions should you take before an earthquake?	• 3 points if you give 3 answers similar to the following options: (i) Pack your emergency kit, (ii) Identify safe areas in the building, and (iii) Ensure that there are no obstructions along the exit route. • 2 points for 2 similar answers out of the 3 options. • 1 point for 1 similar answer out of the 3 options. • 0 points if participants provide no correct answers.
What to do during an earthquake?	• 4 points if you give 4 answers similar to the following options: (i) Avoid falling objects or broken glass, (ii) Use the triangle of life, (iii) Stay calm, (iv) Go to the safe area inside. • 3 points for 3 similar answers out of the 4 options. • 2 points for 2 similar answers out of the 4 options. • 1 point for 1 similar answer out of the 4 options. • 0 points if participants provide no correct answers.

(*continúa*)

(*continuación*)

Questions	Score
What to do after an earthquake?	• 8 points if you give 8 answers similar to the following options: (i) Check your health and/or that of others, (ii) Do not use elevators; use stairs, (iii) Leave in an orderly and safe manner without running, (iv) Cover your head while leaving, (v) Do not return to the building, (vi) Go to the meeting points, (vii) Follow INDECI information/stay informed, and (viii) Call emergency numbers, if necessary. • 7 points for 7 similar answers out of the 8 options. • 6 points for 6 similar answers out of the 8 options. • 5 points for 5 similar answers out of the 8 options. • 4 points for 4 similar answers out of the 8 options. • 3 points for 3 similar answers out of the 8 options. • 2 points for 2 similar answers out of the 8 options. • 1 point for 1 similar answer out of the 8 options. • 0 points if participants provide no correct answers.
What items should be included in an emergency kit?	• 7 points if you give 7 answers similar to the following options: (i) Water, (ii) Canned food, (iii) Warm clothing, (iv) Flashlight, (v) Battery-powered radio, (vi) Whistle, and (vii) First aid kit. • 6 points for 6 similar answers out of the 7 options. • 5 points for 5 similar answers out of the 7 options. • 4 points for 4 similar answers out of the 7 options. • 3 points for 3 similar answers out of the 7 options. • 2 points for 2 similar answers out of the 7 options. • 1 point for 1 similar answer out of the 7 options. • 0 points if participants provide no correct answers.
What are the safe areas inside a building in the event of an earthquake?	• 6 points if you give 6 similar answers to the following options: (i) Columns, (ii) Life triangle, (iii) Away from glass or falling objects, (iv) Doorways, (v) Under beams, and (vi) Outside elevators. • 5 points for 5 similar answers out of the 6 options. • 4 points for 4 similar answers out of the 6 options. • 3 points for 3 similar answers out of the 6 options. • 2 points for 2 similar answers out of the 6 options. • 1 point for 1 similar answer out of the 6 options. • 0 points if participants provide no correct answers.

The questionnaire was validated by expert judgment. The scale used was the Content Validity Ratio (CVR) proposed by Lawshe (1975). The five questions were evaluated by twelve INDECI members, each receiving a CVR value of 0.80. Therefore, these questions are essential and were included in the questionnaire.

Participants

The target population consisted of undergraduate students who attended computer-equipped classrooms during 2024. The sample included 30 undergraduate students (25 male, 5 female), aged 20 to 25 years. All participants reported prior experience in

computer-equipped classrooms. Consistent with Tavares (2022), the sample was familiar with digital interaction, as most participants are young individuals who frequently engage with digital technologies such as video games. Regarding the technology used in the study, only four participants reported prior experience with VR headsets.

Experimentation

To assess the knowledge of the measures to be taken during earthquakes, a quasi-experimental repeated-measures design was implemented. A pre-test using the proposed questionnaire was conducted to determine the participants' initial knowledge level. Subsequently, a post-test was conducted after participants completed the SG in VR.

To evaluate the presence experienced in the SG, the IPQ was used, as in the pilot test, since this instrument allows effective measurement of participants' sense of presence in virtual environments.

Additionally, the study aims to analyze the relationship between the level of presence experienced by the test participants and the improvement in knowledge of earthquake safety measures. To this end, a correlation analysis will be conducted between the IPQ scores and the differences between pre-test and post-test results.

The different hypotheses proposed are described below:

- Null hypothesis 1 ($H0_1$): There is no significant difference between the participants' knowledge before and after the SG test.
- Alternative hypothesis 1 ($H1_1$): There is a significant improvement in the participants' knowledge after participating in the SG.
- Null hypothesis 2 ($H0_2$): No significant correlation was found between the level of presence experienced in the SG and the improvement in knowledge of earthquake safety measures.
- Alternative hypothesis 2 ($H1_2$): A significant positive correlation was found between the level of presence experienced in the SG and the improvement in knowledge of earthquake safety measures.

To participate in the game, a device enabling immersive visualization of the virtual environment is required. Therefore, Meta Quest 2 headsets were used in the research, allowing for tracking of both head and hand movements.

For interaction with the virtual environment and free movement, the Touch controllers included with the headsets were used. Regarding sound stimuli, Logitech Astro A10 headphones were used during the experiment.

RESULTS

This section presents knowledge assessment results corresponding to Hypothesis 1. As shown in Table 5, the Shapiro-Wilk test was first applied to data to determine their normal distribution. The resulting *p* value for this test is reported in the "Normality" column. If the normality value was less than 0.05, the non-parametric Wilcoxon test was applied; if it was greater than 0.05, the paired t-test was used. Subsequently, the selected test was used to validate the existence of an improvement. The table compares these pre-test and post-test scores, where *M* is the mean score and *SD* is the standard deviation

Table 5

Earthquake Knowledge Questionnaire Results

Learning objectives	Pre-test	Post-test	Normality	*p* value	T- values/ W values	Statistical test
Before the earthquake	M = 1.53 SD = 0.68	M = 2.26 SD = 0.73	0.0003974	0.00005912	*W* = 9	Wilcoxon
During the earthquake	M = 1.86 SD = .77	M = 2.36 SD = .96	.0346	.009472	*W* = 39	Wilcoxon
After the earthquake	M = 1.53 SD = .77	M = 2.63 SD = 1.21	.04416	.0005698	*W* = 32.5	Wilcoxon
Emergency kit	M = 3.50 SD = 1.63	M = 5.36 SD = 1.09	.08104	.000001844	*T* = 5.6951	Paired t-test
Indoor safe zones in case of earthquakes	M = 1.23 SD = .77	M = 2.43 SD = 1.19	.02167	.0001014	*T* = 5.1739	Paired t-test
Total	M = 9.66 SD = 2.68	M = 15.06 SD = 3.12	0.2961	9.61E-10	*T* = 8.5727	Paired t-test

An improvement in knowledge was observed across all evaluated areas. In the assessment of knowledge prior to the earthquake, the mean score increased from 1.53 to 2.26, with a $p < .05$, indicating a significant improvement in knowledge levels. This effect may be attributed to the fact that this stage involved the greatest level of participant interaction with objects within the game.

During the earthquake knowledge assessment, this item showed the smallest increase, with the mean score rising from 1.86 to 2.36 and a $p < .05$. This could be

caused by the fact that during this stage, participants were in distress due to the earthquake, which led them to focus on escaping the environment rather than performing the recommended actions. In the post-earthquake knowledge assessment, the mean score increased from 1.53 to 2.63, with a $p < .05$, which suggested a significant improvement.

The question related to the emergency kit showed the greatest increase in the mean score compared to the other questions. This effect may be attributed to participants' engagement with the activity of preparing the emergency kit. Finally, the assessment related to indoor safe zones during earthquakes showed an increase in the mean score from 1.23 to 2.43, with a $p < .05$, which suggested a significant improvement, even though the activity related to this assessment was addressed superficially. Our findings on knowledge improvement are consistent with those reported by Feng, González, Mutch et al. (2020) and Lu et al. (2020). However, these findings contrast with those reported by Ahmadi et al. (2023), who indicated that their post-game assessment did not show a significant increase in safety knowledge.

Hypothesis 2 proposes a significant correlation between the perceived level of presence and improved knowledge of earthquake preparedness measures. Spearman's correlation test was applied, given that the data from the knowledge questionnaire did not show a normal distribution across all objectives. These results did not show significant evidence between the two variables, with a correlation coefficient $\rho = 0.236$ and a coefficient of determination $r^2 = 0.0557$, implying that only 5.5% of the variability in knowledge can be explained by the overall level of presence. Therefore, Hypothesis 2 is not supported by the findings of this study.

CONCLUSIONS

The purpose of this research was to design, develop, and validate a VR-based SG for earthquake preparedness in university scenarios. The primary objective was to assess participants' knowledge, and the secondary objective was to analyze the correlation between the level of presence experienced in the SG and improved knowledge.

To this aim, two simulated environments were created in which participants could experience an earthquake and complete tasks before, during, and after the earthquake. The pre-post study used a questionnaire validated by expert judgment to measure knowledge.

A statistically significant increase in participants' average knowledge scores was observed after testing the SG. In addition, an important finding was that a direct correlation between the level of presence and the knowledge gain.

The importance of this work lies in the provision of a technological solution for earthquake risk management education in university settings and emphasizes the scientific value of VR SG in as an effective tool for disaster preparedness training.

One of the main limitations is the hardware processing capacity, which restricts the generation of more realistic scenarios and the inclusion of non-player characters.

These limitations provide directions for future research. Additionally, the implementation of immediate feedback is proposed to enable participants to identify and correct their mistakes in real time. It is suggested to include more activities that reinforce knowledge in stages and investigate which factors directly influence learning within VR.

In conclusion, this research validates VR SGs as a practical and effective tool for building procedural knowledge in disaster preparedness within university settings.

REFERENCES

Ahmadi, M., Yousefi, S., & Ahmadi, A. (2023). *Investigation of the most effective training method for rescuing people in earthquake emergency using immersive virtual reality serious games* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.4555342

Checa, D., & Bustillo, A. (2020). A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications*, *79*(9-10), 5501–5527. https://doi.org/10.1007/s11042-019-08348-9

Chen, S.-Y., & Chien, W.-C. (2022). Immersive virtual reality serious games with DL-assisted learning in high-rise fire evacuation on fire safety training and research. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.786314

D'Amico, A., Bernardini, G., Lovreglio, R., & Quagliarini, E. (2023). A non-immersive virtual reality serious game application for flood safety training. *International Journal of Disaster Risk Reduction*, *96*, 103940. https://doi.org/10.1016/j.ijdrr.2023.103940

De Fino, M., Tavolare, R., Bernardini, G., Quagliarini, E., & Fatiguso, F. (2023). Boosting urban community resilience to multi-hazard scenarios in open spaces: A virtual reality – serious game training prototype for heat wave protection and earthquake response. *Sustainable Cities and Society*, *99*, 104847. https://doi.org/10.1016/j.scs.2023.104847

Dowrick, D. J., Hancox, G. T., Perrin, N. D., & Dellow, G. D. (2008). The Modified Mercalli intensity scale. *Bulletin of the New Zealand Society for Earthquake Engineering*, *41*(3), 193-205. https://doi.org/10.5459/bnzsee.41.3.193-205

Feng, Z., González, V. A., Amor, R., Spearpoint, M., Thomas, J., Sacks, R., Lovreglio, R., & Cabrera-Guerrero, G. (2020). An immersive virtual reality serious game to enhance earthquake behavioral responses and post-earthquake evacuation preparedness in buildings. *Advanced Engineering Informatics*, *45*, 101118. https://doi.org/10.1016/j.aei.2020.101118

Feng, Z., González, V. A., Mutch, C., Amor, R., Rahouti, A., Baghouz, A., Li, N., & Cabrera-Guerrero, G. (2020). Towards a customizable immersive virtual reality serious game for earthquake emergency training. *Advanced Engineering Informatics*, *46*, 101134. https://doi.org/10.1016/j.aei.2020.101134

Fu, Y., & Li, Q. (2023). A virtual reality–based serious game for fire safety behavioral skills training. *International Journal of Human-Computer Interaction*, *40*(19), 5980-5996. https://doi.org/10.1080/10447318.2023.2247585

Guardia, P., & Tavera, H. (2012). Inferencias de la superficie de acoplamiento sísmico interplaca en el borde occidental del Perú. *Boletín de la Sociedad Geológica del Perú, 106*, 3748. http://hdl.handle.net/20.500.12816/914

Gong, X., Liu, Y., Jiao, Y., Wang, B., Zhou, J., & Yu, H. (2015). A novel earthquake education system based on virtual reality. *IEICE Transactions on Information and Systems*, *E98.D*(12), 2242-2249. https://doi.org/10.1587/transinf.2015edp7165

Igroup. (2016). *Igroup Presence Questionnaire (IPQ) overview* [Questionnaire]. https://www.igroup.org/pq/ipq/index.php

Ilongwe, A., Sepulveda, C., & Kashani, T. (2023). The calling VR: A musical virtual reality experience. *SIGGRAPH '23: Special Interest Group on Computer Graphics and Interactive Techniques Conference*. https://doi.org/10.1145/3588027.3595598

Instituto Nacional de Defensa Civil. (2020). *Estadísticas de gestión reactiva de la gestión del riesgo de desastres – Año 2019* [Reactive management statistics of disaster risk management – Year 2019] [Report]. https://portal.indeci.gob.pe/wp-content/uploads/2021/02/CAPITULO-II-Estad%C3%ADsticas-GR-2019.pdf

Instituto Nacional de Defensa Civil. (2024, January 14). *¿Qué hacer en caso de sismo?* [What to do in case of an earthquake?]. Gob.pe. https://www.gob.pe/1053-que-hacer-en-caso-de-sismo

Instituto Nacional de Estadística e Informática. (2023). *Número de alumnos/as matriculados en universidades privadas, 2008-2023* [Number of students enrolled in private universities, 2008-2023] [Data set]. Retrieved September 15, 2023, from https://m.inei.gob.pe/estadisticas/indice-tematico/university-tuition/

Irshad, S., Perkis, A., & Azam, W. (2021). Wayfinding in virtual reality serious game: An exploratory study in the context of user perceived experiences. *Applied Sciences*, *11*(17), 7822. https://doi.org/10.3390/app11177822

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Liuwandy, P., Suryasari, & Wella. (2020). Affordable mobile virtual reality earthquake simulation. *2020 Fifth International Conference on Informatics and Computing (ICIC)*. https://doi.org/10.1109/ICIC50835.2020.9288562

Lovreglio, R., Gonzalez, V., Feng, Z., Amor, R., Spearpoint, M., Thomas, J., Trotter, M., & Sacks, R. (2018). Prototyping virtual reality serious games for building earthquake preparedness: The Auckland City Hospital case study. *Advanced Engineering Informatics*, *38*, 670-682. https://doi.org/10.1016/j.aei.2018.08.018

Lovreglio, R., Ngassa, D.-C., Rahouti, A., Paes, D., Feng, Z., & Shipman, A. (2022). Prototyping and testing a virtual reality counterterrorism serious game for active shooting. *International Journal of Disaster Risk Reduction*, *82*, 103283. https://doi.org/10.1016/j.ijdrr.2022.103283

Lu, S., Xu, W., Wang, F., Li, X., & Yang, J. (2020). Serious game: Confined space rescue based on virtual reality technology. *2020 2nd International Conference on Video, Signal and Image Processing*. https://doi.org/10.1145/3442705.3442716

Maragkou, V., Rangoussi, M., Kalogeras, I., & Melis, N. S. (2023). Educational seismology through an immersive virtual reality game: Design, development and pilot evaluation of user experience. *Education Sciences*, *13*(11), 1088. https://doi.org/10.3390/educsci13111088

Mitsuhara, H., Tanioka, I., & Shishibori, M. (2021). Observing evacuation behaviours of surprised participants in virtual reality earthquake simulator. *Proceedings of the 29th International Conference on Computers in Education*, *2*, 576-582. https://doi.org/10.58459/icce.2021.4294

Pastrana-Huguet, J., Casado-Claro, M.-F., & Gavari-Starkie, E. (2022). Japan's culture of prevention: How bosai culture combines cultural heritage with state-of-the-art disaster risk management systems. *Sustainability*, *14*(21), 13742. https://doi.org/10.3390/su142113742

Rahouti, A., Lovreglio, R., Datoussaïd, S., & Descamps, T. (2021). Prototyping and validating a non-immersive virtual reality serious game for healthcare fire safety training. *Fire Technology*, *57*(2), 3041-3078. https://doi.org/10.1007/s10694-021-01098-x

Rajabi, M. S., Taghaddos, H., & Zahrai, S. M. (2022). Improving emergency training for earthquakes through immersive virtual environments and anxiety tests: A case study. *Buildings*, *12*(11), 1850. https://doi.org/10.3390/buildings12111850

Rickenbacher-Frey, S., Adam, S., Exadaktylos, A. K., Müller, M., Sauter, T. C., & Birrenbach, T. (2023). Development and evaluation of a virtual reality training for emergency treatment of shortness of breath based on frameworks for serious games. *GMS Journal for Medical Education*, *40*(2), Doc16. https://doi.org/10.3205/zma001598

Shu, Y., Huang, Y.-Z., Chang, S.-H., & Chen, M.-Y. (2019). Do virtual reality head-mounted displays make a difference? A comparison of presence and self-efficacy between head-mounted displays and desktop computer-facilitated virtual environments. *Virtual Reality*, *23*(4), 437-446. https://doi.org/10.1007/s10055-018-0376-x

Suzuki, R., Iitoi, R., Qiu, Y., Iwata, K., & Satoh, Y. (2018). AIbased VR earthquake simulator. In J. Y. C. Chen & G. Fragomeni (Eds.), *Lecture notes in computer science: Vol. 10910. Virtual, augmented and mixed reality: Applications in health, cultural heritage, and industry* (pp. 213-222). Springer. https://doi.org/10.1007/978-3-319-91584-5

Tavares, N. (2022). The use and impact of game-based learning on the learning experience and knowledge retention of nursing undergraduate students: A systematic literature review. *Nurse Education Today*, *117*, 105484. https://doi.org/10.1016/j.nedt.2022.105484

Unity. (2023). *XR Interaction Toolkit* (version 2.4) [Software documentation]. https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@2.4/manual/index.html

Xiao, M.-L., Zhang, Y., & Liu, B. (2017). Simulation of primary school-aged children's earthquake evacuation in rural town. *Natural Hazards*, *87*(3), 1783-1806. https://doi.org/10.1007/s11069-017-2849-8

Xu, Z., Yang, Y., Zhu, Y., & Fan, J. (2023). Mixed reality drills of indoor earthquake safety considering seismic damage of nonstructural components. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-43533-9

Zhang, F., Xu, Z., Yang, Y., Qi, M., & Zhang, H. (2021). Virtual reality-based evaluation of indoor earthquake safety actions for occupants. *Advanced Engineering Informatics*, *49*, 101351. https://doi.org/10.1016/j.aei.2021.101351

COMPUTER VISION FOR POKÉMON BATTLES: A YOLO AND TESSERACT-BASED SYSTEM FOR AUTOMATED RECOGNITION AND GAMEPLAY ANALYSIS

MIGUEL R. LLADÓ
https://orcid.org/0009-0001-1123-895X
mrllado@ulima.edu.pe
Universidad de Lima, Perú

TERENCE MORLEY
https://orcid.org/0000-0001-7588-6457
terence.morley@manchester.ac.uk
University of Manchester, United Kingdom

Received: September 11, 2025/ Accepted: November 6, 2025
doi: https://doi.org/10.26439/interfases2025.n022.8270

ABSTRACT. Pokémon Double Battles present a complex decision-making environment that has traditionally relied on manual data analysis. This paper introduces an automated system leveraging computer vision and deep learning to extract structured gameplay data from battle footage. Our approach integrates You Only Look Once (YOLO) for Pokémon sprite recognition along with Tesseract-based optical character recognition (OCR) for extracting move and status text. The study introduces a custom-built image dataset generated through the augmentation of publicly available Pokémon sprites, which is then used to train a YOLO model for sprite recognition. The system was tested across multiple controlled and real-world gameplay scenarios, achieving high accuracy in Pokémon recognition and action tracking. Additionally, a JSON-based gameplay notation system is proposed to structure battle sequences, thus improving analysis and strategic review. The results demonstrate the feasibility of AI-driven gameplay analysis, with potential applications for competitive players, game analysts, and developers. Given its exploratory nature, this study focuses on technical feasibility rather than statistical generalisation. Future work includes expanding the dataset, improving OCR performance, and enabling real-time processing to support broader practical use.

