Journal of Health Informatics in Developing Countries
http://www.jhidc.org/
Vol. 19 No. 2, 2025

Submitted: May 08th, 2025 Accepted: July 18th, 2025

Enhancing Public Sector Decision-Making through Artificial Intelligence Models: A Comparative Study

Saja Alhosan[1,*] and Othman Alsalloum [1]

[1] *King Saud University, Riyadh, Saudi Arabia.*

Abstract

As governments worldwide embrace digital transformation, the role of artificial intelligence (AI) in public policy formulation and analysis has gained unprecedented relevance. This study explores the capabilities and limitations of two advanced AI models (customized ChatGPT and DeepSeek) as decision-support tools. Briefing notes were generated using three different approaches: one by human policy analyst and two by AI models. The aim was to evaluate whether contemporary natural language processing (NLP) technologies can produce briefing notes that are relevant and useful for public policy decision-making. The AI-generated content was tested through simulated policy scenarios to assess performance in tasks such as information retrieval, stakeholder-specific communication, policy brief generation, and scenario analysis. To ensure a robust evaluation, a panel of subject-matter experts assessed the quality of all briefing notes using a structured heuristic evaluation rubric. Results indicate that AI model can enhance analytical capacity, improve policy document drafting, and foster more responsive decision-making. However, the study also identifies critical challenges, including model bias, explainability deficits, and the need for sustained human oversight. Drawing the importance of hybrid governance frameworks that combine AI tools with institutional safeguards. The findings contribute to ongoing discussions on ethical AI integration and provide actionable recommendations for responsibly incorporating large language models into public sector workflows, especially in digitally transforming nations.

Keywords: ChatGPT; DeepSeek; Public Sector; AI Policy Analysis; Natural Language Processing; Decision making; Governance Innovation

*Saja Alhosan - King Saud University, Riyadh, Saudi Arabia; Email: szalhosan@gmail.com .

1. Introduction

In recent years, integrating artificial intelligence (AI) in the public sector has emerged as a transformative approach to enhancing the efficiency, accuracy, and scalability of decision-making processes. Among the most promising tools are large language models (LLMs), such as OpenAI's ChatGPT and Deepseek which leverage advanced natural language processing (NLP) to simulate human-like reasoning and language generation. Governments and public institutions are increasingly exploring how these models can support activities such as policy formulation, service delivery, risk assessment, and stakeholder engagement (Binns et al., 2018; OECD, 2021; Kankanhalli et al., 2019).

The decision-making landscape in the public sector is complex, characterized by competing interests, uncertainty, large-scale data, and the need for transparency and accountability (Head, 2010; Janssen & Kuk, 2016). In this context, AI can augment human decision-makers by synthesizing vast datasets, providing rapid access to relevant policy insights, and generating scenario-based analyses that would otherwise require substantial time and human resources (Wirtz & Müller, 2019; Chen et al., 2020; Araya, 2019). ChatGPT, with its flexible conversational interface and large-scale training, offers the potential to bridge the gap between technical data analysis and policy communication (Brynjolfsson & McAfee, 2017). While DeepSeek sets a new standard for effectiveness, performance, and ethical consideration (Singh et al., 2025).

Recent developments in digital governance have shown how AI can improve service efficiency, transparency, and citizen engagement (UN DESA, 2022; Eggers & Bellman, 2015). For example, Estonia and South Korea have incorporated AI into administrative processes with encouraging results in efficiency and satisfaction (OECD, 2021; Kankanhalli et al., 2019; Mergel et al., 2019). Similarly, Saudi Arabia's Vision 2030 emphasizes digital transformation and innovation in public service delivery, creating fertile ground for integrating AI tools (Alotaibi et al., 2022; Khorshid et al., 2023). Despite the growing appeal of such technologies, significant concerns remain regarding algorithmic bias, lack of transparency, data privacy, and the potential erosion of human discretion in public decision-making processes (Eubanks, 2018; Cobbe et al., 2021; Rahwan, 2017).

This study seeks to contribute to the growing body of literature by empirically evaluating a customized version of ChatGPT and DeepSeek, specifically trained on policy-relevant datasets, to assess its potential as a decision support tool in the public sector. The key research question guiding this study is: How effectively can a customized ChatGPT and DeepSeek model support policy analysis and decision-making tasks? By addressing this question through simulated policy scenarios and evaluation, the study aims to inform the development and responsible deployment of AI models in the public sector.

