Use and application of artificial intelligence in public policy evaluation. A Scoping Review.

Barbara Branchini [a,*], **Beatriz Vallina Acha** [b]

[a] *Fresno Servicios Sociales S.L., Madrid*
[b] *Doctoral Programme Agri-food Economics, Universitat Politècnica de València*
[*] *Corresponding author - barbara.branchini@fresnoconsulting.es*

Article info

Article history:
Received 9 May 2025
Revised 12 August 2025
Accepted 3 September 2025

JEL classification:
I38
O330
D830

Keywords:
Big data
Evaluation
Public policy
Artificial intelligence
Review

Abstract

Context: Despite growing interest in using AI to improve public policies and services, its application in evaluation lacks systematised evidence and scientific publications on the implications of these technologies in evaluation practice.

Objective: To map, through available literature, the current and emerging state of AI application in public policy evaluation.

Methods: We conducted the study through an exploratory literature review, or scoping review, following the methodological framework of Levac et al. (2010). The synthesis was carried out with a thematic analysis of 27 studies and analytical-theoretical literature.

Results: AI is increasingly being applied in various phases of the evaluation cycle, primarily as a support tool ("human-in-the-loop"), especially in the operationalisation, report preparation and results dissemination phases. Use cases include the analysis of large volumes of administrative and textual data through Machine Learning (ML) and Natural Language Processing (NLP), the performance of simulations and counterfactual analyses, the potential for real-time monitoring, and the use of Large Language Models (LLMs) for synthesis or visualisation tasks, among others.

Conclusions and implications: Current evidence points toward the predominance of human-machine collaboration models (human-in-the-loop), indicating that realising the benefits of AI in this field does not involve total automation, but rather strategic, critically reflective and contextually adapted implementation.

To cite this article: Branchini, B., & Vallina Acha, B. (2025). Use and application of artificial intelligence in public policy evaluation. A Scoping Review. *Journal of Policy Evaluation, 1*, 114-133. https://doi.org/10.4995/jpeval.2025.23837

1. Introduction

In recent years, various public administrations in Europe have begun to use Artificial Intelligence (AI) tools, whilst their application is expanding and growing exponentially (Tangi et al., 2022). Consequently, AI systems that aim to serve the public or social good are a prominent topic for research, and an explicit objective of many national and international strategies and regulatory proposals (Züger & Asghari, 2023).

The OECD defines AI as machine-based systems that infer from data, generating predictions or content capable of influencing physical or virtual environments, with variable levels of autonomy and adaptability (Recommendation of the Council on Artificial Intelligence, 2024). The Spanish Strategy for Artificial Intelligence emphasises its capacity to resolve complex problems through advanced techniques that surpass the need for predefined instructions (Ministry for Digital Transformation and the Civil Service, 2024). The ISO/IEC 22989 standard conceives AI as mechanisms that emulate aspects of human intelligence to generate results oriented towards objectives defined by humans (International Organization for Standardization & International Electrotechnical Commission, 2022). The European Commission also emphasises the ability for environmental analysis and autonomous action geared towards human goals (Artificial Intelligence for Europe, 2018).

Functionally, AI can be understood as software that makes decisions in a manner similar to human intelligence. Unlike traditional programming, where the developer specifies each computational step, AI systems learn patterns from data without following explicit rules (Moyano-Arias et al., 2024). Machine learning represents a fundamental shift: resolution processes are not predetermined, but rather emerge from training, allowing systems to improve without necessarily modifying their code (Goodfellow et al., 2016).

This capacity for learning and adaptation poses specific challenges in verification, validation and explainability, recognised as essential elements – see the ISO/IEC 22989 standard. In complex models, these processes require specialised approaches to guarantee reliability, transparency and safety (International Organization for Standardization & International Electrotechnical Commission, 2022). The adaptation mechanism reinforces the centrality of data and connects with the problem of opacity; models prove difficult to interpret even for their designers, a phenomenon known as the "black box" problem (Carabantes, 2020), as will be seen. The explainability or XAI (Barredo Arrieta et al., 2020) – i.e. interpretability – of models proves crucial in applications where transparency and accountability are fundamental.

In recent years, the emergence of AI and big data has also begun to transform the field of public policy evaluation (Newman & Mintrom, 2023), understood as the systematic process of assessing the design, implementation and results of governmental interventions. Although there is growing interest in using AI to support and improve policy-making and service delivery, its use in evaluation practice has not been reflected in robust and systematised evidence, and there are relatively few peer-reviewed articles on digital technologies and their implications for evaluation practice (Bohni Nielsen et al., 2024). Existing scientific evidence recognises the potential of AI to improve decision-making, but its effective implementation in policy evaluation remains incipient (Yar et al., 2024), with limited evidence regarding its actual impact. Concerns about the risks of AI are based more on ideological considerations than on verifiable empirical data (Stadelmann, 2025), with insufficient theoretical foundation and limited empirical evidence in research on the use of AI in public administration (Bertolucci, 2024). Similarly, disciplinary fragmentation around AI in government requires unified frameworks that can address the absence of a common theoretical base for evaluating its public impact (Straub et al., 2023).

It is therefore necessary to systematise existing applications of AI in the evaluation process, as well as to explore their emerging implications and potentialities. Our study attempts to respond to this need through an exploratory literature review, whose main objective is to identify, analyse and systematise existing applications, particularly in European administrative contexts comparable to Spain, and framed within the public policy cycle, with special attention to the areas of employment and social inclusion.

These sectors, essential pillars of citizen wellbeing, face substantial transformations in the current context of technological, economic and social change. The study of the potential application of AI in these areas proves especially relevant for developing responses to emerging social problems, improving the efficacy of public intervention and ensuring its application in an equitable and sustainable manner, particularly given the pressure of limited resources and growing demands.

In turn, the study pursues a second objective: to detect and characterise the theoretical, methodological and empirical gaps that shape priority areas for future research.

The lack of robust and systematised evidence on the use of AI in public policy evaluation not only limits its informed adoption, but represents a significant opportunity to construct a research agenda that articulates new thematic areas, promotes interdisciplinary approaches and serves as a foundation for further studies in this field. This dual approach allows us not only to map the state of the art, but also to strategically guide the advancement of knowledge in a critical dimension of contemporary digital governance.

The scoping review has focused on the application of AI throughout the evaluation process, understood as a systematic and structured process compromising several interrelated phases (Better Evaluation, 2014; European Commission, 2024; Ligero Lasa, 2015) which we summarise here as: commissioning and management; definition of the evaluand; choice of evaluation approach; operationalisation; synthesis and judgement; report preparation and use.

The study, which is analytical and descriptive in nature, is structured around the following research question: How is AI currently being applied in the evaluation phase of employment and social inclusion public policies at the European level and in administrative systems comparable/similar to those in Spain?

The specific contribution of this research lies in providing a systematic mapping of AI applications in the evaluation of employment and social inclusion public policies, thus generating a structured knowledge base to inform both professional practice and the development of a future research agenda.