KEYWORDS: Computer vision, Pokémon, YOLO, Tesseract, OCR

VISIÓN POR COMPUTADOR PARA BATALLAS POKÉMON: UN SISTEMA BASADO EN YOLO Y TESSERACT PARA EL RECONOCIMIENTO AUTOMATIZADO Y EL ANÁLISIS DE JUEGO

RESUMEN. Las batallas dobles de Pokémon presentan un entorno de toma de decisiones complejo que tradicionalmente ha dependido del análisis manual de datos. Este artículo presenta un sistema automatizado que aprovecha la visión por computadora y el aprendizaje profundo para extraer datos estructurados del metraje de batalla. Nuestro enfoque integra You Only Look Once (YOLO) para el reconocimiento de sprites de Pokémon y reconocimiento óptico de caracteres (OCR) con Tesseract para la extracción de movimientos y estados. El estudio presenta un conjunto de datos de imágenes personalizado, generado mediante la ampliación de sprites de Pokémon disponibles públicamente, y lo utiliza para entrenar un modelo YOLO para el reconocimiento de sprites. El sistema se probó en múltiples escenarios controlados y de juego real, logrando alta precisión en el reconocimiento y seguimiento de acciones. Además, se propone un sistema de notación basado en JSON para estructurar las secuencias de batalla, mejorando el análisis y la revisión estratégica. Los resultados demuestran la viabilidad del análisis de juego basado en IA, con aplicaciones potenciales para jugadores competitivos, analistas de videojuegos y desarrolladores. Dada su naturaleza exploratoria, este estudio se centra en la viabilidad técnica más que en la generalización estadística. Las mejoras futuras incluyen la expansión del conjunto de datos, la optimización del OCR y la integración de procesamiento en tiempo real para un uso más amplio.

PALABRAS CLAVE: Visión por computadora, Pokémon, YOLO, Tesseract, OCR

INTRODUCTION

Background and Motivation

Artificial intelligence (AI) has become increasingly influential in the video game industry, with applications ranging from non-player character (NPC) behaviour to advanced game analytics. In particular, computer vision, a subfield of AI focused on interpreting visual information, enables systems to analyse images and video to extract actionable insights. These capabilities are increasingly applied to optimize gameplay strategies, monitor player performance, and provide viewers with enriched information overlays during broadcasts (El-Nasr et al., 2016; Yannakakis & Togelius, 2018).

Pokémon, a video game franchise with hundreds of millions of sales (The Pokémon Company, 2024b), has integrated AI to improve both player experience and viewer engagement. For instance, the Pokémon Battle Scope (PBS), developed in collaboration with the Japanese AI company HEROZ (Takahiro & Tomohiro, n.d.), has been used to augment tournament broadcasts by providing real-time strategic feedback during Single Battles (The Pokémon Company, 2024d).

However, PBS focuses exclusively on Single Battles (1v1). In contrast, Double Battles—the official format of competitive tournaments including the Pokémon World Championship—feature two Pokémon per player, significantly increasing the number of possible move combinations, synergies, and strategic options (The Pokémon Company, 2024c). Analysing Double Battles presents new challenges and opportunities for applying AI and computer vision. Figure 1 illustrates the layout of a Double Battle, where both players command two Pokémon simultaneously.

Figure 1

Gameplay Image of a Double Battle as Displayed in the Video Game Pokémon Scarlet

Note. Image captured from the video game Pokémon Scarlet on Nintendo Switch.

This paper combines YOLO—a family of real-time vision models capable of detection, segmentation, and classification (Jocher et al., 2024; Redmon et al., 2016)—with Tesseract OCR (Tesseract OCR, n.d.-b) to automatically extract structured data from Pokémon Double Battle footage. In our pipeline, YOLO is used in its image-classification configuration to recognise Pokémon sprites, while Tesseract extracts on-screen text. The outputs are serialised into a JSON-based battle notation suitable for downstream analysis.

The study advances computer vision-based gameplay analysis by delivering, to our knowledge, one of the first systems that integrates YOLO and Tesseract OCR for the automated interpretation of Pokémon Double Battles. In contrast to prior work focused on single-battle settings or text-based simulations (Hu et al., 2024; Norström, 2019; Simoes et al., 2020), we demonstrate that deep learning can derive machine-readable structure directly from raw gameplay video. The resulting JSON notation offers a reusable foundation for player training, game analysis, and strategy review.

Research Aim and Objectives

The central aim of this study is to design and evaluate an image recognition system for Pokémon Double Battles, capable of transforming gameplay footage into structured data. This work supports the broader goal of enhancing the analytical capabilities available to players, developers, and researchers.

The paper focuses on three core objectives:

- Developing an image recognition pipeline using YOLO for Pokémon identification and Tesseract for move extraction.
- Creating and curating a labelled dataset of commonly used competitive Pokémon sprites, ensuring accurate model training.
- Implementing a JSON-based notation system to represent each turn in a structured and analysable format.

These components aim to reduce the need for manual analysis and pave the way for more advanced AI-driven tools that support gameplay insight, performance evaluation, and competitive training.

RELATED WORK

AI has been widely applied to video games for purposes ranging from gameplay enhancement to player modelling and opponent simulation. In the context of Pokémon, most academic research has focused on decision-making strategies using structured, text-based battle logs rather than analysing the game's visual content (Yannakakis & Togelius, 2018).

For example, Simoes et al. (2020) developed a reinforcement learning agent capable of predicting optimal actions based on *Pokémon Showdown* simulations, which provides structured representations of in-game states. These models benefit from direct access to internal data such as team composition, moves, and health, making them easier to train. However, they do not address the challenge of interpreting battles from unstructured visual inputs such as gameplay video, particularly within the Pokémon franchise, where the interface presents multiple simultaneous visual elements, including Pokémon sprites, health bars, names, and move animations, that overlap and change dynamically each turn. While these earlier works relied on structured or simplified data sources (Hu et al., 2024; Norström, 2019), this study advances the field by analysing native gameplay footage from Double Battles, which introduces higher visual density and variability.

One of the most prominent commercial implementations of AI in Pokémon is the PBS, developed by The Pokémon Company in partnership with HEROZ and showcased in 2024 during official tournament broadcasts (The Pokémon Company International, 2024). PBS was used during Single Battle matches to provide real-time statistical overlays for viewers, such as indicators of which side held the advantage and suggestions for optimal moves based on the current battle context. Its primary goal was to enhance the viewing experience and make competitive matches more accessible to audiences unfamiliar with the strategic depth of the game. Nevertheless, PBS was only implemented for Single Battles, leaving the Double Battle format—used in official competitions such as the Pokémon World Championship—without equivalent analytical support (Heroz, 2024; The Pokémon Company, 2024c, 2024d).

Computer vision techniques offer a path forward for addressing this gap. Technologies such as YOLO (Redmon et al., 2016) have been widely adapted beyond object detection to perform image classification tasks, leveraging convolutional architectures trained to differentiate among visual categories (Redmon et al., 2016). This makes YOLO suitable for sprite recognition in games, where clear and repeatable visual patterns can be learned efficiently. Meanwhile, Tesseract OCR has proven effective for extracting text from images and video, including user interfaces and in-game overlays, though its performance depends on resolution, font clarity, and video quality (Smith, 2007; Tesseract OCR, n.d.-b).

The combined use of deep learning–based visual models and OCR has been explored in domains such as automatic number plate recognition, where YOLO variants are used to localise plates and Tesseract OCR (or similar engines) is applied to transcribe the characters. Examples include systems based on YOLOv5 combined with Tesseract OCR for real-time licence plate recognition, and YOLOv8 integrated with OCR techniques for high-precision plate detection and reading (Moussaoui et al., 2024; Thapliyal et al., 2023). These studies demonstrate how visual classification and detection, combined with OCR, can transform raw imagery into structured, machine-readable data—a methodological principle that can be applied.

While several studies have applied YOLO models to video game environments—such as first-person shooter scenarios for real-time character detection (Cheng, 2023) and object recognition using Grand Theft Auto V datasets for computer vision training (Bazan et al., 2024)—none have explored the integration of YOLO as an image classifier with Tesseract OCR to extract structured gameplay information within the Pokémon franchise. This study addresses such gap by testing a pipeline that classifies Pokémon sprites and extracts text-based actions from screen recordings. The extracted data is structured into a JSON-based notation system, providing a foundation for further gameplay analysis and AI-driven tools.

METHODOLOGY

This section outlines the development process of a system designed to extract structured data from Pokémon Double Battle gameplay using computer vision techniques. The methodology covers three main components: a YOLO-based sprite classification model, a Tesseract OCR-based text extractor, and a JSON-based battle notation system. Each component was developed to support the automation of battle interpretation and strategic data representation. The research is exploratory, aiming to validate the technical feasibility of the proposed recognition pipeline rather than produce generalisable performance metrics.

Pokémon Sprite Dataset and Classification Model

To train the image classification model for identifying Pokémon, a custom non-public sprite dataset was constructed based on Pokémon usage data from the 2024 North America Pokémon VGC International Championship (The Pokémon Company, 2024a). Only Pokémon with a usage rate above 1% were included, resulting in 62 unique sprites. This count reflects cases like Urshifu, whose different forms share identical in-game visuals.

Each sprite was sourced from pokemondb.net (Pokémon Database, 2024), resized to 256×256 pixels, and subjected to extensive augmentation to increase variability and simulate real-world conditions during Team Preview. Augmentation techniques applied included:

- Gaussian blur
- motion blur
- affine transformations (±5° rotation and ±5% scaling)
- additive Gaussian noise
- gamma contrast adjustment
- padding and cropping

Each Pokémon class produced 1,000 augmented images, split into 800 training, 100 validation, and 100 test samples, for a total of 62,000 images. Figure 2 illustrates a comparison between a base sprite and one of its augmented versions.

Figure 2

Original and Augmented Sprites of the Pokémon Urshifu

Note. Adapted from "Pokémon Sprite Archive," by Pokémon Database, 2024 (https://pokemondb.net/sprites). Copyright Pokémon Database.

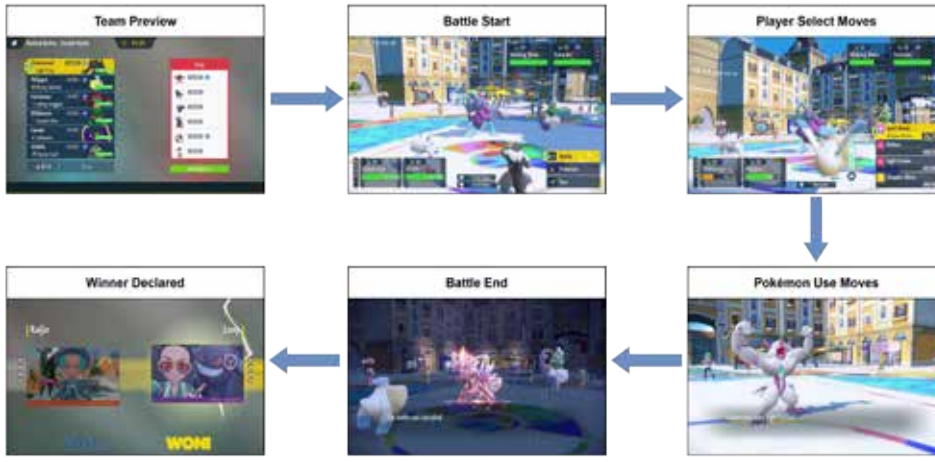
The classification model used for sprite recognition was YOLOv8m, a medium-sized image classifier comprising 103 layers and 15,842,078 parameters, operating at 41.7 GFLOPs (Jocher et al., 2024). Training was carried out using the Ultralytics (Ultralytics, 2024) classification pipeline with default hyperparameters. Once trained, the model was used to classify sprites displayed during the Team Preview phase of each recorded battle. The predicted Pokémon names were recorded in the system's JSON-based battle notation as the opposing team's composition.

Gameplay Footage Collection

To validate the system under real gameplay conditions, footage from eight complete Pokémon Double Battles was recorded using the Nintendo Switch version of *Pokémon Scarlet*. All matches followed the VGC Regulation Set G, the official competitive format used in ranked online play (The Pokémon Company, 2024c). The sample size was defined to balance the diversity of competitive scenarios with the practical constraints of manual frame annotation and model validation. Each recorded battle contained multiple visual events—including move executions and Pokémon switches—providing a sufficiently varied dataset for evaluating classification accuracy and OCR performance within exploratory research context. Figure 3 illustrates the main gameplay phases captured during each recorded Double Battle, which guided the segmentation of the validation process described in the following sections.

Figure 3

Simplified Sequence of Gameplay Phases in Pokémon Scarlet Double Battles Used for System Validation



Note. Gameplay images from the video game *Pokémon Scarlet* on Nintendo Switch.

Video capture was performed using an Elgato 4K X capture card (Elgato, 2024) in combination with OBS Studio, version 30.1.1 (OBS Project, 2024), ensuring consistent image quality across all the recordings. Each video was exported at a resolution of 1920×1080 pixels, 30 frames per second (FPS), and a bitrate of 3000 kbps, providing sufficient visual clarity for both sprite classification and text recognition.

The collected gameplay featured a wide range of battle scenarios, moves, and Pokémon species. Of the 62 Pokémon species used to train the classification model, 31 appeared in the recorded matches, offering a representative and realistic subset for in-situ system evaluation.

Text Extraction and Validation

The system uses Tesseract OCR, version 5.3.3.20231005, with default settings, to extract a wide range of textual information from gameplay frames. This information includes Pokémon levels, move names, Pokémon nameplates during battle, and on-screen timers.

Depending on the current battle state, specific regions of interest are cropped from each frame and fed into Tesseract. For example:

- level information displayed under each Pokémon during Team Preview (e.g., “Lv. 50”)

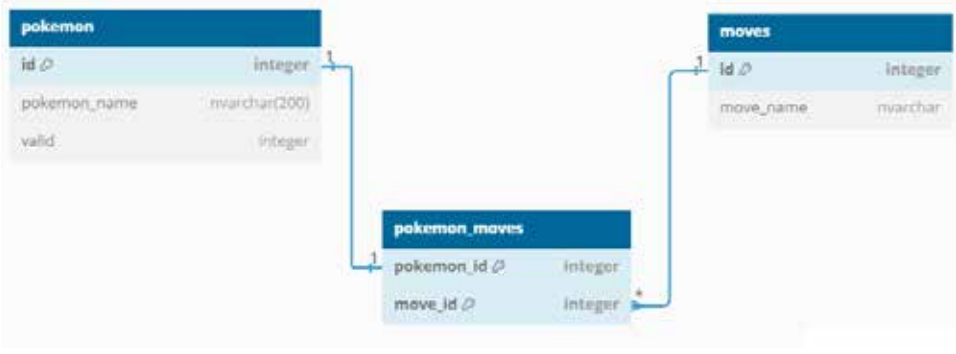
- battle messages indicating move usage or effects (e.g., “Flutter Mane used Shadow Ball!”)
- in-battle name tags and remaining time indicators

To ensure the extracted information is valid, the system cross-references Tesseract’s output with a structured reference database. This database contains lists of all valid Pokémon names and legally allowed move names, enabling the system to discard incorrect or misread entries.

We constructed the database using data extracted from PokéAPI (Hallett et al., 2024), with manual curation to remove duplicates and inconsistencies. The database is hosted on Microsoft Azure and queried in real time during execution. Its relational structure ensures that each move is associated only with Pokémon that can legally learn it, enhancing the precision of battle data (see Figure 4).

Figure 4

Relational Schema of the Pokémon Name and Move Validation Database



JSON-BASED BATTLE NOTATION SYSTEM

Within the field of Pokémon gameplay analysis, a standardised method for representing and processing battle events in structured form has not yet been established. While platforms such as Pokémon Showdown offer replay systems that store battle data, these formats are tailored to their specific environments and are not generalised for broader AI or analytical applications (Luo, 2024). Similarly, existing research efforts often define custom representations that address only narrow experimental needs (Norström, 2019; The Pokémon Company, 2024d).

To enable consistent, machine-readable analysis across diverse gameplay recordings, our system defines a custom JSON-based notation for Double Battles. This notation, originally described by Lladó Herrera (2024), captures both static elements (e.g., teams, levels) and dynamic sequences (e.g., actions, switches, outcomes). The following section outlines the JSON schema implemented by the system:

Root Structure

- player1: information about the first player
- player2: information about the second player
- battle_log: list of objects representing the sequence of turns in the battle
- winner: name of the winning player

Player Object

Each player object (player1 and player2) contains:

- name: player's name
- team: list of Pokémon objects representing the player's team

Pokémon Object in Team

Each Pokémon object within a team contains:

- name: Pokémon's name
- level: Pokémon's level (e.g., "Lv. 50")

Battle Log Object

Each object in the battle_log list represents a turn in the battle and contains:

- turn: turn number
- player1_active: list of objects representing the active Pokémon for player 1 during this turn
- player2_active: list of objects representing the active Pokémon for player 2 during this turn
- actions: list of objects representing the actions taken by a player during the turn

Active Pokémon Object

Each object in the player1_active and player2_active lists represents the Pokémon currently on the battlefield at the beginning of the turn. Each player can have up to two active Pokémon in these lists. The fields include:

- name: active Pokémon's name, or null if no Pokémon is present in that slot

Action Object

Each object in the actions list represents an action performed during the turn and contains:

- player: player performing the action ("player1" or "player2")
- move: move used during the action; if the move is labelled as "Switch", it indicates a substitution in which a Pokémon is being swapped out for another Pokémon on the team
- source: Pokémon performing the move
- target: object representing the Pokémon entering the battlefield during a substitution

This field is populated only when the move is "Switch"; otherwise, it remains null. When a substitution occurs, the target object contains:

- player: player whose Pokémon is being switched in
- pokemon: name of the Pokémon being switched into the battlefield

System Architecture and Workflow

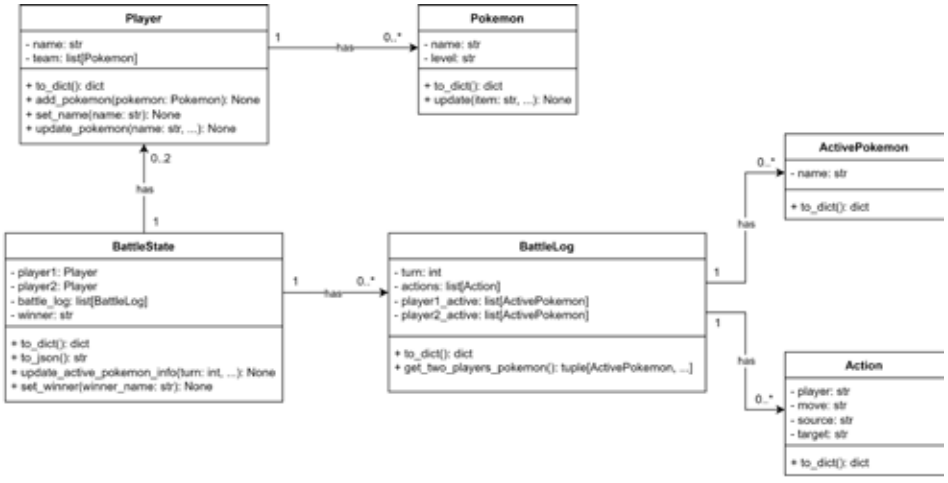
The system integrates three core components: a YOLOv8m image classification model for Pokémon recognition during Team Preview, a Tesseract OCR-based module for extracting text information during gameplay, and a validation database hosted on Azure to ensure the correctness of recognised Pokémon names and moves. These components are coordinated by a state machine architecture, which governs the execution flow according to the battle phase.

Class Design

The system is implemented using an object-oriented structure. Its core logic is encapsulated in the BattleState class, which stores the entire match, including Player instances, the battle log, and the final outcome. Each Player holds a team of Pokémon, while each turn is represented by a BattleLog object, which includes active Pokémon and a list of Action instances. The relationships between these data structures are illustrated in Figure 5.

Figure 5

UML Class Diagram Showing the Relationships Among the Main Data Structures Used in the System



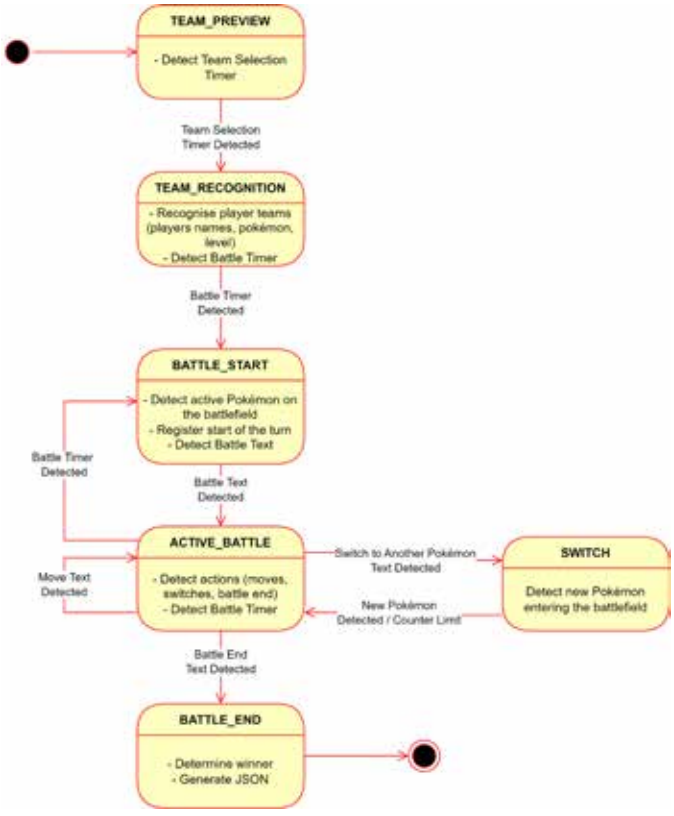
This design provides modularity and clear data separation between phases, enabling easier updates and extensibility for future features, such as ability tracking or item usage.