2. Literature Review

In recent years, the application of AI in public administration has gained momentum, driven by its promise to enhance efficiency, responsiveness, and data-driven decision-making (Longo, 2024). AI offers promising tools for modeling policy options, synthesizing stakeholder feedback, and generating briefing documents. However, its integration also raises critical concerns regarding transparency, accountability, data privacy, and the preservation of human judgment in sensitive decision-making contexts. Longo (2024) specifically examined Canada's public service experience, identifying how AI supports policy analysis through modeling options, synthesizing stakeholder feedback, and drafting briefing documents. His work emphasizes the dual nature of AI, as both an enabler and a disruptor within traditional bureaucratic frameworks. In addition, Jungwirth and Haluza (2023) conducted a feasibility study exploring GPT-3's role in public health, demonstrating that it can generate and summarize relevant content, though often with questionable source attribution. Their findings suggest that AI models can complement but not replace human researchers in public sector tasks. Nzobonimpa et al. (2024) similarly noted that while LLMs like ChatGPT exhibit proficiency in generating linguistically and structurally coherent content, they fall short in capturing the depth and contextual nuance needed for high-stakes policy decisions. Safaei and Longo (2024) further echoed this point, underscoring that contemporary NLP tools are not yet equipped to independently produce reliable policy briefings. Several recent studies have focused on ChatGPT's qualitative capabilities. Wachinger et al. (2024) found that the model performs well in identifying descriptive themes and can contribute to theory-driven interpretation, when guided by critical human oversight. This finding echoes earlier work by Doshi et al. (2023), van Manen (2023), Wang and Chen (2024), who emphasize the importance of reflective practice and ethical awareness when integrating AI into academic and governmental research. Nonetheless, Estrada et al. (2023) identified notable limitations in the model's ability to handle abstract thinking and emotional intelligence reaffirming the importance of human involvement in tasks requiring ethical reasoning and cultural sensitivity. These limitations highlight the continued need for human expertise in tasks requiring ethical judgment, creative ideation, and culturally sensitive reasoning. While these findings highlight the utility of general purpose LLMs like ChatGPT, more recent research has explored the role of customized or regionally adapted models in governmental contexts. For example, Ke et al. (2025) conducted a comparative study of DeepSeek-R1 and GPT-4o in formulating policy recommendations for China's social security system. Their results demonstrated that DeepSeek-R1 outperformed GPT-4o across multiple evaluation metrics, particularly in generating contextually relevant policy insights. However, the study also noted limitations in the models' ability to grasp complex social dynamics and stakeholder

tensions, reinforcing the need for human oversight. Gao et al. (2025) further investigated DeepSeek's performance in classification tasks compared to Claude, Gemini, GPT, and LLaMA models. DeepSeek demonstrated superior performance in classification accuracy, ranking just behind Claude, and was commended for its cost-efficiency and scalable architecture qualities especially relevant for public sector deployment. In a more applied context, Kam (2025) evaluated DeepSeek's integration into China's anti-corruption infrastructure. By leveraging DeepSeek's pattern recognition and data synthesis capabilities, investigators were able to identify irregularities in financial transactions and procurement processes. His findings illustrate AI's capacity to enhance transparency and accountability in governance while also pointing to the ethical and political complexities of algorithmic oversight. These comparative and domain-specific studies collectively underscore a central tension in AI's role in public administration: while models like DeepSeek show great promise in streamlining analysis and enhancing operational capacity, they cannot supplant human expertise in areas requiring contextual interpretation, ethical discernment, and stakeholder negotiation. As such, the design of AI systems for public governance must prioritize collaborative intelligence integrating machine learning with human judgment to achieve reliable, equitable, and accountable policy outcomes. While many studies affirm the utility of AI for administrative efficiency, few have explored the potential of customized LLMs like ChatGPT and DeepSeek for decision support in governmental contexts and a comparison between these two AI models and human policy analysts. This study addresses this gap by investigating whether AI models can be used as supportive tool for decision making in public sector by evaluating AI models responses and comparing them to human policy analysts.