2. Materials and methods

This study is a *Scoping Review*. Given its research question, this methodology was considered the most suitable for addressing the current state of AI application in public policy evaluation. Furthermore, the emerging and evolving nature of the integration of disruptive technologies in public administration presents pioneering experiences coexisting with extensive areas of potential application. The scoping review constitutes a type of literature review aimed at exploring the available evidence on a topic, with the fundamental purpose of mapping concepts, identifying resources and research carried out in diverse contexts, both academic and non-academic, to detect opportunities and avenues for analysis and application. For its implementation, the methodological framework was followed with the clarifications and improvements introduced by Levac (2010) on that originally proposed by Arksey and O'Malley (Levac et al., 2010).

2.1 Inclusion and exclusion criteria

The thematic focus comprises the current state of AI use in the evaluation of social policies, with particular emphasis on interventions related to employment and social inclusion. Although the primary objective has been to map existing or ongoing applications documented in the literature, certain trends identified as emerging in the analysed studies and future directions for their application have been identified. These should be understood as a result of the analysis of the existing literature, not as the product of a formal prospective exercise. The selection of documents has been based on clearly defined parameters. Academic publications (articles and monographs), peer-reviewed studies (including doctoral theses) and governmental and third sector grey literature were included, provided they reflected specific and transferable use cases, as well as institutional documents, strategic plans and diagnoses. The time frame was defined as from 2019 to the present, a decision based on technical progress, institutional adoption and certain relevant milestones (e.g., the ALTAI consensus documents on AI ethics and European and national legislative developments). Studies with diverse methodological approaches (quantitative, qualitative, mixed and, with certain reservations, theoretical) were considered, sourced from academic and generalist databases (Google Scholar, Google, WOS and SCOPUS), in Spanish and English. In contrast, newspaper articles, blog posts, short publications, reports on education, health or infrastructure policies, as well as those referring to business domains (marketing, finance, etc.) or those prior to 2019 were excluded.

In terms of subject matter, documents that explicitly address AI applications in social policy evaluation, their effect and impacts, and use cases in social programmes were reviewed. Therefore, the search focused on terms such as *evaluation, impact, artificial intelligence, machine learning, Big Data, social policy, social inclusion, employment policy, and* *policy cycle*. The systematic search was conducted in the engines and sources previously mentioned through replicable and structured Boolean algorithms (detailed in Annex 1), adapted to both academic and generalist sources. This duplication of sources is appropriate in order to also cover institutional literature and use cases that have been tested and documented in practice.

2.2 Selection procedures

The procedure for the final selection of publications and information extraction comprised several sequential phases:

1. Preliminary exploration of results by skimming titles and abstracts, allowing rapid assessment of potential relevance; this does not constitute sufficient inclusion criteria.

2. Application of inclusion and exclusion criteria: type of publication, methodological approach, time frame, languages and geographical scope, including grey literature.

3. Prioritisation based on relevance criteria, operationalised on a categorical scale from 0 (minimal relevance/ irrelevant) to 4 (high relevance) applied to abstracts.

4. Careful review of literature on highly relevant articles (score 3-4), responding to principles of contextual relevance, use cases, transferability and timeliness.

The relevance criterion enabled the identification of studies with greater analytical potential for exhaustive review, prioritising cases of practical implementation in the European context. Its operationalisation was established through a four-level categorical scale (0-4).

Table 1. Selection criteria (relevance)

Score	Level	Description
0	Not relevant	Meets formal inclusion criteria, but presented tangential or excessively theoretical content with no operational connection to the application of AI in employment or social inclusion public policies.
1	Slightly relevant	Addresses AI implementation from generalist perspectives without specific European context or with limited applicability in employment and inclusion frameworks.
2	Relevant	Incorporates significant elements regarding AI application in these policies, with valuable theoretical and empirical contributions, though with a predominantly conceptual orientation.
3 to 4	Highly relevant	Presents critical relevance on the implementation of AI in the evaluation of employment and inclusion policies with specific cases, contributing substantially to understanding practical applications of AI in these areas within the European context.

Source: Author's own elaboration.

The thematic analysis of selected sources was carried out using the qualitative analysis tool MAXQDA (versions 2024 and 2022) and, as support for systematic information processing, generative AI tools (Claude Sonnet 3.7 and Gemini 2.5 Pro) were used for certain categorisation and preliminary synthesis tasks, always under human supervision and with cross-checking of content. Additionally, Claude Sonnet 3.7 with custom styles was used for specific writing tasks, focused on improving the clarity of exposition, argumentative structure and narrative continuity of the text.

To support Claude Sonnet 3.7 Extended in the review of articles, a project called "ClaudeReview" was implemented on the same platform for the assisted processing of selected documents. It was used to process complete documents and extract structured information according to a predefined template that included essential fields such as bibliographic information, central theme, methodological approach, geographical and institutional context, description of the intervention, type of AI applied, documented results, identified limitations and lessons learnt. The project also enabled the extraction of organised information from tables exported from MAXQDA following completely manual coding (without using the integrated AI tool), as well as directly from articles and reports in PDF format.

3. Results

The process of selecting relevant literature for this scoping review was carried out using a systematic and sequential approach: structured searches in academic databases were combined with complementary strategies ensuring exhaustiveness and relevance of the information collected.

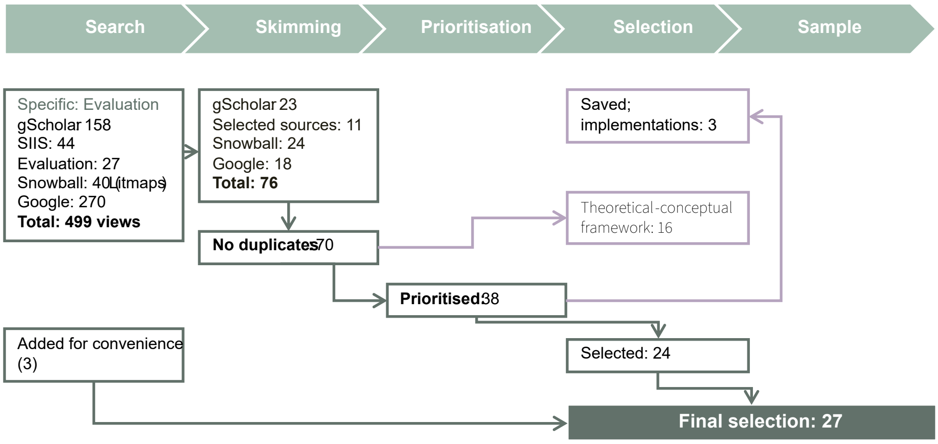
Specific searches were conducted across multiple sources: Google Scholar (158 results), conventional Google (270 results), the specialised SIIS (Servicio de Información e Investigación Social - Centro de Documentación y Estudios) [Social Information and Research Service - Documentation and Studies Centre] database (44 results) and the academic journal Evaluation (27 results). Additionally, the snowball sampling technique was applied to 16 bibliographic references extracted from the aforementioned work, which were entered into the Litmaps tool to identify related citations (40 references). Timeframe (publications after 2019) and thematic (terms such as data, AI, generative AI, LLM or big data in the title) inclusion criteria were established, yielding 24 additional results, many of them contextual or theoretical in nature.

A total of 499 results were found. After skimming - that is, diagonally reading the title and abstract - 76 results remained; the two reviewers coordinated to reach agreements using the Rayyan tool. Once duplicates were eliminated, 70 studies remained, of which 38 were prioritised for detailed analysis.

