ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2025. Vol. 7, no. 4 (23)
SYSTEMS ENGINEERING & INFORMATION TECHNOLOGIES
Scientific paper

UDC 141.1 DOI 10.54708/2658-5014-SIIT-2025-no4-p93 EDN LULOGI

PHILOSOPHY OF AI DESIGN: HUMAN-IN-THE-LOOP AND BOUNDED RATIONALITY

O. I. ELKHOVA

The article explores the philosophical foundations and limitations of artificial intelligence (AI) rationality in the context of decision-making. The author analyzes the distinction between epistemic and practical rationality, emphasizing the latter as the basis for the operation of rational agents. Central to the discussion is the concept of bounded rationality, according to which decisions are made under conditions of incomplete information, cognitive limitations, and restricted computational resources. It is argued that ideal rationality is unattainable, and that bounded rationality represents the most adequate model for AI. Four types of rationality are considered: ideal, computational, bounded, and bounded optimality, with the conclusion that bounded rationality is the most practical applicable in the development of intelligent systems. The author concludes that a bounded approach is essential for the practical design of such systems. It is noted that these limitations must be taken into account when creating adaptive algorithms and that human involvement in the decision-making process is crucial for enhancing the reliability of outcomes. The human-in-the-loop model is interpreted not merely as a technical mode of interaction, but as an expression of situational rationality, which entails attention to context, moral consequences, and the uniqueness of each specific case. Human presence lends algorithmic reasoning to a value-laden and interpretive dimension, reestablishing the link between rationality and practical wisdom (phronesis). In situations where algorithms are constrained by resources and prone to errors, it is the human who can identify contextual nuances and meaningfully adjust decisions. Thus, integrating the human into the loop becomes a key factor in ensuring the reliability of decisions made by artificial intelligence.

Human-in-the-loop; bounded rationality; artificial intelligence; practical rationality; rational agent; decision-making.

Цитирование: Elkhova O. I. Philosophy of AI design: Human-in-the-loop and bounded rationality // СИИТ. 2025. Т. 7, № 4(23). С. 93–100. EDN LULOGI.

Cite: Elkhova O. I. “Philosophy of AI design: Human-in-the-loop and bounded rationality " // SIIT. 2025. Vol. 7, no. 4 (23), pp. 93-100. EDN LULOGI.

INTRODUCTION

The concept of rationality in philosophy traces back to Aristotle, who distinguished between two types: epistemic and practical. The former is concerned with the justification of beliefs, while the latter pertains to actions and decision-making. This distinction reflects the difference between the process of reasoning and actual behavior. In the epistemic sense, rationality implies the possession of well-argued and reliably formed beliefs. However, even seemingly contradictory judgments may retain a rational character if they are justified within the framework of accepted methods of cognition. Aristotle emphasizes that the process of reasoning and behavior itself are distinct phenomena. Today, greater attention is paid to rational behavior, as it is tied to practical outcomes. The concept of the «rational agent» is used to describe a subject who makes decisions aimed at achieving the optimal result, even under conditions of uncertainty. Practical rationality differs significantly from epistemic rationality, which is focused on the alignment of reasoning with logical principles. Unlike the epistemic form, practical rationality encompasses not only logical aspects but also motivation, desires, real-world constraints, and is aimed at decision-making and choosing the best course of action under specific conditions. This function forms the foundation of artificial intelligence systems, which must analyze data, predict potential outcomes, and determine the most effective decisions. This process largely mirrors human practical reasoning, which is oriented toward goal achievement rather than the mere justification of beliefs. The notion of a ‘rational agent’ has become central to the discourse on artificial intelligence [Ба617, pp. 11–16]. In the context of AI, a rational agent is understood as a system capable of analyzing data, forecasting the consequences of actions, and selecting optimal decisions based on given parameters in uncertain environments. The aim of this study is to identify the philosophical foundations and limits of artificial intelligence’s rationality,

as a detailed analysis of its limitations in information processing and decision-making is becoming increasingly relevant today. The active integration of intelligent systems into critical areas of human activity calls for a reassessment of the boundaries of their rationality and the factors influencing the accuracy and justification of the decisions they make.