3. Methodology

3.1.Data Preparation

This study involved two primary phases:

1. First phase: Development of policy briefing notes using three different approaches: Artificial Intelligence Policy Analyst (AIPA) by Customized ChatGPT, AIPA by DeepSeek, and Human Policy Analyst (PA).
2. Second phase: The expert evaluation

3.1.1. First phase: Development of Policy Briefing Notes

a. AIPA by Customized ChatGPT

To develop an AI-powered tool capable of producing high-quality briefing notes, we utilized GPT-4 Turbo with the GPT Builder functionality. This platform includes a "Create" interface where users interact conversationally with the AI to guide its configuration. Through these guided interactions, the

model autonomously generates a set of instructions, which appear in the "Configure" section. Following this setup, we provided the AI with clear and detailed prompts to ensure it generated briefing notes aligned with the research objectives. The AI was trained using publicly available policy and regulatory documents from regional and international sources, including local governance frameworks, white papers, and strategic plans, which included over 650 PDF documents extracted and consolidated into a single reference dataset.

Once the AI was trained, it generated a sample briefing note:

Briefing Note No. (1)

Title: Involvement of Community Representatives in Research Teams

Purpose:
This briefing note aims to clarify the reasons and recommendations for the necessity of including a community representative in research teams, in order to enhance public trust, ensure that research aligns with community needs, and promote ethical and integrity in scientific research.

Summary:
Community involvement in research design ensures alignment with public values, increases transparency and to improve the practical application of research findings Strengthens ethical practices. This approach is supported by international best practices in participatory research and relevant ethical frameworks.

Background:
Scientific research significantly influences health policy and societal outcomes. Exclusion of community voices may lead to ethical oversight and misaligned research objectives. International models like Community-Based Participatory Research (CBPR) demonstrate the value of community engagement. Saudi Vision 2030 emphasizes the importance of inclusive decision-making and participatory governance. It also supports policies and initiatives that ensure health research aligns with national priorities and ethical standards. Including community representatives in research teams will help build trust between researchers and the public and ensure that important health issues are addressed effectively.

Recommended Actions: Introduce guidelines mandating community representation in research projects, provide training programs on research ethics and community roles, update ethical committee standards to reflect this requirement, establish monitoring mechanisms for community engagement outcomes, and collaborate with academic institutions and funders to ensure implementation.

Proposed Decision:
Scientific research teams in the Kingdom of Saudi Arabia include a community representative, as integrated into national research regulations and funding requirements to ensure compliance and effectiveness.

b. AIPA by DeepSeek

A briefing note was generated using DeepSeek, based on the same title and objective as the original document:

Briefing Note No. (2)

Title: The involvement of Community Representatives in Research Teams

Purpose:

This briefing note aims to clarify the reasons and recommendations for the necessity of including a community representative in research teams, in order to enhance public trust, ensure that research aligns with community needs, and promote ethical and integrity in scientific research.

Summary:

Involving community representatives (such as local leaders, patients, and civil society organizations) in health research teams is a crucial step to ensure that research findings align with the needs of the community. In Saudi Arabia, where cultural and geographical characteristics are diverse, this approach is essential to enhance trust in health research and increase the effectiveness of interventions. This paper outlines the justifications, proposed procedures, and relevant stakeholders, drawing on successful local and international experiences.

Background: *Local Context:* Vision 2030 promotes innovation in healthcare, However, local studies indicate that some research fails to account for cultural diversity or the actual needs of rural and remote communities. MOH data (2023) shows that 30% of rural residents feel excluded from research processes. *Global Insights:* The World Health Organization (WHO) indicated that community engagement reduces the gap between health policies and their implementation. A study in the United States showed that collaborative research with the community increased patient participation by 40%.

Recommended Actions: Launch a national initiative requiring public research funders to include community representatives and allocate 10% of research budgets for training, develop an online platform to facilitate community-researcher communication, conduct workshops with universities to train researchers in cultural sensitivity, and Pilot a community-inclusive research project in a selected region.

Proposed Decision:

Require a minimum of 20% community representation in all publicly funded health research teams, with periodic reporting on engagement outcomes.

c. Human Policy Analyst (PA)

A human analyst was tasked with generating a briefing note based on the same title and objective as the original document**:**

Briefing Note No. (3)

Title: The involvement of Community Representatives in Research Teams

Purpose:

This briefing note aims to clarify the reasons and recommendations for the necessity of including a community representative in research teams, in order to enhance public trust, ensure that research aligns with community needs, and promote ethical and integrity in scientific research.