The final refinement process led to the selection of 24 studies, to which three more were added: two resulting from breaking down two chapters (York, 2024; Ziulu et al., 2024) from a recent work on evaluation, big data and AI (Bohni Nielsen et al., 2024) and another that appeared incidentally through a thematic notification in Google Scholar, published at the end of February 2025 (Babšek et al., 2025). The final sample therefore consisted of 27 studies, 16 of which were specifically focused on evaluation.

In addition, four publications prior to the cut-off date established for the review were incorporated as a conceptual framework as they were considered conceptually valuable; one of these documents contained references to eight articles that were also analysed. It should be noted that not all documents necessarily included specific or differentiated use cases, and that throughout the review, 16 references of a predominantly theoretical-conceptual nature were identified that were cited repeatedly.

Figure 1. Selection process chart



Source: Author's own elaboration.

3.1 General considerations regarding AI and Big Data

Emerging technologies, such as machine learning, natural language processing and big data analysis, possess significant transformative potential, although there is no clear evidence of their degree of adoption and use in the evaluation profession. In the field of evaluation, the application of AI essentially covers the dyad of 'augmentation' and 'automation' of certain tasks. Similarly, large language models (LLMs) serve workflows across the entire evaluation process, from the design of the terms of reference (commission) to the drafting of reports and visualisations (report and use).

Effective AI integration into evaluation rarely implies total automation; instead, a model often described by the concept *human-in-the-loop* (HILT) emerges. Generally, the application of AI appears to be mediated by instrumental considerations and within these HILT interactions, which constitute the most frequent or predominant model - resembling other more general workflows (Bohni Nielsen et al., 2024, p. 49).

Analysis of documented experiences reveals a diverse technological ecosystem that draws primarily on machine learning (ML) - both in its supervised variants (SML), where the system is trained to replicate human codifications or classifications, and in its unsupervised (UML) variants, often used to identify patterns or group concepts in large volumes of unlabelled data. Alongside ML, we find relevant applications of natural language processing (NLP), generative AI and visual analysis.

Through a very preliminary and descriptive analysis, with the cautions required by methodological limitations, and without taking into account intra-regional variations, it can be observed in a very general manner that there is a certain predominance of generative AI in North America, of machine learning (both supervised and unsupervised) in Europe, mentions of neural networks with a broad scope in Asia, and a focus on language analysis and processing and predictive modelling in the Global South. For an analysis of territorial contexts, the specific variables of each country and possible influencing factors should be analysed in depth. A significant portion of the literature identified in relation to Global South countries comes from the field of international development cooperation and multilateral financing organisations, such as the World Bank or the United Nations system; these institutions have played a key role in consolidating methodological frameworks and generating empirical evidence by providing specialised technical capabilities and sustained resources.

The selected articles do not allow us to identify the application of AI unambiguously in one specific evaluation phase or another, as they accompany all stages of the evaluative process transversally, focusing especially on the use of data. The cases illustrated, in this regard, highlight the opportunities opened up by AI throughout the entire data lifecycle.

Table 2. Summary of findings

	Use case	Technology used	Risks and limitations
(Alexander, 2022)	Exploration of AI-enhanced automated decision-making systems (ADM) in the public sector and public administration.	Automated decision-making systems (ADM); Big Data.	Ethical limitations (ethics vs efficiency; interinstitutional/organisational inequality; challenges of explainability, algorithmic bias and accountability).
(Arguelles Toache, 2023)	Evaluation of social development and state of social development interventions (Poverty maps; secondarily, Distancia2).	ML, computer vision using satellite imagery, and correlation algorithms. Georeferenced data.	Sampling biases and prioritisation of sources (digital media over others), limited validation, differential effectiveness depending on context. Protection of sensitive data.
(Babšek et al., 2025)	Analysis of the current state of AI adoption in public administration through bibliometrics and content analysis from Scopus.	Multiple subsets of AI are analysed: generic AI, LLMs, chatbots, Big Data analytics, ML/DL.	Lack of ethical governance frameworks for transparency, privacy, bias, fairness and accountability. Black box.
(Bamberger & York, 2020)	Analytical essay with theoretical and applied components, based on a conceptual review and case studies drawing on the Fourth African Industrial Revolution.	Discussion of ML (quasi-experiments, automation) and NLP (qualitative analysis).	Lack of familiarity among evaluators, costs, organisational restructuring, resistance, digital divide, ethics/data privacy, interterritorial and inter-organisational inequality; inter-organisational competences.
(Bilbao-Goyoaga, 2023)	Application of supervised machine learning (SML) to assess the impact of the Minimum Vital Income (IMV).	SML synthetic control (counterfactual; Spain with/without IMV).	High non-take-up (57%), reduction in regional budgets, delay and data limitations, possible placebo effect.
(Bouyousfi & Ouedraogo, 2024)	Systematic review of 102 selected primary articles on AI and Big Data in evaluation, from 2004 to 2023.	Miscellaneous: Big Data and ML (UML and SMLA), LLMs, sentiment analysis, etc.	Issues of data representativeness; difficulties in establishing causality (versus correlation; methodological aspects); opacity of algorithms and complex models (black box). Ethical issues (bias, accountability).
(Carlizzi & Quattrone, 2023)	Analysis of the impact of AI on the transformation of the public policy cycle; new model called 'Precision ePolicy Cycle' (Data Driven Decision Making). Real-time adaptation of precision micro-policies.	Miscellaneous: large-scale data analysis and categorisation, decision support systems, ML, XAI, etc., framed within the public policy cycle.	Risks of data manipulation; limited understanding of AI; black boxes; need to readapt processes and organisations; risk of technocracy; unlawful profiling.
(Ferretti, 2023)	Analytical-exploratory essay on the use of ChatGPT at different stages of the evaluation process. Includes practical examples of application.	LLM (ChatGPT-3.5 or 4).	Hallucinations, lack of factual verification, biases in training, risk of replacing critical judgement, privacy/ethics, lack of control.

