MAKTABGACHA VA MAKTAB TA'LIMI VAZIRLIGI

O'ZBEKISTON RESPUBLIKASI OLIY TA'LIM, FAN VA INNOVATSIYALAR VAZIRLIGI

O'zbekiston Milliy Pedagogika Universiteti

BARQAROR HAYOT
xalqaro ijtimoiy-ma'rifat markazi

"ZAMONAVIY PSIXOLOGIYANING DOLZARB MUAMMOLARI"

mavzusida xalqaro miqyosda ilmiy-amaliy anjuman materiallari

“Zamonaviy psixologiyaning dolzarb muammolari” mavzusida xalqaro miqyosda ilmiy-amaliy anjuman materiallari
660 b.

O‘zbekiston Respublikasi Maktabgacha va maktab ta’limi vaziri E.Karimova tahriri ostida chop etildi.

Nizomiy nomidagi O’zbekiston milliy pedagogika universiteti 2025-yil 10-noyabrda bo‘lib o‘tgan ilmiy texnik kengashda muhokama qilingan va nashrga tavsiya etilgan

CONTENTS

MUNDARIJA СОДЕРЖАНИЕ CONTENTS

MUNDARIJA СОДЕРЖАНИЕ CONTENTS

MUNDARIJA СОДЕРЖАНИЕ CONTENTS

MUNDARIJA СОДЕРЖАНИЕ CONTENTS

MUNDARIJA СОДЕРЖАНИЕ CONTENTS

MUNDARIJA СОДЕРЖАНИЕ CONTENTS

MACHINE PSYCHOLOGY: BRIDGING HUMAN LEARNING PRINCIPLES AND ARTIFICIAL GENERAL INTELLIGENCE DEVELOPMENT

Farkhod Alisherovich Alisherov
PhD, Vice-Rector for Research and Innovation of the
National Pedagogical University of Uzbekistan named after Nizami

Nigora Ruzikulova Shukhratovna
PhD, associate professor, leading specialist of the department
“Scientific-Pedagogical Personnel Training”
of the National Pedagogical University of Uzbekistan named after Nizami

Abstract: The pursuit of Artificial General Intelligence (AGI) represents one of the most ambitious goals in artificial intelligence research, yet current approaches often overlook the foundational principles that govern biological intelligence. This paper introduces and examines Machine Psychology as an interdisciplinary framework that systematically integrates principles from learning psychology-particularly operant conditioning and Relational Frame Theory-with adaptive reasoning systems to advance AGI development. We propose a bidirectional learning model wherein psychological principles inform AI architecture while AI systems provide novel insights into cognitive mechanisms. Through analysis of recent implementations using the Non-Axiomatic Reasoning System (NARS), we demonstrate how core psychological constructs such as reinforcement learning, derived relational responding, and functional equivalence can be computationally realized to produce flexible, context-sensitive artificial cognition. This framework addresses critical limitations in contemporary AI systems, including brittleness in novel contexts, inability to generalize across domains, and lack of metacognitive capabilities. The paper further explores implementation challenges specific to developing economies, using Uzbekistan’s accelerated digital transformation as a case study for culturally-adapted AGI development strategies. We conclude that Machine Psychology offers a principled pathway toward human-level artificial intelligence while simultaneously enriching our understanding of natural cognition.

Key words: Machine Psychology, Artificial General Intelligence, Operant Conditioning, Relational Frame Theory, Cognitive Architecture, NARS, Adaptive Systems.

Annotatsiya: The pursuit of Artificial General Intelligence (AGI) represents one of the most ambitious goals in artificial intelligence research, yet current approaches often overlook the foundational principles that govern biological intelligence. This paper introduces and examines Machine Psychology as an interdisciplinary framework that systematically integrates principles from learning psychology-particularly operant conditioning and Relational Frame Theory-with adaptive reasoning systems to advance AGI development. We propose a bidirectional learning model in which psychological principles inform AI architecture while AI systems provide novel insights into cognitive mechanisms. Through analysis of recent implementations using the Non-Axiomatic Reasoning System (NARS), we demonstrate how core psychological constructs such as reinforcement learning, derived relational responding, and functional equivalence can be computationally realized to produce flexible, context-sensitive artificial cognition. This framework addresses critical limitations in contemporary AI systems, including brittleness in novel contexts, inability to generalize across domains, and lack of metacognitive capabilities. The paper further explores implementation challenges specific to developing economies, using Uzbekistan’s accelerated digital transformation as a case study for culturally-adapted AGI development strategies. We conclude that Machine Psychology offers a principled pathway towards human-level artificial intelligence while simultaneously enriching our understanding of natural cognition.