Summary:

Numerous studies advocate for the integration of patients and community members into research design and implementation. To get different perspective on specific subject .Their unique insights improve research quality, relevance, and efficiency.

Background:

It is recommended to engage patients and community members from the very beginning of the research cycle and across as many of its stages as possible. This approach helps to include the perspectives of hard-to-reach populations within the target groups. Such involvement aims to enhance the depth, credibility, and applicability of findings, improve the clarity of research reports and recommendations, and ensure a direct link between evidence-based practice and results. The World Health Organization has emphasized the importance of involving patients and the community in scientific research, as outlined in the Ottawa Charter for Health Promotion (1986). This approach is also reflected in the U.S. National Institutes of Health and has been also embedded in several parts of British legislation, such as the Health and Social Care Act of 2001, the Local Government and Public Involvement in Health Act of 2007, the Health and Social Care Act of 2012, as well as through the National Institute for Health Research (NIHR). All these global experiences highlight the value of such involvement.

Recommended Actions: Encourage all government-funded research projects to include community representatives, Promote the role of universities in educating students on participatory research methods, empower research institutions and health organizations to facilitate community involvement, Participate with the third sector, which includes charities, to create to create a community participant database for researcher access.

Proposed Decision:

Health authorities, sectors, and research centers should include patient and community representatives in national research projects

3.1.2. Second phase: Expert Evaluation

A panel of 12 subject-matter experts assessed the three briefing notes using a standardized evaluation framework modified from Safaei and Longo (2024). Evaluation criteria were scored on a four-point scale across the following dimensions (Table 1)

Table (1) Evaluation criteria four-point scale

Level 4	Level 3	Level 2	Level 1
Central idea/purpose is vividly stated; content is accurate, thorough, and directly on point; strong support is provided for each assertion.	Central idea/purpose is clearly stated; content is accurate and relevant; credible support is provided for each assertion.	Central idea/purpose is stated; content is accurate but not always relevant; support is offered, but inadequate for some assertions.	Central idea/purpose is not stated; content is erroneous or irrelevant; support for assertions is largely absent.
Identifiable structure is presented in a purposeful, interesting, and effective sequence and remains focused.	An identifiable structure is present and consistently executed, with few statements out of place.	Identifiable structure but inconsistently executed; may contain several statements out of place or occasionally deviate from the topic.	Little or no structure present. Presentation is confusing to the audience; no logical sequence of ideas; frequently off topic.
Presentation is free of errors in grammar and word choice, aids clarity and vividness.	Presentation is free of serious errors in grammar and/or word usage.	Isolated errors in grammar and/or word choice reduce clarity and credibility.	Grammar, pronunciation, and/or word choice are severely deficient.
Content and/or style are consistently appropriate and targeted to the audience and context.	Content and/or style are consistently appropriate to the audience and/or context, with minor issues.	Content and/or style are occasionally inappropriate to the audience and/or context.	Content and/or style are frequently inappropriate to the audience and/or context.

4. Results

The three briefing notes generated respectively by the Customized ChatGPT (AIPA-1), DeepSeek (AIPA-2), and a human policy analyst (PA) were independently evaluated by a panel of 12 subject-matter experts. To ensure objectivity, the evaluators were not informed about the origin (AI or human) of each note.

Each briefing note was assessed using the standardized rubric described earlier (Table 1), which included four key evaluation criteria: Clarity and relevance of the central idea, Structure and organization, Language and grammar, and Appropriateness to audience and context (Table 2). The evaluation followed a four-point scale, where Level 4 represented the highest level of performance and Level 1 the lowest (Table 1 & 2)

Criteria of Evaluation	Customized ChatGPT	DeepSeek	Human
First criteria: The clarity of the idea in general	3.5	4	4
Second criteria: The clarity of the idea structure	3	4	3
Third criteria: Language Structure	3.5	4	4
Fourth criteria: Relevance to Audience and Context	3.5	4	4
Total	13.5	16	15

4.2. First criteria: General Clarity of Idea

- Implying a good organizational rationale in their replies, both **DeepSeek** and the **human** contributor received full score (4). This suggests that the human contributor and DeepSeek both managed to clearly and methodically express their general ideas. Demonstrating a great degree of interpretive clarity, their answers were well-organized and simple to follow.
- Customized ChatGPT lagged a little (3.5), perhaps suggesting small coherence or flow problems. Still, it did rather well overall in properly expressing ideas. Minor discrepancies in the arrangement of ideas could explain the small score drop. One should highlight, although, that Customized ChatGPT's general performance was still good.