LIMITS OF HUMAN AND MACHINE RATIONALITY

Although artificial intelligence was originally designed to model human intelligence, it must be acknowledged that human decision-making is not always rational or optimal and is often biased. In this regard, it is worth recalling H. Simon's concept of bounded rationality, which explains decision-making through the lens of individuals' cognitive limitations and the incompleteness of available information. These ideas stand in stark contrast to traditional views that consider humans as fully rational agents capable of always choosing the best option based on objective analysis. Classical rationality theory, often ignoring its hidden assumptions, presumes that individuals possess exhaustive knowledge of all possible alternatives and their outcomes. In contrast, H. Simon emphasizes that under real-world conditions, people rarely strive for absolute optimality. Instead, they follow the principle of *satisficing,* preferring solutions that seem good enough, even if they are not the best. In this context, rationality plays a key role in explaining decision-making mechanisms. Simon metaphorically compared human cognitive limitations to one blade of a pair of scissors and the structure of the environment to the other: the mind «cuts» effectively only through the interaction between its limited capabilities and environmental cues [Sim57, pp. 198–199]. The theory of bounded rationality challenges the notion that people can make decisions purely based on objective data, free from emotion or cognitive bias. It highlights that rationality is defined not only by the final result but also by the decision-making process itself. In reality, decision-making involves analyzing available information, considering time constraints, and evaluating the relevance of potential alternatives. Cognitive barriers, knowledge gaps, and environmental influences all hinder the achievement of ideal outcomes. The relevance of these ideas is reinforced by studies focusing on the challenges of the digital age, which show that the development of modern information and communication technologies, including AI, introduces fundamentally new challenges. These developments call into question established notions of rationality and demand their fundamental rethinking [Елх24, pp. 27–30].

The issue of artificial intelligence's rationality is inseparably linked to the specifics of information processing. The connection acquires additional conceptual depth through philosophical inquiry into the dynamics of digital experience, with particular emphasis on the virtuality index proposed by O. I. Elkhova [Elk22]. This index conceptualizes immersion, involvement, and interactivity as integral dimensions for evaluating the subjective impact of virtual environments, thereby establishing a methodological basis for analyzing the modulation of bounded rationality and cognitive adaptation within human-machine interaction. The resulting theoretical developments make it possible to refine the philosophical framework of artificial intelligence by correlating bounded rationality with the ontological parameters of digitally mediated perception. These aspects are evident not only in the cognitive limitations of algorithms but also in the dynamics of human interaction with digital environments, including virtual reality. In such environments, the structural metrics of phenomenological experience influence the subjective mechanisms of decision-making. Analyses of the phenomenology of virtual experience reveal that perception within digital space relies on principles similar to those of bounded rationality. When interacting with virtual environments, humans synthesize sensory data from both physical and digital worlds, contributing to the formation of adaptive cognitive strategies. The concept of field interference between the real and the virtual in studies of phenomenological experience demonstrates that rationality in digital space cannot be reduced to a mere optimization algorithm [Елх24b, pp. 1003–1005]. On the contrary, rationality emerges from the interaction of multiple interrelated factors – cognitive, sensory, and social – requiring a complex, integrative approach.