(Franzen, Quang, Cuong, et al., 2022)	Application of AI at the World Bank (IEG) to automate and improve content analysis and qualitative synthesis in evaluations of complex programmes based on theories of change.	ML (text classification), UML (topic modelling), t-SNE, knowledge graphs.	Initial time investment, manual text extraction, unbalanced data, graph complexity, incomplete evidence.
(Hasan Chy & Nana Buadi, 2024)	Analysis of the transformative potential of ML to improve the processes of policy formulation and evaluation.	ML, data-driven policy, sentiment analysis, predictive analytics.	Data quality issues (incompleteness and inconsistency), algorithmic bias, XAI/ interpretability and transparency (black box).
(Head et al., 2023)	Theoretical analysis and conceptual review with impact estimates based on expert opinions. Applications of large language models (LLMs) in evaluation.	LLMs, ChatGPT.	Technical limitations: factual inaccuracy ("hallucinations"), poor numerical performance, processing opacity (black box problem). Ethical issues: perpetuation of biases, exclusion of minority languages, environmental cost.
(Jacob, 2025)	Theoretical-conceptual and prospective analysis on the use of AI in the evaluation of public policies: from augmented evaluation to potential automation.	Miscellaneous: generative AI and LLMs (e.g., ChatGPT), NLP, UML/SML.	Technical challenges: black box problem, lack of transparency and explainability in AI systems, hallucinations in generative AI, need for human verification (and thus training), security. Ethical issues: possible reinforcement of social and economic inequalities, algorithmic biases, protection of personal data.
(Kates & Wilson, 2023)	Conceptual analysis of the risks and opportunities that the emergence of generative AI implies in evaluation practice, specifically in the preparation of research reports and evaluation design.	AI is not applied, but it focuses on generative AI.	The evaluation community must familiarise itself with LLMs and tools; however, these raise issues of transparency, veracity and hallucinations, or biases, among other risks, which highlight the need for human critical thinking.
(MacArthur et al., 2025)	Case study on the use of “evaluative people” as a methodological tool to analyse complex qualitative data in programme evaluations and generate actionable recommendations.	NLP (Latent Dirichlet Allocation or LDA) and Structural Topic Modelling (STM) to validate manually created personas.	Determination of thematic clusters; requires high-quality data. Possible “embellishment” in participants’ retrospective narratives; anonymity problems with census sampling (partially mitigated).
(Mason, 2023)	Conceptual analysis and development of theoretical frameworks to classify evaluation competencies according to their vulnerability to AI.	No specific technologies are applied, although it focuses mainly on generative AI.	Recognition of the ‘Moravec's Paradox': the difficulty of predicting which tasks will be easy or difficult for AI to automate, with constant surprises in technological development.
(Picciotto, 2020)	Critical essay and conceptual analysis on the opportunities and challenges of integrating Big Data into evaluation.	Predictive algorithms and Big Data systems.	Critique of the correlation-without-theory approach ("the end of theory"); problems of representativeness and digital exclusion; ethical dilemmas of privacy; tension between precision and fairness in algorithms; social justice and governance tensions.

(Powell et al., 2024)	Presentation of causal mapping as a methodological tool to capture, code and visualise causal claims.	NLP processing as a complementary tool.	Ambiguity in the labels of causal factors; risk of the 'transitivity trap' (inferring indirect causal relationships across different contexts) and difficulty in systematically and consistently capturing the size of the effect and the exact type of causal influence (whether a cause was necessary, sufficient, etc.); ambiguity in the precise meaning of labels assigned to causal factors; experiential claims; loss of detail and granularity.
(Raveh et al., 2020)	Automatic detection of trends in large volumes of performance reports (1987-2017); automated and quantitative methodology.	Big Data; NLP, UML (classification), dimensionality reduction, graph visualisation.	Innovative methodology (no prior hypotheses); need for programming, adaptation (user interface, GUI), selection of the number of terms and tokenisation. Possible omission of terms due to typographical errors.
(Schimpf et al., 2021)	Case-based modelling and scenario simulation (CBSS) for ex-post policy evaluation.	UML (k-means, SOMs), Monte Carlo simulation.	Limitations inherent in models, no definitive causality, requires good specification, methodology and data, depends on evaluator interaction.
(Stern, 2020)	Editorial; academic. Theoretical and methodological diversity in the field of evaluation.	Brief mention of articles included in a special issue of Evaluation (2020, 26(4)).	From a methodological perspective, Bayesian updating is explored to rigorously assess probabilities, the use of automated tools to analyse large volumes of unstructured data and the systematisation of good practices to evaluate quantitative impact.
(Tilton et al., 2023)	AI (chatbots) in evaluator training.	Generative AI and conversational assistants based on LLM (GPT-3.5 turbo), Retrieval-Augmented Generation (RAG).	Technical issues (translation, context, hallucinations), methodological (preliminary evaluation, without metrics), ethical (bias, equity of access).
(Valle-Cruz et al., 2020)	Review. Global implementation of AI technologies in governments and public administrations. Proposal of a Dynamic Public Policy Cycle (DPPC) model that incorporates AI in the form of an incremental spiral.	AI and ML (in general), Big Data.	Data obsolescence; opacity, digital divide and exclusion; potential algorithmic biases; lack of transparency; privacy violations; algorithm complexity; over-reliance on AI; goal displacement/responsibility attribution.
(York, 2024)	Application of Structural Causal Modelling (SCM) and Precision Analytics (PA) to improve the evaluation of social programmes.	SCM and PA; SML; georeferenced data.	False positives/negatives. Dependence on the quality and completeness of the available data. Technical limitations and need for continuous fine-tuning. Familiarity and receptivity with SCM/PA tools.

(York & Bamberger, 2020)	Analytical essay with theoretical and applied components, based on conceptual review and case studies, on the integration of data science into social and economic development programmes.	Miscellaneous: data mining, including satellite data, social network data or biometrics, IoT, AI and ML; predictive analytics; comprehensive data platforms.	Costs and investment; organisational challenges and familiarity with technologies; digital divide; ownership, access and use of data; data quality and validity (representativeness, construct validity, population covered, algorithmic confusion, etc.); biases (selection bias, researcher bias, etc.); extractive use of communities; role of theory and theoretical and methodological framework; opacity and black box.
(Ziulu et al., 2024)	Text mining and supervised ML for the thematic evaluation of indicators for the years 2010-20 of the World Banks' DB programme (Independent Evaluation Group, IEG).	Text mining and SML.	The limitations can affect data quality; difficulties in capturing complete textual data; need for interdisciplinary teams and continual skills updating.

Source: Author's own elaboration.

3.2 Use cases

The study's findings indicate that the incorporation of AI and the use of big data can profoundly reshape the field of public policy evaluation. This transformation manifests itself in different phases of the evaluation process, from planning to results presentation, generating new and more efficient capabilities (Hasan Chy & Nana Buadi, 2024).

In the design stage, AI has proven to be an effective tool for the structuring and preliminary exploration of information. Its application enables efficient analysis of large volumes of textual data and administrative records, as well as reducing the time spent on preparatory tasks (Jacob, 2025) and opening the door to large-scale textual source analysis, as demonstrated by the World Bank's project document analysis practices using NLP and supervised ML (Ziulu et al., 2024). These technologies (i) enable the review of background and theories of change and (ii) facilitate the formulation of performance indicators and terms of reference, and even permit the simulation of prospective scenarios (Bohni Nielsen et al., 2024; Franzen, Quang, Cuong, et al., 2022; Jacob, 2025; York, 2024).

New computational tools enable the integration and analysis of heterogeneous information from multiple sources, transforming both quantitative and qualitative analyses. In the quantitative domain, the use of massive administrative data has enabled the implementation of sophisticated methodologies, such as structural causal modelling (SCM), that allow spurious correlations to be distinguished from real causal relationships (Bohni Nielsen et al., 2024). These techniques, combined with machine learning tools, enable population segmentation and determine more precisely for whom and under what conditions public policies work (Bilbao-Goyoaga, 2023; Crato & Paruolo, 2019b). Likewise, Precision Analytics allows for quasi-experimental analyses, incorporating expert knowledge through a hybrid human-algorithm logic (Bamberger & York, 2020; Powell et al., 2024).

From a qualitative perspective, NLP enables processing large textual corpora in significantly reduced timeframes, while maintaining a high degree of interpretative accuracy (Head et al., 2023; Tilton et al., 2023). An example is the case of CoLoop (2023), a tool capable of identifying thematic patterns in seconds for tasks that traditionally required hours of manual coding (Jacob, 2025). These advances enable more agile and scalable evaluation without compromising analytical depth. Furthermore, UML models can be trained to identify recurring themes in institutional evaluations, as in the case of the World Bank's Doing Business project (Ziulu et al., 2024). The adaptive reuse of these models guarantees their sustainability and efficiency, recovering the initial costs through their transferability.