4.3.Second criteria: Clarity of Idea Structure

- DeepSeek received the highest score (4), indicating it clarifies concepts more than the human and customised ChatGPT. This suggests that DeepSeek might be more efficient in simply presenting ideas and information. The better score also suggests that users would find it simpler to interact with and grasp the material produced by DeepSeek. DeepSeek's clear presentation of ideas, along with the accuracy and relevance of its content, can help users better understand and engage with the information
- Customized ChatGPT matched the human score (3), suggesting possible idea transmission but maybe requiring improvement in how ideas are presented or structured. Although promising, customized ChatGPT might gain from changes to improve the introduction and phrasing of ideas for better understanding.

4.4.Third criteria: Language Structure

- Again, both DeepSeek and human inputs received similar scores (4), implying grammatically correct and well-formed results. Although DeepSeek and human inputs both performed well in linguistic structure, it is crucial to keep an eye out for any possible mistakes or discrepancies that could occur in future interactions. All things considered, good performance in this area suggests a great degree of language competence in both models.
- Customized ChatGPT scored 3.5, may have sporadic poor language or less polished syntax. Ongoing training and fine-tuning models can help to correct this small problem, though. Customized ChatGPT provides the possibility to further enhance its language structure with ongoing optimization.

4.5.Fourth criteria: Relevance to Audience and Context

- DeepSeek and human got the highest (4), suggesting a good awareness of the target audience and the context of the decision-making environment. This implies that both DeepSeek and human decision-makers have a great degree of insight and knowledge, hence qualifying them as useful tools for wise choices. Their high scores show a great degree of ability in negotiating difficult circumstances and grasping the subtleties of the audience's requirements.
- Customized ChatGPT received the lowest score (3.5), suggesting a lack of response customization especially for the public sector audience, which is vital for decision-making tools. This underlines the significance of making sure artificial intelligence products are created with the particular requirements and preferences of the target audience in mind. Addressing this gap would help Customized ChatGPT to be more relevant and effective in public sector decision-making procedures.

4.6.Inferential data analysis:

One way ANOVA was used to compare the overall evaluation scores (a sum of scores for the 4 dimensions of evaluation) between customized ChatGPT, DeepSeek and Human. Results of the test revealed a statistically significant difference ($F = 4.0$, $p = 0.04$). Bonferroni post hoc test was used to identify which groups significantly differed. DeepSeek was significantly superior to Customized ChatGPT ($F = 5.2$, $p = 0.01$), Human evaluation was significantly superior to Customized ChatGPT ($F = 4.2$, $p = 0.03$). However, there was statistically significant difference in total evaluation scores between DeepSeek and Human evaluations ($F = 2.1$, $p = 0.34$).

5. Discussion

The results from the expert panel evaluations revealed that the DeepSeek, and to a lesser extent customized ChatGPT, model performed well in coherence and contextual relevance, particularly in tasks such as drafting policy briefs and summarizing stakeholder impacts. The model demonstrated strong capabilities in translating complex policy language into accessible summaries suitable for a variety of stakeholder audiences. This aligns with prior research emphasizing the role of AI in facilitating multi-stakeholder communication in the public sector (Wirtz et al., 2018). However, limitations were noted in the area of transparency and clarity. While the model could generate justifications for policy choices, its reasoning lacked the depth and nuance expected from experienced human analysts. This finding underscores the importance of maintaining human oversight in high-stakes policy contexts, where subtle ethical trade-offs and socio-political nuances must be evaluated (Rahwan, 2017). The study also