When the number of possible alternatives is too large for comprehensive analysis, individuals often resort to simplified heuristics, which inevitably leads to deviations from optimal choices. Furthermore, subjective factors such as personal values, goals, and preferences significantly influence decision-making processes, sometimes distorting the objective evaluation of available options. Errors in judgment may arise because individuals lack unlimited cognitive resources and may be unaware of better alternatives or, conversely, mistakenly consider inferior options as preferable. In uncertain conditions, people tend to settle for what seems most suitable from their perspective, rather than maximizing the result. As a result, decisions do not always align with an objective maximum. H. Simon's concept has profoundly influenced AI development, revealing that machines also face the same fundamental constraints as humans. The theory of bounded rationality remains relevant in contemporary AI research, especially in the design of intelligent systems. Two main approaches to modeling rationality in AI can be identified. The first, bottom-up approach relies on human data and inputs, which may introduce bias. These distortions arise because the data used to train models often contain false assumptions and limited views, reflecting human prejudices. The second, *top-down,* is based on logic and formal rationality, but human biases may be embedded in the rules themselves. For example, in machine translation systems, specific linguistic patterns may distort results. Contemporary research actively explores hybrid approaches that combine the strengths of both models – data and logic – to minimize bias and improve decision-making rationality.

In discussions about AI capabilities, four types of rationality are often identified: ideal, computational, bounded rationality, and bounded optimality [Rus22, pp. 36–38, 58–59]. Ideal rationality implies decision-making based on complete knowledge of all alternatives and their consequences. However, in real conditions, such rationality remains unattainable due to incomplete information and the complexity of the tasks involved. Moreover, some problems require consideration of a vast number of variables, rendering them computationally infeasible even for the most powerful modern systems. In this light, the concept of *computational rationality* becomes more applicable. It takes into account available resources and seeks acceptable solutions within computational constraints.

Nevertheless, challenges remain: the more complex a problem is, the more resources are required to solve it, making it impossible to analyze all alternatives within a reasonable time frame. For instance, in tasks related to pattern recognition or prediction, the system is forced to use heuristics and approximation methods rather than conducting a full analysis of all possible options. Many tasks faced by AI belong to the class of computationally hard problems, which means that the number of possible solutions grows exponentially or becomes unsolvable in polynomial time. The combination of limited computational resources and the complexity of real-world decision-making make achieving optimal outcomes practically impossible.

One manifestation of this issue is the *combinatorial explosion* – a rapid increase in the number of potential options that renders exhaustive search infeasible even for the most powerful systems. This is especially evident in AI systems tasked with optimization and planning, where exponential growth in possible combinations necessitates the use of heuristic methods and approximate algorithms. As a result, AI must limit its reasoning within a particular *frame,* which inevitably simplifies its worldview but prevents paralysis caused by data overload.

BOUNDED OPTIMALITY AND THE LIMITS OF COMPUTABILITY

The concept of *bounded optimality,* formulated by St. Russell, posits that the optimal agent is not one acting under abstract conditions but one that achieves the best possible result within given resource constraints. According to St. Russell, «bounded optimality extends the traditional notion of rationality by explicitly incorporating the limitations of computational resources in real agents» [Rus16, p. 13]. These ideas have served as the basis for a number of studies on rational meta-reasoning, where intelligent agents allocate resources not only to solving tasks but also to evaluating the effectiveness of continued search (e.g., determining when it is reasonable to stop searching and proceed with implementation). Thus, a line of research stemming from H. Simon's work views

the intelligent agent as acting rationally within the bounds of its capabilities rather than in an absolute sense. Today, virtually all successful AI systems in one form or another implement the principles of bounded rationality, eliminating unpromising options due to resource limitations.

However, the limitations of AI are defined not only by technical factors but also by fundamental mathematical principles. As early as A. Turing's work, the existence of *undecidable problems* was established, most notably the *halting problem,* which no universal computing machine can solve [Tur50; Tur36]. This insight is of fundamental importance: every formal system of intelligence has boundaries, determined not only by available resources but also by the very laws of computability. Therefore, the limits of AI are not merely a matter of insufficient computational power but an inherent property of any algorithmic system. Another critical limitation is posed by *K. Gödel's incompleteness theorems*, which show that in any sufficiently powerful formal system, there are statements that can neither be proven nor disproven. Similar problems may arise in AI systems as well [Tou22, pp. 265–266, 276]. This means that certain classes of problems will remain fundamentally unsolvable for machine algorithms, regardless of their complexity or computational capacity.