Kalit so‘zlar: Machine Psychology, Artificial General Intelligence, Operant Conditioning, Relational Frame Theory, Cognitive Architecture, NARS, Adaptive Systems.

Аннотация: Создание общего искусственного интеллекта (ОИИ) представляет собой одну из самых амбициозных целей в исследованиях искусственного интеллекта, однако современные подходы часто упускают из виду основополагающие принципы, лежащие в основе биологического интеллекта. В данной статье рассматривается машинная психология как междисциплинарная структура, систематически интегрирующая принципы психологии обучения, в частности, оперантное обусловливание и теорию реляционных фреймов, с адаптивными системами рассуждений для содействия развитию ОИИ. Мы предлагаем двунаправленную модель обучения, в которой психологические принципы формируют архитектуру ИИ, а системы ИИ предоставляют новые знания о когнитивных механизмах. Анализируя недавние реализации с использованием неаксиоматической системы рассуждений (NARS), мы демонстрируем, как основные психологические конструкции, такие как обучение с подкреплением, производные реляционные реакции и функциональная эквивалентность, могут быть реализованы вычислительным путем для создания гибкого, контекстно-зависимого искусственного познания. Эта структура учитывает критические ограничения современных систем ИИ, включая хрупкость в новых контекстах, невозможность обобщения между предметными областями и отсутствие метакогнитивных возможностей. В статье дополнительно рассматриваются проблемы внедрения, характерные для развивающихся экономик, на примере ускоренной цифровой трансформации Узбекистана в качестве примера для разработки культурно адаптированных стратегий развития искусственного интеллекта. Мы приходим к выводу, что машинная психология предлагает принципиальный путь к созданию искусственного интеллекта человеческого уровня, одновременно обогащая наше понимание естественного познания.

Ключевые слова: машинная психология, искусственный интеллект общего назначения, оперантное обусловливание, теория реляционных фреймов, когнитивная архитектура, NARS, адаптивные системы.

INTRODUCTION

Artificial General Intelligence has been characterized as the Holy Grail of artificial intelligence research since the field’s inception in the 1950s. Unlike narrow AI systems optimized for specific tasks, AGI aims to replicate the breadth and flexibility of human intelligence across diverse domains and novel situations (Morris et al., 2024). Despite remarkable advances in machine learning, particularly with large language models and deep reinforcement learning systems (Zhao et al., 2023), the creation of truly general artificial intelligence remains elusive. Contemporary AI systems, while impressive in their specialized domains, exhibit fundamental limitations when confronted with tasks requiring genuine understanding, flexible reasoning, or transfer of knowledge to unfamiliar contexts.

The challenge lies not merely in scaling existing approaches but in identifying the foundational principles that underpin general intelligence itself. Recent work by Johansson (2024a, 2024b) has proposed Machine Psychology as a novel interdisciplinary framework that addresses this challenge by systematically integrating principles from learning psychology with computational reasoning systems. This approach diverges from prevailing paradigms that either attempt to simulate neural architectures through brain emulation or rely solely on scaling transformer-based models. Instead, Machine Psychology adopts a principle-based methodology, seeking to identify and implement the psychological mechanisms that enable organisms to learn, adapt, and reason across contexts.

This paper contributes to the growing discourse on AGI development by proposing a comprehensive bidirectional learning framework grounded in Machine Psychology principles. Our central thesis is that the relationship between psychology and AI should be mutually informative: psychological theories provide blueprints for artificial cognitive architectures, while computational implementations offer testable models that advance our understanding of natural cognition. We specifically examine how operant conditioning principles (Skinner, 1938; Sutton & Barto, 2018), Relational Frame Theory (Hayes et al., 2001, 2021), and adaptive reasoning systems (Wang, 2006, 2013) can be synthesized to create more flexible and human-like artificial intelligence.

The timing of this research is particularly relevant given the accelerated global investment in AI infrastructure. Nations across Central Asia, including Uzbekistan, Kazakhstan, and Kyrgyzstan, have unveiled ambitious national AI strategies targeting economic transformation through digital technologies (Sayfiddinov, 2024; Nurgazina et al., 2025). Uzbekistan’s Strategy for the Development of Artificial Intelligence until 2030 exemplifies this trend, projecting a $1.5 billion AI software market and committing to train one million developers in partnership with international technology firms. Understanding how psychological principles can inform AGI development becomes critical not only for advancing the science but also for ensuring that AI systems deployed in diverse cultural and economic contexts exhibit appropriate flexibility and adaptability.

