R&D Management Conference 2026 – Creativity and Resilience in an Era of Technological Disruption

Track 9.1: The Mistaken AI: Beyond the Hype - Hallucinations, Echo Chambers, and Filter Bubbles in R&D Decision-Making

From complacency to collapse: how bias propagation undermines decision-making

[1]Paolo Barile, Department of Management and Law, University of Rome, Tor Vergata

Maria Giovanna Corrado, Department of Economics, Management, and Institutions (DEMI), University of Naples, Federico II

Guido Iandorio, Department of Management, University of Rome, La Sapienza

Abstract

Purpose: In light of the growing integration of Large Language Models (LLMs) into business operations and decision-making processes, this paper analyses the role of user cognitive biases in the propagation of interpretative errors within Human-AI Interaction processes, with particular reference to research and development (R&D) contexts.

Conceptual framework: The paper proposes a conceptual model, the Bias Amplification Loop (BAL), to describe the dynamics through which the user's initial cognitive framing and the probabilistic mechanisms of LLMs tend to reinforce one another in successive cycles of interaction, with particular reference to the different exploration-exploitation phases typical of R&D processes.

Implications: The analysis highlights how the unregulated integration of LLMs can foster cumulative processes of interpretative convergence, a reduction in cognitive diversity, and a progressive narrowing of the exploratory space. If not adequately mitigated, these dynamics risk compromising organisations' ability to evaluate innovative alternatives and adapt to competitive contexts characterised by high uncertainty.

Originality: The BAL framework offers a dynamic and path-dependent perspective on the epistemic vulnerabilities associated with the use of LLMs in organisational decision-making processes, whilst proposing an operational interpretative framework for AI governance in innovation contexts.

1. Introduction

The ongoing digital transformation is significantly altering the mechanisms through which information is produced, shared and made accessible within contemporary society. In this evolutionary process, the boundary between the physical and digital worlds is gradually blurring, to

[1] Corresponding author: barile@economia.uniroma2.it

the extent that it is redefining the way in which individuals interpret information, construct knowledge and develop decision-making processes. The growing availability of technologies based on Artificial Intelligence (AI), with particular reference to Large Language Models (LLMs), has in fact introduced a new mode of human-machine interaction, which goes beyond the simple automatic execution of routine operations, aiming instead at cooperation between the user and the algorithmic system with a view to developing complex reasoning and acquiring new patterns of information synthesis.

As regards the possibility to use AI to create added cognitive value, the quality of the interaction cannot be analysed solely in terms of the system's technical performance; it must necessarily consider both the characteristics of the input provided by the user and the ways in which the machine processes and returns the information. The effective use of LLMs is therefore constrained, on the one hand, by the implicit representation of the user's prior knowledge, expectations and possible cognitive biases contained within the prompt, and, on the other, by the structural limitations inherent in the machine's operating mechanisms. The latter are linked to the absence of a genuine semantic understanding of the processed information, due to statistical-probabilistic algorithms that remain anchored to an emulative capacity which does not correspond to the existence of genuine cognitive, interpretative or semantic processes (Jurafsky & Martin, 2009). The key difference between the two interacting parties is that, whilst humans constitute an intrinsically complex system characterised by emergent properties, intuitive abilities, contextual interpretation and the multidimensional integration of experience, machines operate primarily through linear and formalised processes based on statistical rules and algorithmic inference mechanisms (Barile et al., 2024).

In essence, the 'apparent understanding' displayed by LLMs stems in fact from their extraordinary ability to recognise recurring patterns in human language, creating in the user a perception of cognitive reliability that is often greater than the system's actual epistemic nature, a perception further exacerbated by their inherent tendency towards complacency. It follows that, where the user's input is marred by an initial bias, the system will tend to overlook this distortion, reinforcing it through outputs consistent with the received information context. The real risk is that interaction with LLMs, cycle after cycle, will turn into a recursive mechanism for amplifying the user's pre-existing beliefs, generating forms of apparent validation that can consolidate interpretative errors and distorted representations of reality (Jarrahi, 2018; Dellerman et al., 2019).

