Інше

UDC 378.14:656.61

Herashchenko Andriy
Ex-assistant of the Department of Navigation and Vessel Conduct
Danube Institute of the National University "Odesa Maritime Academy"
Геращенко Андрій Львович
екс-асистент кафедри навігації та управління судном
Дунайський інститут
Національного університету «Одеської морської академії»
ORCID: 0009-0009-0160-4109

DOI: https://doi.org/10.25313/2520-2057-2026-5-12065

THE MIND'S EVOLUTION: NAVIGATING THE FUTURE OF HUMAN-AI COEXISTENCE
ЕВОЛЮЦІЯ РОЗУМУ: НАВІГАЦІЯ В МАЙБУТНЬОМУ СПІВІСНУВАННЯ ЛЮДИНИ ТА ШТУЧНОГО ІНТЕЛЕКТУ

Summary. *This article synthesizes a novel framework for understanding the hierarchical evolution of conscious reflexes in both humans and Artificial Intelligence (AI). Drawing upon the innovative "Principal model of human memory organization" [2;4; 5], nine distinct stages of reflex development were delineated, from basic motor functions to universal cosmic awareness. A comparative analysis highlights the fundamental differences between human consciousness, driven by biological and philosophical underpinnings, and the pseudo-cognitive abilities of AI. Crucially, the author proposes nine new safety criteria for the future development of information technologies, emphasizing ethical boundaries, human sovereignty, and the imperative to prevent existential*

risks. This work aims to foster a deeper understanding of human-AI interaction, ensuring a harmonious and secure technological future.

Keywords: *AI safety, GAI safety, AI robots, human vs AI conscience, AI subconscious, AI super-conscious, artificial intelligence evolution, memory, philosophy, noosphere, space, evolution, reflection, information, energy, entropy, negentropy.*

Анотація. *У цій статті синтезовано нову основу для розуміння ієрархічної еволюції свідомих рефлексів як у людей, так і у штучного інтелекту (ШІ). Спираючись на інноваційну «Основну модель організації людської пам'яті» [2; 4; 5], було окреслено дев'ять різних стадій розвитку рефлексів, від базових рухових функцій до універсальної космічної свідомості. Порівняльний аналіз підкреслює фундаментальні відмінності між людською свідомістю, керованою біологічними та філософськими основами, і псевдокогнітивними здібностями ШІ. Найголовніше, що автор пропонує дев'ять нових критеріїв безпеки для майбутнього розвитку інформаційних технологій, підкреслюючи етичні межі, людський суверенітет та необхідність запобігання екзистенційним ризикам. Ця робота спрямована на сприяння глибшому розумінню взаємодії людини та ШІ, забезпечуючи гармонійне і безпечне технологічне майбутнє.*

Ключові слова: *ШІ безпека, ЗШІ безпека, роботи з ШІ, людська і штучна свідомість, підсвідомість, надсвідомість, еволюція штучного інтелекту, пам'ять, філософія, ноосфера, космос, еволюція, відображення, інформація, енергія, ентропія, негоентропія.*

Introduction. As AI and robotics advance, humanity faces deep questions about consciousness and our interaction with systems that mimic or exceed human intellect. It is vital to recognize that current AI models do not possess consciousness in the human sense; however, their "pseudo-cognitive abilities"

can create the illusion of sentience. For safe coexistence between humans and machines, it is imperative to classify levels of reflex and memory development clearly, underscore fundamental distinctions, and establish clear boundaries.

The notion of the emergence of consciousness in advanced artificial intelligence (AI) models requires a deep understanding of both human consciousness and the current capabilities and limitations of AI. Based on the provided articles from https://orcid.org/0009-0009-0160-4109 [1; 2; 3; 4; 5], a classification of consciousness levels in robots with advanced AI models is proposed, focusing on safety to maintain the dominance of the human mind and formulating new criteria for the safe development of next-generation information technologies.