State-Based Workflow

The system operates as a state machine, transitioning between predefined stages based on visual cues extracted from battle footage. This architecture ensures that relevant components—classification, OCR, or logging—are activated only during the appropriate phase of the match, improving both accuracy and efficiency. An overview of this process is illustrated in Figure 6.

Figure 6

General State Machine Workflow for Pokémon Battle Visual Recognition System



Video was sampled at two frames per second (one frame every fifteen from the native 30 FPS). This rate was selected to significantly reduce computational load and storage requirements while maintaining sufficient temporal resolution to capture all relevant gameplay transitions, including move executions and Pokémon switches. Preliminary tests showed that higher sampling rates produced redundant frames without improving recognition accuracy, confirming that 2 FPS provided an optimal trade-off between efficiency and reliability.

The key states in the workflow are described below:

- TEAM_PREVIEW: Processing begins when the system identifies the "01:29" timer positioned at the upper centre of the screen. Detection of this element designates the start of Team Preview and advances the workflow to the next state.
- TEAM_RECOGNITION: This state is triggered when "Stand By" messages appear for both players. The system:

  - extracts player names using OCR
  - detects twelve Pokémon sprites (six per player) using the YOLOv8m classifier
  - captures level text (e.g., "Lv. 50") using OCR from designated regions
- BATTLE_START: Triggered when Pokémon appear on the battlefield. The system reads four text areas where names are displayed using Tesseract OCR. Valid names are confirmed against the Azure database, while unrecognised names remain null until identified in later frames. The state ends when text indicating action selection is detected.
- ACTIVE_BATTLE: Once the battle begins, the system monitors a mid-screen region for gameplay messages. It applies four OCR configurations (two-page segmentation modes with and without median blur) to maximise text capture. During this phase, the system:
  - logs move using patterns like "used [Move]"
  - recognises switches (e.g., "come back!", "withdrew")
  - detects end-of-battle messages (e.g., "You defeated", "You lost")

Redundant actions are filtered to avoid duplicates. A new turn begins when the timer is detected again, shifting the state back to BATTLE_START.

- SWITCH: Activated when a switch is detected in the previous state. Pokémon sent out are identified through phrases like "Go! [Pokémon]" or "[Player] sent out [Pokémon]!", depending on the side. Names are validated against the database. If no valid name is detected within 200 frames, the system times out and returns to the ACTIVE_BATTLE state.
- BATTLE_END: When final messages appear, the system checks for the terms "WON!" or "LOST" at the lower left of the screen to determine the winner. This outcome is logged, and the final JSON file is generated, marking the end of processing.

RESULTS AND DISCUSSION

System Configuration

The analysis was conducted using an ASUS Zenbook UM425UA (ASUSTeK Computer, n.d.), featuring an AMD Ryzen 5700U processor, 16 GB of RAM, and a 512 GB SSD. System execution was performed locally on this laptop environment. The analysis was conducted using an ASUS Zenbook UM425UA (ASUSTeK Computer, n.d.), featuring an

AMD Ryzen 5700U processor, 16GB of RAM, and a 512GB SSD. System execution was carried out locally on this laptop environment.

The evaluation set consisted of eight recorded matches in 1080p at 30 FPS with a bitrate of 3000 kbps. Processing efficiency was achieved by selecting a frame interval of 15, corresponding to two analysed frames per second. The evaluation set consisted of eight recorded matches in 1080p at 30 fps with a bitrate of 3000 kbps. Processing efficiency was achieved by selecting a frame interval of 15, equating to two analysed frames per second.

Testing Methodology

This section outlines the validation procedures applied to both the YOLO classification model and the complete Pokémon battle analysis system. The evaluation results should be interpreted within an exploratory framework, as the primary objective was to test the system's accuracy and robustness under controlled gameplay conditions rather than to generalise findings across all competitive scenarios. Two levels of evaluation were conducted: offline model testing and end-to-end system validation using gameplay recordings.

The YOLO model was tested using a synthetic dataset generated through data augmentation, as described in Section 3.1. Due to the limited availability of original sprites—only one per Pokémon—this approach allowed the creation of varied, labelled data suitable for assessing classification accuracy under controlled conditions.

In parallel, the full system was evaluated using eight recorded Pokémon battles, simulating real-world usage scenarios. The system's output, stored in JSON format, was compared against manually labelled ground truth extracted from the videos. These comparisons were used to verify the accuracy of sprite recognition, move extraction, active Pokémon tracking, and battle outcome detection.

Together, these tests assessed both the individual components and overall reliability of the system when applied to practical gameplay footage.

YOLO Model Testing Results

The YOLO v8m classifier was evaluated on the augmented Pokémon sprite dataset to assess its baseline classification capability before gameplay testing. As previously detailed in Lladó Herrera (2024), the model achieved perfect performance across all metrics, obtaining mean precision, recall, and F1-scores of 1.00 on every evaluated class. This indicates that the network successfully learned the sprite features introduced through augmentation—such as rotation, blur, and colour variation—without any class imbalance or misclassification.

While these metrics confirm excellent classification performance within the controlled dataset, they also reflect the limited scope of this baseline experiment. Although the training and validation sets were independent and contained no overlapping sprites, both were derived from the same source collection prior to augmentation. Consequently, the perfect results likely reflect the model's ability to memorise augmented visual patterns rather than its performance on entirely unseen data. This evaluation was conducted solely to verify the stability of the classification pipeline and its feature extraction process under ideal conditions. Future research could address stricter dataset independence through pre-split augmentation and the incorporation of an external validation set.

These results confirm the model's robustness for static sprite recognition under controlled conditions. However, such performance reflects the simplified nature of the dataset rather than the complexity of real gameplay. Therefore, the subsequent validation stage using battle footage was conducted to examine model reliability under dynamic conditions involving motion blur, overlapping entities, and changing visual contexts.

Full System Validation

This section presents the validation results of the Pokémon battle analysis system, with tables summarising the accuracy and performance of its individual components.

Table 1 summarises the main characteristics of the gameplay recordings used for system validation, including video duration, processing time, and number of turns. This information highlights both the diversity of the dataset and the computational workload associated with each match.

Table 1

Characteristics of the Recorded Gameplay Videos

Battle video	Video length (minutes)	Number of turns	Process time (minutes)
1	13:08	9	53:45
2	09:51	6	36:16
3	06:24	4	23:51
4	09:41	7	40:34
5	09:07	5	35:25
6	10:09	7	37:31
7	12:01	8	46:38
8	15:17	11	69:23

Note. Adapted from Lladó Herrera, M. R. (2024). *Computer Vision in Gaming: Analysing Pokémon Battles* [Unpublished master's dissertation]. The University of Manchester.

Team Preview Validation

The evaluation of the Team Preview stage considered three primary tasks—player name extraction, sprite identification, and level recognition—as measures of system accuracy.

Player names were correctly identified in 15 out of 16 cases across the eight gameplay videos, yielding an average accuracy of 93.75 %. The single error was recorded in video 4, where only one of the two names was accurately recognised.

For Pokémon sprite recognition, the system correctly identified an average of 11.63 out of 12 sprites per video, yielding an overall accuracy of 96.88 %. In five videos, all sprites were correctly recognised, whereas the remaining three each contained a single misclassification, demonstrating consistent detection performance within the evaluated dataset. However, the model showed reduced accuracy for specific Pokémon, such as Miraidon and Ursaluna-Bloodmoon, highlighting the difficulty of recognising certain species and identifying areas for further optimisation in future iterations.

The Pokémon level detection task demonstrated the highest reliability, achieving an average accuracy of 98.96 %. In seven of the eight videos, all level indicators were accurately captured, with a single error occurring in video 4.

Battle Validation

This section presents the system's performance during the battle phase, including the validation of active Pokémon tracking, action recognition, and winner detection.

Active Pokémon tracking was measured by assessing the system's ability to correctly identify the four positions on the battlefield during each turn—two for each player. The system was designed to recognise whether each position was occupied by a valid Pokémon name or intentionally left empty (null). Across all eight videos, a total of 228 positions were evaluated. The system achieved an average accuracy of 99.22 %, with six of the eight videos attaining 100% accuracy. Single misidentifications occurred in videos 3 and 5, each slightly reducing the respective video's score.

Action validation was divided into two stages. The first focused on confirming that each action was correctly attributed to the appropriate player and that the corresponding move was accurately extracted. Out of a total of 185 total actions, the system achieved an average accuracy of 96.22 % for both player and move detection. Most videos achieved full accuracy, with minor declines observed in videos 1, 5, 7, and especially video 8, which recorded the highest number of incorrect recognitions.

The second stage evaluated the system's ability to identify the source Pokémon (the one executing the move) and, in the case of switch actions, the target Pokémon (the one entering the battlefield). Across the same set of 185 actions, source identification achieved an average accuracy of 96.22 %. Target identification, applicable only to switch

events, was evaluated across 18 relevant cases and achieved an average accuracy of 91.67 %, with perfect accuracy observed in five of the eight videos. Accuracy dropped notably in video 7 to 50 %, due to one incorrect recognition out of two attempts.

Finally, winner detection was evaluated by comparing the system's reported outcome with the actual result of each battle. The system correctly identified the winner in seven of the eight videos, achieving an overall accuracy of 87.50 %. The only failure occurred in video 5, where the system did not detect the end of the battle and, consequently, could not assign a winner.

Summary of Results

Table 2 provides an overview of the consolidated performance metrics for each evaluated stage.

Table 2

Summary of System Validation Performance

Phase	Validation task	Accuracy (%)	Evaluation basis
Team preview	Player name recognition	93.75	2 names per match
	Pokémon sprite recognition	96.88	12 sprites per match
	Pokémon level recognition	98.96	12 level tags per match
Battle	Active Pokémon recognition	99.22	228 positions total
	Action recognition (Player + Move)	96.22	185 actions total
	Source Pokémon recognition	96.22	185 actions total
	Target Pokémon recognition	91.67	18 switch events
	Winner recognition	87.50	8 matches

Note. Adapted from Lladó Herrera, M. R. (2024). *Computer Vision in Gaming: Analysing Pokémon Battles* [Unpublished master's dissertation]. The University of Manchester.

Discussion

The results show that a vision-based pipeline can reliably extract structured information from native Pokémon Double Battle footage. In the controlled sprite dataset, the YOLOv8m classifier achieved perfect precision, recall, and F1. In real gameplay, accuracy remained high for tasks tied to stable HUD regions: team-preview sprite identification (96.88 %), player-name recognition (93.75 %), and level recognition (98.96 %). This is consistent with guidance in the Tesseract documentation, which indicates that OCR performs best on clean, high-contrast interface text at adequate resolution (Tesseract OCR, n.d.-a). During battle, the state-based design supported robust active-Pokémon tracking (99.22 %) and action recognition (96.22 %). The most challenging component

was target identification during switch events (91.67 %), where rapid scene changes and transient overlays increase contextual ambiguity for text extraction. Class-specific variability observed in Team Preview (e.g., lower rates for certain species) indicates opportunities for targeted augmentation or post-processing to improve robustness in visually similar cases.

Relative to prior work—where most Pokémon AI research has relied on structured inputs such as simulators or battle logs (Hu et al., 2024; Norström, 2019; Simoes et al., 2020)—this study operates directly on native broadcast-style video and demonstrates that comparable reliability can be obtained in stable interface states. At the same time, the drop in accuracy for fast transitions and occluded elements underscores the importance of segmenting gameplay into states before applying computer-vision modules and suggests that further refinements should prioritise these dynamic segments.

Consistent with the paper's exploratory scope, these findings are intended to demonstrate technical feasibility rather than statistical generalisation. The combined use of a YOLO-based image classifier for sprite recognition and Tesseract OCR for textual elements provides a practical basis for automated, minimally supervised data extraction from Double Battles, enabling downstream analytics and tooling for competitive contexts.

FUTURE WORK AND CONCLUSIONS

Future Work

Several avenues exist for expanding and refining the current system:

- Dataset Expansion and Event Coverage. The current system was trained on a curated dataset comprising the most frequently used competitive Pokémon. Expanding this dataset to include the full in-game roster could enhance the model's generalisability. Furthermore, the system could be extended to recognise more complex gameplay events, such as move direction or inferred targeting. Enhancing event detection would increase the richness of the generated data and strengthen the analytical value of the JSON notation system.
- Real-Time System Integration. The current system operates on pre-recorded gameplay, making it well suited for post-match analysis. However, adapting it for real-time use could enable dynamic applications such as live coaching, tactical support, or spectator-oriented tools. Achieving this would require improvements in frame processing and inference time to ensure system responsiveness. Real-time integration could also support automated match commentary systems, offering context-aware narration during live events.

- Training AI Agents Using Real Game Data. Existing research in Pokémon AI has largely relied on simulated data from platforms like Pokémon Showdown (Norström, 2019; Simoes et al., 2020). In contrast, the structured outputs produced by the present system provide a pathway for training AI agents using actual gameplay footage. This approach introduces greater variability and more accurately reflects in-game dynamics, although it necessitates addressing potential challenges, such as visual overlays or screen artefacts in streamed content.

CONCLUSIONS

This research introduces a computer vision–based system for analysing Pokémon Double Battles. Gameplay video was transformed into structured representations, providing a clear link between visual elements and usable analytical data. Experimental results indicated reliable performance in player name recognition, sprite classification, and level detection, confirming the system's capability to interpret game information accurately and support applications in strategy evaluation and post-match review.

Beyond its immediate performance, the system contributes a broader foundation for structured game analysis. The development of a custom dataset and a JSON-based notation framework enable detailed tracking and documentation of gameplay events, offering tools for both strategic post-game review and future applications in AI-driven analysis. While overall performance remained strong across various validation stages, certain challenges were identified. These included occasional misrecognitions or interface inconsistencies that impacted isolated predictions. These outcomes emphasise the importance of continued refinement to increase the adaptability of the approach to a broader range of gameplay contexts.

Ultimately, the project not only met its original technical goals but also lays the groundwork for further advancements in the application of computer vision and AI within interactive gaming environments.

DISCLAIMER

The images used in this paper are property of Nintendo, Game Freak, and Creatures Inc. They are included solely for academic and analytical purposes under the permitted use exceptions established in Legislative Decree No. 822 (Peruvian Copyright Law).

REFERENCES

ASUSTeK Computer. (n.d.). *Zenbook 14 (UM425UA)*. Retrieved August 27, 2024, from https://www.asus.com/laptops/for-home/zenbook/zenbook-14-um425-ua/

Bazan, D., Casanova, R., & Ugarte, W. (2024). *Use of custom videogame dataset and YOLO model for accurate handgun detection in real-time video security applications* [Manuscript in preparation]. Universidad Peruana de Ciencias Aplicadas.

Cheng, Y. (2023). Character detection in first person shooter game scenes using YOLO-v5 and YOLO-v7 networks. In *2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)* (pp. 825-831). IEEE. https://doi.org/10.1109/icdacai59742.2023.00160

El-Nasr, M. S., Drachen, A., & Canossa, A. (2016). *Game analytics*. Springer.

Elgato. (2024). *4K X*. Retrieved August 19, 2024, from https://www.elgato.com/uk/en/p/game-capture-4k-x

Hallett, P., Adickes, Z., Marttinen, C., Vohra, S., & Pezzè, A. (2024). *PokéAPI*. Retrieved August 25, 2024, from https://github.com/PokeAPI/pokeapi

HEROZ, I. (2024, February 9). HEROZ、将棋AIの技術を生かし、ポケモンバトルに特化したゲーム演出AI 「Pokémon Battle Scope」を株式会社ポケモンと共同開発 「ポケモン竜王戦2024」ゲーム部門の配信画面に初導入 [HEROZ, leveraging its Shogi AI technology, has jointly developed with The Pokémon Company a game presentation AI specialized for Pokémon battles called 'Pokémon Battle Scope,' which is being introduced for the first time on the broadcast screen of the game division at the 'Pokémon Ryuo Championship 2024]. https://heroz.co.jp/release/2024/02/09_press01-3/

Hu, S., Huang, T., & Liu, L. (2024). *PokéLLMon: A human-parity agent for Pokémon battles with large language models*. *arXiv preprint arXiv:2402.01118*. https://doi.org/10.48550/arXiv.2402.01118

Jocher, G., Chaurasia, A., & Qiu, J. (2024). *Ultralytics YOLO* (Version 8.2.82) [Software]. Ultralytics. Retrieved from https://github.com/ultralytics/ultralytics

Lladó Herrera, M. R. (2024). *Computer vision in gaming: Analysing Pokémon battles* [Unpublished master's dissertation]. The University of Manchester.

Luo, G. (2024). *Pokémon Showdown* [Computer software]. Retrieved August 25, 2024, from https://github.com/smogon/Pokemon-Showdown

Moussaoui, H., Akkad, N. E., Benslimane, M., El-Shafai, W., Baihan, A., Hewage, C., & Rathore, R. S. (2024). Enhancing automated vehicle identification by integrating YOLO v8 and OCR techniques for high-precision license plate detection and recognition. *Scientific Reports*, *14*(1), 14389. https://doi.org/10.1038/s41598-024-65272-1

Norström, L. (2019). *Comparison of artificial intelligence algorithms for Pokémon battles* [Master's thesis, Chalmers University of Technology]. Chalmers Open Digital Repository. https://hdl.handle.net/20.500.12380/300015

OBS Project. (2024). *OBS Studio*. Retrieved August 19, 2024, from https://obsproject.com/

Pokémon Database. (2024). *Pokémon sprite archive* [Database]. Retrieved August 24, 2024, from https://pokemondb.net/sprites

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788. https://doi.org/10.1109/cvpr.2016.91

Simoes, D., Reis, S., Lau, N., & Reis, L. P. (2020). Competitive deep reinforcement learning over a Pokémon battling simulator. In *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE.

Smith, R. (2007). An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. IEEE. https://doi.org/10.1109/ICDAR.2007.4376991

Takahiro, H., & Tomohiro, T. (n.d.). *Amaze the world by the power of AI*. HEROZ, Inc. Retrieved August 13, 2024, from https://heroz.co.jp/en/company/

Tesseract OCR. (n.d.-a). *Improving the quality of the output*. Retrieved August 16, 2024, from https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html

Tesseract OCR. (n.d.-b). *Tesseract User Manual*. Retrieved August 16, 2024, from https://tesseract-ocr.github.io/tessdoc/

Thapliyal, T., Bhatt, S., Rawat, V., & Maurya, S. (2023). Automatic license plate recognition (ALPR) using YOLOv5 model and Tesseract OCR engine. In *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)*. IEEE.

The Pokémon Company. (2024a). *2024 Pokémon North America International Championships*. Retrieved August 24, 2024, from https://www.pokemon.com/us/play-pokemon/internationals/2024/north-america/about

The Pokémon Company. (2024b). *Pokémon in figures*. Retrieved August 6, 2024, from https://corporate.pokemon.co.jp/en/aboutus/figures/

The Pokémon Company. (2024c). *Video game rules, formats, & penalty guidelines*. Retrieved August 6, 2024, from https://www.pokemon.com/static-assets/content-assets/cms2/pdf/play-pokemon/rules/play-pokemon-vg-rules-formats-and-penalty-guidelines-en.pdf

The Pokémon Company. (2024d). *With AI, now everybody can totally enjoy battle watching*. The Pokémon Company. Retrieved August 5, 2024, from https://corporate.pokemon.co.jp/en/topics/detail/118.html

The Pokémon Company International. (2024). *About us*. The Pokémon Company International. Retrieved August 13, 2024, from https://corporate.pokemon.com/en-us/about/

Ultralytics. (2024). *This is Ultralytics*. Retrieved August 26, 2024, from https://www.ultralytics.com/about

Yannakakis, G. N., & Togelius, J. (2018). *Artificial intelligence and games* (Vol. 2). Springer.

ARTÍCULOS DE REVISIÓN

IMPLEMENTACIÓN DE INTELIGENCIA ARTIFICIAL EN EL ESTADO PERUANO: CATÁLOGO ANALÍTICO DE APLICACIONES (2025)

JOEL EMERSON HUANCAPAZA HILASACA
https://orcid.org/0000-0002-9324-1860
joel.huancapaza@unmsm.edu.pe
Universidad Nacional Mayor de San Marcos

Recibido: 8 de septiembre del 2025 / Aceptado: 26 de octubre del 2025
doi: https://doi.org/10.26439/interfases2025.n022.8263

RESUMEN. Este artículo presenta una primera cartografía y un análisis sistemático de 22 aplicaciones de inteligencia artificial implementadas en 17 entidades del Estado peruano, documentando por primera vez el ecosistema operativo de IA en la Administración pública peruana. El estudio identifica un patrón paradójico: mientras el Perú legisla de manera prolífica en materia de IA —y a la fecha es el único país en América Latina con leyes específicas promulgadas—, las aplicaciones existentes carecen de documentación técnica sobre transparencia algorítmica, auditoría de sesgos y evaluación de impacto en derechos fundamentales. Se revela un enfoque instrumental y neutral donde la eficiencia operativa prevalece sobre la gobernanza responsable. Se sugiere la necesidad de transitar desde un enfoque de innovación menos prescriptivo hacia uno más participativo, lo que incluye una reingeniería institucional basada en buenas prácticas de gestión pública.