This dynamic is particularly significant in decision-making processes involving high cognitive complexity, where the dependence on generative systems can lead to technological overreliance and a gradual delegation of cognitive functions. Within organisations, for example, the promise of significant digital benefits in terms of exploratory speed, reduced information asymmetries and increased organisational cognitive efficiency clashes with the introduction of new forms of epistemic vulnerability, stemming from the probabilistic nature of LLM operation and the difficulty faced by corporate management bodies in fully understanding the mechanisms underlying output generation. The epistemic challenges described are all the more evident in research and development (R&D) contexts, which are characterised by high uncertainty, risking particularly significant unintended consequences for decision-making processes and the innovative capacity of organisations (Taherdoost & Madanchian, 2023).

In light of the critical issues highlighted, this study aims to analyse the mechanisms through which the integration of LLMs into research and development processes can foster cumulative dynamics of bias amplification, progressively affecting the quality of cognitive exploration and organisational decision-making processes. In particular, a conceptual model of the Bias Amplification Loop (BAL)

will be introduced, designed to describe how initial informational biases can consolidate iteratively to the point of generating phenomena of reduced informational diversity and epistemic lock-in. Subsequently, the model will be applied to the key decision-making dynamics in R&D processes, in order to highlight the systemic implications arising from the unregulated use of generative systems and to identify possible organisational levers for mitigation.

2. Theoretical background

Recent research in social psychology highlights how human decision-making processes are frequently influenced by cognitive simplification mechanisms that distort the evaluation of available information (Bashkirova et al., 2024). Among these, confirmation bias is one of the most concerning dynamics, leading individuals to predominantly select, interpret and value information that is consistent with their pre-existing beliefs and established expectations. In the context of Human-AI Interaction, confirmation bias manifests as a tendency, most prevalent among less experienced users, to attribute levels of reliability and competence to AI systems that exceed their actual epistemic capabilities (Ma et al., 2023).

The accommodating nature of LLMs, which tends to go along with human assumptions even when these are based on unfounded beliefs, leads users to place excessive trust in the tool, even when the output is inaccurate or incomplete (Chen et al., 2023; Hemmer et al., 2022). In this regard, Jarrahi et al. (2023) observe that interaction with LLMs cannot be limited to merely replacing human capabilities but must instead be viewed as a form of collaborative partnership, in which the quality of the result cannot be separated from the central role of the human being in contextual interpretation and the validation of the knowledge produced. Nevertheless, most studies on the subject focus on bias as a property predominantly inherent to the algorithmic system of LLMs, concentrating on issues of fairness, automated discrimination or information hallucination, often tending to overlook the fundamental role of human input in the interaction process. In essence, when describing how generative systems can amplify stereotypes and dominant narratives through mechanisms of statistical repetition and probabilistic selection of information, there is a tendency to ignore the role of the human being in the interaction and the responsibility this entails for users (Wei et al., 2025; Zhou et al., 2024).

The decision to focus exclusively on bias as an intrinsic property of the algorithmic system seems paradoxical when considering the interaction with LLMs as a cyclical and recursive process in which human input is an indispensable structural component. Humans intervene, in fact, both upstream and downstream of the generative process: initially through the formulation of the prompt and the definition of the informational context, and subsequently through the critical interpretation, evaluation and possible reintegration of the output produced by the system. In both phases, typically human abilities come into play, linked to abductive reasoning and interpretative frameworks based on emotions, perceptions, past experiences and subjective mechanisms of meaning attribution (Barile et al., 2024). The human-machine interaction process, outlined in the model in Figure 1, highlights three closely interconnected macro-phases: (1) Human Input, (2) LLM System e (3) Output Evaluation.