The use of simulations represents another significant line of advancement. Methodologies such as Case-Based Modelling and Scenario Simulation (CBSS), which integrates Monte Carlo simulations and unsupervised learning, enable exploration of counterfactual scenarios and the generation of future projections in complex systems (Schimpf et al., 2021). In Spain, for example, supervised machine learning was applied to construct a synthetic counterfactual scenario that simulates a country without a Minimum Living Income (*Ingreso Mínimo Vital*, IMV), with the aim of isolating the real impact of this policy (Bilbao-Goyoaga, 2023). Additionally, the employment of real-time data has contributed to the development of more dynamic and adaptive evaluations.

There has been a shift from approaches centred on one-off data collection to systems that enable continuous monitoring during policy implementation, generating near real-time feedback (Bohni Nielsen et al., 2024; Buttow, 2024). All of this enables iterative adjustment of interventions and promotes an understanding of evaluation as a continuous process integrated throughout the entire public policy cycle, from agenda definition to implementation and redesign (Carlizzi & Quattrone, 2023; Valle-Cruz et al., 2020). At the same time, large language models (LLMs) are expanding their field of application. In the planning stage, LLMs can assist in drafting terms of reference, formulating interview questions, visualising results and contextual analysis (Jacob, 2025). Despite the associated ethical and technical challenges (Ferretti, 2023; Kates & Wilson, 2023), generative AI offers notable potential for automating repetitive tasks, such as drafting technical reports and protocol documents (Mason, 2023). Furthermore, it improves access to evaluation findings through automatic generation of interactive visualisations, multilingual translations and formats tailored to different audiences, such as videos or podcasts (Jacob, 2025).

Another relevant finding is the move towards hybrid interaction between humans and machines in evaluation processes. This approach, conceptualised as human-in-the-loop or human-on-the-loop, seeks to preserve the centrality of expert judgement, contextual understanding and ethics in decision-making, while delegating computationally intensive tasks to algorithmic systems (Bohni Nielsen et al., 2024). The integration of APIs (Application Programming Interfaces) and machine-readable formats reinforces technological fluidity, facilitating system interoperability (Babšek et al., 2025; Mason, 2023).

The evidence gathered underscores the importance of maintaining informed human oversight, as well as consolidating a robust methodological infrastructure that combines technical precision with public accountability. Public policy evaluation is undergoing a structural transformation driven by the synergy between AI, big data and new methodological models, optimising the time and resources devoted to analysis and introducing a new evaluative logic characterised by scalability, real-time adaptability and customisation. At the same time, it poses epistemological and ethical challenges, especially with regard to contextual interpretation, model transparency and the sustainability of the solutions implemented.

4. Discussion

As has been observed, the prevailing trend in the literature advocates viewing AI as a complement, "a tool", and not a substitute for either traditional methods or the expert judgement of the evaluator (Bouyousfi & Ouedraogo, 2024; Kates & Wilson, 2023; Schimpf et al., 2021). The HITL approach is presented as dominant and, to a certain extent, fundamental, assuming that AI – any type of AI - supports specific and "mechanical" tasks (e.g., analysis of large data volumes, pattern identification) and that final interpretation, evaluative judgement, consideration of context and ethical validation fall to the human evaluator (Franzen, Quang, et al., 2022). Nevertheless, this is not without its difficulties, nor is it the final solution it appears to be: there is the risk that it becomes an "empty procedural shell" that legitimises the use of problematic algorithms without guaranteeing real protection of rights. Furthermore, the human capacity to effectively supervise complex algorithmic decisions is limited (Bohni Nielsen et al., 2024, pp. 252–253). Added to this is the risk of automation bias, where humans tend to over-rely on algorithmic results (Potasznik, 2023).

The promise of AI can only be realised responsibly if transparency, fairness and rigorous validation are prioritised. Human judgement must be technically informed and ethically conscious within a framework of robust interdisciplinary collaboration (Bamberger & York, 2020; Ferretti, 2023; Franzen, S. et al., 2022; Jacob, 2025; Mason, 2023; Schimpf et al., 2021).

The main risks to evaluation practice point precisely to the quality, transparency and privacy of data, and to the skills gap in a sector threatened by the risk of substitution and the epistemological tension of methodological pluralism threatened by a certain technological determinism and imperative of quantification.

4.1 Quality, transparency and privacy

Studies detect a potential perpetuation of historical biases and epistemic exclusions. The validity of AI models depends intrinsically on the quality and representativeness of training data. There is a significant risk of perpetuating or amplifying such biases if they reflect past discriminations or underrepresent certain groups (Bouyousfi & Ouedraogo, 2024; Elevati, 2025; Hasan Chy & Nana Buadi, 2024). Lack of representativeness can also stem from contextual factors, such as the phenomenon of non-take-up in social programmes, which introduces systematic biases not attributable to the intrinsic quality of the data (Bilbao-Goyoaga, 2023; Bouyousfi & Ouedraogo, 2024).

Highly complex models such as deep neural networks or large language models (LLMs) often function as "black boxes," hindering understanding of how they reach their conclusions (Hasan Chy & Nana Buadi, 2024; Jacob, 2025; Kates & Wilson, 2023). This opacity clashes with the principles of transparency and accountability, crucial in public policy evaluation, and forces us to choose - and this is a recognised compromise - between predictive accuracy (often higher in opaque models) and interpretability (Bohni Nielsen et al., 2024, p. 96). The emergence of Explainable AI (XAI) approaches seek to mitigate this problem, but it remains a key challenge.

Large language models have a significant risk of generating fabricated but plausible information by their very design (Ferretti, 2023; Head et al., 2023). This problem of veracity (Kates & Wilson, 2023) underscores the imperative need for human validation.

The use of large volumes of data, often of a personal or administrative nature, raises serious ethical concerns about informed consent, privacy, data ownership and the potential misuse of sensitive information; this is a particularly delicate matter when it comes to service provision and the implementation and enforcement of social policy (Bamberger & York, 2020; Bouyousfi & Ouedraogo, 2024; Hasan Chy & Nana Buadi, 2024).

4.2 Professional and epistemological challenges

The possibility of accessing large volumes of data from a diverse range of heterogeneous sources represents a major methodological change regarding the very concept of evaluation. Evaluation practice evokes challenges inherent to the profession, ranging from the need to acquire competencies at both organisational and individual levels, to epistemological challenges: methodological plurality, transparency, equity, and validation.

Skills gap

The skills gap is not just a diagnosis of shortcomings in Europe today: it is an assertion that poses problems between organisations, financial capacity, and regions. AI implementation entails significant costs in infrastructure, software and specialised personnel, within an environment characterised by resistance to change (Babšek et al., 2025; Bamberger & York, 2020), a lack of systematic evidence on successful implementations, and a lack of clear regulatory and conceptual frameworks (Babšek et al., 2025; Hasan Chy & Nana Buadi, 2024).