2. Background and Related Work

2.1 The AGI Challenge and Current Approaches

The definition of AGI itself remains contested within the research community. Google DeepMind’s framework proposes levels of AGI based on performance breadth and generality, ranging from narrow competence in specific tasks to systems that surpass human capabilities across all cognitive domains (Morris et al., 2024). This gradation acknowledges that AGI development is not a binary achievement but rather a continuum of increasingly general capabilities. Recent assessments suggest that contemporary large language models demonstrate emergent properties that some researchers classify as preliminary AGI characteristics (Dillon et al., 2023), though significant debate surrounds these claims.

Current approaches to AGI can be broadly categorized into three paradigms. The first, exemplified by scaled transformer models such as GPT-4 and its successors, relies on massively parallel training on diverse datasets to achieve broad task competence through pattern recognition and statistical inference (Zhao et al., 2023). The second approach pursues whole brain emulation, attempting to replicate neural structures and dynamics through detailed computational models of biological systems. The third paradigm, which Machine Psychology represents, adopts a principle-based methodology that identifies fundamental cognitive mechanisms and implements them computationally without necessarily replicating biological substrates (Johansson, 2024a).

Each approach confronts distinct challenges. Scaled language models, while demonstrating impressive capabilities in text generation and comprehension, exhibit brittleness when operating outside their training distribution and lack genuine understanding of causal relationships. Brain emulation efforts face the dual challenges of insufficient knowledge regarding neural computation and prohibitive computational requirements. The principle-based approach, while theoretically appealing, requires accurate identification of the relevant cognitive principles and successful translation of these principles into computational architectures (Goertzel et al., 2010).

2.2 Psychological Foundations: Learning Theory and Cognition

The foundations of Machine Psychology rest upon decades of research in behaviorist and post-behaviorist psychology. Operant conditioning, pioneered by Skinner (1938), describes how organisms modify behavior based on consequences, with positive reinforcement increasing the likelihood of repeated actions and punishment decreasing such likelihood. This fundamental learning mechanism has direct computational analogs in reinforcement learning algorithms, which have proven remarkably successful in training artificial agents for game-playing, robotics, and control tasks (Mnih et al., 2015; Sutton & Barto, 2018). The connection between operant conditioning and reinforcement learning illustrates how psychological principles can inform algorithmic design.

However, operant conditioning alone cannot account for the full complexity of human cognition, particularly our capacity for language and abstract reasoning. Relational Frame Theory, developed by Hayes and colleagues (2001), extends behavioral psychology by proposing that human symbolic cognition emerges from a learned capacity for Arbitrarily Applicable Relational Responding (AARR). Unlike simple associative learning observed in other species, AARR enables humans to derive novel relationships between stimuli based on contextual cues rather than physical properties. For instance, understanding that if A is larger than B, and B is larger than C, then A must be larger than C demonstrates combinatorial entailment—a form of relational reasoning that emerges without direct training on the A-C relationship.

RFT identifies three core properties of AARR: mutual entailment (bidirectional relationships), combinatorial entailment (deriving new relationships from existing ones), and transformation of function (transferring behavioral functions across related stimuli). Research demonstrates that training individuals in relational framing skills correlates with improvements in fluid intelligence and scholastic performance (Cassidy et al., 2016), suggesting that these capacities form a cognitive foundation for higher-order thinking. The implications for AI are profound: if relational reasoning constitutes a fundamental building block of intelligence, then computational systems implementing RFT principles might achieve more flexible and generalizable cognition.

2.3 Computational Implementations: NARS and Cognitive Architectures

The Non-Axiomatic Reasoning System represents a computational framework specifically designed for reasoning under conditions of insufficient knowledge and resources—constraints that characterize natural intelligence (Wang, 2006, 2013). Unlike classical logic systems that assume complete and consistent knowledge bases, NARS operates on non-axiomatic logic, which allows for revision of beliefs based on new evidence and handles contradictory information gracefully. This design philosophy aligns well with psychological theories emphasizing adaptive learning in uncertain environments.

Recent implementations have demonstrated NARS’s capacity to exhibit behavior consistent with operant conditioning principles (Johansson, 2024b). In experimental tasks involving simple discrimination, changing contingencies, and conditional discrimination, NARS successfully adapted its responses based on reinforcement patterns, achieving accuracy rates comparable to or exceeding those of biological organisms in analogous

tasks. Critically, the system demonstrated flexibility when contingencies reversed midway through tasks, rapidly adjusting behavior to maximize reinforcement under new conditions. These results suggest that NARS provides a suitable computational substrate for implementing psychological learning principles.