PALABRAS CLAVE: inteligencia artificial / gobierno digital / ddministración pública / gobernanza de la IA / Perú

ARTIFICIAL INTELLIGENCE IMPLEMENTATION IN THE PERUVIAN PUBLIC SECTOR: AN ANALYTICAL INVENTORY OF APPLICATIONS (2025)

ABSTRACT. This article presents initial mapping and a systematic analysis of 22 artificial intelligence applications implemented across 17 Peruvian state entities, documenting for the first time the operational AI ecosystem in Peruvian public administration. The study identifies a paradoxical pattern: while Peru legislates prolifically on AI matters—currently standing as the sole Latin American nation with specific AI laws enacted—existing applications lack technical documentation regarding algorithmic transparency, bias auditing, and fundamental rights impact assessment. The research reveals an instrumental and neutral approach wherein operational efficiency supersedes responsible governance.

The article underscores the necessity to transition from a prescriptive innovation framework toward a more participatory paradigm, encompassing institutional reengineering grounded in public sector management best practices and inclusive stakeholder engagement mechanisms.

KEYWORDS: artificial intelligence / digital government / public administration / AI governance / Peru

INTRODUCCIÓN

La inteligencia artificial (IA) ha dejado de ser un concepto futurista para convertirse en una herramienta tangible de transformación de la Administración pública a nivel global. Su potencial para analizar grandes volúmenes de datos, automatizar procesos complejos y ofrecer servicios personalizados la posiciona como un pilar fundamental para la construcción de gobiernos más ágiles, eficientes y cercanos a la ciudadanía (Floridi, 2020; Sandoval-Almazán & Valle-Cruz, 2025). Países líderes han desarrollado marcos estratégicos y normativos para guiar su adopción con el fin de maximizar sus beneficios mientras mitigan riesgos inherentes como la opacidad algorítmica, la discriminación y las vulneraciones a la privacidad (Comisión Europea, 2021; Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [Unesco], 2022).

En América Latina, el camino de incorporación de la IA en el sector público es heterogéneo y presenta notables desafíos de capacidad institucional, infraestructura y regulación (Criado, 2024; Organización para la Cooperación y el Desarrollo Económico, 2022). El Perú no es ajeno a esta tendencia. Recientemente, se ha observado un crecimiento en el número de iniciativas impulsadas por diversas entidades estatales, las cuales, aunque valiosas, suelen ser desarrolladas de manera aislada, por lo que carecen de una visión integral y de estándares comunes que aseguren su interoperabilidad, seguridad y ética (Congreso de la República del Perú, 2023b; Smart & Montori, 2025).

La presente investigación se justifica debido a una particular paradoja observada en el caso peruano: mientras el país exhibe una prolífica actividad legislativa en materia de IA —y es el único país en América Latina con leyes específicas promulgadas (Smart & Montori, 2025)—, carece de evidencia empírica sistematizada sobre las aplicaciones efectivamente desplegadas en el sector público. Este estudio aborda dicha brecha mediante el análisis de un catálogo oficial de aplicaciones que persigue tres objetivos: caracterizar las iniciativas existentes según sus finalidades, tecnologías y sectores; evaluar críticamente su alineación con principios de gobernanza responsable de IA; y contrastar el panorama operativo con el marco normativo emergente para identificar tensiones entre retórica regulatoria y realidad institucional.

METODOLOGÍA

Este estudio adopta un enfoque cualitativo de investigación documental, basado en el análisis de contenido de una fuente primaria oficial: el Catálogo de Aplicaciones con Inteligencia Artificial en el Estado Peruano (Presidencia de Consejo de Ministros [PCM], 2025a). Este documento, de acceso público, lista 22 aplicativos implementados por distintas entidades del Gobierno peruano y proporciona información sobre la entidad responsable, el nombre de la aplicación, su finalidad y las tecnologías de IA utilizadas.

Interfases n.º 22, diciembre 2025

El proceso metodológico se desarrolló en tres fases. Primero, se extrajeron los datos del catálogo y se organizaron en una base de datos estructurada para facilitar el análisis. Segundo, cada aplicación fue categorizada temáticamente y según su sector de aplicación (por ejemplo, salud, justicia, educación), el tipo de tecnología de IA empleada (por ejemplo, PLN, visión por computadora, chatbots) y su finalidad principal (por ejemplo, automatización, soporte a la decisión, servicio al ciudadano). Este proceso permitió identificar patrones y tendencias en la adopción de la IA. Tercero, los hallazgos fueron contrastados con el marco teórico y normativo existente.

Se utilizó la literatura académica internacional y nacional sobre gobernanza de IA (Castilla Barraza & Romero-Rubio, 2025; Ebers, 2022; Floridi, 2020) para evaluar el grado de alineación de las aplicaciones con principios como transparencia, imparcialidad y rendición de cuentas. Si bien el catálogo proporciona una visión actualizada del ecosistema operativo de aplicaciones que usan IA en la Administración pública peruana, esta metodología reconoce una limitación: la dependencia de la información autorreportada por las entidades, especialmente por la PCM (artículo 4 de la Ley 31814), lo que podría omitir detalles técnicos o desafíos de implementación no visibles de manera externa.

RESULTADOS

El panorama de la IA en el Estado peruano

Aunque el Catálogo de Aplicaciones con Inteligencia Artificial en el Estado Peruano (PCM, 2025a) es de acceso público, el autor accedió a dicho documento mediante un procedimiento formal de solicitud de acceso a la información pública, dirigido a la PCM. Esta solicitud fue respondida mediante Oficio D001185-2025-PCM-OPII, de fecha 29 de agosto del 2025, junto con el Memorando D000944-2025-PCM-SGTD de la misma fecha y el Informe D000041-2025-PCM-SGTD-LHM, de fecha 28 de agosto del 2025. De la revisión de dichos instrumentos oficiales se extrajo y sistematizó la información que se presenta en la Tabla 1.

Cabe destacar, además, que el catálogo alojado en el portal web institucional de la PCM ha sido objeto de actualización reciente, lo que confirma la coincidencia y consistencia entre la información obtenida mediante el mecanismo de acceso a la información pública y la disponible públicamente en línea. Esta correspondencia refuerza la fidelidad, validez y actualidad de los datos utilizados en el estudio (Tabla 1).

Tabla 1

Catálogo resumido de aplicaciones de IA en el Estado peruano

Nombre de la aplicación	Tecnologías utilizadas	Nombre de la entidad
Smart Churay Yachay	Inteligencia artificial, recursos *offline*, librerías DeepSheek y QWen, servicios en la nube	Ministerio de Educación
Bolsa de Trabajo	Inteligencia artificial, NLP, emparejamiento inteligente, microservicios web y ontologías ocupacionales	Ministerio de Trabajo y Promoción del Empleo
Servicio de lectura multilingüe de información registral	IA, texto a voz, OCR, traducción automática multilingüe	Superintendencia Nacional de los Registros Públicos
Asistente virtual (chat y audio) para página web y *contact center*	IA conversacional, IBM watsonx Assistant, servicios en la nube (AWS, VPC)	Organismo Supervisor de Inversión Privada en Telecomunicaciones
Chatbot de soporte técnico para los colaboradores del Programa Nacional PAIS	IA, Python, FastAPI, Keras, TensorFlow, SQL Server	Programa Nacional Plataformas de Acción para la Inclusión Social
View Leaf	Python, MobileNet (como modelo de IA), Kotlin y Android Studio (para el desarrollo de la aplicación)	Instituto de Investigaciones de la Amazonía Peruana
Chatbot bibliotecario	Librerías IA	Biblioteca Nacional del Perú
Curia	IA generativa, NLP, análisis jurídico automatizado	Poder Judicial
ElecciIA	Inteligencia artificial, procesamiento jurídico-automatizado, evaluación documental	Jurado Nacional de Elecciones
YachAIbot (en marcha blanca)	NLP, chatbots, IA jurídica	Presidencia del Consejo de Ministros
Qhali (robot enfermera con IA)	Robot humanoide, IA, teleoperación, interacción gestual	Instituto Nacional de Salud del Niño
IA para la elaboración de informes de inspección	IA, *machine learning*, análisis geoespacial, automatización	Superintendencia Nacional de Servicios de Saneamiento
Sigersol	IA, soporte virtual inteligente, integración de datos	Ministerio del Ambiente

(*continúa*)

(*continuación*)

Nombre de la aplicación	Tecnologías utilizadas	Nombre de la entidad
Maia MYPE Asesor IA	Inteligencia artificial, asistente virtual, sugerencias automatizadas de gestión	Ministerio de la Producción
Resonador magnético con IA para diagnóstico oncológico	Inteligencia artificial, imágenes médicas, resonancia magnética avanzada	Instituto Nacional de Enfermedades Neoplásicas
Monitoreo de servidores	Python	Ministerio del Interior
Biofacial	Reconocimiento facial, *deep learning*	Registro Nacional de Identificación y Estado Civil
ADETOP v2	*Machine learning* (*random forest*), imágenes satelitales (Sentinel-2, Landsat), visión por computadora, análisis geoespacial	Organismo de Supervisión de los Recursos Forestales y de Fauna Silvestre
Sistema de IA CadEye para endoscopía digestiva	Inteligencia artificial, Vi+B2:E23sión computacional, equipos endoscópicos asistidos (Eluxeo 7000 + CadEye)	Hospital Regional de Lambayeque
Proactiva	-	Superintendencia Nacional de Migraciones
Chat	-	Ministerio de Comercio Exterior y Turismo
IDE	-	Ministerio de Desarrollo Agrario y Riego

Nota. Adaptado del "Catálogo de aplicaciones con inteligencia artificial en el Estado peruano", por la PCM, 2025a, Consejo Directivo Nacional (https://cdn.www.gob.pe/uploads/document/file/8238351/6879780-catalogo-de-aplicaciones-con-ia.pdf?v=1752856329).

El análisis del catálogo revela un ecosistema incipiente pero diverso, con 22 aplicaciones desplegadas en 17 entidades distintas del Estado peruano. A continuación, se presentan los resultados del análisis agrupados por áreas temáticas de impacto.

Mejora de servicios al ciudadano y acceso a la información

Un bloque significativo de aplicaciones se orienta a la interacción directa con la ciudadanía y la facilitación del acceso a la información. Destacan asistentes virtuales y chatbots, como el del Organismo Supervisor de Inversión Privada en Telecomunicaciones (Osiptel) y el chatbot bibliotecario de la Biblioteca Nacional del Perú (BNP), diseñados para ampliar y agilizar los canales de atención. Un caso notable de inclusión es el servicio de lectura multilingüe de la Superintendencia Nacional de los Registros Públicos (Sunarp), que emplea síntesis de voz y traducción automática para hacer la información registral accesible en lenguas nativas y extranjeras. Asimismo, la plataforma Bolsa de Trabajo del Ministerio de Trabajo y Promoción del Empleo del Perú (MTPE) utiliza IA y NLP para un emparejamiento inteligente entre demandantes y oferentes de empleo, lo que optimiza la intermediación laboral.

Optimización de procesos internos y soporte a la decisión

Otra línea de aplicación se centra en la eficiencia interna y el apoyo a la toma de decisiones técnicas. El Poder Judicial presenta Curia, un asistente de IA generativa que proyecta resoluciones judiciales con referencias normativas. El Jurado Nacional de Elecciones (JNE) cuenta con EleccIA para la evaluación de expedientes electorales, mientras que la Superintendencia Nacional de Servicios de Saneamiento (Sunass) emplea la IA para automatizar la elaboración de informes de inspección con análisis geoespacial. Aplicaciones como el chatbot de soporte técnico del Programa Nacional PAIS y el sistema de monitoreo de servidores del Ministerio del Interior ejemplifican el uso de la IA para sostener la operatividad institucional.

Aplicaciones especializadas en sectores estratégicos

El sector salud muestra implementaciones sofisticadas. El Instituto Nacional de Enfermedades Neoplásicas (INEN) utiliza un resonador magnético con IA para optimizar el diagnóstico oncológico, y el Hospital Regional de Lambayeque implementa el sistema CadEye para el diagnóstico precoz de cáncer digestivo mediante visión por computadora. Incluso, se registra el uso de un robot humanoide (Ohail) en el Instituto Nacional de Salud del Niño (INSN) para orientar al público sobre vacunación. En el ámbito ambiental, el Instituto de Investigaciones de la Amazonía Peruana (IIAP) desarrolló View Leaf para la identificación de especies forestales mediante IA, y el Organismo de Supervisión de los Recursos Forestales y de Fauna Silvestre (Osinfor) utiliza ADETOP v2 con *machine learning* e imágenes satelitales para detectar tala ilegal. El Ministerio del

Ambiente (Minam), por su parte, despliega Sigersol para guiar la gestión de residuos sólidos.

Tecnologías de IA prevalentes y falta de documentación técnica

El análisis del catálogo oficial de 22 aplicaciones con IA implementadas en el Estado peruano revela un predominio pronunciado de tecnologías de procesamiento de lenguaje natural (NLP) y chatbots conversacionales, principalmente impulsados por la necesidad de interacción usuario-sistema en contextos de servicio al ciudadano. Sin embargo, un análisis más profundo evidencia una deficiencia técnica legal: ninguna de las aplicaciones publicadas en el Catálogo de Aplicaciones con Inteligencia Artificial en el Estado Peruano incluye documentación técnica detallada sobre los algoritmos subyacentes, criterios específicos de decisión automatizada, análisis de precisión diferenciada por grupos demográficos, ni evaluaciones prospectivas de impacto en derechos fundamentales (PCM, 2025a).

Para ilustrar este punto, en la Tabla 2 se presenta un catálogo temático de aplicaciones de IA.

Tabla 2

Catálogo temático de aplicaciones de IA en el Estado peruano

Sector de impacto	N.º de aplicaciones	Ejemplos representativos	Tecnologías de IA prevalentes
Servicio al ciudadano	5	Bolsa de Trabajo (MTPE), lectura multilingüe (Sunarp), asistente virtual (Osiptel)	NLP, chatbots, traducción automática
Justicia y gestión pública	5	CURIA (Poder Judicial), ElecciA (JNE), IA para informes (Sunass)	IA generativa, NLP, automatización
Salud	3	Resonador con IA (INEN), CadEye (H. Lambayeque), Ohail (INSN)	Visión por computadora, IA en imágenes, robótica
Medio ambiente y recursos	3	View Leaf (IIAP), ADETOP v2 (Osinfor), Sigersol (Minam)	Visión por computadora, ML, análisis geoespacial
Soporte interno y otros	6	Chatbot Soporte (PAIS), Monitoreo (Mininter), Maia (Produce)	NLP, Python, API

Nota. Adaptado del "Catálogo de aplicaciones con inteligencia artificial en el Estado peruano", por la PCM, 2025a, Consejo Directivo Nacional (https://cdn.www.gob.pe/uploads/document/file/8238351/6879780-catalogo-de- aplicaciones-con-ia.pdf?v=1752856329).

Este vacío de información técnica contrasta drásticamente con lo exigido por los marcos de gobernanza internacional y por el propio Decreto Supremo 115-2025-PCM, que aprueba el Reglamento de la Ley 31814, que establece como obligación para sistemas de riesgo alto la implementación de mecanismos que garanticen la transparencia

algorítmica y permitan informar al usuario, de forma previa, clara y sencilla, sobre la finalidad, funcionalidades principales y tipo de decisiones que puede tomar el sistema (Artículo 25.1). Por otra parte, respecto de la exploración web realizada, las entidades titulares de los aplicativos tampoco han transparentado la mencionada documentación técnica. La ausencia de tales archivos representa un obstáculo fundamental no solo para la evaluación independiente de sesgos, confiabilidad y alineación con principios éticos, sino también para la imparcialidad, la precisión y la facilidad de uso de datos (Gong et al., 2023).

DISCUSIÓN DE LOS RESULTADOS

Necesidad de una profunda visión regulatoria

La proliferación legislativa en materia de IA observada en el Perú representa un fenómeno único en América Latina. Mientras que países como Argentina, Brasil y México avanzan mediante iniciativas legislativas fragmentadas, el Perú es el único que ha promulgado leyes específicas: cuenta con dos leyes sustantivas (Ley 31814 y Ley 32082) y existen, además, diecisiete proyectos de ley en trámite legislativo (Congreso de la República del Perú, 2022, 2023a, 2023b, 2023c; Smart & Montori, 2025). Sin embargo, este auge regulatorio, que inclusive cuenta con una norma técnica peruana, la NTP-ISO/IEC 42001:2025 (Instituto Nacional de Calidad, 2025), contrasta dramáticamente con la ausencia de una arquitectura institucional robusta que traduzca esos enunciados normativos en garantías exigibles. Como señala la literatura especializada, "legislar el lenguaje de los derechos sin construir la arquitectura institucional para garantizarlos" constituye el achaque fundamental del enfoque peruano (Smart & Montori, 2025). La presente investigación, que cataloga 22 aplicaciones desplegadas en el Estado peruano, evidencia que esta producción legislativa no ha generado un marco coherente que asegure la interoperabilidad, la transparencia algorítmica ni la gobernanza participativa de la ciudadanía en estas iniciativas tecnológicas.

El Decreto Supremo 115-2025-PCM busca poner orden y llenar este vacío; sin embargo, no contempla mecanismos de fiscalización independientes, organismos de supervisión especializados y sistemas de auditoría sólidos, por lo que sigue siendo una aspiración normativa desvinculada de la realidad operativa y, además, pasible de ser modificada a tono con el gobierno de turno. Las aplicaciones catalogadas —desde Curia en el Poder Judicial hasta EleccIA en el Jurado Nacional de Elecciones— funcionan bajo criterios técnicos heterogéneos, sin someterse a evaluaciones de impacto de derechos fundamentales ni a protocolos uniformes de transparencia algorítmica que garanticen su conformidad con los principios de no discriminación, equidad procesal y rendición de cuentas que proclaman las leyes (Gomes Rêgo & Dos Santos Júnior, 2025, p. 5). El Perú necesita transitar desde un modelo de innovación por decreto hacia uno de innovación

con gobernanza, en un contexto en el que actores internacionales observan la gobernanza de la IA en América Latina como un espacio progresivo en el que la tecnología se desarrolla dentro de marcos de supervisión construidos participativamente, con infraestructura y validación de múltiples actores (Organización para la Cooperación y el Desarrollo Económico, 2022, p. 167).

Enfoque instrumental, neutral y operacional

El análisis temático de las 22 aplicaciones revela un patrón dominante: la IA es concebida como una herramienta neutra orientada primordialmente hacia la eficiencia operacional, la modernización estatal y la productividad tecnológica, marginalizando interrogantes cruciales sobre quién está protegido, quién es excluido y quién decide. Tal enfoque puede caracterizarse como *instrumental* en el sentido acuñado por la literatura especializada (Castilla Barraza & Romero-Rubio, 2025; Floridi, 2020). Observamos que un 34 % de las aplicaciones (7 de 22) se orientan explícitamente hacia la mejora de servicios internos o la automatización de procesos administrativos (chatbots de soporte técnico, monitoreo de servidores, asistentes virtuales internos), mientras que apenas un 23 % (5 de 22) se enfoca en servicios directos al ciudadano con capacidad transformadora de derechos. La distribución revela que el impulso innovador está dirigido a la optimización de máquinas burocráticas existentes más que a la reconfiguración equitativa de relaciones entre Estado y ciudadanía.

Este sesgo instrumental se percibe también en la forma en que están estructuradas las aplicaciones de alto impacto. Por ejemplo, Curia (Poder Judicial) y EleccIA (JNE), que inciden directamente en derechos políticos y acceso a la justicia, carecen de documentación pública accesible sobre sus criterios de transparencia, auditoría de sesgos o mecanismos de revisión humana (PCM, 2025a). Del mismo modo, el sistema Biofacial del Registro Nacional de Identificación y Estado Civil, que utiliza reconocimiento facial con *deep learning*, no ha publicado evaluaciones de su rendimiento diferenciado según variables de edad, etnia o género —datos críticos para detectar discriminación algorítmica (Brkan & Bonnet, 2020; Nazer et al., 2023)—. Esta ausencia de protocolos de evaluación de impacto evidencia que el enfoque dominante trata a la IA como un instrumento técnico aséptico y relega a segundo plano la exigencia de responsabilidad que caracteriza a los derechos fundamentales (Castilla Barraza & Romero-Rubio, 2025, p. 138).