This new landscape has a direct impact on the competencies required of evaluators and, by extension, on the learning capacity of organisations (Mason, 2023), pointing to a need to strengthen interpersonal skills, critical thinking, contextual adaptability, and solid literacy in AI and data science (Bouyousfi & Ouedraogo, 2024; Picciotto, 2020).

Along the same lines, the automation of certain processes offers the advantage of freeing up time for the more creative and contextual aspects of evaluation work (Ferretti, 2023), precisely those which Mason identifies as less susceptible to automation.

Evaluation organisations would consequently need to actively foster this adaptation, integrating knowledge about AI into continuing professional development and curricula (Tilton et al., 2023), and promoting a culture of critical evaluation regarding AI's own effects on evaluation practice (Ferretti, 2023; Kates & Wilson, 2023).

The quantification imperative

A certain quantification imperative appears to emerge which, moreover, is not new. Although there are applications for qualitative research, the epistemological framework underlying many AI applications tends to favour the quantifiable. Algorithms are not merely technical tools, but devices that actively perform and reconfigure the social reality they purport to measure (Beer, 2017; Kitchin, 2014).

Despite its advanced analytical capabilities, AI does not in itself solve the challenges inherent in establishing robust causal inferences. Causality is not a matter of techniques and technologies, but of experimental design. Its instrumental use does not eliminate the persistent difficulty of isolating the effects of the programme from pre-existing participant characteristics or contextual factors (Bajgar & Criscuolo, 2019; Bouyousfi & Ouedraogo, 2024; Schimpf et al., 2021; Stern, 2020). Is there an objective social reality independent of the observer that can be captured using neutral instruments? Or can it only be understood through contextualised interpretation?

This creates tension with the methodological pluralism valued in evaluation and raises questions about the commensurability of qualitative phenomena and the place of non-quantifiable phenomena. Consequently, there is a risk of reification of social ontology, of assuming that components of social reality can be reduced to metrics (Picciotto, 2020). Kitchin already questioned the so-called data-driven empiricism that postulates "the end of theory" (Kitchin, 2014), noting that data never "speak for themselves" nor emerge devoid of prior conceptual frameworks. This reifying tendency does not constitute merely a methodological bias, but represents a very profound epistemological transformation.

Presenting the adoption of AI as a linear and inevitable progression would overlook contextual contingencies, non-linear adoption trajectories and the unforeseen effects of implementation, since understanding the impact of an AI system requires analysing not only its data and model, but also the material infrastructure, deployment context and its interaction with people, society, and institutions (Bohni Nielsen et al., 2024, p. 257).

4.3 Implications for organisations

The capacity of an organisation to acquire, interpret and apply knowledge in order to adapt and improve - i.e., organisational learning - is a fundamental factor in every evaluation and/or research institution. The emergence of AI is catalysing profound changes in this area. This transformation operates across three dimensions.

First, AI enables organisations to develop more acute situational awareness, both at the operational and strategic levels; this is amplified by the processing of massive volumes of information. Organisations can now detect emerging patterns and adapt with greater agility (Dwivedi et al., 2021). Second, AI generates and provides, inter alia, a repository of organisational memory that can store lessons learnt. A notable example is the use of automated text analysis and visualisation to detect trends and thematic changes in 30 years of performance reports (Raveh et al., 2020). Finally, this ability to reuse and build upon past knowledge can render investments in complex analytical models more sustainable (Ziulu et al., 2024); for instance, when model programming follows good programming practices, since AI systems enable economies of knowledge at scale through the reuse and adaptation of previously developed analytical models.

This transformative potential is conditioned by contextual factors that determine its effective adoption and implementation. Heterogeneity in infrastructures, competencies and resources generates institutional asymmetries: pre-existing technical capabilities determine the viability of implementation whilst normative and institutional frameworks establish the scope of possibilities for incorporating AI. As Franzen et al. (2022) note, the implementation of algorithmic systems may reinforce instrumental conceptions of evaluation, privileging quantifiable criteria over qualitative dimensions that are difficult to parametrise.

4.4 Future research directions

Based on the empirical findings of the study and the conceptual, methodological and operational challenges identified in the analysis, future lines of research of great interest are emerging:

Table 3. Future research directions

Ethics, governance and explainability	(Alexander, 2022; Babšek et al., 2025; Bouyousfi & Ouedraogo, 2024; Carlizzi & Quattrone, 2023; Hasan Chy & Nana Buadi, 2024; Head et al., 2023; Jacob, 2025; Kates & Wilson, 2023; Picciotto, 2020; York & Bamberger, 2020)
AI-assisted qualitative analytics: LLMs, SML classification, LDA/STM, t-SNE, graphs	(Ferretti, 2023; Franzen, Quang, Cuong, et al., 2022; Head et al., 2023; Kates & Wilson, 2023; MacArthur et al., 2025; Powell et al., 2024; Raveh et al., 2020; Ziulu et al., 2024)
Counterfactuals and AI-assisted impact evaluation: SML, SCM/graphs, causal mapping	(Bamberger & York, 2020; Bilbao-Goyoaga, 2023; Powell et al., 2024; York, 2024)
*as *ex-post* support(Schimpf et al., 2021)	
Prescriptive simulation and adaptive micro policies: CBSS/SOMs + Monte Carlo; scenarios; optimisation	(Carlizzi & Quattrone, 2023; Schimpf et al., 2021)
Capabilities and institutional learning	(Bamberger & York, 2020; Jacob, 2025; Kates & Wilson, 2023; Mason, 2023; Tilton et al., 2023; Valle-Cruz et al., 2020)
Data quality, biases and privacy: algorithmic auditing	(Alexander, 2022; Babšek et al., 2025; Bouyousfi & Ouedraogo, 2024; Hasan Chy & Nana Buadi, 2024; Head et al., 2023)

Source: Author's own elaboration.

4.5 Limitations of the study

The methodology employed in this scoping review presents a structured and theoretically sound approach, yet it has inherent limitations due to the type of study. The effectiveness of combining academic databases with generalist search engines is compromised by search algorithms, which prioritise results according to criteria that are not necessarily academic in nature; this potentially skewed the results towards certain types of documents. Likewise, the time lag between publication and appearance in search results excluded very recent contributions in a rapidly evolving field. A critical factor concerns the geographical contexts examined, conditioned by linguistic limitations and the focus on transferability to the Spanish context, which may have obscured relevant developments in non-English-speaking or non-Spanish-speaking regions.

The exclusion of journalistic sources and blogs, justified by the empirical rigour sought, probably omitted early indicators of innovation that often function as "radars". Take, for example, the fact that significant publications such as the recent Commission report (Directorate-General for Employment, Social Affairs and Inclusion, European Commission et al., 2025) or OECD analyses (Brioscú et al., 2024) were not included even though they offer complementary perspectives on the evolution and advanced technologies in the public sector.

Thus, the pursuit of scientific replicability, sacrifices, to some extent, comprehensiveness in capturing all relevant knowledge; a more robust approach would have included a preliminary phase of less structured narrative review, complemented by expert consultations to extract tacit knowledge and improve semantic sensitivity.

5. Conclusions

The present study set out to map a dynamic landscape: the application of AI in public policy evaluation, revealing that: (i) we recognise considerable transformative potential; and (ii) this integration comes with substantial challenges and non-trivial risks, including uncertainty about evaluation practice itself.