Extensions to NARS have begun incorporating RFT principles through mechanisms for acquired relations and relational naming (Hammer & Lofthouse, 2020; Johansson, 2025). These additions enable the system to derive symbolic relationships from sensorimotor experiences and generalize these relationships to novel stimuli. Preliminary results demonstrate successful modeling of stimulus equivalence, mutual entailment, and combinatorial entailment—core phenomena in RFT research (Sidman, 1994). The integration of operant conditioning with relational reasoning within a unified computational framework represents a significant step toward human-like cognitive flexibility.

Other cognitive architectures, including ACT-R, SOAR, and CLARION, have incorporated reinforcement learning modules to model human decision-making and skill acquisition (Laird, 2012; Langley, 2006). These architectures typically separate procedural knowledge (skill-based operations) from declarative knowledge (factual information), mirroring psychological theories of memory systems. ACT-R, for instance, uses production rules that strengthen or weaken based on reinforcement-like mechanisms, capturing both the acquisition of expertise and the persistence of cognitive biases. While these architectures have successfully modeled specific cognitive phenomena, they have not yet achieved the general flexibility that Machine Psychology aims to produce through systematic integration of learning principles.

3. The Machine Psychology Framework: A Bidirectional Learning Model

3.1 Core Principles and Architecture

Machine Psychology rests on three foundational principles that distinguish it from alternative AGI approaches. First, adaptation constitutes the core characteristic of both biological and artificial intelligence (Wang, 2006). Intelligence is not a static property but rather emerges from dynamic interaction with environments across varying timescales—from milliseconds during sensorimotor processing to years during developmental maturation. Second, operant conditioning provides a universal learning mechanism applicable to both simple stimulus-response associations and complex cognitive operations (Skinner, 1938; Sutton & Barto, 2018). The generality of reinforcement-based learning makes it a suitable foundation for artificial systems operating across diverse domains. Third, relational reasoning, as characterized by RFT, represents a learnable skill that amplifies cognitive capabilities exponentially once acquired, enabling language, mathematics, and abstract thought (Hayes et al., 2001).

The bidirectional learning model proposed here establishes a symbiotic relationship between psychological science and artificial intelligence development. In the forward direction, psychological principles inform the design of AI systems by identifying the minimal necessary components for intelligent behavior. Theories of memory, attention, learning, and reasoning derived from centuries of psychological research constrain and guide architectural decisions in artificial systems (Rumelhart et al., 1986; Werbos, 1974). This principle-based approach contrasts with purely empirical machine learning methods that discover statistical regularities without theoretical foundation.

In the reverse direction, computational implementations serve as precise, manipulable models that test and refine psychological theories. Traditional psychological research faces limitations in experimental control and measurement precision when studying natural cognition. Artificial systems implementing psychological principles can be subjected to systematic manipulations impossible with human subjects, providing insights into causal mechanisms underlying intelligent behavior (Chollet, 2019). This bidirectional flow creates an iterative refinement process: psychological theories inform computational designs, implementations reveal inadequacies or ambiguities in theories, refined theories guide improved architectures, and so forth.

3.2 Integration of Operant Conditioning with Reasoning Systems

The integration of operant conditioning principles with NARS exemplifies how Machine Psychology synthesizes psychological and computational approaches. Operant conditioning traditionally describes how organisms learn through trial and error, with behavior frequencies modulated by consequences (Skinner, 1938). In computational terms, this maps naturally onto reinforcement learning, where agents learn policies (mappings from states to actions) that maximize cumulative reward (Sutton & Barto, 2018). However, standard reinforcement learning algorithms typically operate on fixed state and action spaces with predefined reward functions—constraints that limit flexibility and generalization.

NARS addresses these limitations through its treatment of goals and sensorimotor experience (Wang, 2013; Hammer & Lofthouse, 2020). Rather than operating on fixed state representations, NARS constructs beliefs about the world based on sensory input and updates these beliefs as new information arrives. Actions are selected based on predicted consequences given current beliefs and goals, with the system learning which actions lead to desired outcomes through experience. This approach more closely resembles how organisms adaptively construct representations of their environments rather than operating on predetermined feature spaces.