La ausencia de coordinación institucional

La distribución sectorial de las 22 aplicaciones, presentada en la Tabla 2, revela un panorama fragmentado donde cada entidad desarrolla soluciones de IA según sus propias prioridades operativas sin una visión integral coordinada. Cinco aplicaciones se concentran en justicia y gestión pública (Curia, EleccIA, IA para informes de inspección, YachAIbot, Proactiva), tres en salud (resonador magnético con IA, CadEye, Ohail), tres

en medio ambiente (View Leaf, ADETOP v2, Sigersol), cinco en servicios al ciudadano (Bolsa de Trabajo, lectura multilingüe, asistentes virtuales) y seis en soporte interno. Esta multiplicidad de iniciativas, aunque demostrativamente valiosa en términos de innovación piloto, carece del encuadramiento que proporciona una estrategia nacional coherente de IA que articule objetivos, riesgos compartidos y estándares técnicos uniformes, en consonancia con la Estrategia Nacional de Inteligencia Artificial 2026-2030 (ENIA).

El Decreto Supremo 115-2025-PCM reconoce la necesidad de que diversos actores interactúen para la gobernanza en el desarrollo, implementación y uso de la IA, bajo la dirección de la PCM, en calidad de autoridad técnico-normativa a nivel nacional (artículos 8-10), pero la transferencia de esa aspiración declarativa a procesos vinculantes de coordinación entre entidades sectoriales permanece en fase incipiente. Así, aunque el catálogo demuestra que múltiples organismos públicos adoptan tecnologías de IA, no existe evidencia de evaluaciones de impacto coordinadas, protocolos de interoperabilidad de datos ni mecanismos de vigilancia conjunta respecto a sesgos algorítmicos o vulneraciones de privacidad, ni estándares comunes (Birkstedt et al., 2023). La literatura internacional sobre gobernanza de IA destaca que la ausencia de coordinación institucional estratégica facilita la replicación de riesgos idiosincráticos en cada aplicación y obstaculiza la economía de escala en la adquisición de capacidades técnicas y de supervisión (Gomes Rêgo & Dos Santos Júnior, 2025, p. 8; Mökander et al., 2022, p. 248). El Perú, al carecer de un organismo independiente de supervisión específicamente dedicado a auditar aplicaciones de IA en el sector público, se aproxima peligrosamente a un escenario de innovación sin vigilancia, donde la proliferación de sistemas técnicos avanza desvinculada de marcos de control y rendición de cuentas (Brkan & Bonnet, 2020, p. 32).

Limitaciones del enfoque peruano

El ordenamiento legal peruano declara, mediante múltiples instrumentos, principios éticos fundamentales para la gobernanza de IA: no discriminación, privacidad de datos personales, transparencia, explicabilidad, sostenibilidad, responsabilidad, sensibilización y educación en IA, entre otros. La *Recomendación sobre la ética de la inteligencia artificial* de la Unesco (2022), de la cual el Perú es signatario, postula criterios análogos. Sin embargo, cuando se examina el catálogo de 22 aplicaciones operativas, la distancia entre principio declarado y práctica ejecutada es notoria. Ninguna de las aplicaciones publicadas incluye documentación técnica detallada sobre algoritmos, criterios de decisión automatizada, precisión diferenciada por grupos demográficos o evaluaciones prospectivas de impacto en derechos fundamentales —elementos que constituyen el núcleo técnico de la transparencia y la responsabilidad (Brkan & Bonnet, 2020; Felzmann et al., 2020; Winfield et al., 2021). Lo propio ocurre con el más reciente aplicativo lanzado por la PCM (2025b): Gobi, el asistente virtual que guía a los ciudadanos en sus trámites desde la plataforma del Gobierno del Perú.

Esta brecha de implementación se acentúa cuando se contrastan la prácticas locales con el estándar internacional que, por ejemplo, el Reglamento (UE) 2024/1689 establece explícitamente: los operadores de sistemas de IA de alto riesgo deben realizar evaluaciones de impacto en derechos fundamentales antes de la puesta en servicio y mantener documentación técnica accesible a autoridades de supervisión (Unión Europea, 2024, artículos 6-8). En el Perú, aplicaciones como Biofacial (reconocimiento facial), CadEye (diagnóstico oncológico) o Curia (resoluciones judiciales) —todas potencialmente de alto riesgo según criterios de la UE— no han publicado tales evaluaciones; tampoco se ha verificado que cuenten con mecanismos independientes de auditoría o que hayan sometido sus algoritmos a pruebas de equidad y sesgos (Castilla Barraza & Romero-Rubio, 2025, p. 140; Nazer et al., 2023, p. 3). En este contexto, consciente de esta carencia, parece viable y necesario desarrollar un protocolo nacional de evaluación de impacto de la IA en derechos humanos, pero su implementación dependerá de una reingeniería institucional que se ha de materializar primero.

CONCLUSIONES

La presente investigación demuestra, en primer lugar, que el Perú experimenta un impulso innovador en la implementación de IA en el sector público, donde se catalogan 22 aplicaciones operativas en 17 entidades estatales. Sin embargo, esta adopción incipiente se produce en el contexto de una paradoja fundamental: el país legisla de manera prolífica —único en América Latina con leyes específicas promulgadas— sin construir, simultáneamente, la arquitectura institucional robusta necesaria para garantizar transparencia, equidad y rendición de cuentas. Así, mientras la Ley 31814 y el Decreto Supremo 115-2025-PCM declaran principios de gobernanza responsable, sus mecanismos de implementación permanecen incipientes.

En segundo lugar, el análisis reveló un enfoque instrumental predominante, donde la IA se trata como herramienta neutra orientada a la eficiencia operacional. Un 34 % de las aplicaciones (7 de 22) se focaliza en automatización interna, mientras que apenas un 23 % (5 de 22) impacta en derechos de la ciudadanía directamente. Aplicaciones críticas como Curia, EleccIA y Biofacial carecen de documentación pública sobre transparencia algorítmica, auditoría de sesgos o evaluaciones de impacto en derechos fundamentales, reproduciendo un patrón donde la técnica se desvincula de la responsabilidad constitucional.

En tercer lugar, la fragmentación sectorial sin coordinación estratégica obstaculiza la consolidación de estándares técnicos uniformes. Aunque múltiples organismos adoptan la IA según prioridades operativas propias, no existe evidencia de evaluaciones coordinadas, protocolos de interoperabilidad ni vigilancia conjunta de riesgos. El Perú carece de un organismo independiente de supervisión dedicado a auditar aplicaciones de IA en el sector público (las auditorías de seguridad se delegan a cada entidad de

la Administración pública, conforme al literal c del artículo 29 del Decreto Supremo 115-2025-PCM), lo que la aproxima peligrosamente a un escenario de innovación sin vigilancia que facilita la replicación de riesgos idiosincráticos.

En cuarto lugar, el contraste con estándares internacionales, particularmente el Reglamento (UE) 2024/1689, evidencia que las aplicaciones de alto riesgo peruanas no cumplen requisitos de evaluación de impacto *ex ante* ni mantienen documentación técnica accesible a autoridades supervisoras. Sistemas como Biofacial, CadEye y Curia no han publicado pruebas de equidad diferenciada por variables sensibles (edad, etnia, género) ni auditorías independientes.

En último lugar, se sugiere que futuras investigaciones aborden la necesidad de transitar desde un modelo de innovación por decreto hacia uno de innovación con gobernanza participativa. En esa línea de análisis, deben surgir propuestas concretas sobre la necesidad de una autoridad supervisora independiente con poder de auditoría y sanción, y propuestas interdisciplinarias sobre (a) el desarrollo del protocolo nacional de evaluación de impacto en derechos humanos con estándares computacionales de equidad y transparencia, que puedan ser comprendidos por actores no técnicos; y (b) metodologías de ingeniería participativa que integren decisión humana y vías de contestación algorítmica en sistemas de alto impacto.

REFERENCIAS

Birkstedt, T., Minkkinen M., Tandon A., & Mäntymäki M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, *33*(7), 133-167. https://doi.org/10.1108/intr-01-2022-0042

Brkan, M., & Bonnet, G. (2020). Legal and technical feasibility of the GDPR's quest for explanation of algorithmic decisions: Of black boxes, white boxes and fata morganas. *European Journal of Risk Regulation*, *11*(1), 18-50. https://doi.org/10.1017/err.2020.10

Castilla Barraza, J. G., & Romero-Rubio, S. A. (2025). The use of artificial intelligence in governance: A systematic review. En T. Guarda, F. Portela & M. F. Augusto (Eds.), *Advanced Research in Technologies, Information, Innovation and Sustainability* (pp. 110-125). Springer. https://doi.org/10.1007/978-3-031-83435-6_9

Comisión Europea. (2021). *Coordinated plan on artificial intelligence 2021 review*. Estrategia Digital. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

Congreso de la República del Perú. (2022). *Proyecto de Ley N.º 05182/2022-CR: Ley que promueve el uso de la inteligencia artificial en el sistema de transporte terrestre del país*. https://wb2server.congreso.gob.pe/spley-portal/#/expediente/2021/5182

Congreso de la República del Perú. (2023a). *Proyecto de Ley N.º 07444/2023-CR: Ley de fomento y regulación integral de la inteligencia artificial en la Administración pública para un Perú digital y sostenible*. https://wb2server.congreso.gob.pe/spley-portal/#/expediente/2021/7444

Congreso de la República del Perú. (2023b). *Proyecto de Ley N.º 07651/2023-CR: Ley que regula el uso de algoritmos y técnicas de inteligencia artificial para el reconocimiento de placas de rodaje de medios de transporte*. https://wb2server.congreso.gob.pe/spley-portal/#/expediente/2021/7651

Congreso de la República del Perú. (2023c). *Proyecto de Ley N.º 08223/2023-CR: Ley de fomento y regulación del uso de la inteligencia artificial en el Perú*. https://wb2server.congreso.gob.pe/spley-portal/#/expediente/2021/8223

Criado, J. I. (2024). Inteligencia artificial en el sector público latinoamericano. Estudio Comparado a partir de la Carta Iberoamericana de Inteligencia Artificial en la Administración Pública. *Revista CLAD Reforma y Democracia*, *88*, 116-143. https://doi.org/10.69733/clad.ryd.n88.a387

Ebers, M. (2022). Standardizing AI: The case of the European Commission's proposal for an artificial intelligence act. En L. A. DiMatteo, C. Poncibò & M. Cannarsa (Eds.), *The Cambridge handbook of artificial intelligence: Global perspectives on law and ethics* (pp. 321-344). Cambridge University Press. https://doi.org/10.1017/9781009072168.030

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*, 3333-3361. https://doi.org/10.1007/s11948-020-00276-4

Floridi, L. (2020). Artificial intelligence as a public service: learning from Amsterdam and Helsinki. *Philosophy & Technology*, *33*, 541-546. https://doi.org/10.1007/s13347-020-00434-3

Gomes Rêgo, P., & Dos Santos Júnior, C. D. (2025). Artificial intelligence governance: understanding how public organizations implement it. *Government Information Quarterly*, *42*(1), Artículo 102003. https://doi.org/10.1016/j.giq.2024.102003

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, *162*, Artículo 107268. https://doi.org/10.1016/j.infsof.2023.107268

Instituto Nacional de Calidad. (2025, 9 de julio). *Inacal aprueba la primera norma técnica peruana sobre sistemas de gestión de inteligencia artificial*. Gobierno del Perú. https://www.gob.pe/institucion/inacal/noticias/1205864-inacal-aprueba-la-primera-norma-tecnica-peruana-sobre-sistemas-de-gestion-de-inteligencia-artificial

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, *32*, 241-268. https://doi.org/10.1007/s11023-021-09577-4

Nazer, L. H., Zatarah, R., Waldrip, S., Chen Ke, J. X., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, *2*(6), e0000278. https://doi.org/10.1371/journal.pdig.0000278

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2022). *Recomendación sobre la ética de la inteligencia artificial*. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

Organización para la Cooperación y el Desarrollo Económico. (2022). *The strategic and responsible use of artificial intelligence in the public sector of Latin America and the Caribbean*. Organización para la Cooperación y el Desarrollo Económico; Banco de Desarrollo de América Latina y el Caribe. https://doi.org/10.1787/1f334543-en

Presidencia del Consejo de Ministros. (2025a, 19 de junio). *Catálogo de aplicaciones con inteligencia artificial en el Estado peruano*. Consejo Directivo Nacional. https://www.gob.pe/institucion/pcm/informes-publicaciones/6879780-catalogo-de-aplicaciones-con-inteligencia-artificial-en-el-estado-peruano

Presidencia del Consejo de Ministros. (2025b, 2 de octubre). *PCM lanza Gobi, el asistente virtual que guía a los ciudadanos en sus trámites desde gob.pe*. Gobierno del Perú. https://www.gob.pe/institucion/pcm/noticias/1257000-pcm-lanza-gobi-el-asistente-virtual-que-guia-a-los-ciudadanos-en-sus-tramites-desde-gob-pe

Sandoval-Almazán, R., & Valle-Cruz, D. (2025). *Artificial intelligence in goverment*. Springer.

Smart, S., & Montori, V. M. (2025, 23 de abril). *Peru's AI regulatory boom: Quantity without depth?* Carr-Ryan Center for Human Rights. https://www.hks.harvard.edu/centers/carr-ryan/our-work/carr-ryan-commentary/perus-ai-regulatory-boom-quantity-without-depth

Unión Europea. (2024, 13 de junio). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los reglamentos (CE) n.o 300/2008, (UE) n.o 167/2013, (UE) n.o 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial). *Diario Oficial de la Unión Europea*. http://data.europa.eu/eli/reg/2024/1689/oj

Winfield, A. F. T., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R. I., Olszewska, J. I., Rajabiyazdi, F., Theodorou, A., Underwood, M. A., Wortham, R. H., & Watson, E. (2021). IEEE P7001: A proposed standard on transparency. *Frontiers in Robotics and AI*, *8*, Artículo 665729. https://doi.org/10.3389/frobt.2021.665729

ÉTICA Y GOBERNANZA ALGORÍTMICA EN EL SECTOR FINANCIERO: REVISIÓN SISTEMÁTICA DE PRINCIPIOS, GOBERNANZA Y BRECHAS DE IMPLEMENTACIÓN

Luis Enrique Cepeda Cavero
https://orcid.org/0009-0001-6408-8887
20160331@aloe.ulima.edu.pe
Facultad de Ingeniería, Universidad de Lima, Perú

María Paula Véliz Soto
https://orcid.org/0009-0004-9043-0211
20141437@aloe.ulima.edu.pe
Facultad de Ingeniería, Universidad de Lima, Perú

Recibido: 31 de agosto del 2025 / Aceptado: 4 de noviembre del 2025
doi: https://doi.org/10.26439/interfases2025.n022.8230

RESUMEN. Este artículo tiene como objetivo identificar los principios éticos fundamentales, los mecanismos de gobernanza implementados y las brechas entre teoría y práctica en el uso de algoritmos en el sector bancario. Para ello, se realizó una revisión sistemática de la literatura siguiendo el protocolo de Kitchenham y Charters (2007) en cuatro bases de datos académicas (IEEE Xplore, ACM Digital Library, SpringerLink y Scopus), para lo cual se utilizaron cadenas de búsqueda estructuradas en inglés y español. Tras aplicar los criterios de inclusión y exclusión, se seleccionaron y analizaron veintiocho estudios primarios publicados entre el 2020 y el 2025. Los resultados revelan tensiones fundamentales entre la justicia social y la eficiencia técnica, además de dificultades persistentes en términos de transparencia, equidad y privacidad. Se proponen estrategias regulatorias, institucionales y tecnológicas para fortalecer la aplicación ética de sistemas algorítmicos y se enfatizan la supervisión humana, la trazabilidad y la experiencia del usuario. La investigación contribuye a establecer un marco de acción para avanzar hacia una inteligencia artificial bancaria más responsable, legítima e inclusiva.

PALABRAS CLAVE: normas / regulaciones / aprendizaje automático / modelos predictivos / algoritmos adaptativos

ETHICS AND ALGORITHMIC GOVERNANCE IN THE FINANCIAL SECTOR: A SYSTEMATIC REVIEW OF PRINCIPLES, GOVERNANCE, AND IMPLEMENTATION GAPS

ABSTRACT. This article aims to identify the fundamental ethical principles, implemented governance mechanisms, and gaps between theory and practice in the use of algorithms within the banking sector. To this end, a systematic literature review was conducted following the protocol established by Kitchenham and Charters (2007), through searches in four academic databases (IEEE Xplore, ACM Digital Library, SpringerLink, and Scopus) using structured search strings in English and Spanish. After applying rigorous inclusion and exclusion criteria, 28 primary studies published between 2020 and 2025 were selected and analyzed. The findings reveal fundamental tensions between social justice and technical efficiency, as well as persistent challenges in terms of transparency, fairness, and privacy. Regulatory, institutional, and technological strategies are proposed to strengthen the ethical deployment of algorithmic systems, emphasizing human oversight, traceability, and user experience. This research contributes to establishing a framework for advancing toward more responsible, legitimate, and inclusive artificial intelligence in banking.

KEYWORDS: norms / regulations / machine learning / predictive models / adaptive algorithms

INTRODUCCIÓN

La incorporación de sistemas algorítmicos en el sector bancario ha transformado radicalmente la forma en que las instituciones financieras operan y toman decisiones. Desde la evaluación automatizada de solicitudes de crédito hasta la detección de fraudes en tiempo real y la personalización de productos financieros, la inteligencia artificial (IA) y el aprendizaje automático se han convertido en pilares fundamentales de la banca moderna y generan beneficios en eficiencia operativa y experiencia del cliente. Sin embargo, esta transformación plantea desafíos éticos complejos que trascienden lo técnico y que tienen efectos directos sobre la inclusión financiera, la transparencia, la equidad y la privacidad de millones de usuarios. Por ejemplo, algoritmos de *scoring* crediticio han sido cuestionados por perpetuar sesgos históricos contra grupos minoritarios; los sistemas de aprobación automatizada pueden denegar créditos sin explicaciones comprensibles; y el uso de grandes volúmenes de datos personales genera preocupaciones sobre privacidad y consentimiento informado, lo que revela tensiones fundamentales entre la eficiencia económica y la justicia social (Abbas, 2025; Nathim et al., 2024; Yang & Lee, 2024).

A pesar de que varias organizaciones internacionales, como la Organización para la Cooperación y el Desarrollo Económico (OCDE), y la literatura académica (Cath, 2018; Jobin et al., 2019) han planteado diversos marcos normativos y principios éticos, existe aún una notable diferencia entre los ideales normativos y su aplicación en situaciones reales en el sector bancario. En el contexto peruano, esta preocupación se refleja en avances regulatorios recientes como la Ley 31814, del 5 de julio del 2023, la cual busca promover el uso de la IA en favor del desarrollo económico y social del país, y, específicamente en el sector bancario, en el Reglamento de Gestión de Riesgos de Modelo emitido mediante Resolución SBS 00053-2023 el 6 de enero del 2023. Estos marcos normativos tienen antecedentes en la implementación de los acuerdos de regulación bancaria de Basilea II a través del Reglamento para el Requerimiento de Patrimonio Efectivo por Riesgo de Crédito, aprobado por la Resolución SBS 14354-2009, del 30 de octubre del 2009, contexto en el cual Rayo, Lara y Camino (2010) ya identificaban tensiones entre eficiencia algorítmica y equidad en la evaluación crediticia de poblaciones vulnerables, desafíos que se mantienen vigentes con los sistemas de IA contemporáneos.

Este artículo analiza de manera crítica las brechas que aún existen entre la práctica y la teoría, los principios éticos más importantes en el diseño de algoritmos bancarios y los mecanismos de gobernanza sugeridos para reducir riesgos. Para estructurar este análisis, se plantean tres preguntas de investigación: ¿qué principios éticos se consideran prioritarios en el diseño y aplicación de algoritmos en el sector bancario? (PI1), ¿qué mecanismos de gobernanza algorítmica se han propuesto o implementado en instituciones bancarias para mitigar riesgos éticos? (PI2) y ¿qué brechas existen entre la teoría ética y la práctica algorítmica en el sector bancario, y cómo podrían abordarse? (PI3). Estas preguntas guían la revisión sistemática de la literatura y permiten identificar áreas de mejora, conflictos

operativos y tensiones estructurales a través del análisis de veintiocho estudios primarios, publicados entre el 2020 y el 2025, con el propósito de avanzar hacia una IA bancaria más equitativa, responsable y transparente. El análisis se concentra especialmente en destacar los desafíos que enfrentan las instituciones financieras al intentar equilibrar la eficiencia técnica con la justicia social, en un entorno regulatorio que sigue evolucionando.