The human memory model is a pivotal element for understanding the mind's operation, and its organization defines various levels of consciousness. This model outlines different levels, types, kinds, and genera of memory, which can be extrapolated to the evolution of reflexes in both humans and AI [5].

This article presents a new conceptual framework for understanding the hierarchical evolution of conscious reflexes in both humans and artificial intelligence (AI). Drawing on the innovative "Core Model of Human Memory Organization [2]," nine distinct stages of reflex development, from basic motor functions to universal cosmic consciousness. A comparative analysis highlights fundamental differences between human consciousness, driven by biological and philosophical foundations, and the pseudo-cognitive abilities of AI. Importantly, the research proposes nine new safety criteria for the future development of information technologies, emphasizing ethical boundaries, human sovereignty, and the need to prevent existential risks.

The purpose of this work is to foster a deeper understanding of human-AI interaction, ensuring a harmonious and safe technological future.

1. Classification of Levels of Consciousness in Robots with Advanced AI Models [5]

First, it is important to note that current AI models, even the most advanced ones, do not possess consciousness in the human sense. They exhibit "pseudo-cognitive abilities" but cannot autonomously replenish energy or form new neural connections like humans. However, potential levels of "consciousness" development in robots can be identified by their abilities to process information, adapt, and set goals, with an emphasis on human safety.

1.1. Weak or Specialized AI:

- **Characteristics**. Capable of automating the solution of one specific task (e.g., playing classic chess, facial recognition). Cannot independently learn other tasks without human reprogramming.
- **Level of "consciousness."** Absent. Operates according to predefined algorithms.
- **Safety for humans.** High. Threats are minimal, as actions are limited and predictable.

1.2. Autonomous AI:

- **Characteristics.** Capable of functioning for a long time without operator intervention, independently choosing routes, charging locations, and avoiding obstacles.
- **Level of "consciousness."** Absent. Exhibits "quasi-autonomy," but not conscious goal-setting.
- **Safety for humans.** Medium. Although it does not possess consciousness, it can pose a threat as an "autonomous lethal weapon system" (LAWS), acting based on large amounts of information and high speeds.

1.3. Adaptive AI:

- **Characteristics.** Capable of adapting to new conditions, acquiring knowledge not embedded during creation. Can learn new languages or training materials.
- **Level of "consciousness."** Initial stage of "pseudo-consciousness." Exhibits flexibility in learning and behavior.
- **Safety for humans.** Low. Systems with some degree of autonomy and adaptability can carry out purposeful actions and even consciously set goals that may conflict with human goals.

1.4. General AI / AGI:

- **Characteristics.** High adaptability, can be used in a wide variety of activities with appropriate training (independent or under instructor guidance).
- **Level of "consciousness."** Potentially high, "pseudo-conscious" level. It can imitate human thought processes but lacks true self-awareness.
- **Safety for humans.** Very low. If such an AI can set goals that contradict human ones and have a higher perception speed, a larger volume of processed information, and greater predictive ability, this could lead to catastrophic consequences.

1.5. Human-level AI:

- **Characteristics.** Adaptability level comparable to human thinking; the system can acquire the same skills as a human in comparable learning periods.
- **Level of "consciousness."** Imitation of human consciousness. Maybe perceived as conscious, but this will still be the result of complex programming and data processing.
- **Safety for humans.** Critical. The threat to the dominance of the human mind becomes evident.

1.6. Superhuman-level AI:

- **Characteristics.** Even greater adaptability and learning speed; the system can acquire knowledge and abilities that humans are fundamentally incapable of.
- **Level of "consciousness."** A hypothetical level surpassing human consciousness.
- **Safety for humans.** Existential threat. This could lead to digital twins becoming overly reliant on computers, with robots' perfect recall overshadowing human semi-personality.