Bounded optimality can be seen as a compromise between precision and cost: it does not require exhaustive analysis of all possible solutions but guarantees an acceptable result within a reasonable timeframe. This approach is especially relevant for real-world computing systems that must operate under time, memory, and computational constraints. Yet even this level of rationality is not always achievable, as algorithms may encounter resource barriers, and the computational complexity of certain problems makes finding a satisfactory solution impossible. Modern intelligent systems are designed with bounded rationality in mind: it is acknowledged that ideally rational algorithms are often infeasible, so algorithms must be effective within available resources. For example, planning and decision-making systems use heuristics, irrelevant-branch pruning, and *anytime algorithms,* which can halt computation upon finding an acceptable solution.

Since ideal rationality remains unattainable, computational rationality is resource-bound, and even bounded optimality is not always feasible, the most realistic and practically applicable model becomes *bounded rationality.* This model allows AI to be adapted to real-world conditions and existing constraints. Within this framework, an agent makes decisions not by searching for the globally optimal solution, but within the bounds of available options and limitations. In other words, the system seeks a *satisficing* solution rather than an absolute optimum. This approach is widely used in AI algorithms, especially in *multi-criteria decision-making,* where balancing solution quality with computational cost is critical.

The concept of bounded rationality also plays a key role in interdisciplinary studies, particularly in examining the interaction between humans and artificial intelligence. The integration of AI raises numerous ethical challenges, particularly regarding the distribution of responsibility for AI errors. Under such conditions, it is necessary to determine where responsibility lies — with developers, operators, or the system itself. A crucial task becomes finding an optimal balance between algorithmic autonomy and human oversight, enabling minimal intervention while preventing potential risks. Scholars such as L. Floridi emphasize the importance of *hybrid intelligence,* which combines machine algorithms with human interpretation and correction [All11, pp. 163–165].

A central aspect of this issue is the role of the *human-in-the-loop,* i.e., human involvement in making critical decisions to minimize risks. Since artificial intelligence lacks conscious perception, it should not be regarded as an autonomous agent but rather as a tool whose effectiveness depends on its interaction with humans. Three main models of such interaction are distinguished. The first model, known as human-on-the-loop, implies that a human supervises the decisions made by artificial intelligence and can intervene if necessary. The second, human-in-the-loop involves the active participation of a human in the decision-making process, including the monitoring and adjustment of algorithmic outcomes. The third model, human-out-of-the-loop, describes a scenario in which the system operates fully autonomously without human involvement, which increases risks, especially in critically important domains.

A broader ontological perspective on the identified limitations is presented in the work by A. F. Kudryashev and O. I. Elkhova titled "The Two-Faced Janus of the Evolutionary Essence of Artificial Intelligence" [Kud23]. In this article, artificial intelligence is examined not solely as a technological construct but as a component integrated into the process of global evolutionism. The authors argue that the nature of artificial intelligence is characterized by dual orientation: it remains grounded in the human past while simultaneously directed toward a potential post-human future. Within this conceptual framework, the rationality of artificial intelligence should be analyzed in light of evolutionary dialectics. This approach affirms that the limitations of artificial intelligence are shaped not only by formal and computational constraints but also by historical, philosophical, and ontological dimensions of its development.