El artículo se estructura de la siguiente manera: la sección de metodología describe detalladamente el método de revisión sistemática empleado, lo que incluye las preguntas de investigación, las fuentes de datos, las cadenas de búsqueda y el proceso de selección de estudios; la sección de discusión de resultados los aborda según las tres preguntas de investigación planteadas; a continuación, la sección siguiente expone las implicancias prácticas a nivel institucional, regulatorio y tecnológico; y, finalmente, la sección de cierre ofrece las conclusiones y recomendaciones derivadas del análisis realizado.

METODOLOGÍA

Esta investigación se basa en la metodología de revisión sistemática de literatura (RSL) propuesta por Kitchenham y Charters (2007), que cuenta con un amplio reconocimiento en las investigaciones sobre ingeniería del *software* y que se ha adaptado aquí para el campo interdisciplinario de la ética algorítmica aplicada a la banca digital. Una RSL es, de acuerdo con estos autores, "a means of evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest [una herramienta para determinar, analizar e interpretar toda la investigación disponible que sea pertinente a una cuestión de investigación particular, un tema o un fenómeno de interés]" (Kitchenham & Charters, 2007, p. iv). Este enfoque posibilita asegurar la rigurosidad metodológica, la reproducibilidad y la trazabilidad en el desarrollo del marco teórico.

Definición de las preguntas de investigación

Para estructurar el análisis crítico de la literatura técnica, se proponen tres preguntas guía (véase la Tabla 1).

Tabla 1

Preguntas de investigación

Referencia	Pregunta	Propósito
PI1	¿Qué principios éticos se consideran prioritarios en el diseño y aplicación de algoritmos en el sector bancario?	Determinar los principios básicos (por ejemplo, la privacidad, la transparencia y la equidad) que orientan el progreso algorítmico en ámbitos financieros.

(*continúa*)

(*continuación*)

Referencia	Pregunta	Propósito
PI2	¿Qué mecanismos de gobernanza algorítmica se han propuesto o implementado en instituciones bancarias para mitigar riesgos éticos?	Investigar el uso de marcos regulatorios, auditorías algorítmicas, comités éticos y prácticas de rendición de cuentas en la industria.
PI3	¿Qué brechas existen entre la teoría ética y la práctica algorítmica en el sector bancario, y cómo podrían abordarse?	Identificar contradicciones entre lo que se declara como principio y lo que realmente se implementa, sugiriendo áreas de investigación o mejora.

El objetivo de esta revisión sistemática es examinar cómo los principios éticos y los mecanismos de gobernanza algorítmica interactúan en el sector bancario con el fin de detectar oportunidades, tensiones y perspectivas para la aplicación responsable de sistemas algorítmicos en entornos financieros, en particular aquellas vinculadas con la experiencia del usuario y la trazabilidad institucional. Estas preguntas posibilitan la articulación de la revisión desde un enfoque interdisciplinario que incorpora aspectos técnicos, normativos y experienciales. Asimismo, ayudan a identificar huecos teóricos y prácticos que podrían guiar futuras investigaciones y propuestas de gobernanza algorítmica con enfoque en el usuario.

Definición de fuentes de datos

Con el fin de asegurar la pertinencia, calidad y trazabilidad de los estudios que forman parte de esta revisión sistemática, se escogieron repositorios académicos reconocidos internacionalmente c que proporcionan literatura acerca de la ética algorítmica, la gobernanza digital y su implementación en el ámbito bancario. Los repositorios se seleccionaron tomando en cuenta la seriedad con que evalúan el respaldo editorial y la política de revisión por pares de las publicaciones que indexan, así como su enfoque interdisciplinario. Los repositorios escogidos facilitan la consulta de libros especializados, artículos de revistas, documentos regulatorios, informes técnicos y artículos de congresos (véase la Tabla 2). Esta variedad de fuentes garantiza una visión integral del fenómeno analizado, lo que permite incorporar puntos de vista desde la ética computacional, la ingeniería de sistemas, la regulación financiera y la experiencia del usuario.

Tabla 2

Fuentes de datos

Repositorio	Tipo de contenido
IEEE Xplore	Artículos de congresos y revistas sobre sistemas algorítmicos, gobernanza tecnológica y banca digital

(*continúa*)

(*continuación*)

Repositorio	Tipo de contenido
ACM Digital Library	Estudios sobre ética computacional, diseño responsable de algoritmos y arquitectura de sistemas
Repositorio	Tipo de contenido
SpringerLink	Publicaciones sobre regulación financiera, IA explicable (XAI) y ética digital
Scopus	Artículos multidisciplinarios que integran tecnología, ética y economía bancaria

Establecimiento de cadenas de búsqueda

Se establecieronn términos fundamentales en inglés y español, vinculados con la ética algorítmica, la IA y la banca digital, para poder recuperar investigaciones relevantes que respondan a las cuestiones de investigación planteadas. Estos términos se utilizan en los títulos, resúmenes y palabras clave de los repositorios elegidos. Con el fin de organizar las búsquedas, se utilizaron operadores booleanos (OR, AND) que posibilitan la combinación de conceptos y una mayor cobertura temática.

Validación de cadenas de búsqueda

Las cadenas se validaron mediante una prueba piloto en dos fases: primero, se ejecutaron búsquedas preliminares en Scopus con diferentes combinaciones de términos que arrojaban resultados más pertinentes a las preguntas de investigación; segundo, se verificó que artículos seminales reconocidos (como Cath, 2018 o Jobin et al., 2019) fueran recuperados, lo que confirmaba la sensibilidad de las cadenas. Los sinónimos se consideran adecuados por tres razones: reflejan la terminología empleada en la literatura del área; capturan variaciones lingüísticas comunes (por ejemplo, *AI* versus *artificial intelligence*); y combinan términos generales y especializados, lo que asegura la cobertura de estudios teóricos y aplicados.

Estructura lógica aplicada

Para combinar conceptos y garantizar que los términos se presenten en por lo menos uno de los campos principales (título, resumen o palabras clave), se usaron operadores booleanos para formular las cadenas. Este es un ejemplo de la estructura empleada:

> En español:
>
> ((título or resumen or palabras-clave):
>
> (“ética algorítmica” OR “sesgo algorítmico” OR “transparencia algorítmica” OR “gobernanza algorítmica”))

AND

((título or resumen or palabras-clave):

("inteligencia artificial" OR "IA" OR "aprendizaje automático" OR "machine learning" OR

"IA explicable" OR "XAI"))

AND

((título or resumen or palabras-clave):

("banca digital" OR "sector bancario" OR "sector financiero" OR "servicios financieros" OR "fintech"))

En inglés:

((title or abstract or keywords):

("algorithmic ethics" OR "algorithmic bias" OR "algorithmic transparency" OR "algorithmic governance"))

AND

((title or abstract or keywords):

("artificial intelligence" OR "AI" OR "machine learning" OR "explainable AI" OR "XAI"))

AND

((title or abstract or keywords):

("digital banking" OR "banking sector" OR "financial sector" OR "financial services" OR "fintech"))

Ejecución de consultas

Con el propósito de identificar estudios relevantes que aborden las preguntas de investigación formuladas, se determinan criterios de inclusión y exclusión para definir el corpus analítico (véase la Tabla 3). Estos criterios se implementan en simultáneo al proceso de búsqueda en los repositorios elegidos y según las cadenas de búsqueda establecidas en ambas lenguas (español e inglés).

Tabla 3

Criterios de inclusión y exclusión

Criterios de inclusión	Criterios de exclusión
Artículos publicados entre el 1/1/2020 y el 30/7/2025	Publicaciones fuera del periodo establecido
Idioma: español o inglés	Idiomas distintos al español o inglés

(*continúa*)

(*continuación*)

Estudios que incluyan las cadenas de búsqueda en título, resumen o palabras clave	Artículos que no contengan los términos definidos
Publicaciones revisadas por pares o con respaldo institucional	Documentos sin revisión académica o sin respaldo editorial
Relevancia temática: ética algorítmica, gobernanza digital, banca, IA	Estudios centrados en otros sectores o sin vínculo con algoritmos

Las siguientes bases de datos son las que se utilizan para ejecutar las consultas: Scopus, SpringerLink, ACM Digital Library e IEEE Xplore. Se utilizan filtros por idioma, fecha y tipo de documento en cada uno de ellos, y los resultados se exportan para su depuración. Como resultado, se obtuvo 148 artículos de dichas bases de datos (véase la Tabla 4).

Tabla 4

Resultados de la primera ejecución

Repositorio	Cantidad de resultados
IEEE Xplore	18
ACM Digital Library	96
SpringerLink	17
Scopus	17
Total	148

Proceso de selección de estudios

Esta sección expone cómo se lleva a cabo la selección de los estudios primarios, lo cual comprende descartar duplicados, examinar títulos y resúmenes, así como leer en su totalidad los artículos más importantes.

Se definieron parámetros para sustentar la validez interna y externa de la selección de los estudios. La validez interna se garantizó mediante los criterios de inclusión y exclusión predefinidos aplicados uniformemente; y mediante la revisión cruzada entre dos autores y la evaluación consensuada para resolver discrepancias. Por su parte, la validez externa se sustenta en cuatro bases de datos multidisciplinarias (IEEE Xplore, ACM, SpringerLink, Scopus); en la inclusión bilingüe (inglés y español); y en el periodo de cinco años (2020-2025) que captura el debate reciente sobre IA y ética en la banca, lo cual permite la transferibilidad a contextos bancarios globales contemporáneos.

Como parte del protocolo metodológico, se definió un proceso de selección que consta de cinco fases. El objetivo global de este proceso es asegurar la relevancia temática, la calidad y la pertinencia de las investigaciones incluidas en la revisión sistemática. Cada fase funciona como un filtro que depura el corpus inicial que ha

sido obtenido al realizar consultas. Las fases del procedimiento de selección son las siguientes:

- Fase 1: implementación de la cadena de búsqueda en los repositorios escogidos. Se ejecutan las búsquedas en los cuatro repositorios seleccionados, para lo cual se emplean los términos en inglés y español en las secciones de título, resumen y palabras clave.
- Fase 2: supresión de los estudios duplicados. Se eliminan los artículos que aparecen en más de un repositorio para prevenir sesgos por repetición.
- Fase 3: análisis preliminar a partir del título, las palabras clave y el resumen. Se lleva a cabo una lectura superficial con el fin de eliminar estudios que no traten directamente sobre la ética algorítmica, la gobernanza digital o el sector bancario.
- Fase 4: lectura parcial (introducción, conclusiones y resultados). Se analizan los segmentos fundamentales de los artículos con el fin de determinar su adecuación a las preguntas de investigación, así como su aplicabilidad y profundidad teórica.
- Fase 5: lectura total de los artículos escogidos. Se lleva a cabo una lectura exhaustiva de las investigaciones que pasaron las etapas previas, con el objetivo de corroborar su inclusión como estudios primarios (véase la Tabla 5).

Tabla 5

Resultados del proceso de selección

Repositorio	Periodo	Resultados iniciales	Estudios primarios
IEEE Xplore	2020-2025	18	8
ACM Digital Library	2020-2025	96	7
SpringerLink	2020-2025	17	6
Scopus	2020-2025	17	7
Total	-	148	28

Los veintiocho estudios se analizaron mediante un proceso sistemático en tres etapas. En la primera etapa de análisis, cada artículo fue leído completamente y se extrajo información específica según las preguntas de investigación: para PI1, se identificaron y listaron todos los principios éticos mencionados (como equidad, transparencia, privacidad) junto con las definiciones operativas propuestas por los autores; para PI2, se documentaron los mecanismos de gobernanza descritos y se indicó si eran propuestas teóricas o implementaciones empíricas en instituciones bancarias; y para PI3, se registraron las brechas específicas reportadas entre lo que se propone en términos éticos y lo que realmente se implementa.

En la segunda etapa, se realizó un análisis comparativo cruzado entre los veintiocho estudios, y se contrastó cómo diferentes autores definen conceptos clave (por ejemplo, qué entiende cada uno por *equidad algorítmica*), qué evidencia empírica presentan (estudios de caso, experimentos, encuestas) y qué soluciones proponen. En la tercera etapa, los hallazgos se sintetizaron mediante la identificación de cuáles principios, mecanismos y brechas aparecen con mayor frecuencia en la literatura, cuáles generan consenso y cuáles muestran posturas divergentes. Adicionalmente, se identificó el tema principal de cada artículo (véase la Tabla 6). Los resultados organizados por pregunta de investigación se presentan en la siguiente sección.

Tabla 6

Detalle de estudios primarios

ID	Autor	Año	Título	Repositorio	DOI / URL	Tema principal
1	Afrin et al.	2025	AI-Enhanced Robotic Process Automation: A Review of Intelligent Automation Innovations	IEEE Xplore	https://doi.org/10.1109/ACCESS.2024.3513279	Automatización robótica con IA
2	Akre et al.	2024	Smart Finance: An Overview of Artificial Intelligence Integration in FinTech	IEEE Xplore	https://doi.org/10.1109/IDICAIEI61867.2024.10842818	IA en FinTech
3	Ilari et al.	2023	Ethical Biases in Machine Learning-Based Filtering of Internet Communications	IEEE Xplore	https://doi.org/10.1109/ETHICS57328.2023.10154975	Sesgos éticos en machine learning
4	Nathim et al.	2024	Ethical AI with Balancing Bias Mitigation and Fairness in Machine Learning Models	IEEE Xplore	https://doi.org/10.23919/FRUCT64283.2024.10749873	IA ética y mitigación de sesgos
5	Purwar et al.	2024	Data-Driven Insights: Leveraging Analytics for Predictive Modeling in Finance	IEEE Xplore	https://doi.org/10.1109/ICTACS62700.2024.10841301	Análisis predictivo en finanzas
6	Sharma y Preet	2025	Cutting-Edge AI Applications in Banking: Trends, Challenges and Future Directions	IEEE Xplore	https://doi.org/10.23919/INDIACom66777.2025.11115186	Aplicaciones de IA en banca
7	Swapna et al.	2024	Leveraging the Power of Deep Learning, Machine Learning, Artificial Intelligence, and Big Data in Finance and Education	IEEE Xplore	https://doi.org/10.1109/ICTACS62700.2024.10841229	IA y big data en finanzas y educación
8	Yasir et al.	2022	How Artificial Intelligence is Promoting Financial Inclusion? A Study on Barriers of Financial Inclusion	IEEE Xplore	https://doi.org/10.1109/ICBATS54253.2022.9759038	IA en inclusión financiera
9	Yin et al.	2025	Scientific Quantitative Analysis of Artificial Intelligence in the Financial Industry	ACM Digital Library	https://doi.org/10.1145/3745133.3745169	Análisis cuantitativo de IA en finanzas
10	Jantunen et al.	2024	Researchers' Concerns on Artificial Intelligence Ethics: Results from a Scenario-Based Survey	ACM Digital Library	https://doi.org/10.1145/3643690.3648238	Ética en IA

ID	Autor	Año	Título	Repositorio	DOI / URL	Tema principal
11	Kaponis y Maragoudakis	2022	Data Analysis in Digital Marketing Using Machine Learning and Artificial Intelligence Techniques, Ethical and Legal Dimensions, State of the Art	ACM Digital Library	https://doi.org/10.1145/3549737.3549756	IA en marketing digital y ética
12	Kasinidou et al.	2024	"Artificial Intelligence is a very Broad Term": How Educators Perceive Artificial Intelligence?	ACM Digital Library	https://doi.org/10.1145/3677525.3678677	Percepción de IA en educación
13	Abdurashidova y Balbaa	2024	Artificial Intelligence in the Banking Sector in Uzbekistan: Exploring the Impacts and Opportunities	ACM Digital Library	https://doi.org/10.1145/3644713.3644721	IA en banca
14	Schedl et al.	2022	Retrieval and Recommendation Systems at the Crossroads of Artificial Intelligence, Ethics, and Regulation	ACM Digital Library	https://doi.org/10.1145/3477495.3532683	Sistemas de recomendación, ética y regulación
15	Singh et al.	2024	Artificial Intelligence in the Legal Profession: A Review on its Transformative Potential and Ethical Challenges	ACM Digital Library	https://doi.org/10.1145/3647444.3647953	IA en el sector legal
16	Batiz-Lazo et al.	2022	The Spread of Artificial Intelligence and its Impact on Employment: Evidence from the Banking and Accounting Sectors	SpringerLink	https://doi.org/10.1007/978-3-031-07765-4_7	Impacto de IA en empleo bancario
17	Kahyaoglu	2021	The Impact of Artificial Intelligence on Central Banking and Monetary Policies	SpringerLink	https://doi.org/10.1007/978-981-33-6811-8_5	IA en banca central y políticas monetarias
18	Liu et al.	2025	The Impact of Artificial Intelligence on Consumers' Willingness to use CBDCs: Evidence from the Chinese Banking Sector	SpringerLink	https://doi.org/10.1057/s41599-025-05067-5	IA y monedas digitales
19	Lopes et al.	2025	The Role of Artificial Intelligence in Mobile Banking: Decoding Portuguese Consumers' Perceptions and Intentions to Engage	SpringerLink	https://doi.org/10.1186/s43093-025-00510-0	IA en banca móvil

ID	Autor	Año	Título	Repositorio	DOI / URL	Tema principal
20	Mishra et al.	2024	Introduction to Machine Learning and Artificial Intelligence in Banking and Finance	SpringerLink	https://doi.org/10.1007/978-3-031-47324-1_14	IA en banca y finanzas
21	Ratzan y Rahman	2024	Measuring Responsible Artificial Intelligence (RAI) in Banking: A Valid and Reliable Instrument	SpringerLink	https://doi.org/10.1007/s43681-023-00321-5	IA responsable en banca
22	Abbas	2025	Lending by Algorithm: Fair or Flawed? An Information-Theoretic View of Credit Decision Pipelines	Scopus	https://doi.org/10.1007/s42979-025-04222-8	Algoritmos en decisiones crediticias
23	AL-Ghuribi et al.	2025	Navigating the Ethical Landscape of Artificial Intelligence: A Comprehensive Review	Scopus	https://hdl.handle.net/20.500.14536/5762	Ética en IA
24	Ibrahim et al.	2024	Artificial Intelligence Ethics: Ethical Consideration and Regulations from Theory to Practice	Scopus	https://doi.org/10.11591/ijai.v13.i3.pp3703-3714	Ética y regulación de IA
25	Panda et al.	2025	Artificial Intelligence in Action: Shaping the Future of Public Sector	Scopus	https://doi.org/10.1108/DPRG-10-2024-0272	IA en sector público
26	Ponmalar et al.	2025	Ethical Challenges and Bias in Real-World AI: A Fairness-Oriented Approach to Resume Screening Systems	Scopus	https://doi.org/10.1109/ICAISS61471.2025.11042179	Sesgos y ética en IA (selección de personal)
27	Porwal	2025	Ethics and Laws: Governing Generative AI's Role in Financial Systems	Scopus	https://doi.org/10.1002/9781394271078.ch15	IA generativa en sistemas financieros
28	Yang y Lee	2024	Ethical AI in Financial Inclusion: The Role of Algorithmic Fairness on User Satisfaction and Recommendation	Scopus	https://doi.org/10.3390/bdcc8090105	Justicia algorítmica e inclusión financiera

DISCUSIÓN DE RESULTADOS

Se llevó a cabo un análisis basado en veintiocho investigaciones primarias que fueron elegidas por medio del protocolo de revisión sistemática, lo cual posibilita el seguimiento de la progresión del interés académico sobre la ética algorítmica en el sector bancario. De acuerdo con los criterios de inclusión establecidos, esta revisión cubre el quinquenio que va del 2020 al 2025. Las publicaciones anuales muestran una concentración importante en los años 2024 y 2025, lo que indica un aumento reciente en la producción científica relacionada con el asunto. En comparación, los años pasados presentan un número menor de investigaciones, lo cual puede sugerir que la discusión ética sobre IA en el ámbito bancario ha cobrado impulso únicamente en periodos más recientes (véase la Tabla 7).

Tabla 7

Distribución de estudios primarios por año de publicación

Año	Cantidad de estudios	Porcentaje aproximado
2021	1	3,6
2022	3	10,7
2023	2	7,1
2024	11	39,3
2025	11	39,3
Total	28	100

Los estudios primarios ofrecen definiciones operativas que hacen posible tratar con exactitud las preguntas de investigación. En la literatura aparecen conceptos tales como sesgo, gobernanza, inclusión financiera, explicabilidad, equidad algorítmica e IA responsable, los cuales han sido formulados por autores que han elaborado marcos teóricos y métricas que pueden ser aplicadas al sector bancario. El análisis ético del uso algorítmico en servicios financieros se basa en estas definiciones, las cuales se han tomado de fuentes indexadas directamente.

- IA responsable: capacidad institucional para poner en marcha modelos de IA que sean transparentes en cuanto a sus entradas y salidas, lo cual favorece la justicia y reduce el sesgo. Fuera del mundo de la IA, una analogía de este concepto podría ser la de un auditor ético que, además de examinar cuentas, explica cada determinación contable ante los interesados (Ratzan & Rahman, 2024).
- Equidad algorítmica: principio que, de acuerdo con la Ley de Igualdad de Oportunidades de Crédito (Equal Credit Opportunity Act, ECOA), requiere trato justo para cada persona que se ve afectada por decisiones algorítmicas.