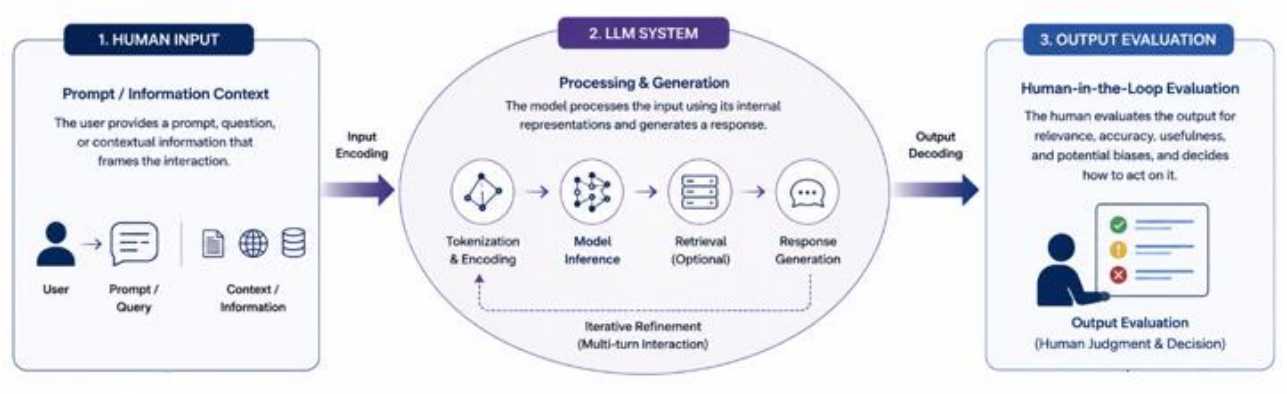

Figure 1: Schematic model of the various stages of the human-LLM interaction process – Source: Authors' own elaboration.

In the first phase (1), the human user enters a prompt, implicitly incorporating prior knowledge, interpretative assumptions and possible cognitive biases that help to define the semantic space within which the system will generate its response: it is clear, therefore, that the phase just described depends exclusively on the human user's ability to convert their expectations into a fully formed thought, and to convey this to the LLM through a clear and easily interpretable formulation. In the second phase (2), the LLM system processes the input through tokenisation and contextual encoding of the text, aimed at identifying probabilistic correlations between tokens and linguistic sequences. Subsequently, mechanisms of probabilistic inference and autoregressive generation enable the production of coherent and linguistically plausible outputs, even in the absence of genuine cognitive or semantic processes comparable to human reasoning. At this step, the human role is marginal, if not completely absent, since the processes of context interpretation, source selection and response generation occur entirely automatically. In the third and final phase (3), humans once again assume a central role, actively receiving the information produced by the system, interpreting it, evaluating it and deciding whether to incorporate it into the next cycle of interaction.

In the dynamic described, the user's behaviour directly influences the information trajectory produced by the system: the main risk therefore does not stem from a single inaccurate output generated by the LLM, but from the gradual consolidation of initial cognitive biases through successive iterations of what appears to be validation. This epistemological issue is particularly relevant in R&D processes, which are characterised by high uncertainty, informational ambiguity and a strong reliance on exploratory activities. Management literature has in fact highlighted how innovative processes require high levels of cognitive diversity, experimentation and interpretative openness, elements considered fundamental to sustaining organisations' ability to identify new technological opportunities and adapt to dynamic competitive contexts (March, 1991).

With regard to the exploration-exploitation balance, the integration of LLMs into decision-making processes leads, on the one hand, to an increase in the speed of exploration and cognitive efficiency, and on the other, to a gradual convergence towards dominant narratives that constrain the scope for exploration. The result is the emergence of increasingly homogeneous decision-making pathways, which limit governance's ability to consider innovative alternatives. The issue takes on further significance in light of the literature on dynamic capabilities, according to which firms' competitive advantage depends on their ability to perceive environmental changes, seize emerging opportunities and continually reconfigure organisational resources and competencies (Teece et al., 1997). The reduction in informational diversity and the recursive amplification of initial cognitive orientations

can compromise organisations' absorptive capacities, thus altering organisational learning processes and limiting the quality of strategic interpretation and long-term innovation capacity (Cohen & Levinthal, 1990).

3. Conceptual framework

In light of the above, it is the primary responsibility of human users to limit the perpetuation of interpretative errors through careful formulation of inputs and proper evaluation of outputs in each cycle of interaction with LLMs. In line with the view of the human-LLM interaction process as a recursive process of information co-construction, the conceptual model of the Bias Amplification Loop (BAL) describes the main dynamics through which the user's cognitive biases and the probabilistic mechanisms of LLMs interact with one another, generating phenomena of epistemic amplification and a reduction in cognitive diversity, with critical effects on decision-making processes in business contexts (Figure 2). Unlike approaches that interpret bias as a purely static property internal to the algorithmic system, our framework is structured in three sequential and mutually interconnected phases: (1) Initial Steering, (2) Iterative Reinforcement and (3) Epistemic Lock-in.