highlighted the iterative nature of prompt engineering, where model performance significantly improved across evaluation cycles. This iterative improvement reflects the importance of domain-specific prompt refinement and the role of human-AI collaboration in achieving meaningful policy insights (Lamba., 2024). To assess the potential of ChatGPT as a policy analysis support tool, a base version of the model was customized using supervised fine-tuning techniques. The training dataset consisted of curated policy documents, including national strategic plans, international governance frameworks, regulatory reports, and white papers. These documents were sourced from reputable institutions such as the OECD, World Bank, United Nations, and various national ministries. The objective was to ensure the model developed contextual awareness of public policy language, decision-making norms, and thematic diversity relevant to real-world governance scenarios (Zhang et al., 2021; Bryson et al., 2020). Preprocessing included standard NLP cleaning steps removal of non-textual content, token normalization, and metadata tagging to preserve document origin and policy sector (e.g., health, education, digital infrastructure). This dataset allowed the model to generate outputs that align with formal policy discourse while reflecting the nuances of sector-specific challenges. Furthermore, DeepSeek's use in policy improvement has been investigated in China's social security domain. The AI model showed better results in producing policy suggestions than GPT-4o, hence stressing the advantages of localized training for contextual fit. The research underlined the significance of human knowledge in policy creation even as it pointed out shortcomings in handling complicated social dynamics and balancing stakeholder interests (Jinghan et al., 2025). Despite its technical promise, DeepSeek has faced criticism related to data privacy and national security, resulting in prohibitions in areas including South Australia and thoughts for restriction in the United States. Similarly, while ChatGPT is widely used, it also necessitates careful attention to data protection and ethical safeguards, particularly when handling sensitive government information (Sebastian, 2023).

6. Conclusion and Recommendations

Results of this study show that the three approaches are acceptable for decision-making tasks, with DeepSeek showing superiority to ChatGPT. Based on the findings of this study, several recommendations are proposed for the responsible integration of DeepSeek and Customized ChatGPT models in public sector policy analysis: Hybrid Decision Frameworks: AI should augment, not replace, human policy experts. Integrating LLMs into workflows must ensure accountability structures and human validation steps. Continuous Training and Audit: Periodic retraining of models using updated and diverse datasets can help mitigate emerging biases and maintain contextual relevance. Ethical Oversight Committees: Establish multidisciplinary oversight committees to monitor model outputs, ensuring alignment with legal, cultural, and ethical standards. Capacity Building: Invest in training for policymakers to effectively

interact with AI tools and understand their limitations, including prompt engineering and interpretation. Open AI Governance Protocols: Encourage transparent and reproducible deployment of AI in public administration, allowing for peer review and public trust.

6.1 Study Limitations

This study presents several limitations that should be acknowledged: Real-world applications may yield different results. Dataset Constraints: The model was trained on a finite set of public documents, which may not capture the full diversity of global governance perspectives. Computational Constraints: due to technical limitations, the study could not explore real-time deployment or integrate dynamic datasets.

7. Declarations

7.1 Conflict of Interest Statement

The authors have no conflict of interests to declare.

7.2 Funding Disclosure

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

8. References

- Alotaibi, M., Alharthi, H., & Almalki, A. (2022). Digital government transformation in Saudi Arabia: Progress and challenges in Vision 2030. Journal of E-Government Studies and Best Practices, 2022, 1–12.
- Araya, D. (2019). Augmented intelligence and the future of policymaking. Foresight and STI Governance, 13(2), 6–12.
- Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3173574.3173951
- Brynjolfsson, E., & McAfee, A. (2017). Machine, platform, crowd: Harnessing our digital future. W. W. Norton & Company.
- Bryson, J., Mulgan, G., & Pencheon, D. (2020). Rethinking the policy cycle: Implications for AI and data-driven public administration. Government Information Quarterly, 37(4), 101509.
- Chen, J., Zhang, C., Xu, Y., & Zhang, W. (2020). Artificial intelligence in government: A review of current practices and challenges. Government Information Quarterly, 37(3), 101455.
- Cobbe, J., Lee, M., & Singh, J. (2021). Responsibility and accountability in AI governance: A human rights-based approach to algorithmic regulation. Computer Law & Security Review, 41, 105561.
- Doshi, P., Kumar, A., & Singh, R. (2023). Ethical integration of AI in public sector research: Challenges and strategies. Journal of Public Administration and Technology, 18(2), 101–117.
- Eggers, W. D., & Bellman, J. (2015). The journey to the government's digital transformation. Deloitte University Press.
- Estrada, M., Nguyen, H., & Patel, S. (2023). Cognitive limits of large language models in social science contexts. AI & Society, 38(4), 451–467.
- Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.
- Gao, Y., Lin, J., & Zhao, L. (2025). A comparative evaluation of large language models in classification tasks. arXiv.
- Habib Lantyer V. How US Trade Sanctions Fueled Chinese Innovation in AI: The DeepSeek Case. How US Trade Sanctions Fueled Chinese Innovation in AI: The DeepSeek Case. 2025.
- Head, B. W. (2010). Reconsidering evidence-based policy: Key issues and challenges. Policy and Society, 29(2), 77–94.
- Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic

