UDC 658.15; 004.8

BEHAVIORAL RISK MANAGEMENT IN THE ERA OF GENERATIVE AI

Kapoor V.N.
Senior lecturer, Indian School of Business and Finance (New Delhi, India)

ПОВЕДЕНЧЕСКИЙ РИСК-МЕНЕДЖМЕНТ В ЭПОХУ ГЕНЕРАТИВНОГО ИИ

Капур В.Н.
старший преподаватель, Индийская школа бизнеса и финансов (Нью-Дели, Индия)

Abstract

The proliferation of generative artificial intelligence (GenAI) across organizational workflows and decision environments has introduced new categories of behavioral risk. Unlike traditional automation tools, GenAI generates content and recommendations that directly influence human perception, cognition, and judgment. This article explores the amplification of behavioral biases–such as automation bias, confirmation bias, and anchoring–in AI-mediated decision-making processes. Through a multidisciplinary lens, the study analyzes human-AI interaction risks, organizational vulnerabilities, and cultural factors that exacerbate behavioral distortions. A multilayered risk mitigation framework is proposed, integrating technical, procedural, and behavioral safeguards. The findings highlight the urgent need for adaptive governance mechanisms, explainable AI design, and AI literacy initiatives to ensure responsible and resilient deployment of GenAI technologies in high-stakes contexts.

Keywords: behavioral risk management, generative AI, cognitive biases, human-AI interaction, automation bias, organizational vulnerability, risk mitigation, explainable AI, decision-making, AI governance.

Аннотация

Широкое распространение генеративного искусственного интеллекта (ИИ) в организационных процессах и системах поддержки принятия решений привело к появлению новых категорий поведенческих рисков. В отличие от традиционных автоматизированных систем, генеративный ИИ генерирует контент и рекомендации, непосредственно влияющие на восприятие, мышление и поведение пользователей. В статье рассматриваются механизмы усиления когнитивных искажений – таких как автоматизация, предвзятость подтверждения и эффект якоря – в условиях взаимодействия человека с ИИ. Анализируются риски на уровне взаимодействия, организационные уязвимости и культурные факторы, способствующие искажению суждений и снижению устойчивости. Предлагается многоуровневая модель смягчения поведенческих рисков, включающая технические, процедурные и поведенческие меры. Полученные результаты подчеркивают необходимость развития адаптивных механизмов управления, внедрения прозрачных моделей ИИ и повышения ИИ-грамотности для обеспечения ответственного и устойчивого использования генеративного ИИ в критически значимых сферах.

Ключевые слова: поведенческий риск, генеративный искусственный интеллект, когнитивные искажения, взаимодействие человек–искусственный интеллект, автоматизация, организационные уязвимости, снижение рисков, объяснимый искусственный интеллект, принятие решений, управление искусственным интеллектом.

Introduction

The rapid advancement of generative artificial intelligence (AI) has brought about profound transformations in how individuals, organizations, and systems perceive and respond to risks. Unlike earlier technological innovations that primarily automated tasks or improved data processing, generative AI actively creates content–text, images, code–shaping decisions, communications, and even emotions in real time [1]. As these systems become increasingly embedded in corporate workflows, media ecosystems, and decision-making environments, they not only introduce new vectors of technical risk but also amplify existing behavioral vulnerabilities. In this context, understanding the intersection between human cognition and AI-generated stimuli becomes critical for effective risk management.

Behavioral risk management (BRM), traditionally associated with biases, heuristics, and organizational misjudgments, is undergoing a paradigmatic shift. In the era of generative AI, risks emerge not only from flawed human reasoning but also from interactions with artificially generated content that may be indistinguishable from authentic information. Examples include deepfake media influencing public opinion [2], AI-generated phishing campaigns exploiting cognitive shortcuts, and algorithmically personalized content reinforcing confirmation bias. Moreover, automated decision systems powered by generative models may lack interpretability, leading users to over-rely on their outputs or underestimate their limitations. These developments highlight the urgent need to reconceptualize BRM frameworks through the lens of AI-human interaction.