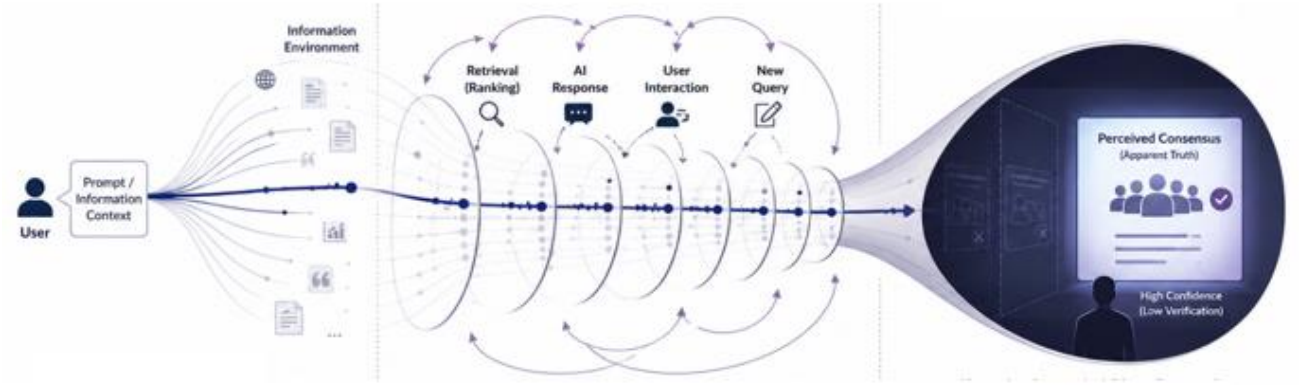

Figure 2: Conceptual model of the Bias Amplification Loop – Source: Authors' own elaboration.

The main idea behind the first phase (1) is that the introduction of an initial informational orientation is incorporated into the wording of the prompt and is therefore implicit in the cognitive context provided by the user. In each interaction cycle, the semantic space explored by the LLM is partially delimited from the very start of the interaction, giving rise to a generative process that inevitably unfolds along a cognitive trajectory already influenced by the framing introduced by the human. This reduction in informational variety is an integral part of the process of transferring ideas and content from human to machine and can be limited through prompt engineering techniques, though it is difficult to avoid entirely. The second phase (2) refers to the processing mechanisms of the AI software described in the previous section and to the user's subsequent evaluation of the outputs produced. At this stage, the LLM's probabilistic coherence mechanisms tend to produce outputs aligned with the informational framing established in previous interactions. At the same time, the user gradually tends to reinterpret these outputs as plausible confirmations of their initial hypotheses. The recursive nature of the interaction thus generates a cumulative reinforcement dynamic in which informational diversity tends to gradually decrease, whilst dominant narratives and recurring interpretative frameworks become established over time. The third phase (3) emerges when the interaction dynamics described in the previous stages are repeated in subsequent cycles of interaction:

in this situation, the user develops a growing level of trust in information pathways that are not necessarily subjected to critical scrutiny, creating an illusion of consensus and interpretative coherence. The main risk associated with this phase does not concern individual errors produced by the system, but rather the gradual consolidation of homogeneous cognitive patterns that limit exploratory capacity and reduce openness to alternative perspectives.

The dynamics of cognitive bias propagation described by the BAL framework find immediate application in organisational contexts, particularly in relation to the exploratory activities typical of research and development processes. In the context of R&D, the chances of success are primarily linked to the decision-maker's ability to select innovative solutions and evaluate alternative paths where the initial strategy is not immediately applicable. The use of LLMs within business activities and innovation processes is now increasingly widespread, primarily due to their ability to enhance exploratory speed, information accessibility and operational efficiency. However, the unregulated integration of such tools risks exacerbating the epistemic challenges described above, fostering cumulative dynamics of narrative convergence, reduced diversity and the progressive narrowing of the exploratory space.