The critical innovation lies in treating reinforcement not merely as an algorithmic updating rule but as a genuine environmental consequence that the system reasons about (Johansson, 2024b). When NARS receives positive reinforcement after an action, it doesn’t simply increment a value function as in Q-learning; instead, it forms beliefs about causal relationships between actions and outcomes, which it can then use for planning and reasoning. This distinction enables the system to generalize learned relationships to novel situations where surface features differ but underlying causal structures remain similar—a hallmark of human cognitive flexibility.

3.3 Relational Reasoning and Symbolic Cognition

While operant conditioning provides a mechanism for learning through interaction, RFT explains how humans transcend simple associative learning to achieve symbolic reasoning (Hayes et al., 2001, 2021). The key innovation of AARR is that relationships between stimuli become operants themselves—behaviors that can be reinforced and shaped. A child who learns that receiving a cookie after saying thank you exemplifies a specific coordination relation. However, once the child has acquired the general pattern of relating through coordination, they can apply this pattern to novel stimuli: if told that a schnauzer is a type of dog, they immediately infer that dogs include schnauzers, without requiring this reverse relationship to be explicitly taught. This bidirectional inference (mutual entailment) emerges from the relational frame itself rather than from direct training on each specific relationship.

Combinatorial entailment extends this capacity to chains of relationships: if A coordinates with B, and B coordinates with C, then A coordinates with C, even if this final relationship was never directly trained or observed (Sidman, 1994). More complex relational frames beyond simple coordination—such as opposition, comparison, spatial relations, and temporal relations—combine to produce the full richness of human symbolic cognition. Understanding negation, relative size, causality, and counterfactual reasoning all emerge from various combinations and applications of relational frames.

Implementing AARR computationally requires several components. First, the system must learn to respond not to stimulus properties themselves but to relations between stimuli. Second, it must generalize these relational patterns to arbitrary stimulus pairs based on contextual cues. Third, it must derive novel relations through mutual and combinatorial entailment. Fourth, it must transfer behavioral functions across related stimuli. Recent NARS extensions address these requirements by introducing acquired relations mechanisms that enable the system to learn same and opposite relationships from minimal training and generalize these relationships to novel stimuli (Johansson, 2025). Critically, the system demonstrates transfer of function, where behavioral significance learned in one context automatically applies to related stimuli identified through derived relations.

The potential impact of successfully implementing AARR in artificial systems extends far beyond academic interest. Language acquisition, mathematical reasoning, and analogical problem-solving all depend on relational reasoning capacities (Cassidy et al., 2016). An artificial system capable of genuine AARR would exhibit flexible intelligence resembling human cognition, able to understand explanations, follow complex instructions, and solve problems in domains far removed from its training experience. This represents a qualitative advance beyond current AI systems that excel at pattern matching within their training distributions but struggle with genuine understanding and transfer.

4. Cultural Adaptation and Regional Implementation: The Uzbekistan Case Study

4.1 Digital Transformation in Central Asia

Central Asia has emerged as a region of intensive AI investment and digital transformation. Uzbekistan’s 2030 AI Strategy, approved in 2024, commits $1.5 billion to AI software development, establishment of ten specialized AI research laboratories, and training of one million AI developers through international partnerships (Sayfiddinov, 2024). Kazakhstan aims to position itself as a regional AI hub with its International Center for Artificial Intelligence, while Kyrgyzstan has proposed establishing a Regional AI Hub coordinating efforts across the Eurasian Economic Union. These ambitious national strategies reflect recognition of AI as a critical technology for economic development and competitiveness in the 21st century.

The rapid pace of digital transformation in these contexts presents both opportunities and challenges for AGI development. On one hand, substantial government support, international partnerships, and motivated populations create favorable conditions for AI innovation. Uzbekistan’s IT Park hosts over 2,000 technology companies with fiscal incentives, while the Unified Portal of Interactive Public Services provides digital access to 60% of government services, demonstrating infrastructure readiness (Sayfiddinov, 2024). On the other hand, limited existing AI research capacity, shortage of experienced researchers, and nascent regulatory frameworks create implementation barriers.

Machine Psychology principles offer particular advantages for developing economies pursuing AGI development. The framework’s emphasis on fundamental learning principles rather than massive computational resources aligns well with contexts where access to computing infrastructure may be constrained. Training one million developers in psychological principles underlying intelligence creates a workforce capable of contributing

meaningfully to AGI research rather than merely applying existing tools. Moreover, the bidirectional learning model positions local researchers to contribute novel insights to global AI development while simultaneously building indigenous capacity.