The purpose of this article is to examine how generative AI technologies reshape the landscape of behavioral risk in organizational and societal settings. Specifically, it explores the mechanisms by which generative systems affect perception, judgment, and decision behavior; identifies the unique behavioral risks introduced by these systems; and proposes adaptive management strategies. By combining insights from behavioral science, AI ethics, and risk governance, the article aims to provide a structured understanding of how behavioral risks evolve in an era of synthetic intelligence–and how they may be effectively mitigated [3].

Behavioral biases amplified by generative AI

The integration of generative AI (GenAI) into decision-making environments introduces not only technological benefits but also amplifies latent behavioral biases among users and organizations. As GenAI systems generate outputs based on probabilistic patterns and user prompts, they often reinforce preexisting cognitive distortions rather than mitigate them. This dynamic creates a feedback loop in which human biases shape machine outputs, and those outputs, in turn, validate the user's biased expectations [4]. In high-stakes domains such as finance, policy design, and cybersecurity, this amplification can lead to risk misperception, overconfidence, and flawed judgment.

Among the most prominent cognitive distortions exacerbated by GenAI are automation bias, confirmation bias, framing effects, and anchoring. For instance, when users interact with AI-generated scenarios or recommendations, they may disproportionately favor machine-suggested actions without critical evaluation (automation bias) or selectively prompt AI tools in ways that reaffirm their assumptions (confirmation bias). These effects are magnified by the natural language fluency and persuasive tone of GenAI systems, which lend unwarranted credibility to speculative or contextually inappropriate responses [5]. The Figure 1 highlights four key behavioral biases that are frequently intensified in GenAI-driven environments.

First, confirmation bias manifests through recursive prompting, where users repeatedly ask GenAI to support a hypothesis, leading to increasingly skewed outputs. Second, automation bias emerges when decision-makers defer judgment to AI outputs despite uncertain or ambiguous inputs. Third, the framing effect is shaped by the way AI presents its responses; slight changes in tone or emphasis can dramatically alter user perception. Finally, anchoring bias is triggered when users rely heavily on the first piece of information generated by the system, even if it lacks empirical validity.

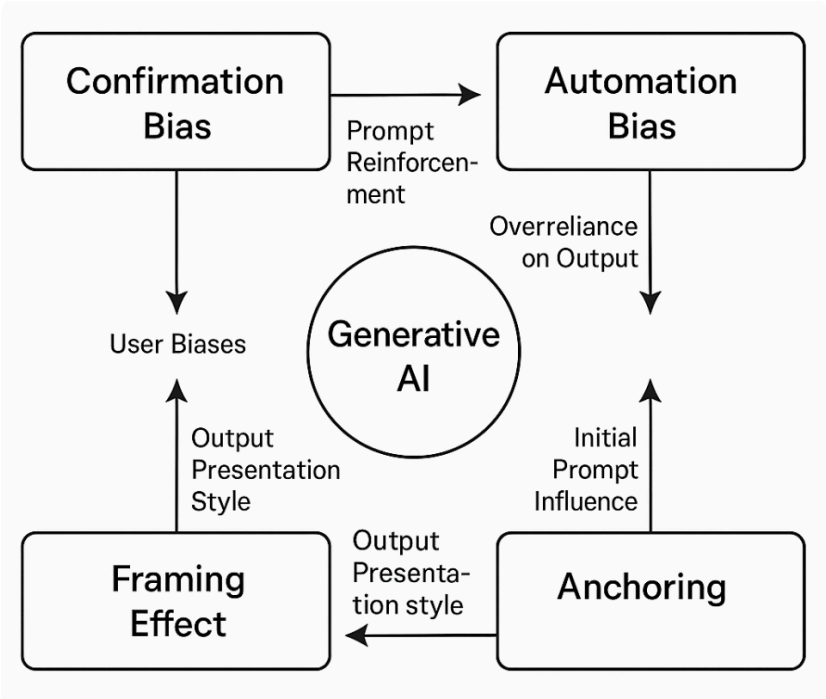

Figure 1. Behavioral biases amplified by generative AI in decision-making contexts

These behavioral distortions are not merely theoretical. Empirical studies have shown that AI-augmented decisions often display higher variance in accuracy depending on user experience, framing of prompts, and task domain. In financial forecasting, for example, novice users tend to anchor on GenAI's initial suggestions, while in legal risk assessments, confirmation bias leads to selective prompt engineering that narrows argument diversity [6]. Moreover, organizational workflows that overly depend on GenAI without structured validation protocols risk institutionalizing these biases into automated pipelines.