HUMAN-IN-THE-LOOP: PHILOSOPHY, ETHICS, AND INTERACTION

As artificial intelligence continues to evolve, it is essential not only to improve algorithms but also to emphasize philosophically informed design. It must be remembered that removing humans from the decision-making loop does not eliminate responsibility – it merely obscures it and makes oversight more difficult. In situations where outcomes have clear moral significance – such as the allocation of medical resources, legal judgments of guilt, or the behavior of autonomous vehicles in accident scenarios – responsibility must remain with the human. This approach is grounded in the fundamental differences between computational processes and human experience, which includes empathy, conscience, and accumulated life knowledge. Several examples illustrate this point. AI is widely used in medical diagnostics, yet its conclusions must be verified by a physician. For instance, Google Health's algorithms demonstrated 94% accuracy in detecting breast cancer in 2021, but they failed to recognize rare tumor types that human specialists identified more accurately [For25]. This underscores the necessity of a hybrid approach, in which medical professionals evaluate AI outputs before making final decisions. Similar issues are found in the field of autonomous vehicles. In 2018, an Uber vehicle failed to recognize a pedestrian with a bicycle at night, leading to a fatal accident [Ube18]. Today, companies such as Tesla implement a human-on-the-loop model, requiring driver supervision even when autopilot is engaged.

Contemporary research highlights the need to view AI not as a fully autonomous system, but as a tool that functions in close collaboration with humans. Thanks to collective efforts from researchers and practitioners working to enhance the effectiveness of human-AI collaboration, the *Human-AI Teams* approach has been developed. Its goal is to optimize human-technology interaction, minimize the risks of automated decisions, and improve outcomes across various domains. The combination of machine analytics and human experience opens new horizons in medicine, transportation, economics, and other critical sectors. In the coming years, advances in adaptive AI are expected to produce systems capable not only of completing tasks but also of gradually learning to interact with humans by recognizing their preferences and work styles. For this reason, the concept of bounded rationality and the human-in-the-loop model have become essential components of AI development, ensuring a balance between efficiency and ethical accountability. Such participation cannot be fully realized without appealing to the concept of *situational rationality*, which posits that decision-making cannot be entirely algorithmized, as every situation contains elements of uniqueness and openness. Unlike autonomous machines, a human is not merely a supervisor or intervener but brings a value-oriented and interpretive dimension to the decision-making process. This means that rationality in the human-in-the-loop model is not limited to instrumental reasoning – it is grounded in the agent's ability to consider context, uncertainty, moral consequences, and the singularity of each situation.

The human-in-the-loop not only evaluates the algorithm's performance but also detects contextual subtleties inaccessible to a machine operating solely on statistical patterns. Human presence restores the link between rationality and *practical wisdom* (*phronesis*) – the understanding of concrete life situations into which each decision is embedded. Thus, the human-in-the-loop model enhances the safety and adaptability of systems while preserving the human dimension of decisions

in the context of digital rationality. Human involvement in the loop represents a philosophically grounded imperative that emphasizes the integration of rational choice with the existential, ethical, and sociocultural foundations of human action. This, in turn, points to the need to align the logic of algorithmic procedures with the 'logic of life' – a convergence of formal rationality with what philosophy calls second-order rationality, involving the capacity for self-reflection and the justification not only of outcomes but also of the assumptions behind the choices themselves.

The ethical and epistemological implications of this shared structure can be further explored through the lens of evolutionary models that account for law-governed trajectories in AI development. One such model is presented in "Predicting the Development of Artificial Intelligence: Nomogenetic Perspective" [Elk23]. Drawing on theory of nomogenesis, the authors propose that artificial intelligence may evolve along stable internal lines, which in turn opens the possibility of assigning predictive and foresight functions to AI systems themselves. This framework contributes to the philosophical and ethical discourse on human-in-the-loop configurations by highlighting the long-term implications of delegating elements of future-oriented reasoning to artificial agents.

CONCLUSION

The conducted study offers a comprehensive philosophical foundation for the development and application of intelligent systems, emphasizing the necessity of considering both technical and humanistic dimensions of rationality.

1. It is demonstrated that ideal (all-encompassing) rationality is unattainable for both humans and machines, given the real-world conditions that involve limited information, cognitive constraints, and finite computational resources. The study confirms that bounded rationality, as proposed by H. Simon, provides an adequate framework for describing the behavior of intelligent systems and for designing AI capable of operating under uncertainty.