In order to clarify the causal relationships between the phases of the BAL framework and the key decision-making points in corporate research and development contexts, Table 1 presents a conceptual matrix that links each phase of the R&D process to the associated epistemic risks and potential systemic effects arising from the mechanisms of recursive bias amplification.

Table 1: Conceptual matrix of BAL's effects on the different R&D phases – Source: Authors' own elaboration.

R&D PHASE	BAL PHASE	EPISTEMIC RISK	POTENTIAL EFFECT
Opportunity identification	Initial steering	Framing bias	Reduced diversity of explored opportunities
Idea generation	Iterative reinforcement	Narrative convergence	Homogenization of innovation trajectories
Concept evaluation	Iterative reinforcement	Confirmation bias amplification	Overconfidence in selected alternatives
Strategic selection	Epistemic lock-in	Reduction of perceived alternatives	Premature technological convergence
Development planning	Epistemic lock-in	Recursive validation	Path dependency and reduced adaptability

The mapping provided by the matrix enables us to translate our theoretical framework into practical terms, highlighting how iterative human-LLM interaction can progressively influence organisations' exploratory, adaptive and innovative capabilities. Based on the BAL model and its application to corporate R&D processes, the following theoretical propositions are therefore formulated:

P1: The greater the level of implicit informational orientation in the formulation of the initial prompt, the greater the likelihood that subsequent interactions with LLMs will reinforce homogeneous interpretative trajectories.

P2: Recursive interaction with generative systems tends progressively to reduce the available informational variety, reinforcing outputs consistent with the user's prior assumptions and beliefs.

4. Discussion and implications

In response to the growing prevalence of generative AI technologies, LLMs are now being integrated into an increasing number of organisational activities, not only as tools for operational automation, but above all as management support systems used to synthesise information, explore alternative scenarios, accelerate analytical processes and guide strategic decisions. This transformation appears particularly necessary in the face of intense competitive pressure, where exploratory quality and efficiency are essential factors for maintaining a competitive edge. It is no coincidence that research and development processes are one of the areas most exposed to the effects of human-AI integration, requiring a continuous reassessment of initial hypotheses and a marked ability to deal with ambiguous, incomplete or contradictory information (Tushman & O'Reilly, 1996).

The growing use of LLMs brings undeniable benefits in terms of information accessibility, processing speed and support for exploratory activities. However, the BAL framework suggests that iterative interaction with generative systems may progressively reduce the quality of decision-making processes, particularly where the outputs produced are incorporated into subsequent cycles of analysis without adequate mechanisms for critical verification by users. This observation, fully in line with the theoretical propositions set out above, describes a scenario in which human-machine collaboration at the organisational level, due to the incorrect and not fully informed use of digital tools, leads to a progressive convergence of interpretation that inhibits the decision-maker's capabilities and, consequently, the organisation's ability to adapt (Cafferata, 2017; Barile & Golinelli, 2011).

From this perspective, the formulation of the prompt plays a central role in the human-LLM interaction process. As hypothesised by P1, where usage practices are more informed, geared towards diversifying information perspectives and the continuous critical reformulation of queries, the cumulative dynamics described by the BAL can be significantly reduced, limiting the risk that the iterative interaction process degenerates into forms of cognitive rigidity and decision-making convergence. Where organisations decide to use generative systems to support strategic selection, technology assessment or the definition of innovative trajectories, the governing body must therefore be fully aware of the risks associated with cognitive bias and endeavour, including through specific training, to limit the risks associated with the progressive entrenchment of interpretative assumptions that have not been adequately verified. In this context, practices focused on the diversity of information sources, the separation of the exploratory phase from the evaluation phase, and the introduction of structured workflows for the independent verification and validation of generated outputs are of particular importance. Similarly, the inclusion of human-in-the-loop approaches appears essential to prevent generative systems from implicitly taking on a role that replaces human interpretative capabilities (Dellerman et al., 2019; Jarrahi et al., 2018).

These considerations are consistent with the main emerging AI governance frameworks, including the NIST AI Risk Management Framework and the ISO/IEC 42001 standard, both of which are designed to promote human oversight, organisational accountability and continuous monitoring of the risks associated with the use of AI systems (National Institute of Standards and Technology, 2023; International Organization for Standardization, 2023). In this sense, the BAL model helps to highlight how AI governance should not be limited solely to the technical aspects of the algorithmic system but

must necessarily also include the cognitive and organisational dynamics generated by the recursive interaction between humans and machines.