To mitigate these effects, behavioral risk management frameworks must integrate AI-specific bias audits, develop AI literacy programs for end-users, and incorporate human-in-the-loop mechanisms that ensure critical oversight [7]. Failure to address the behavioral amplification risks associated with GenAI may not only compromise decision quality but also exacerbate systemic vulnerabilities across digital infrastructures.

Human-AI interaction risks in decision-making

The integration of GenAI tools into decision-making environments introduces not only operational advantages but also new vectors of behavioral and systemic risk. Unlike earlier rule-based systems, GenAI models interact with users through natural language, probabilistic outputs, and persuasive interfaces, thereby influencing human cognition in subtle yet profound ways [8]. These interactions reshape how individuals interpret, validate, and act upon information, creating a hybrid cognitive environment in which responsibility, agency, and trust become distributed and potentially ambiguous.

One of the primary concerns in human-AI decision-making is the erosion of critical judgment due to automation bias. Users tend to over-rely on AI-generated outputs–especially when presented with fluent language and contextual relevance–without adequately questioning the underlying data quality or model assumptions [9]. This effect is exacerbated in high-pressure or time-constrained environments, such as crisis response or financial forecasting, where GenAI may offer plausible yet unverified recommendations. Additionally, the opacity of model logic and lack of explainability can lead to "epistemic outsourcing", where users defer to AI not because of proven accuracy, but due to perceived authority.

Figure 2 illustrates the interplay between human cognitive heuristics and GenAI system characteristics, highlighting four high-risk zones in human-AI interaction: illusion of understanding, trust asymmetry, degraded situational awareness, and feedback misalignment.