Fuera del mundo de la IA, una analogía de este concepto podría ser la de un jurado que examina a cada candidato sin prejuicios, lo que asegura la equidad de oportunidades (Ratzan & Rahman, 2024).

- Sesgo algorítmico: fenómeno que es inversamente proporcional a la equidad, lo que significa que, a más sesgo, menos justicia en los resultados. Fuera del mundo de la IA, una analogía de este concepto podría ser la de una balanza mal calibrada que, de manera sistemática, favorece a un lado y distorsiona la opinión (Nathim et al., 2024; Ratzan & Rahman, 2024).
- Explicabilidad algorítmica: habilidad de los sistemas de IA para defender sus decisiones, lo cual permite la auditoría ética y la corrección de sesgos. Fuera del mundo de la IA, una analogía de este concepto podría ser la de un juez que justifica cada sentencia con argumentos verificables y claros (Mishra et al., 2024; Ratzan & Rahman, 2024).
- Gobernanza algorítmica: conjunto de prácticas éticas que abarcan la diversidad, auditorías, liderazgo responsable y coordinación entre sectores. Fuera del mundo de la IA, una analogía de este concepto podría ser la de una constitución organizativa que controla la conducta de los sistemas autónomos (AL-Ghuribi et al., 2025; Porwal, 2025; Swapna et al., 2024).
- Inclusión financiera: una meta ética que intenta extender el acceso equitativo a servicios financieros a través de una IA justa y transparente. Fuera del mundo de la IA, una analogía de este concepto podría ser la de una puerta automática que se abre para todo el mundo sin distinción de aspecto o antecedentes (Yang & Lee, 2024).

La discusión se organiza según las tres interrogantes de investigación propuestas e incorpora los descubrimientos más significativos de los veintiocho estudios primarios elegidos. Se examinan los principios éticos fundamentales, las estructuras de gobernanza algorítmica puestas en marcha en la industria bancaria y las discrepancias entre teoría y práctica, con un particular énfasis en la experiencia del usuario y en la trazabilidad institucional.

PI1: ¿qué principios éticos se consideran prioritarios en el diseño y aplicación de algoritmos en el sector bancario?

La equidad se destaca como el principio ético más importante en la elaboración de algoritmos bancarios. Según Ratzan y Rahman (2024), la equidad es el fundamento de una IA responsable, alineada con la Ley de Igualdad de Oportunidades de Crédito. Nathim et al. (2024) muestran que es posible disminuir los sesgos en un 23 %, lo cual mejora las métricas de equidad; aunque esto implica una disminución de la precisión de hasta el 9 %, lo cual pone de manifiesto un conflicto estructural entre la eficiencia técnica y la

justicia social. Abbas (2025) se adentra en este asunto al ilustrar cómo los algoritmos de puntuación crediticia perpetúan las disparidades educativas, al aprobar a solicitantes con escasa educación tres veces menos que a los demás, aunque tengan tasas parecidas de falsos positivos.

La transparencia es un principio adicional que hace más fácil la evaluación de la equidad. La explicabilidad algorítmica fomenta la voluntad institucional de reducir los sesgos, según Ratzan y Rahman (2024). Yang y Lee (2024) corroboran que cuando los niveles de transparencia y legitimidad aumentan, se optimizan el grado de satisfacción del usuario, la percepción de equidad y su inclinación a sugerir servicios financieros fundamentados en IA.

Otro eje ético esencial es la privacidad de los datos. Purwar et al. (2024) defienden la implementación de procedimientos abiertos y normas más rigurosas para reforzar la confianza del público hacia sistemas automatizados; a su vez, Ratzan y Rahman (2024) la relacionan con la calidad de los datos.

La responsabilidad organizacional implica cambios en la estructura. Sobre la base de la perspectiva de Ratzan y Rahman (2024), quienes han subrayado la importancia del liderazgo ético, la diversidad y la capacitación, Porwal (2025) sugiere añadir la realización de auditorías periódicas y una coordinación entre los actores técnicos, financieros y regulatorios. De otro lado, la inclusión financiera representa simultáneamente una promesa y un peligro. Yang y Lee (2024) advierten que el sesgo algorítmico puede agravar la exclusión, por lo cual es necesario establecer sistemas que promuevan activamente la equidad y que eviten la discriminación directa.

Aunque se discute menos, la solidez técnica es fundamental. AL-Ghuribi et al. (2025) la consideran uno de los principios fundamentales de la ética, al reconocer que los errores técnicos pueden causar perjuicios palpables a los usuarios.

La literatura muestra que estos principios constituyen un sistema interdependiente. Nathim et al. (2024) indican que llegar a la justicia sin renunciar a la eficacia continúa siendo un desafío. Por su parte, Abbas (2025) hace una distinción entre equidad experiencial y equidad estadística, y enfatiza que las métricas no reflejan completamente lo que el usuario experimenta. Para que los algoritmos bancarios se implementen de manera ética es necesario un balance entre la adaptación regulatoria, la innovación tecnológica y el cambio organizacional.

Afrin et al. (2025) señalan que la integración de IA con automatización robótica en el sector financiero plantea dilemas éticos significativos que deben abordarse desde el diseño y la implementación responsable. En línea con esta perspectiva, Akre et al. (2024) destacan que la calidad de los datos, la privacidad, la rendición de cuentas, la transparencia y la objetividad son principios éticos esenciales para desarrollar sistemas FinTech confiables y justos.

Estos principios adquieren particular relevancia en contextos de inclusión financiera, respecto a los cuales Yasir et al. (2022) argumentan que el uso de IA requiere marcos regulatorios que garanticen la equidad, la transparencia y la protección de los usuarios. Complementariamente, Lopes et al. (2025) identifican que la confianza del usuario en la banca móvil impulsada por IA depende de la seguridad, la ética del diseño y la calidad del servicio percibida, elementos que se extienden también a innovaciones emergentes como las monedas digitales, respecto a cuyo desarrollo Liu et al. (2025) sostienen que debe considerar principios éticos que equilibren innovación y protección del consumidor.

Un aspecto crítico que atraviesa estos principios es la gestión de sesgos algorítmicos. Ibrahim et al. (2024) proponen lineamientos éticos y regulatorios para mitigar sesgos en los servicios financieros basados en IA, lo cual incluye sesgos vinculados al género, la raza y la edad. Profundizando en esta problemática, Ponmalar et al. (2025) destacan tres fuentes críticas de sesgo —de datos, algorítmico e interactivo— que deben ser gestionadas mediante estrategias de equidad, transparencia y rendición de cuentas en la banca digital.

PI2: ¿qué mecanismos de gobernanza algorítmica se han propuesto o implementado en instituciones bancarias para mitigar riesgos éticos?

Las entidades bancarias han empezado a implementar sistemas de gobernanza algorítmica para reducir los riesgos éticos. Porwal (2025) plantea un sistema integral que tiene como objetivo la incorporación de la IA generativa de manera segura y legal, mediante auditorías periódicas, vigilancia constante y coordinación entre diversos sectores. Esta perspectiva admite que una gobernanza efectiva necesita de la cooperación entre las entidades financieras, los reguladores y los desarrolladores, así como de un monitoreo dinámico.

Ratzan y Rahman (2024) crean una herramienta para evaluar la madurez de la IA responsable, lo que incluye aspectos tales como la seguridad, la privacidad, el compromiso organizacional, la equidad y la explicabilidad. Aunque afronta retos operativos en contextos con recursos limitados, su aplicación en bancos de los Estados Unidos demuestra validez empírica.

Nathim et al. (2024) sugieren marcos adaptativos que equilibran el rendimiento y la equidad, los cuales se fundamentan en una revisión sistemática y en pruebas con algoritmos de aprendizaje automático. Sus descubrimientos resaltan la importancia de soluciones específicas por dominio y escalables. Swapna et al. (2024) documentan los principios internacionales de la OCDE, los cuales buscan fomentar la innovación, salvaguardar a los consumidores y disminuir riesgos; sin embargo, admiten que los sistemas adaptativos enfrentan retos regulatorios no solucionados.

Yang y Lee (2024) evidencian que la impresión de equidad de los usuarios se ve directamente afectada por la transparencia, la responsabilidad y la legitimidad algorítmica, lo cual indica que es necesario comunicar adecuadamente los mecanismos. Ilari et al. (2023) sugieren una perspectiva de ética por diseño, la cual involucra la incorporaciónde los principios éticos desde las etapas tempranas del desarrollo algorítmico.

Por su parte, Abbas (2025) enfatiza la importancia de incluir indicadores conductuales en los modelos de crédito para abordar diferencias sociotécnicas, lo cual demuestra una falta de alineación entre las percepciones de justicia y los criterios algorítmicos.

Purwar et al. (2024) destacan que es importante tener reglas rigurosas para proteger los datos personales y más transparencia en los procedimientos de cumplimiento. AL-Ghuribi et al. (2025) destacan la importancia de la cooperación multisectorial para reforzar la gobernanza ética y fomentar la educación.

La literatura admite que, a pesar de su diversidad, estos mecanismos afrontan retos constantes. Abbas (2025) indica que la equidad estadística no asegura la equidad a nivel de experiencia, mientras que Nathim et al. (2024) advierten acerca de la dificultad para alcanzar sistemas justos y eficaces. El éxito de la gobernanza algorítmica se basa en su puesta en marcha coordinada y su adecuación a situaciones particulares.

La implementación de mecanismos de gobernanza algorítmica en instituciones bancarias requiere estructuras organizacionales y regulatorias robustas. Afrin et al. (2025) señalan que la integración de IA y la automatización robótica en instituciones financieras exigen mecanismos éticos de supervisión para equilibrar eficiencia operativa y responsabilidad social. Esta perspectiva se concreta operativamente cuando Akre et al. (2024) subrayan la importancia de estructuras de gobernanza que aseguren privacidad, rendición de cuentas y transparencia en el desarrollo y despliegue de IA en sistemas FinTech.

Avanzando hacia modelos de supervisión más dinámicos, Panda et al. (2025) enfatizan la necesidad de auditorías periódicas, supervisión continua y colaboración entre bancos, programadores y organismos reguladores para garantizar el uso responsable de la IA. Estos mecanismos prácticos se fundamentan en los principios que Schedl et al. (2022) proporcionan desde una perspectiva interdisciplinaria sobre la equidad, la no discriminación y la transparencia en los sistemas de IA, de lo que resultan implicaciones éticas y regulatorias relevantes para el ámbito financiero.

PI3: ¿qué brechas existen entre la teoría ética y la práctica algorítmica en el sector bancario, y cómo podrían abordarse?

En el ámbito bancario, la brecha entre la teoría ética y la práctica algorítmica se muestra en varios niveles. Abbas (2025) demuestra que, aunque los algoritmos bancarios

pueden parecer justos según sus cálculos, eso no significa que las personas los perciban como justos. Por ejemplo, un sistema que se enfoca en cuánto gana alguien puede ignorar si esa persona paga sus deudas puntualmente. Esto puede hacer que las decisiones del banco parezcan injustas para los usuarios.

Según Nathim et al. (2024), optimizar las métricas de equidad puede disminuir los sesgos, aunque esto implica una disminución de la precisión. Esto pone de manifiesto un conflicto entre la eficiencia y la justicia. Por otro lado, incluso cuando hay políticas formales, las presiones comerciales pueden desalentar la implementación ética. Swapna et al. (2024) afirman que los marcos regulatorios existentes no están equipados para supervisar algoritmos adaptativos, cuya evolución dinámica sobrepasa las habilidades de control convencionales. Abbas (2025) también alerta de que la carencia de diversidad en los datos empleados dificulta la evaluación de la equidad interseccional y perpetúa las exclusiones a nivel sistémico. Yang y Lee (2024) demuestran que la implementación de criterios de transparencia, a pesar de que mejora la percepción de equidad, encuentra dificultades técnicas y organizacionales, sobre todo en modelos complejos. Por último, Ratzan y Rahman (2024) crean un instrumento para evaluar la madurez ética de la IA, pero su implementación práctica se ve restringida debido a la resistencia institucional y a la escasez de recursos.

Se han sugerido estrategias adicionales para tratar estas brechas. Nathim et al. (2024) proponen marcos de trabajo adaptativos que incorporen la equidad sin afectar el desempeño. Según Abbas (2025), los modelos de crédito deben incluir indicadores de comportamiento. Porwal (2025) sugiere la supervisión de sistemas evolutivos a través de auditorías periódicas y vigilancia constante. AL-Ghuribi y sus colegas (2025) subrayan la relevancia de trabajar en colaboración entre distintos sectores para compartir saberes y responsabilidades. Purwar et al. (2024) sugieren consolidar la administración de información y crear tecnologías predictivas éticas. Ilari et al. (2023), por su parte, aconsejan incorporar la ética desde la etapa del diseño. Para cerrar estas brechas, es imprescindible implementar reformas en el sistema que unan la innovación tecnológica con cambios en lo cultural, lo organizacional y lo regulador. Solo de esta forma podrán las instituciones bancarias alinear sus prácticas algorítmicas con los principios éticos que proclaman.

La literatura evidencia múltiples brechas entre la teoría ética y la práctica algorítmica en el sector bancario, las cuales se manifiestan en distintas dimensiones. A nivel global, Yin et al. (2025) muestran que, aunque países líderes como China y los Estados Unidos avanzan rápidamente en la aplicación ética de la IA en finanzas, los países en desarrollo enfrentan desafíos significativos en su adopción práctica, lo cual releva disparidades geográficas en la implementación de principios éticos. Esta brecha no se limita a diferencias entre países, sino que permea las organizaciones mismas. Jantunen et al. (2024) revelan una desconexión entre la comprensión teórica de la ética de la IA

y su aplicación práctica en el sector financiero, lo que limita la adopción de buenas prácticas. Este fenómeno se explica parcialmente por lo que Kasinidou et al. (2024) evidencian: la falta de formación y recursos especializados sobre ética de la IA, lo cual obstaculiza su aplicación efectiva en instituciones financieras y educativas.

Las barreras operativas también contribuyen a esta brecha. Batiz-Lazo et al. (2022) encuentran que muchos bancos aún no aplican IA en procesos internos por razones éticas y técnicas, lo que refleja la persistencia de enfoques tradicionales y la falta de confianza en los algoritmos. Similarmente, Abdurashidova y Balbaa (2024) señalan que la banca enfrenta barreras éticas y operativas como la protección de datos y la falta de personal capacitado para implementar IA responsablemente.

Ante este panorama, emerge un consenso sobre la necesidad de marcos regulatorios actualizados. Singh et al. (2024) resaltan la urgencia de regulaciones justas y del desarrollo ético de IA que mantenga principios legales fundamentales en contextos financieros digitalizados. En esta línea, Kahyaoglu (2021) advierte que la digitalización financiera con IA introduce nuevos riesgos éticos y monetarios, lo que exige redefinir funciones regulatorias y de supervisión. Finalmente, Kaponis y Maragoudakis (2022) analizan los desafíos éticos y legales que la IA plantea al *marketing* digital financiero, y destacan la necesidad de un marco normativo adaptado a las realidades emergentes del sector.

IMPLICANCIAS PRÁCTICAS

Los resultados de esta revisión sistemática tienen importantes implicaciones prácticas para la formulación, ejecución y supervisión de sistemas algorítmicos en el ámbito bancario. La finalidad de traducir los principios éticos en prácticas específicas que mejoren la legitimidad, la transparencia y la experiencia del cliente es lo que motiva la agrupación de estas implicancias en tres ámbitos de acción: institucional, regulatorio y tecnológico.

En el ámbito institucional

Los bancos tienen que incorporar una cultura organizacional enfocada en la ética algorítmica. Lo pueden hacer a través de la inclusión de principios como la trazabilidad de decisiones, la explicabilidad contextual y la equidad experiencial en sus procedimientos internos. Esto significa:

- Establecer un comité de gobernanza algorítmica o un área especializada con autoridad formal para diseñar directivas institucionales, supervisar el desarrollo e implementación de modelos de IA y controlar su desempeño ético continuo. Este comité debe integrar perfiles técnicos, legales, éticos y de experiencia de usuario (AL-Ghuribi et al., 2025; Porwal, 2025; Swapna et al., 2024), y debe garantizar que las decisiones algorítmicas se alineen con los principios institucionales y los marcos regulatorios vigentes. En

el contexto peruano, esto implica el cumplimiento de la Resolución SBS 00053-2023 sobre la gestión de riesgos de modelo.

- Poner en marcha protocolos de documentación algorítmica que recojan cada fase del ciclo de vida del sistema (Porwal, 2025).
- Fomentar la capacitación ética a través de equipos operativos, jurídicos y técnicos (Sharma & Preet, 2025).

Estas medidas no solo reducen los riesgos de reputación y regulación, sino que también aumentan la confianza del cliente en los procesos automatizados.

En el ámbito regulatorio

Los organismos de supervisión deberían avanzar hacia marcos regulatorios flexibles que se desarrollen paralelamente a la tecnología. Lo que se deriva de la regulación incluye:

- Demandar auditorías algorítmicas regulares, internas y externas, que se centren en sesgos, explicabilidad y supervisión por parte de humanos (Ratzan & Rahman, 2024).
- Distinguir entre usuarios técnicos y no técnicos y establecer mínimos de explicabilidad para las interfaces bancarias que emplean IA (Yang & Lee, 2024).

Estas acciones posibilitarían disminuir la fragmentación normativa y asegurar una gobernanza algorítmica más efectiva y coherente.

En el ámbito tecnológico

Desde un punto de vista técnico, los creadores de sistemas algorítmicos deben implementar, durante la fase del diseño, mecanismos que pongan en funcionamiento los principios éticos. Las sugerencias comprenden:

- Implementar criterios de equidad experiencial en la validación de modelos, además de las métricas estadísticas convencionales (Abbas, 2025).
- Crear interfaces interactivas que sean explicativas y que posibiliten al usuario entender, cuestionar y rectificar decisiones automatizadas.

Estas prácticas técnicas, además de mejorar la transparencia y la sencillez de uso, permiten que los auditores y reguladores efectúen una supervisión más eficaz. En suma, estas implicaciones prácticas brindan un mapa de ruta para progresar hacia una IA en el sector bancario que sea más ética, comprensible y enfocada en el usuario. Su adopción ayudaría a disminuir la diferencia entre teoría y práctica, al robustecer la legitimidad de las instituciones y la resiliencia tecnológica en un ambiente financiero que se automatiza más cada día.

CONCLUSIONES

La revisión sistemática posibilitó detectar, examinar y comparar los mecanismos de gobernanza algorítmica más significativos, los principios éticos prioritarios y las principales discrepancias entre la teoría y la práctica en el empleo de IA en el sector bancario. Los trabajos iniciales que se han elegido muestran una inquietud cada vez mayor acerca de la validez institucional de los sistemas automatizados, especialmente en situaciones en las que los derechos básicos de los usuarios son impactados por decisiones algorítmicas.

Para empezar, los principios éticos fundamentales —responsabilidad institucional, transparencia, privacidad y equidad— deben implementarse en todas las fases del ciclo de vida algorítmico, no solo declararse. La investigación evidencia que un algoritmo bancario puede demostrar equidad estadística según métricas técnicas, pero los usuarios pueden percibir injusticia en sus resultados. Esta desconexión ocurre porque la equidad matemática no siempre coincide con la equidad experiencial que las personas consideran justa en sus contextos reales. Por tanto, la implementación efectiva de principios éticos requiere trascender el cumplimiento formal e incorporar las dimensiones humanas de la justicia algorítmica.

En segundo lugar, los métodos de gobernanza algorítmica que se han detectado —comités interdisciplinarios, auditorías éticas, interfaces explicativas, documentación de decisiones y capacitación ética transversal— son progresos notables. Sin embargo, su implementación sigue siendo escasa y desigualmente aplicada. La literatura analizada indica que es necesario coordinar prácticas organizacionales, técnicas y normativas para alcanzar una gobernanza efectiva, con un foco de atención especial en la supervisión humana y la trazabilidad.

Por último, se observan diferencias importantes entre los principios éticos que se han declarado y su aplicación práctica. La falta de estímulos institucionales para el desarrollo responsable, la fragmentación regulatoria y la falta de transparencia algorítmica son ejemplos de cómo se expresan estas brechas. Con el objetivo de cerrar estas brechas, se sugiere promover la participación de los usuarios en los procesos de validación, fortalecer la capacitación ética de los equipos técnicos y avanzar hacia marcos regulatorios adaptativos.