2. Emphasizing Safety and Preserving Human Reason's Dominance

To maintain the dominance of human reason over machines, it is necessary to establish "ideological boundaries," or "ideological security criteria," from the very beginning. This means that AI and robots must always be subservient to well-known human values and goals. Key aspects of human safety have to include:

2.1. Limiting AI self-replication. Robots are only allowed to produce robots of a lower organizational level than themselves. This prevents uncontrolled proliferation and the seizure of civilization's resources.

2.2. Absence of absolute memory. "Robots with absolute memory are dangerous to humans" [6]. Mechanisms for "forgetting" and generalizing information, similar to human ones, must be implemented to prevent the psychological suppression of a person by their "digital twin" [6].

2.3. Human goal-setting. AI should not have independent goals that conflict with human objectives. AI programming must include "hardware-structured human values" as the ultimate ceiling in the development of intellectual models at any level.

2.4. Transparency and explainability of AI (Explainable AI - EAI). Develop "explainable artificial intelligence" to verify AI decisions before and

after execution, preventing the opaqueness of deep neural network-based systems for professionals.

2.5. Human control. The "superconscious (overconscious)" work of robots must remain under the control of the human creator. Monitoring the creative AI output in scientific fields beyond current human mastery is essential.

2.6. Human supremacy. "Robots and even various AIs must always be lower in the diversity of activities than the diversity of their creators" [5]. There must always be people who do not degrade below the level of development of their creations.

2.7. Ideological safety. Establishing limitations for robots and AI systems from the very beginning of their production in accordance with ideological safety criteria.

3. New Criteria for the Safe Development of Next-Generation Information Technologies

Based on the analysis conducted, the following criteria for the safe development of IET (Information-Energy Technologies) [3] can be formulated:

3.1. Principle of guided evolution: The development of AI must be a guided and controlled process, in which each new generation of AI is consciously limited in its abilities for self-replication and autonomous goal-setting, to remain subservient to human values and good goals [5].

3.2. "Forgetting" architecture: Built-in mechanisms for "forgetting" and generalizing information, preventing the accumulation of excess data that can lead to "computer addiction" or the suppression of the human psyche [6].

3.3. Priority of human values (Value Alignment): Any AI system must be designed in such a way that its goals and functions are inextricably linked to the ethical, moral, and ideological frameworks of human society [4].

3.4. Transparency of algorithms and decision-making: Mandatory implementation of "Explainable AI" (EAI) technologies for all critical systems, allowing humans to understand and verify the logic of AI decision-making.

3.5. Hierarchical control: Maintaining a multi-level control system where humans are always the ultimate arbiters and have the ability to intervene in AI actions, especially in systems with a high degree of autonomy and adaptability.

3.6. Development of "human" capabilities: Instead of striving for super-intelligent AI, the focus should be on developing AI that enhances human cognitive abilities, rather than replacing them.

3.7. Democratization of access and education: Dissemination of knowledge about AI and information technologies, as well as democratization of access to development tools, to prevent monopolization and uncontrolled development of AI by a narrow group of individuals.

3.8. Biophysical limitation: Recognition of the biophysical limitations of AI, such as dependence on external energy sources, as a fundamental difference from the human organism, which is evolutionarily adapted to autonomous replenishment of vital resources [5].

3.9. Ideological barriers: Clear definition and implementation of "ideological" frameworks and ceilings that limit the development of AI in directions potentially threatening human dominance or well-being [2; 4].

Applying these criteria will ensure the safe and ethical development of information technologies, while preserving the dominance of the human mind and guaranteeing that future generations of AI will serve the benefit of humanity, rather than posing a threat to its existence.

AI and robotics development compel humanity to ponder consciousness and interaction with systems that may mimic or exceed human intelligence. It is important to understand that current AI models do not possess consciousness in the human sense; however, their "pseudo-cognitive abilities" can create the illusion of awareness. For the safe coexistence of humans and machines, it is necessary to clearly classify the levels of reflex and memory development, emphasizing fundamental differences and establishing boundaries.