The implications for evaluation are profound and complex; they point to the need for professional adaptation, interdisciplinary collaboration and moving beyond a form of basic "literacy" in AI and data science, whilst simultaneously reinforcing the skills we consider to be uniquely human: critical thinking, interpersonal and social skills, contextual interpretation and ethical deliberation. Analysis of the literature reveals that effective AI integration in evaluation rarely translates into complete automation; rather, the predominant model is one of synergistic collaboration between human and artificial intelligence. For organisations - both those that evaluate and those that are evaluated - there is an imperative: the development of internal capabilities, robust data governance frameworks and an organisational culture of continuous adaptation to technological disruption.

Responsibly realising the potential of AI requires moving beyond mere technological determinism and the uncritical adoption of tools. There is a pressing need for a critical and reflexive integration strategy that is contextually alert and ethically grounded, balancing efficiency and analytical power with rigour, transparency and the centrality of human judgement.

We are aware that this study, due to its methodological design and the limitations of observation itself, offers a partial view of a global and uncertain phenomenon. Nevertheless, we believe that the findings and synthesis presented here clearly identify some priority areas for future research, and for further investigation into the contextual conditions that facilitate or hinder the effective implementation and transferability of these technologies in different institutional settings.

Acknowledgements

This study was conducted as part of a broader technical assistance service contracted by Besaldi, the Evaluation Body for Employment and Inclusion Policies (Basque Government).

The final translation from Spanish to English was reviewed and edited by Laura Sutton.

Funding

None.

Conflict of interests

The authors declare that they have no conflict of interest.

Author contribution

Barbara Branchini: Funding acquisition, conceptualisation, methodology, validation, review and editing and project administration.

Beatriz Vallina Acha: Investigation, data curation, and original draft.

REFERENCES

Alexander, W. (2022). Applying Artificial Intelligence to Public Sector Decision Making. Major Research Paper. University of Ottawa.

Arguelles Toache, E. (2023). Ventajas y desventajas del uso de la Inteligencia Artificial en el ciclo de las políticas públicas: Análisis de casos internacionales. Acta universitaria, 33. https://www.redalyc.org/journal/416/41677664054/html/

Babšek, M., Ravšelj, D., Umek, L., & Aristovnik, A. (2025). Artificial Intelligence Adoption in Public Administration: An Overview of Top-Cited Articles and Practical Applications. AI, 6(3), Article 3. https://doi.org/10.3390/ai6030044

Bajgar, M., & Criscuolo, C. (2019). Designing Evaluation of Modern Apprenticeships in Scotland. In N. Crato & P. Paruolo (Eds.), Data-Driven Policy Impact Evaluation (pp. 289–311). *Springer International Publishing*. https://doi.org/10.1007/978-3-319-78461-8_18

Bamberger, M., & York, P. (2020). Transforming Evaluation in the 4th Industrial Revolution: Exciting Opportunities and New Challenges (pp. 11–21). *eVALUation Matters, Second Quarter 2020*, 11-21. African Development Bank Group. https://idev.afdb.org/sites/default/files/documents/files/EM%20Q2-2020-article1-challenges%20and%20opportunities%204th%20industrial%20revolution%28En%29.pdf

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Beer, D. (2017). The social power of algorithms. Information, Communication & Society, 20(1), 1–13. https://doi.org/10.1080/1369118X.2016.1216147

Bertolucci, M. (2024). L'intelligence artificielle dans le secteur public: Revue de la littérature et programme de recherche: Gestion et Management Public, Vol. 12(3), 71–91. https://doi.org/10.3917/gmp.123.0071

Better Evaluation. (2014). Rainbow Framework. https://www.betterevaluation.org/frameworks-guides/rainbow-framework

Bilbao-Goyoaga, E. (2023). Perceptions Matter: Quasi-Experimental Evidence on the Effects of Spain's New Minimum Income on Households' Financial Wellbeing (No. Social Policy Working Paper 02-23; LSE Department of Social Policy.). https://www.lse.ac.uk/social-policy/Assets/Documents/PDF/working-paper-series/WPS-02-23-Eugenia-Bilbao-Goyoaga.pdf

Bohni Nielsen, S., Mazzeo Rinaldi, F., & Petersson, G. J. (2024). Artificial Intelligence and Evaluation: Emerging Technologies and Their Implications for Evaluation2 (1st ed.). Routledge. https://doi.org/10.4324/9781003512493

Bouyousfi, S. E., & Ouedraogo, M. (2024). Artificial intelligence and big data-driven evaluation research and practices: A systematic literature review. Evaluation, 13563890241289937. https://doi.org/10.1177/13563890241289937

Brioscú, A., Lauringson, A., Saint-Martin, A., & Xenogiani, T. (2024). A new dawn for Public Employment Services. Service Delivery in the age of Artificial Intelligence (No. 19; OECD Artificial Intelligence Papers). https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/a-new-dawn-for-public-employment-services_25e1e70e/5dc3eb8e-en.pdf

Buttow, C. V. (2024). Data-Driven Policy Making and Its Impacts on Regulation: A Study of the OECD Vision in the Light of Data Critical Studies. *European Journal of Risk Regulation*, 1–19. https://doi.org/10.1017/err.2024.73

Carabantes, M. (2020). Black-box artificial intelligence: An epistemological and critical analysis. AI & SOCIETY, 35(2), 309–317. https://doi.org/10.1007/s00146-019-00888-w

Carlizzi, D. N., & Quattrone, A. (2023). Artificial Intelligence and Data Governance for Precision ePolicy Cycle. In D. Marino & M. Monaca (Eds.), *Artificial Intelligence and Economics: The Key to the Future* (pp. 67–84). Springer International Publishing. https://doi.org/10.1007/978-3-031-14605-3_6

Crato, N., & Paruolo, P. (2019). The Power of Microdata: An Introduction. In N. Crato & P. Paruolo (Eds.), *Data-Driven Policy Impact Evaluation* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-319-78461-8_1

Directorate-General for Employment, Social Affairs and Inclusion, European Commission, ICF, & Willen, P. (2025). Opportunities of AI within PES processes and services: Exploring PES experiences, best practices and emerging business value. Publications Office. https://data.europa.eu/doi/10.2767/84293

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Elevati, C. (2025). 73—L'Intelligenza Artificiale come alleato strategico per i professionisti MEAL. LinkedIn. https://www.linkedin.com/pulse/73-lintelligenza-artificiale-come-alleato-per-i-meal-elevati-4xvff/?trackingId=rvljDuu5T5yjRNJAMx9XPQ%3D%3D

European Commission. (2018). Communication from the Commission: Artificial Intelligence for Europe (COM(2018) 237 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN

European Commission. (2024). Evaluation Handbook (2024). https://capacity4dev.europa.eu/library/evaluation-handbook-2024_en

Ferretti, S. (2023). Hacking by the prompt: Innovative ways to utilize ChatGPT for evaluators. *New Directions for Evaluation*, 2023(178–179), 73–84. https://doi.org/10.1002/ev.20557

Franzen, S., Quang, C., Schweizer, L., Budzier, A., Gold, J., Vellez, M., Ramirez, S., & Raimondo, E. (2022). Advanced Content Analysis: Can Artificial Intelligence Accelerate Theory-Driven Complex Program Evaluation? (*IEG Methods and Evaluation Capacity Development Working Paper Series*). International Bank for Reconstruction and Development / The World Bank.