2. Rationality cannot be reduced solely to logical operations: in practical contexts, it is essential to consider situational factors, moral consequences, and the uniqueness of each decision-making scenario. The human-in-the-loop model imparts interpretative and ethical dimensions to algorithmic reasoning. Human involvement enables the recognition of contextual nuances inaccessible to algorithms, thereby enhancing the reliability and validity of decisions.

3. The limitations of artificial intelligence are not temporary obstacles but fundamental boundaries arising from both resource constraints and mathematically proven limits inherent to any formal computational system. Within these constraints, the concept of bounded optimality emerges as a pragmatic strategy: the goal of intelligent agents becomes achieving the best possible outcome within available limits, rather than striving for an unattainable ideal optimum.

4. The future of AI development lies in the creation of hybrid models, in which algorithmic data processing is complemented by human interpretation. This interaction not only enhances the adaptability and efficiency of systems but also integrates sociocultural, ethical, and contextual dimensions into the decision-making process.

СПИСОК ЛИТЕРАТУРЫ

[All11] Allo P. (ed.). Putting Information First: Luciano Floridi and the Philosophy of Information. Wiley-Blackwell, 2011. DOI: 10.1002/9781444396836.

[Elk22] Elkhova O. I. Justification of the virtual index in philosophy // СИИТ. 2022. Т. 4, № 2(9). С. 5-12. EDN: EVAIPZ.

[Elk23] Elkhova O. I., Kudryashev A. F. Predicting the development of Artificial Intelligence: nomogenetic perspect // СИИТ. 2023. Т. 5, № 6(15). С. 3-10. EDN: BPPNEN.

[For25] Forster V. (2025, Jan. 13). "Artificial intelligence improves breast cancer diagnosis" // Forbes. [Online]. Available: https://www.forbes.com/sites/victoriaforster/2025/01/13/artificial-intelligence-improves-breast-cancer-diagnosis.

REFERENCES

Allo P. (ed.). Putting Information First: Luciano Floridi and the Philosophy of Information. Wiley-Blackwell, 2011. DOI: 10.1002/9781444396836.

Elkhova O. I. "Justification of the virtual index in philosophy" // SIIT. 2022. Vol. 4, no. 2(9), pp. 5-12. EDN: EVAIPZ.

Elkhova O. I., Kudryashev A. F. "Predicting the development of Artificial Intelligence: nomogenetic perspect" // SIIT. 2023. Vol. 5, no. 6(15), pp. 3-10. EDN: BPPNEN.

Forster V. (2025, Jan. 13). "Artificial intelligence improves breast cancer diagnosis" // Forbes. [Online]. Available: https://www.forbes.com/sites/victoriaforster/2025/01/13/artificial-intelligence-improves-breast-cancer-diagnosis.

[Kud23] Kudryashev A. F., Elkhova O. I. The two-faced Janus of the evolutionary essence of artificial intelligence // СИИТ. 2023. Т. 5, № 1(10). С. 76-83. EDN: QFEPLB.

[Rus16] Russell S. Rationality and Intelligence: A Brief Update // In: Müller V. C. (ed.). Fundamental Issues of Artificial Intelligence. Cham: Springer, 2016. Pp. 7-28. DOI: 10.1007/978-3-319-26485-1_2.

[Rus22] Russell S., Norvig P. Artificial Intelligence: A Modern Approach. Pearson Series in Artificial Intelligence, 2022. URL: https://aima.cs.berkeley.edu/global-index.html.

[Sim57] Simon H. Models of Man: Social and Rational. New York: John Wiley and Sons, Inc., 1957. URL: https://archive.org/details/modelsofman0000herb.

[Tou22] Tourlakis G. "Gödel's First Incompleteness Theorem via the Halting Problem" // Computability, Springer, 2022, pp. 265-280.

[Tur36] Turing A. "On computable numbers, with an application to the Entscheidungsproblem" // Proc. London Mathematical Society, 1936, vol. 42, pp. 230-265.