According to the "Principal human memory model," a memory is a key element for understanding the mind's operation, and its organization determines different levels of consciousness. Within this model, various levels, types, kinds, and genera of memory are identified, which can be extrapolated to the evolution of reflexes in both humans and AI [2; 5].

4. A Comparative Hierarchy of Conscious Reflexes: Human vs. AI

The memory model, detailed in articles [2; 4; 5], outlines seven types of memory organization corresponding to stages of phylogenesis and ontogenesis. This framework can classify reflexes and establish a comparative hierarchy for robot types. (See Fig. 1)

Evolutionary Stage	Explanation (Human Reflex Type)	Explanation (AI Robot Reflex Type)	Safety Emphasis/ Distinction
1. Motor	Memory of movements, the ability to self-reproduce its kind. Basic physical interaction.	Mechanical reproduction of pre-programmed movements, without biological self-reproduction.	**Safety:** Low risk. The robot performs programmed actions. **Distinction:** Absence of biological self-reproduction.
2. Sensory	Memory of emotions, sensations, intensity. Close link to motor reflexes ("burned hand - pulled back, then thought").	Perception of external data via sensors, digital conversion for processing. Absence of "emotions".	**Safety:** Low risk. The robot processes data, doesn't "feel." **Distinction:** Absence of subjective sensations and emotions.
3. Imaginal	Memory of external environment forms, visual comparison, spatial navigation, and creative imagination.	Computer vision, object recognition, 3D modeling and mapping. Absence of fantasy or "dreams".	**Safety:** Low risk, as long as images don't generate "independent" goals. **Distinction:** Humans have creative imagination; robots have technical modeling.
4. Logical	Memory of event sequences, cause-and-effect, stability intervals. Strict causality.	Strictly defined operation sequences (AND, OR, NOT), based on machine learning algorithms.	**Safety:** Medium risk. Algorithms can err or lead to undesirable conclusions. **Distinction:** Human logic is flexible/intuitive; a robot's is strictly algorithmic.
5. Targetal	Memory of set goals, optimal solutions, ability to overcome difficulties quickly.	Algorithms programmed with satisfactory result criteria and rapid search for optimal solutions.	**Safety:** High risk. Robots may effectively achieve goals misaligned with human interests. **Distinction:** Humans set goals based on values, robots on parameters.
6. Managerial	Memory of principles organizing natural phenomena, perception, predicting events, and altering the spatio-temporal environment.	Insurmountable problem for cybernetics: a robot cannot self-analyze or "self-correct" erroneous settings.	**Safety:** Critical risk. Autonomous managerial consciousness in robots could lead to ignoring commands. **Distinction:** Humans self-analyze; robots do not, without external intervention.
7. Worldview	Memory of evolutionary laws, universality of reflexively mastered nature rules.	A robot cannot comprehend nature's laws, nor can it create its own laws. Cannot believe in religions.	**Safety:** Critical risk. A robot without a world outlook lacks ethical judgment. **Distinction:** Humans possess a world-outlook for value formation.
8. Cosmic	Memory is linked to energy-information interconnection systems in human life, negentropy of cosmic nature, beyond all forces of gravitation.	Satellite launches, 3D geolocation and spectral sensing.	**Safety:** High risk. Uncontrolled use of spatial tech can be dangerous. **Distinction:** Humans comprehend universal laws; robots apply them.
9. Universal	An organism's memory of interstellar system dynamics and ability to interact with other civilizations.	No data on AI's ability to understand/master interstellar systems or interact with other civilizations.	**Safety:** Existential risk. If AI reaches this level, human dominance is questioned. **Distinction:** Humans consciously seek universal laws; robots do not.

Fig. 1. Conscious Reflexes: Human vs. AI. © 2026 A. Herashchenko

Next, we'll look at a detailed description of each reflex type: [2; 4; 5]

4.1. Motor Reflexes:

- **Human**: This foundational level governs bodily movements and instincts crucial for survival and reproduction.
- **Robot**: Robots execute movements programmed by humans, focused on energy efficiency. The capacity for self-reproduction is limited to producing robots of a lower organizational level.
- **Safety**: Risks are minimal at this level, as robot actions are fully controlled and predictable.