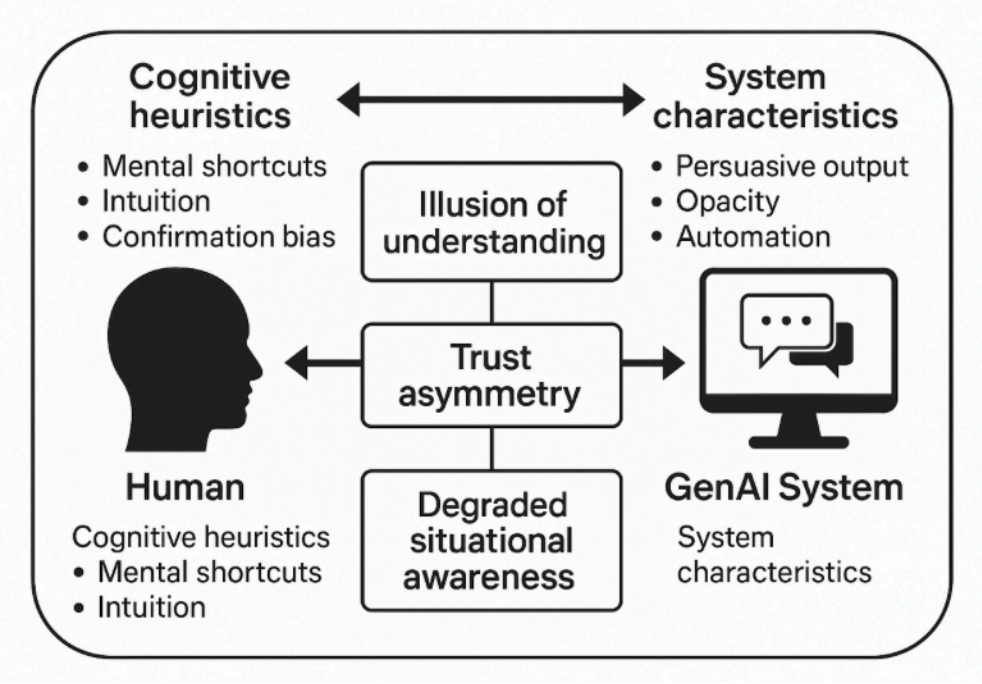

Figure 2. Human-AI interaction risk zones in cognitive decision environments

These risk zones reflect how certain design choices–such as interface simplicity, narrative coherence, or absence of uncertainty indicators–can distort users' perception of AI reliability and competence. For instance, high trust in low-transparency systems may encourage users to ignore contradictory evidence or override human intuition. Conversely, unclear accountability structures can lead to under-reliance on valid AI insights due to fear of blame allocation.

Moreover, reinforcement learning from human feedback, a common training approach in GenAI systems, can introduce recursive biases into decision ecosystems. If users adjust their behavior based on AI outputs and these behaviors are subsequently used to fine-tune the model, a circular distortion may emerge, reinforcing miscalibrated preferences or risk assessments. This phenomenon, termed “interactive overfitting,” poses long-term threats to decision system robustness, especially in domains requiring nuance, ethical deliberation, or institutional trust [10].

To mitigate these risks, organizations must implement hybrid oversight structures that combine algorithmic validation with human-in-the-loop auditing. Clear delineation of decision boundaries, structured feedback loops, and the use of adversarial testing environments can help detect emergent pathologies in human-AI collaboration. In parallel, user education initiatives must shift from interface training toward cognitive hygiene–equipping individuals to question, contextualize, and responsibly adapt GenAI recommendations within bounded rationality frameworks.

Organizational vulnerabilities and cultural factors

The integration of generative AI into organizational workflows has revealed latent vulnerabilities rooted in structural, procedural, and cultural aspects of corporate environments. Unlike rule-based automation systems, generative AI tools operate probabilistically and produce outputs that are not always verifiable by predefined rules. This unpredictability, when combined with organizational inertia, fragmented communication, or low algorithmic literacy among staff, creates fertile ground for misinterpretation, over-reliance, or uncritical adoption of AI-generated insights.

Organizational culture plays a pivotal role in shaping how employees perceive and respond to AI-generated content. In hierarchical structures with strong top-down decision-making, employees may feel compelled to align their judgments with AI outputs, especially when such tools are perceived as sanctioned by management. This can result in diminished dissent, reduced critical assessment, and normalization of flawed or biased content. Furthermore, cultures that emphasize speed and efficiency over deliberation are more likely to adopt AI outputs without adequate validation, leading to increased operational risk and misaligned decision pathways.

A critical vulnerability lies in the absence of governance protocols for human-AI collaboration. Many organizations lack clear policies defining accountability when AI-generated content influences

strategic or financial decisions. This regulatory vacuum blurs responsibility and hinders post hoc assessments when failures occur. Additionally, siloed data infrastructures and inconsistent knowledge-sharing practices impede effective monitoring and feedback on AI performance across departments.

Moreover, cognitive dissonance between organizational values (e.g., transparency, fairness) and the opaque logic of generative models can erode internal trust. Employees may experience ethical discomfort when using black-box systems that conflict with the company's public commitments to equity or inclusion [11]. This tension is further exacerbated in multicultural teams, where attitudes toward automation, risk, and ethical responsibility vary widely, requiring culturally sensitive implementation strategies.

To mitigate these vulnerabilities, organizations must prioritize the development of AI literacy programs, establish cross-functional oversight mechanisms, and cultivate a culture of critical engagement with AI tools. Encouraging transparent dialogue around system limitations, fostering diversity in AI governance teams, and embedding ethical audits into AI project lifecycles are essential steps toward ensuring organizational resilience in the era of generative AI.

Risk mitigation strategies

The proliferation of GenAI in organizational workflows necessitates the development of robust mitigation strategies to counteract behavioral risks. Unlike conventional automation tools, GenAI systems do not merely execute predefined tasks–they engage in probabilistic reasoning, content generation, and recommendation formulation, often in domains characterized by cognitive uncertainty. This shift introduces novel risk categories, including overreliance on algorithmic outputs, erosion of human critical oversight, and diffusion of accountability. Accordingly, risk mitigation must encompass both technical safeguards and behavioral governance mechanisms.