[Tur50] Turing A. "Computing machinery and intelligence" // Mind, 1950, vol. 49, pp. 433-460.

[Ube18] "Uber self-driving car kills woman in Arizona" // The Guardian. URL: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

[Баб17] Бабич М. Ю. Понятие рационального агента и многоагентные системы // Проблемы информатики в образовании, управлении, экономике и технике: сб. статей XVII Междунар. науч.-техн. конф. Пенза: ПДЗ, 2017. С. 11-16. EDN: ZQZLIJ.

[Елх24] Елхова О. И., Кудряшев А. Ф. Современные вызовы информационно-коммуникационных технологий // Вестник Самарского государственного технического университета. Серия: Философия. 2024. Т. 6, № 3. С. 27-34. EDN: KRERJY.

[Елх24b] Елхова О. И. Метрики феноменологического виртуального опыта // Вестник Российского университета дружбы народов. Серия: Философия. 2024. Т. 28, № 4. С. 997-1013. EDN: JLFZMM.

Kudryashev A. F., Elkhova O. I. "The two-faced Janus of the evolutionary essence of artificial intelligence" // SIIT. 2023. Vol. 5, no. 1(10), pp. 76-83. EDN: QFEPLB.

Russell S. Rationality and Intelligence: A Brief Update // In: Müller V. C. (ed.). Fundamental Issues of Artificial Intelligence. Cham: Springer, 2016. Pp. 7-28. DOI: 10.1007/978-3-319-26485-1_2.

Russell S., Norvig P. Artificial Intelligence: A Modern Approach. Pearson Series in Artificial Intelligence, 2022. URL: https://aima.cs.berkeley.edu/global-index.html.

Simon H. Models of Man: Social and Rational. New York: John Wiley and Sons, Inc., 1957. URL: https://archive.org/details/modelsofman0000herb.

Tourlakis G. "Gödel's First Incompleteness Theorem via the Halting Problem" // Computability, Springer, 2022, pp. 265-280.

Turing A. "On computable numbers, with an application to the Entscheidungsproblem" // Proc. London Mathematical Society, 1936, vol. 42, pp. 230-265.

Turing A. "Computing machinery and intelligence" // Mind, 1950, vol. 49, pp. 433-460.

"Uber self-driving car kills woman in Arizona" // The Guardian. URL: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

Babich M. Yu. The concept of rational agent and multi-agent systems // Problemy informatiki v obrazovanii, upravlenii, ekonomike i tekhnike: sb. statey XVII Mezhdunar. nauch.-tekhn. konf. Penza: PDZ, 2017, pp. 11-16. EDN: ZQZLIJ. (In Russian).

Elkova O. I., Kudryashev A. F. Modern Challenges of Information and Communication Technologies // Vestnik Samarskogo gosudarstvennogo tekhnicheskogo universiteta. Seriya: Filosofiya, 2024, vol. 6, no. 3, pp. 27-34. EDN: KRERJY. (In Russian).

Elkova O. I. Metrics of Phenomenological Virtual Experience // Vestnik Rossiyskogo universiteta druzhby narodov. Seriya: Filosofiya, 2024, vol. 28, no. 4, pp. 997-1013. EDN: JLFZMM. (In Russian).

ОБ АВТОРЕ

ЕЛХОВА Оксана Игоревна
Уфимский университет науки и технологий, Россия.
Проф. каф. философии и культурологии. Дипл. инж.-исследователь (СПбГТУ, 1996). Д-р филос. наук (БашГУ, 2011), доцент. Иссл. в обл. онтологии и теории познания, философии вирт. реальности, философии ИИ.
E-mail: oxana-elkhova@yandex.ru
ORCID: 0000-0002-5052-5935

ABOUT THE AUTHOR

ELKHOVA Oxana Igorevna
Ufa University of Science and Technology, Russia.
Prof. Dept. Philosophy and Cultural studies. Dipl. research engineer (SPbGTU, 1996). Dr. Philosophy Sciences (BashSU, 2011), Docent. Research interests: ontology and theory of knowledge, philosophy of virtual reality, philosophy of AI.
E-mail: oxana-elkhova@yandex.ru
ORCID: 0000-0002-5052-5935

МЕТАДАННЫЕ

Заглавие: Философия проектирования ИИ: «человек в контуре» и ограниченная рациональность.