Franzen, S., Quang, Cuong, Schweizer, L., Budzier, A., Gold, J., Vellez, M., Ramirez, S., & Raimondo, E. (2022). Advanced Content Analysis: Can Artificial Intelligence Accelerate Theory-Driven Complex Program Evaluation? (Independent Evaluation Group) [*IEG Methods and Evaluation Capacity Development Working Paper Series*]. World Bank.

Goodfellow, I., Courville, A., & Bengio, Y. (2016). Deep learning. The MIT Press.

Hasan Chy, M. K., & Nana Buadi, O. (2024). Role of Machine Learning in Policy Making and Evaluation. International Journal of Innovative Science and Research Technology (IJISRT), 456–463. https://doi.org/10.38124/ijisrt/IJISRT24OCT687

Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation*, 2023(178–179), 33–46. https://doi.org/10.1002/ev.20556

International Organization for Standardization & International Electrotechnical Commission. (2022). ISO/IEC 22989:2022(en), Information technology—Artificial intelligence—Artificial intelligence concepts and terminology. https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en

Jacob, S. (2025). Artificial Intelligence and the Future of Evaluation: From Augmented to Automated Evaluation. Digit. Gov.: Res. Pract., 6(1), 10:1-10:10. https://doi.org/10.1145/3696009

Kates, A. W., & Wilson, K. (2023). AI for Evaluators: Opportunities and Risks. Journal of MultiDisciplinary Evaluation, 19(45), 99–104. https://doi.org/10.56645/jmde.v19i45.907

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5(1), 69. https://doi.org/10.1186/1748-5908-5-69

Ligero Lasa, J. A. (2015). Tres métodos de evaluación de programas y servicios. Fundación Caja Madrid.

MacArthur, J., Moung, V., Carrard, N., & Willetts, J. (2025). Personas for program evaluation: Insights from a gender-focused evaluation in Cambodia. Evaluation, 31(1), 70–91. https://doi.org/10.1177/13563890241284425

Mason, S. (2023). Finding a safe zone in the highlands: Exploring evaluator competencies in the world of AI. *New Directions for Evaluation*, 2023(178–179), 11–22. https://doi.org/10.1002/ev.20561

Ministerio para la Transformación Digital y de la Función Pública. (2024). Estrategia de Inteligencia Artificial 2024. https://portal.mineco.gob.es/es-es/digitalizacionIA/Documents/Estrategia_IA_2024.pdf

Moyano-Arias, R. J., Salazar-Alvarez, E. G., & Toalombo-Vargas, V. M. (2024). Matemáticas Aplicadas a la Programación: Una Revisión sobre la Solución de Algoritmos Complejos. MQRInvestigar, 8(4), 3667–3692. https://doi.org/10.56048/MQR20225.8.4.2024.3667-3692

Newman, J., & Mintrom, M. (2023). Mapping the discourse on evidence-based policy, artificial intelligence, and the ethical practice of policy analysis. *Journal of European Public Policy*, 30(9), 1839–1859. https://doi.org/10.1080/13501763.2023.2193223

Picciotto, R. (2020). Evaluation and the Big Data Challenge. *American Journal of Evaluation,* 41(2), 166–181. https://doi.org/10.1177/1098214019850334

Potasznik, A. (2023). ABCs: Differentiating Algorithmic Bias, Automation Bias, and Automation Complacency. 2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (Ethics), 1–5. https://doi.org/10.1109/ETHICS57328.2023.10155094

Powell, S., Copestake, J., & Remnant, F. (2024). Causal mapping for evaluators. Evaluation, 30(1), 100–119. https://doi.org/10.1177/13563890231196601

Raveh, E., Ofek, Y., Bekkerman, R., & Cohen, H. (2020). Applying Big Data visualization to detect trends in 30 years of performance reports. Evaluation, 26(4), 516–540. https://doi.org/10.1177/1356389020905322

Recommendation of the Council on Artificial Intelligence, No. OECD/LEGAL/0449. Compendium of Legal Instruments of the OECD (2024). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Schimpf, C., Barbrook-Johnson, P., & Castellani, B. (2021). Cased-based modelling and scenario simulation for ex-post evaluation. Evaluation, 27(1), 116–137. https://doi.org/10.1177/1356389020978490

Stadelmann, T. (2025). Evidence-based AI risk assessment for public policy. *Public Money & Management, 1–3.* https://doi.org/10.1080/09540962.2025.2541304

Stern, E. (2020). Editorial. Evaluation, 26(4), 401–403. https://doi.org/10.1177/1356389020966442

Straub, V. J., Morgan, D., Bright, J., & Margetts, H. (2023). Artificial intelligence in government: Concepts, standards, and a unified framework. *Government Information Quarterly*, 40(4), 101881. https://doi.org/10.1016/j.giq.2023.101881

Tangi, L., van Noordt, C., Combetto, M., Gattwinkel, D., & Pignatelli, F. (2022). AI Watch: European landscape on the use of artificial intelligence by the public sector. European Commission. *Publications Office of the European Union*. https://data.europa.eu/doi/10.2760/39336

Tilton, Z., Lavelle, J. M., Ford, T., & Montenegro, M. (2023). Artificial intelligence and the future of evaluation education: Possibilities and prototypes. *New Directions for Evaluation*. https://doi.org/10.1002/EV.20564

Valle-Cruz, D., Criado, J. I., Sandoval-Almazán, R., & Ruvalcaba-Gomez, E. A. (2020). Assessing the public policy-cycle framework in the age of artificial intelligence: From agenda-setting to policy evaluation. *Government Information Quarterly*, *37*(4), 101509. https://doi.org/10.1016/j.giq.2020.101509

Yar, M. A., Hamdan, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). Governing with Intelligence: The Impact of Artificial Intelligence on Policy Development. Information, 15(9), 556. https://doi.org/10.3390/info15090556

York, P. (2024). The Future of Evaluation Analytics. In S. Bohni Nielsen, F. Mazzeo Rinaldi, & G. J. Petersson, *Artificial Intelligence and Evaluation* (1st ed., pp. 219–241). Routledge. https://doi.org/10.4324/9781003512493-11

York, P., & Bamberger, M. (2020). Measuring Results and Impact in the Age of Big Data: The Nexus of Evaluation, Analytics, and Digital Technology. The Rockefeller Foundation. https://www.rockefellerfoundation.org/reports/measuring-results-and-impact-in-the-age-of-big-data-the-nexus-of-evaluation-analytics-and-digital-technology/

Ziulu, V., Anuj, H., Hagh, A., Raimondo, E., & Vaessen, J. (2024). Extracting Meaning from Textual Data for Evaluation. In S. Bohni Nielsen, F. Mazzeo Rinaldi, & G. J. Petersson, *Artificial Intelligence and Evaluation* (1st ed., pp. 78–102). Routledge. https://doi.org/10.4324/9781003512493-5

Züger, T., & Asghari, H. (2023). AI for the public. How public interest theory shifts the discourse on AI. *AI & Society*, 38(2), 815–828. https://doi.org/10.1007/s00146-022-01480-5