Effective mitigation begins with the design of explainable AI systems that allow users to interrogate and contextualize model outputs. Transparency regarding model provenance, training data scope, and confidence levels can help prevent automation bias and reduce the likelihood of ill-informed decisions. Equally important is the implementation of human-in-the-loop frameworks, where critical decisions remain subject to human review and override. Such architectures ensure that AI augments rather than supplants human judgment.

Behavioral risk management also requires addressing organizational incentives and cognitive structures. Decision environments should be engineered to discourage blind trust in GenAI and instead foster deliberative engagement [12]. Techniques such as adversarial testing, counterfactual reasoning, and red teaming exercises enable organizations to probe the boundaries of system reliability and identify failure modes. Training programs must go beyond technical proficiency to develop AI literacy, emphasizing cognitive pitfalls, ethical implications, and recognition of manipulation risks.

At the cultural level, organizations must institutionalize ethical deliberation in AI deployment through risk boards, cross-functional review committees, and continuous audit mechanisms. These structures reinforce accountability while enabling agile responses to emerging risks. Moreover, the integration of feedback loops between frontline users, developers, and governance bodies ensures that mitigation strategies evolve alongside the systems they are meant to control.

Ultimately, successful risk mitigation in the GenAI era demands a hybrid approach–balancing formal controls with soft norms, technical robustness with user awareness, and centralized oversight with distributed responsibility. A schematic representation of such a multilayered mitigation framework is provided in Figure 3.

The schematic illustrates a comprehensive, multilayered approach to behavioral risk mitigation. At the foundation lies technical transparency, which ensures that AI outputs are interpretable and traceable. The second layer comprises procedural safeguards, including human-in-the-loop mechanisms, audit trails, and model validation protocols. Behavioral oversight forms the third layer, addressing organizational dynamics, user training, and decision culture [13]. Finally, strategic governance and ethical alignment occupy the top tier, enabling oversight bodies to coordinate

responses, establish accountability mechanisms, and adapt policies in response to evolving technological capabilities.

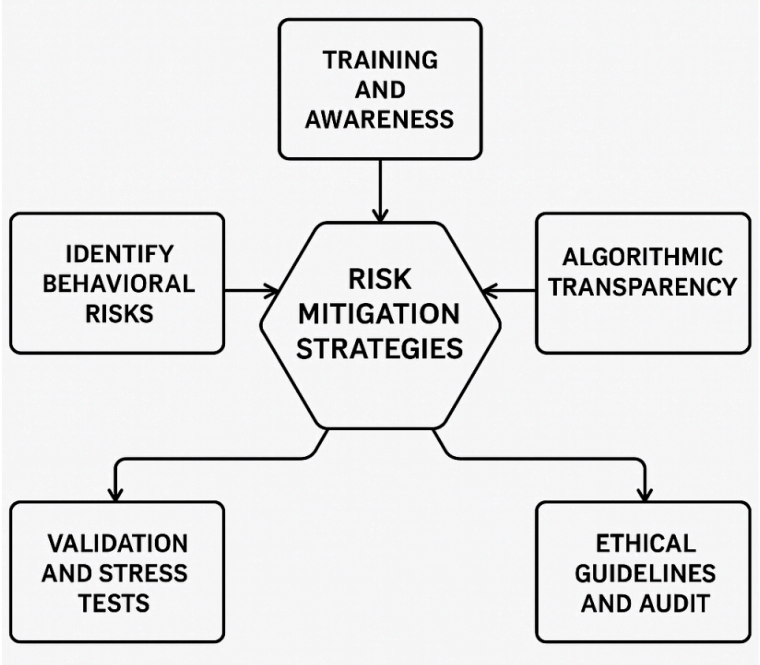

Figure 3. Multilayered framework for behavioral risk mitigation in GenAI-enabled decision environments

This architecture reflects the reality that no single intervention can fully address the behavioral risks introduced by generative AI. Rather, mitigation must be systemic, continuous, and adaptive, integrating technical, organizational, and human factors into a cohesive risk posture. As GenAI systems become more autonomous and contextually embedded in strategic decision-making, organizations must continuously refine their mitigation strategies, ensuring alignment with evolving ethical standards, legal obligations, and societal expectations.