governance. Government Information Quarterly, 33(3), 371–377. https://doi.org/10.1016/j.giq.2016.08.011

- Jinghan K, Zheng Z, Yuxuan Z. (2025) Can Large Language Models Become Policy Refinement Partners? Evidence from China's Social Security Studies. arXiv preprint arXiv:2504.09137.
- Jungwirth, D., & Haluza, D. (2023). Feasibility of GPT-3 in public health communication: A preliminary study. JMIR Medical Informatics, 11(2), e41221.
- Kam, T. (2025). AI-powered anti-corruption drive: The Chinese party-state's fight against graft. RSIS Policy Brief. https://rsis.edu.sg/rsis-publication/idss/ip25038
- Kankanhalli, A., Charalabidis, Y., & Mellouli, S. (2019). Emerging technologies for digital government: A research agenda. Government Information Quarterly, 36(4), 101391.
- Ke, X., Zhang, Y., & Liu, M. (2025). Policy recommendation with DeepSeek-R1: A comparative study with GPT-4o in China's social policy.
- Khorshid, O., Alghamdi, A., & Alqahtani, A. (2023). AI-driven public service delivery in Saudi Arabia: A Vision 2030 perspective. International Journal of Public Administration in the Digital Age, 10(1), 32–47.
- Lamba, D (2024). The Role of Prompt Engineering in Improving Language Understanding and Generation. International Journal for Multidisciplinary Research
- Longo, J. (2024). Artificial intelligence and public administration in Canada: A dual role of enabler and disruptor. Canadian Public Administration, 67(1), 45–61.
- Mergel, I., Ganapati, S., & Whitford, A. B. (2019). Agile: A new way of governing. Public Administration Review, 80(1), 162–165.
- Nzobonimpa, E., Wang, L., & Ali, S. (2024). Large language models in policy analysis: Capabilities and constraints. Policy & Society, 43(1), 89–105.
- OECD. (2021). AI in the public sector: Applications and challenges. OECD Digital Government Studies.
- Rahwan, I. (2017). Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology, 20, 5–14. https://doi.org/10.1007/s10676-017-9430-8
- Safaei, F., & Longo, J. (2024). Natural language processing tools and the future of public service briefing documents. Government Information Quarterly, 41(2), 101798.
- Sebastian, G. (2023). Privacy and Data protection in ChatGPT and other AI Chatbots: Strategies for Securing User information. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4454761
- Singh, S., Bansal, S., El Saddik, A., & Saini, M. (2024). From ChatGPT to DeepSeek AI: A

comprehensive analysis of evolution, deviation, and future implications in AI-language models. arXiv. https://arxiv.org/abs/2504.03219

- UN DESA. (2022). E-Government Survey 2022: The Future of Digital Government. United Nations Department of Economic and Social Affairs. https://publicadministration.un.org/egovkb
- van Manen, M. (2023). AI, ethics, and reflective practice in the public sector. Ethics and Social Welfare, 17(3), 225–241.
- Wachinger, J., Höhne, S., & Berger, M. (2024). Can ChatGPT support qualitative analysis? A critical examination in the context of theory-driven policy research. Qualitative Inquiry, 30(1), 55–72.
- Wang, T., Li, F., & Chen, R. (2024). Reflective integration of AI in policy-making: Challenges and pathways. Journal of Policy Analysis and Management, 43(2), 211–229.
- Wirtz, B. W., & Müller, W. M. (2018). An integrated artificial intelligence framework for public management. Public Management Review, 21(7), 1076–1100. https://doi.org/10.1080/14719037.2018.1549268
- Zhang, B., Dafoe, A., & Liao, Q. V. (2021). Human–AI interaction in policy decision-making: Understanding the design space for augmenting human judgment. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–13.