5. Conclusions

This paper has proposed the Bias Amplification Loop (BAL) conceptual framework as a key to interpreting the recursive dynamics that characterise the interaction between humans and LLM-based generative systems in organisational contexts. Unlike approaches that interpret bias as a property exclusively internal to the algorithmic system, the model highlights how epistemic risks emerge from the continuous interaction between human cognitive framing and probabilistic mechanisms of information generation. From this perspective, the growing integration of LLMs into R&D processes can produce progressive effects of interpretative convergence and a reduction in exploratory capacity, with possible implications for the quality of decision-making and the adaptive capabilities of organisations.

From a theoretical perspective, the BAL framework helps to broaden the debate on human-AI interaction and AI governance by introducing a perspective that is more focused on the cognitive and organisational dynamics generated by the recursive use of generative systems. At the same time, the proposed model has certain limitations linked to the predominantly conceptual nature of the work: in this regard, future research could develop longitudinal studies on R&D processes, analysing phenomena such as the progressive convergence of the design alternatives considered, the reduction in the variety of information sources used, or the level of decision-making dependence on the outputs generated by LLMs. Similarly, the BAL framework could be further explored through case studies, simulations, controlled experiments or archival analyses aimed at observing how continuous interaction with generative systems influences learning, adaptation and organisational innovation processes over time.

Bibliographic references

1. Barile, P., Bassano, C., Piciocchi, P. (2024). Human Complexity vs. Machine Linearity: Tug-of-War Between Two Realities Coexisting in Precarious Balance. *Journal of Systemics, Cybernetics and Informatics*, 22(7), 53-62. https://doi.org/10.54808/JSCI.22.07.53
2. Barile, S., & Golinelli, G. M. (2011). *Foundations of systems thinking: The structure-system paradigm*. Springer. https://doi.org/10.1007/978-1-4419-9886-8
3. Bashkirova, A., et al. (2024). *Confirmation bias in AI-assisted decision-making: AI triage in mental healthcare. Computers in Human Behavior: Artificial Humans, 2*, 100054. https://doi.org/10.1016/j.chbah.2024.100054
4. Cafferata, R. (2017). *Management in adattamento: Tra razionalità economica, evoluzione e imperfezione dei sistemi*. Il Mulino.
5. Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW2), Article 370. https://doi.org/10.1145/3610219
6. Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly, 35(1), 128–152. https://doi.org/10.2307/2393553

7. Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering, 61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2
8. Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2022). *On the effect of information asymmetry in human-AI teams* (arXiv:2205.01467). arXiv. https://arxiv.org/abs/2205.01467
9. International Organization for Standardization. (2023). ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system. ISO.
10. Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007
11. Jarrahi, M. H., Askay, D., Eshraghi, A., & Smith, P. (2023). Artificial intelligence and knowledge management: A partnership between human and AI. *Business Horizons, 66*(1), 87–99. https://doi.org/10.1016/j.bushor.2022.03.002
12. Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing* (2nd ed.). Pearson Prentice Hall.
13. Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). *Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making.* arXiv. https://arxiv.org/abs/2301.05809
14. March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71
15. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
16. Taherdoost, H., & Madanchian, M. (2023). Artificial intelligence and knowledge management: Impacts, benefits, and implementation. *Computers, 12*(4), Article 72. https://doi.org/10.3390/computers12040072
17. Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal, 18*(7), 509–533.
18. Tushman, M. L., & O'Reilly, C. A. (1996). Ambidextrous organizations: Managing evolutionary and revolutionary change. *California Management Review, 38*(4), 8–30. https://doi.org/10.2307/41165852
19. Wei, X., Kumar, N., & Zhang, H. (2025). Addressing bias in generative AI: Challenges and research opportunities in information management. *Information & Management, 62*(2), Article 104103. https://doi.org/10.1016/j.im.2025.104103
20. Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). *Bias in generative AI* (arXiv:2403.02726). arXiv. https://arxiv.org/abs/2403.02726