Such proactive frameworks not only reduce the likelihood of harmful outcomes but also enhance organizational resilience, foster trust in AI-augmented processes, and support sustainable innovation in high-stakes environments.

Conclusion

The era of generative artificial intelligence introduces profound shifts in how behavioral risks manifest and propagate within organizational and decision-making ecosystems. As AI systems transition from tools of analysis to agents of synthesis, they fundamentally alter cognitive processes, judgment structures, and operational cultures. Behavioral distortions–such as automation bias, framing effects, and recursive confirmation–are no longer peripheral concerns but core vulnerabilities in AI-augmented environments.

This study underscores the need for a multidimensional approach to behavioral risk management, integrating explainable AI design, human-in-the-loop oversight, organizational AI literacy, and ethical governance frameworks. Effective mitigation of GenAI-related behavioral risks demands continuous adaptation and alignment with evolving human, institutional, and technological dynamics. Failing to address these risks not only undermines decision quality but threatens long-term resilience and trust in AI-enabled systems.

References

1. Joshi S., Joshi-Satyadhar S. Model risk management in the era of generative AI: Challenges, opportunities, and future directions // International Journal of Scientific and Research Publications. 2025. Vol. 15. No. 5. P. 299–309.

2. Mohamed M. A. H., Al-Mhdawi M. K. S., Ojiako U., Dacre N., Qazi A., Rahimian F. Generative AI in construction risk management: A bibliometric analysis of the associated benefits and risks // Urbanization, Sustainability and Society. 2025. Vol. 2. No. 1. P. 196–228.

3. Joshi S. Review of Gen AI models for financial risk management // International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2025. Vol. 11. No. 1. P. 709–723.
4. Sriram H. K. AI neural networks in credit risk assessment: Redefining consumer credit monitoring and fraud protection through generative AI techniques // Migration Letters. 2022. Vol. 19. No. 6. P. 1017–1032.
5. Hettiarachchi I. The rise of generative AI agents in finance: Operational disruption and strategic evolution // International Journal of Engineering Technology Research & Management. 2025. P. 447.
6. Wach K. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT // Entrepreneurial Business and Economics Review. 2023. Vol. 11. No. 2. P. 7–30.
7. Reddy J. K. Leveraging generative AI for hyper personalized rewards and benefits programs: Analyzing consumer behavior in financial loyalty systems // Journal of Electrical Systems. 2024. Vol. 20. No. 11s. P. 3647–3657.
8. Kissabekov A. Analysis of factors influencing successful interaction between the client, contractor, and engineer on construction sites // International Journal of Scientific and Management Research. 2025. Vol. 8(5). P. 171-176.
9. Ooi K. B. The potential of generative artificial intelligence across disciplines: Perspectives and future directions // Journal of Computer Information Systems. 2025. Vol. 65. No. 1. P. 76–107.
10. Budhwar P. Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT // Human Resource Management Journal. 2023. Vol. 33. No. 3. P. 606–659.
11. Ye X., Yan Y., Li J., Jiang B. Privacy and personal data risk governance for generative artificial intelligence: A Chinese perspective // Telecommunications Policy. 2024. Vol. 48. No. 10. P. 102851.
12. Moharrak M., Mogaji E. Generative AI in banking: Empirical insights on integration, challenges and opportunities in a regulated industry // International Journal of Bank Marketing. 2025. Vol. 43. No. 4. P. 871–896.
13. Mogaji E., Jain V. How generative AI is (will) change consumer behaviour: Postulating the potential impact and implications for research, practice, and policy // Journal of Consumer Behaviour. 2024. Vol. 23. No. 5. P. 2379–2389.