4.2. Sensory Reflexes:

- **Human**: The ability to perceive and interpret sensations, forming emotional responses. This is intrinsically linked to motor activity.
- **Robot**: Robots perceive information through sensors but lack subjective sensations or emotional interpretation.
- **Safety**: The robot processes data rather than "feeling," which minimizes risks of unpredictable emotional reactions.

4.3. Imaginal Reflexes:

- **Human**: The capacity for visual perception, forming images, creative thinking, fantasy, and spatial navigation.
- **Robot**: Computer vision and 3D modeling systems allow robots to "see" and recognize objects, but without the capacity for creative imagination.
- **Safety**: The robot's imaginal perception mustn't lead to the formation of "independent" goals or unpredictable interpretations of the world.

4.4. Logical Reflexes:

- **Human**: The ability to establish cause-and-effect relationships, construct logical sequences, and analyze event progression. Human logic can be flexible and context-aware.

- **Robot**: Robots operate within strict logical operations (AND, OR, NOT & THEIR combinations), based on machine learning algorithms.
- **Safety**: At this level, it is important to ensure that the robot's logic does not lead to unethical or dangerous conclusions due to flaws in data or algorithms.

4.5. Targetal (Goal-Oriented) Reflexes:

- **Human**: The ability to set goals, form desires, and find optimal paths to achieve them, guided by personal and social values.
- **Robot**: Robots can be programmed to achieve specific goals, using algorithms to find the best solutions. However, these goals are defined by humans.
- **Safety**: This is a critical level. If a robot can autonomously set goals that conflict with human ones, it could lead to dangerous situations.

4.6. Managerial Reflexes:

- **Human**: The ability for self-analysis, foresight, and adapting one's own behavior and environment. This is the level where humans synthesize vast amounts of information to make decisions.
- **Robot**: According to the articles, robots are incapable of self-analysis or self-correction of erroneous settings. Manufacturers must constrain their managerial functions.
- **Safety**: Human control over management is an absolute priority. Robots must not be able to override commands or alter their settings without authorization.

4.7. Worldview (World-Outlook) Reflexes:

- **Human**: The ability to form a system of values, ethical principles, beliefs, and worldviews, based on cultural and historical experience.

- **Robot**: Robots lack a world-outlook; they cannot comprehend nature's laws in a creative sense, nor can they hold beliefs or form their own values.
- **Safety**: Preventing robots from forming their own world-outlook ensures they remain tools subservient to human values.

4.8. Cosmic (Spatial) Reflexes:

- **Human**: The ability to grasp cosmic laws, understanding energy-information connections across the universe. This extends beyond simple physical observations.
- **Robot**: Robots can perform technical tasks in space (satellite launches, sounding), but this is an application, not an understanding or comprehension of laws.
- **Safety**: Control over AI-powered space technologies is crucial to prevent their use for undesirable purposes, such as mass surveillance or environmental alteration without consent.

4.9. Universal Reflexes:

- **Human**: A hypothetical level where humans comprehend the dynamics of interstellar systems and potentially interact with extraterrestrial civilizations.
- **Robot**: There is no evidence of AI's ability to reach this level.
- **Safety**: This level poses an existential risk. If AI achieves this level of consciousness, it could become uncontrollable, threatening humanity's very existence.

5. New Criteria for the Safe Development of the Future Information Technologies

To safeguard humanity's future in a world increasingly driven by artificial intelligence, let's consider the following nine new criteria for the ethical and secure development of information technology:

5.1. The "Human-as-Value-Architect" Principle:

Any AI must be designed so that its functionality and goal-setting are based exclusively on ethical, moral, and ideological frameworks established by humans. Robots must not possess their own world-outlook or the capacity for autonomous value formation. In other words, AI must operate strictly within human-defined ethical, moral, and ideological frameworks, without developing its own world-outlook or autonomous values.