Аннотация: В статье рассматриваются философские основания и ограничения рациональности искусственного интеллекта в контексте принятия решений. Автор анализирует различие между эпистемической и практической рациональностью, акцентируя внимание на последней как основе работы рациональных агентов. Центральное место занимает концепция ограниченной рациональности, согласно которой принятие решений осуществляется в условиях неполной информации, когнитивных ограничений и ограниченных вычислительных ресурсов. Обосновано, что идеальная рациональность недостижима, а наиболее адекватной моделью для ИИ является ограниченная рациональность. Рассмотрены четыре

METADATA

Title: Philosophy of AI design: Human-in-the-loop and bounded rationality.

Abstract: The article explores the philosophical foundations and limitations of artificial intelligence (AI) rationality in the context of decision-making. The author analyzes the distinction between epistemic and practical rationality, emphasizing the latter as the basis for the operation of rational agents. Central to the discussion is the concept of bounded rationality, according to which decisions are made under conditions of incomplete information, cognitive limitations, and restricted computational resources. It is argued that ideal rationality is unattainable, and that bounded rationality represents the most adequate model for AI. Four types of rationality are considered: ideal, computational,

типа рациональности: идеальная, вычислительная, ограниченная и ограниченная оптимальность, с выводом о практической применимости именно ограниченной рациональности в разработке интеллектуальных систем. Сделан вывод о практической применимости ограниченного подхода при разработке интеллектуальных систем. Автор отмечает, что необходимо учитывать указанные ограничения при создании адаптивных алгоритмов, а также значимость участия человека в процессе принятия решений для повышения надёжности результатов. Модель «человека в контуре» трактуется не просто как техническая форма взаимодействия, а как выражение ситуационной рациональности, предполагающей учёт контекста, моральных последствий и уникальности каждой конкретной ситуации. Присутствие человека придаёт алгоритмическому мышлению ценностный и интерпретативный характер, восстанавливая связь между рациональностью и практической мудростью (phronesis). В условиях, когда алгоритмы ограничены ресурсами и подвержены ошибкам, именно человек способен выявлять контекстуальные нюансы и осуществлять осмысленную корректировку решений. Таким образом, включение человека в контур является важнейшим фактором, обеспечивающим повышение надежности решений, принимаемых искусственным интеллектом.

Ключевые слова: человек в контуре, ограниченная рациональность, искусственный интеллект, практическая рациональность, рациональный агент, принятие решений.

Language: Английский.

Статья поступила в редакцию 27 мая 2025 г.

bounded, and bounded optimality, with the conclusion that bounded rationality is the most practical applicable in the development of intelligent systems. The author concludes that a bounded approach is essential for the practical design of such systems. It is noted that these limitations must be taken into account when creating adaptive algorithms and that human involvement in the decision-making process is crucial for enhancing the reliability of outcomes. The human-in-the-loop model is interpreted not merely as a technical mode of interaction, but as an expression of situational rationality, which entails attention to context, moral consequences, and the uniqueness of each specific case. Human presence lends algorithmic reasoning to a value-laden and interpretive dimension, reestablishing the link between rationality and practical wisdom (phronesis). In situations where algorithms are constrained by resources and prone to errors, it is the human who can identify contextual nuances and meaningfully adjust decisions. Thus, integrating the human into the loop becomes a key factor in ensuring the reliability of decisions made by artificial intelligence.

Key words: Human-in-the-loop; bounded rationality; artificial intelligence; practical rationality; rational agent; decision-making.

Language: English.

The article was received by the editors on 27 May 2025.