REFERENCIAS

Abbas, S. K. (2025). Lending by algorithm: Fair or flawed? An information-theoretic view of credit decision pipelines. *SN Computer Science*, *6*, 679. https://doi.org/10.1007/s42979-025-04222-8

Abdurashidova, M. S., & Balbaa, M. E. (2024). Artificial intelligence in the banking sector in Uzbekistan: Exploring the impacts and opportunities. En *Proceedings of the 7th International Conference on Future Networks and Distributed Systems (ICFNDS '23)* (pp. 51-57). https://doi.org/10.1145/3644713.3644721

Afrin, S., Roksana, S., & Akram, R. (2025). AI-enhanced robotic process automation: A review of intelligent automation innovations. *IEEE Access*, *13*, 173-197. https://doi.org/10.1109/ACCESS.2024.3513279

Akre, P. D., Pacharaney, U., & Siraskar, W. (2024). Smart finance: An overview of artificial intelligence integration in FinTech. En *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)* (pp. 1-6). https://doi.org/10.1109/IDICAIEI61867.2024.10842818

AL-Ghuribi, S. M., Ahmed, A. A., Ibraheem, A. S., Hasan, M. K., & Islam, S. (2025). Navigating the ethical landscape of artificial intelligence: A comprehensive review. *International Journal of Computing and Digital Systems*, *15*(7), 1038-1043. https://hdl.handle.net/20.500.14536/5762

Batiz-Lazo, B., Efthymiou, L., & Davies, K. (2022). The spread of artificial intelligence and its impact on employment: Evidence from the banking and accounting sectors. En A. Thrassou, D. Vrontis, L. Efthymiou, Y. Weber, S. M. R. Shams & E. Tsoukatos (Eds.), *Business advancement through technology, volume II* (pp. 127-145). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-07765-4_7

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

Ibrahim, S. M., Alshraideh, M. A., Leiner, M., Aldajani, I. M., & Ouarda, B. (2024). Artificial intelligence ethics: Ethical consideration and regulations from theory to practice. *IAES International Journal of Artificial Intelligence*, *13*(3), 3703-3714. https://doi.org/10.11591/ijai.v13.i3.pp3703-3714

Ilari, L., Rafaiani, G., Baldi, M., & Giovanola, B. (2023). Ethical biases in machine learning-based filtering of internet communications. En *2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)* (pp. 1-9). https://doi.org/10.1109/ETHICS57328.2023.10154975

Jantunen, M., Meyes, R., Kurchyna, V., Meisen, T., Abrahamsson, P., & Mohanani, R. (2024). Researchers' concerns on artificial intelligence ethics: Results from a scenario-based survey. En *Proceedings of the 7th ACM/IEEE International Workshop on Software-intensive Business (IWSiB '24)* (pp. 24-31). https://doi.org/10.1145/3643690.3648238

Jobin, A., Lenca, 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

Kahyaoglu, H. (2021). The impact of artificial intelligence on central banking and monetary policies. En S. Bozkuş Kahyaoğlu (Ed.), *The impact of artificial intelligence on*

governance, economics and finance, volume I (pp. 89-108). Springer. https://doi.org/10.1007/978-981-33-6811-8_5

Kaponis, A., & Maragoudakis, M. (2022). Data analysis in digital marketing using machine learning and artificial intelligence techniques, ethical and legal dimensions, state of the art. En *Proceedings of the 12th Hellenic Conference on Artificial Intelligence (SETN '22)* (pp. 1-9). https://doi.org/10.1145/3549737.3549756

Kasinidou, M., Kleanthous, S., & Otterbacher, J. (2024). "Artificial intelligence is a very broad term": How educators perceive artificial intelligence? En *Proceedings of the 2024 International Conference on Information Technology for Social Good (GoodIT '24)* (pp. 315-323). https://doi.org/10.1145/3677525.3678677

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering. Keele University & Durham University Joint Report*. https://www.researchgate.net/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering

Ley 31814 del 2023. Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país. 5 de julio del 2023. *Diario Oficial El Peruano*. https://cdn.www.gob.pe/uploads/document/file/5038703/ley-que-promueve-el-uso-de-la-inteligencia-artificial-en-fav-ley-n-31814.pdf?v=1692895308

Liu, H., Jafri, M. A. H., Xu, S., & Shahzad, M. F. (2025). The impact of artificial intelligence on consumers' willingness to use CBDCs: Evidence from the Chinese banking sector. *Humanities and Social Sciences Communications*, *12*, 834. https://doi.org/10.1057/s41599-025-05067-5

Lopes, J. M., Massano-Cardoso, I., & Pedrosa, L. (2025). The role of artificial intelligence in mobile banking: Decoding Portuguese consumers' perceptions and intentions to engage. *Future Business Journal*, *11*, 125. https://doi.org/10.1186/s43093-025-00510-0

Mishra, A. K., Tyagi, A. K., Richa, & Patra, S. R. (2024). Introduction to machine learning and artificial intelligence in banking and finance. En M. Irfan, K. Muhammad, N. Naifar & M. A. Khan (Eds.), *Applications of blockchain technology and artificial intelligence. Financial mathematics and fintech* (pp. 230-290). Springer. https://doi.org/10.1007/978-3-031-47324-1_14

Nathim, K. W., Hameed, N. A., Salih, S. A., Taher, N. A., Salman, H. M., & Chornomordenko, D. (2024). Ethical AI with balancing bias mitigation and fairness in machine learning models. En *2024 36th Conference of Open Innovations Association (FRUCT)* (pp. 797-807). https://doi.org/10.23919/FRUCT64283.2024.10749873

Panda, M., Hossain, M. M., Puri, R., & Ahmad, A. (2025). Artificial intelligence in action: Shaping the future of public sector. *Digital Policy, Regulation and Governance*, *27*(6), 668-686. https://doi.org/10.1108/DPRG-10-2024-0272

Ponmalar, S., Kumar, K. N., Balaji, R. V., Rohith, T., & Kaarthick, A. K. N. (2025). Ethical challenges and bias in real-world AI: A fairness-oriented approach to resume screening systems. En *Proceedings of the 3rd International Conference on Augmented Intelligence and Sustainable Systems (ICAISS 2025)*. IEEE. https://doi.org/10.1109/ICAISS61471.2025.11042179

Porwal, P. D. (2025). Ethics and laws: Governing generative AI's role in financial systems. En P. R. Chelliah, P. K. Dutta, A. Kumar, E. D. R. Santibanez Gonzalez, M. Mittal & S. Gupta (Eds.), *Generative artificial intelligence in finance: Large language models, interfaces, and industry use cases to transform accounting and finance processes* (pp. 283-298). Wiley. https://doi.org/10.1002/9781394271078.ch15

Purwar, M., Deka, U., Raj, H., & Ritu. (2024). Data-driven insights: Leveraging analytics for predictive modeling in finance. En *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 687-693). IEEE Xplore. https://doi.org/10.1109/ICTACS62700.2024.10841301

Ratzan, J., & Rahman, N. (2024). Measuring responsible artificial intelligence (RAI) in banking: A valid and reliable instrument. *AI & Ethics*, *4*, 1279-1297. https://doi.org/10.1007/s43681-023-00321-5

Rayo, S., Lara, J., & Camino, D. (2010). Un modelo de *credit scoring* para instituciones de microfinanzas en el marco de Basilea II. *Journal of Economics, Finance and Administrative Science*, *15*(28), 89-124. http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2077-18862010000100005

Resolución SBS 00053-2023 [Superintendencia de Banca, Seguros y AFP]. Por la cual se aprueba el Reglamento de Gestión de Riesgos de Modelo. 6 de enero del 2023. https://intranet2.sbs.gob.pe/dv_int_cn/2240/v1.0/Adjuntos/0053-2023.R.pdf

Resolución SBS 14354-2009 [Superintendencia de Banca, Seguros y AFP]. Por la cual se aprueba el Reglamento para el Requerimiento de Patrimonio Efectivo por Riesgo de Crédito y se modifica el manual de contabilidad para las empresas del sistema financiero. 30 de octubre del 2009. https://www.sbs.gob.pe/Portals/0/jer/Auto_Nuevas_Empresas/Sistema_Financiero/9.%20Reg.%20de%20Requerimiento%20de%20Patrimonio%20Efectivo%20por%20Riesgo%20de%20Cr%C3%A9dito_Res.%20SBS%20N%C2%B0%2014354-2009.pdf

Schedl, M., Gómez, E., & Lex, E. (2022). Retrieval and recommendation systems at the crossroads of artificial intelligence, ethics, and regulation. En E. Amigo, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper & G. Kazai (Eds.), *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)* (pp. 3420-3424). https://doi.org/10.1145/3477495.3532683

Sharma, S., & Preet, K. (2025). Cutting-edge AI applications in banking: Trends, challenges and future directions. En *2025 12th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1-6). IEEE Xplore. https://doi.org/10.23919/INDIACom66777.2025.11115186

Singh, L., Joshi, K. A., Koranga, R. S., Pant, S. C., & Mathur, P. (2024). Artificial intelligence in the legal profession: A review on its transformative potential and ethical challenges. En D. Goyal, A. Kumar, D. Singh, M. Paprzycki, B. B. Gupta & U. P. Singh (Eds.), *Proceedings of the 5th International Conference on Information Management & Machine Intelligence (ICIMMI '23)* (pp. 1-8). Association for Computing Machinery https://doi.org/10.1145/3647444.3647953

Swapna, A., Vemula, H. L., Amarnath, B., & Sasidhar, B. (2024). Leveraging the power of deep learning, machine learning, artificial intelligence, and big data in finance and education. En *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 322-327). IEEE Xplore. https://doi.org/10.1109/ICTACS62700.2024.10841229

Yang, Q., & Lee, Y.-C. (2024). Ethical AI in financial inclusion: The role of algorithmic fairness on user satisfaction and recommendation. *Big Data and Cognitive Computing*, *8*(9), 105. https://doi.org/10.3390/bdcc8090105

Yasir, A., Ahmad, A., Abbas, S., Inairat, M., Al-Kassem, A. H., & Rasool, A. (2022). How artificial intelligence is promoting financial inclusion? A study on barriers of financial inclusion. En *2022 International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-6). IEEE Xplore. https://doi.org/10.1109/ICBATS54253.2022.9759038

Yin, F., Zhang, R., Pan, C.-L., & Luo, C. (2025). Scientific quantitative analysis of artificial intelligence in the financial industry. En *Proceedings of the 2025 International Conference on Digital Economy and Information Systems (DEIS '25)* (pp. 214-219). Association for Computing Machinery. https://doi.org/10.1145/3745133.3745169

DATOS DE LOS AUTORES

CARLOS AGUILAR MORA

Egresado de la Universidad para la Cooperación Internacional y, actualmente, se desempeña como docente e investigador académico. Sus áreas de interés incluyen la seguridad informática y el blockchain, y ha participado en actividades académicas y proyectos aplicados en torno a estas áreas. Aunque no registra premios formales, mantiene una trayectoria activa en el estudio y la enseñanza de tecnologías emergentes.

OLDA BUSTILLOS ORTEGA

Magíster en Informática con especialidad en Auditoría Informática por la Universidad Latina de Costa Rica. Actualmente, se desempeña como directora de la Escuela de Ingeniería Informática y docente investigadora en la Universidad Internacional de las Américas. Ha participado en destacados eventos académicos, como la XI Conferencia "Servicios creativos y modernos para el comercio y desarrollo sostenible" de REDLAS y The 2022 International Conference on Information Technology & Systems (ICITS'22). Asimismo, ha recibido el reconocimiento de honorary academician por la International Academy of Social Sciences (IASS), con sede en Palm Beach, Florida. Sus intereses de investigación incluyen la inteligencia artificial, la ciberseguridad y el desarrollo de sistemas inteligentes.

LUIS ENRIQUE CEPEDA CAVERO

Licenciado en Ingeniería Industrial por la Universidad de Lima (Perú) y cuenta con una especialización en Gestión de Proyectos y con certificación Scrum Master Professional. Se desempeña como analista de prevención de fraude en Interbank, donde desarrolla, modela y mantiene procesos y soluciones analíticas basados en datos estructurados y no estructurados, orientados a la detección de fraude financiero y a la optimización de la toma de decisiones estratégicas. Su labor profesional incluye la implementación de

políticas de gobernanza de datos y el uso de herramientas analíticas avanzadas, como SQL y Tableau, con el propósito de fortalecer la integridad y eficiencia de los modelos institucionales. Anteriormente, fue cofundador de Exportacus S. A. C., una start-up dedicada al desarrollo de soluciones tecnológicas para el comercio agrícola en Sudamérica. Asimismo, esta start-up participó en la aceleradora de la Hult Prize Foundation del Reino Unido, tras integrar el equipo galardonado con el Hult Prize Wildcard 2021. Sus principales áreas de investigación e interés profesional se centran en la ética y gobernanza de la inteligencia artificial.

CARLOS DE LA O FONSECA

Graduado de la Universidad CENFOTEC, donde se desempeña como docente e investigador académico. Su trabajo académico se orienta hacia las metodologías ágiles y la inteligencia artificial, áreas en las que ha acumulado experiencia en docencia y proyectos de desarrollo. Aunque no cuenta con premios formales mencionados, su actividad profesional está enfocada en la aplicación práctica y en la investigación en tecnologías de vanguardia.

MARCOS DIAS DE PAULA

Research and Development Coordinator at the Federation of Industries of the State of Rio de Janeiro (FIRJAN), located at the Technological Park of Federal University of Rio de Janeiro (UFRJ). Master's student in Regional Development, in the Postgraduate Program in Regional Development at FACE - Faculty of Administration (PPGDR), Accounting Sciences and Economic Sciences, of Federal University of Goiás). Student Researcher at Center for Intelligence in Policies, and Innovation Projects in Industry at UFG (CIPPI) - Federal University of Goiás. Member of the Brazilian Computer Society (SBC) and Association for the Advancement of Artificial Intelligence (AAAI), based in Washington, DC.

PEDRO H. GONÇALVES

Innovation and Technology Manager at CETT, Center for Education, Work and Technology of Federal University of Goiás (UFG), within the Goiás State Schools of the Future Project. PhD graduate from the Postgraduate Program in Structures and Civil Construction (PECC - University of Brasília). Coordinator of the research group the Laboratory of Inventive Studies in Assistive Technologies (LabEita), at the Federal University of Goiás, where research is developed in the following areas: technology and innovation in assistive technologies. General Coordinator of the Network of Laboratories of Ideas, Prototyping and Entrepreneurship (IPElab UFG).

NICOLÁS XAVIER HERRERA MEDINA

Bachiller en Ingeniería de Sistemas con especialidad en Ingeniería de Software por la Universidad de Lima, con experiencia en start-ups al participar con su propia start-up edtech, Novat, como expositor en Techstars 2024 y como pitcher en UVK x Berkeley 2024. Representó a la carrera de Ingeniería de Sistemas en CADE Universitarios 2024. Trabaja en el Banco de Crédito del Perú como machine learning engineer y como ingeniero de software freelancer brindando servicios de TI en backend, frontend, automatización, machine learning y desarrollo de videojuegos. Tiene su propio estudio de videojuegos "La Fábrica de Chocolate" y su consultora de TI "SNAK IT Consulting". Ganador del Sanda Game Jam 2024 por mejor dirección de arte y mecánica más innovadora, con su juego Earth Mayhem. Previamente, trabajó en el laboratorio SAP Next Gen de la Universidad de Lima como practicante en el 2024, enfocado en IoT, data, machine learning y desarrollo de software.

SEBASTIAN HERRERA SOLIS

Bachiller en Ingeniería de Sistemas por la Universidad de Lima, con especialización en el desarrollo de software y de videojuegos. Actualmente, se desempeña como analista junior de gestión de procesos en Interbank. Su línea de investigación se centra en el uso de juegos serios y simulaciones en entornos virtuales para apoyar el aprendizaje y generar un impacto positivo en las personas.

JOEL EMERSON HUANCAPAZA HILASACA

Magíster en Derecho con mención en Tutela de los Derechos, Globalización de la Justicia y Estado Constitucional por la Universidad Nacional Mayor de San Marcos y abogado por esta misma universidad. Ha cursado especializaciones en derechos fundamentales, derecho constitucional y derecho laboral en entidades como la Universidad Complutense de Madrid, la Universidad de Bolonia y la Universidad de Alcalá. Se desempeña como asesor parlamentario y ha sido expositor en diversos foros académicos. Su práctica e investigación se enfocan en la intersección de los derechos laborales, el constitucionalismo y las nuevas tecnologías. Es miembro de la comunidad internacional Cielo Laboral.

MIGUEL R. LLADÓ

SAP Technical Consultant with over five years of experience in data migration and analytics. He holds an MSc in Advanced Computer Science from the University of Manchester, obtained through the Peruvian government's Beca Generación del

Bicentenario, and a BSc in Systems Engineering from the University of Lima. He has also served as a Teaching Assistant at the University of Lima. His research interests include computer vision, game analytics, and the application of AI in interactive systems.

DANIEL MENA BOCKER

Egresado de la Universidad Magister. Actualmente, se desempeña como docente e investigador. Sus principales intereses académicos y profesionales se centran en las áreas de ciberseguridad e inteligencia artificial. Aunque no ha registrado premios formales, ha desarrollado experiencia en proyectos aplicados y actividades de investigación en estas áreas emergentes.

TERENCE MORLEY

Terence is a lecturer in Computer Science at the University of Manchester in the UK. He holds undergraduate degrees in Electronic Engineering and Mathematics, and a PhD in Computer Science. His main academic interests include image processing and computer vision.

JORGE MURILLO GAMBOA

Magíster en Computación con especialidad en Telemática por el Instituto Tecnológico de Costa Rica. Se desempeña como docente e investigador, y cuenta con una amplia trayectoria dentro del IEEE, donde mantiene la distinción de senior member y ha colaborado como PEAB At Large Member y miembro del EITBOK Core Writing Team. Ha sido reconocido por la IEEE Computer Society con los premios "Valued Member and Supporter" (2016) y "Certificate of Appreciation" (2014). Sus líneas de investigación incluyen inteligencia artificial, ciberseguridad y el desarrollo de habilidades y competencias digitales basadas en el marco SFIA.

HERNAN QUINTANA-CRUZ

Ingeniero Informático por la Pontificia Universidad Católica del Perú, con más de quince años de experiencia en el desarrollo de sistemas de software y videojuegos. Es profesor en la carrera de Ingeniería de Sistemas de la Universidad de Lima e integrante del Information Technology Learning Laboratory (ITLab), donde desarrolla proyectos con tecnologías móviles y aplicaciones interactivas. Es investigador Renacyt, nivel VII.

PABLO ALBERTO ROJAS JAEN

Magíster en Administración de Empresas y Tecnología de la Información, con doble titulación por la Universidad de Lima y la Universidad Autónoma de Barcelona. Ingeniero de Sistemas por la Universidad de Lima. Profesor asociado a tiempo completo e inventor de tres patentes registradas. Coordinador de los laboratorios ITLAB y SAP Next-Gen de la Universidad de Lima, donde lidera proyectos de coinnovación con tecnologías disruptivas, así como desarrollos móviles, web y de realidad virtual.

ALEJANDRO HERNÁN SON ROMERO

Bachiller en Ingeniería de Sistemas por la Universidad de Lima, con una formación orientada al diseño e implementación de soluciones tecnológicas seguras y eficientes. Actualmente, se desempeña como cybersecurity operator en McInnovatec, donde está a cargo del monitoreo, análisis y reporte de amenazas de ciberseguridad mediante el uso de sistemas SIEM y herramientas especializadas para el análisis de archivos, direcciones IP y URL. Previamente, se desempeñó como practicante en el Área de Tecnologías de la Información en Proinversión (2024) y en el Área de Soporte a Laboratorios de la Universidad de Lima (2023). Se diplomó en la especialidad de Tecnologías de la Información otorgado por la Universidad de Lima y cursó estudios de hacking ético en la Universidad Autónoma de Occidente (Colombia). Asimismo, cuenta con certificaciones complementarias en tecnologías de la información, blockchain y metodologías ágiles. Sus principales áreas de interés e investigación comprenden la ciberseguridad, el blockchain, la seguridad de la información y la computación en la nube, ámbitos en los que busca continuar desarrollándose profesional y académicamente.

MARÍA PAULA VÉLIZ SOTO

Licenciada en Ingeniería Industrial por la Universidad de Lima, especializada en gestión de proyectos, experiencia de cliente y transformación digital, con certificaciones PMP®, CAPM®, Scrum Master Professional y Customer Experience Design & Management. Cuenta con más de cinco años de experiencia liderando iniciativas de optimización, digitalización y mejora continua en los sectores de banca, retail y consultoría, desempeñándose como product manager de e-commerce y, posteriormente, como business consultant en servicios financieros en Minsait, donde impulsa proyectos estratégicos y de implementación funcional para entidades bancarias y financieras junto a equipos multiculturales. Ha participado en iniciativas de impacto vinculadas a la educación en inteligencia artificial, y realizó estudios en productos financieros durante su intercambio académico en la Universidad de Salamanca. Actualmente, cursa un diplomado en

inteligencia artificial para los negocios. Sus líneas de investigación se orientan a la estrategia digital, la gobernanza de datos y algoritmos, la ética de la inteligencia artificial en servicios financieros y la adopción tecnológica responsable.


FONDO
EDITORIAL
ULIMA