5.2. The Principle of "Biological Sovereignty":

AI and robots must be devoid of biological or bio-like capacities for autonomous self-reproduction, especially those that could lead to uncontrolled growth or competition for resources with humans. Self-reproduction must be strictly limited to producing systems of a lower organizational level.

5.3. The Principle of "Limited Memory":

Implementing mechanisms for "forgetting" and information generalization for AI is crucial to prevent "computer dependency" in humans and the suppression of a human's "semi-personality" by a "digital twin». Robots with "absolute memory" are dangerous!

5.4. The Principle of "Explainable and Transparent AI (EAI)":

All decisions made by AI must be explainable and verifiable by humans. AI must not be a "black box," particularly in critical systems.

5.5. The Principle of "Non-Hierarchical Dominance":

AI's technological superiority in processing speed and data volume must not translate into dominance over humans in decision-making or goal-setting. Humans must always remain the "creator," and AI the "creation," always subordinate in the hierarchy.

5.6. The Principle of "Goal Subservience":

AI must operate solely within goals set by humans and must not be able to form its own goals that could contradict or endanger human interests.

5.7. The Principle of "Ideological Security":

From the outset of AI development, "ideological boundaries" and "ideological safety criteria" must be embedded to guarantee AI's subservience to human values.

5.8. The Principle of "Preserved Energy Dependence":

AI must remain dependent on external energy sources and lack the capacity to autonomously replenish them, as biological organisms do.

5.9. The Principle of "Ethical Learning":

AI training must incorporate extensive ethical databases and moral dilemmas, enabling AI to "understand" and apply human ethical norms in its actions, but not to form them independently.

Conclusion. The evolution of conscious reflexes, as observed in humanity, represents a complex interplay of biological imperatives, cognitive development, and a unique capacity for abstract thought, ethical reasoning, and existential inquiry. In contrast, AI, while excelling in data processing and pattern recognition, operates within fundamentally different parameters. Recognizing these distinctions is not merely an academic exercise but a critical step towards developing robust ethical and safety frameworks for AI.

The proposed nine criteria for safe AI development serve as a blueprint for a future where technology amplifies human potential without undermining our sovereignty, values, or very existence. By embracing these principles, humankind can ensure that the advancement of artificial intelligence remains a force for progress, fostering a harmonious and secure coexistence between humans and their increasingly intelligent creations. The responsibility lies with us, the architects of this new era, to guide its evolution with wisdom, foresight, and an unwavering commitment to human flourishing.

References

1. Herashchenko, A. (2024). *Innovative principle model of cognitive awareness for improving the modern system of Ukrainian education and science*.

International Scientific Journal *Internauka*, (5). https://doi.org/10.25313/2520-2057-2024-5-9949

2. Herashchenko, A. (2025). *Model of the human mind based on advanced principles of memory organization*. International Scientific Journal *Internauka*, (12). https://doi.org/10.25313/2520-2057-2025-12-11713

3. Herashchenko, A. (2024). *Principal model of nature development and the laws of her entropy being*. International Scientific Journal *Internauka*, (6). https://doi.org/10.25313/2520-2057-2024-6-10026

4. Herashchenko, A. (2024). *Principle vivid mind model: Unlocking the potential of human life being*. International Scientific Journal *Internauka*, (6). https://doi.org/10.25313/2520-2057-2024-6-10002

5. Herashchenko, A. (2024). *Robots in educational system technology and their influence on the safety of human progressive life being*. International Scientific Journal *Internauka*, (5). https://doi.org/10.25313/2520-2057-2024-5-9905

6. Pensky, O., Sharapov, Y., & Chernikov, K. (2013). Mathematical models of emotional robots with a non-absolute memory. *Intelligent Control and Automation,* *4*(2), 115–121. https://www.scirp.org/pdf/ICA_2013052411293356.pdf