4.2 Cultural Considerations in AGI Development

The development of AGI systems that operate effectively across diverse cultural contexts requires attention to how cultural factors shape cognition, communication, and behavior. While fundamental learning mechanisms described by operant conditioning likely exhibit cross-cultural universality (Skinner, 1938), the specific contingencies that shape behavior vary dramatically across cultural contexts. What constitutes appropriate behavior, effective communication, and successful problem-solving differs between individualist and collectivist societies, between high-context and low-context cultures, and between societies with varying power distances and uncertainty avoidance.

Relational frames, while constituting a universal cognitive capacity in humans, acquire culture-specific content through social learning (Hayes et al., 2001). The particular relational networks that individuals construct—concepts, categories, values, and beliefs—reflect cultural transmission of knowledge. An AGI system operating in Central Asian contexts would need to construct relational networks reflecting local conceptual structures rather than simply importing frameworks developed in Western contexts. This implies the necessity of training data that adequately represents regional linguistic, cultural, and conceptual diversity.

Language presents particular challenges for AGI deployment in linguistically diverse regions. Central Asia encompasses Turkic, Persian, and Slavic language families with distinct grammatical structures and semantic organizations (Nurgazina et al., 2025). Current large language models predominantly trained on English and major European languages exhibit performance degradation in low-resource languages. The principle-based Machine Psychology approach offers potential advantages here: if relational reasoning constitutes a language-universal capacity, then systems implementing RFT principles might transfer more readily across linguistic boundaries than purely statistical language models. However, this hypothesis requires empirical validation through multilingual implementation studies.

Ethical considerations specific to AGI development in contexts with emerging AI governance frameworks warrant attention. Uzbekistan’s early adoption of AI regulatory frameworks (among the first in Central Asia) demonstrates proactive engagement with governance challenges (Sayfiddinov, 2024). However, balancing innovation enablement with risk mitigation requires ongoing refinement as AGI capabilities advance. The transparency and interpretability advantages of principle-based approaches may facilitate regulatory oversight compared to opaque neural network systems, though substantial work remains in developing appropriate evaluation and monitoring frameworks.

5. Discussion and Future Directions

5.1 Advantages and Limitations

Machine Psychology offers several distinct advantages over alternative AGI approaches. First, the framework provides theoretical grounding in established psychological science, ensuring that architectural decisions reflect known properties of natural intelligence rather than engineering convenience or computational tractability alone (Johansson, 2024a). This principled foundation increases confidence that resulting systems will exhibit genuine intelligence rather than sophisticated pattern matching. Second, the emphasis on learning mechanisms rather than innate knowledge structures implies systems that can acquire new capabilities through experience rather than requiring complete reprogramming for each new domain.

Third, computational implementations provide unprecedented opportunities for testing psychological theories with precision impossible in human subject research (Chollet, 2019). The ability to manipulate learning histories, systematically vary environmental contingencies, and observe resulting behavioral patterns enables rigorous theory evaluation. Fourth, the relative computational efficiency of principle-based approaches compared to scaling massive neural networks makes AGI development more accessible to research groups and nations with limited resources. While current large language models require billions of parameters and enormous training datasets (Zhao et al., 2023), systems built on psychological principles may achieve comparable flexibility with dramatically reduced computational requirements.

However, significant limitations and challenges remain. The identification of relevant psychological principles, while informed by decades of research, remains incomplete. Human cognition almost certainly involves mechanisms beyond operant conditioning and relational framing, and the relative importance of different components remains unclear. Emotion, motivation, consciousness, and embodied cognition represent aspects of natural intelligence that current Machine Psychology implementations address inadequately if at all (Lundervold, 2025). The extent to which these components prove essential for AGI versus merely correlates of biological implementation remains an open empirical question.

Computational tractability presents another challenge. While NARS demonstrates successful implementation of psychological principles in controlled experimental tasks (Johansson, 2024b), scaling these approaches to

handle the complexity and ambiguity of real-world environments remains undemonstrated. The computational cost of maintaining and reasoning over large belief networks, deriving extensive relational inferences, and integrating multimodal sensory information may prove prohibitive. Hybrid approaches combining principle-based reasoning with neural network components might prove necessary, though this raises questions about maintaining theoretical interpretability.

5.2 Research Directions and Open Questions

Several promising research directions emerge from the Machine Psychology framework. First, systematic evaluation of which psychological principles prove essential for AGI versus merely helpful requires careful ablation studies. Implementations that systematically include or exclude specific mechanisms—episodic memory, attention, meta-cognition—while holding other components constant would clarify the minimal sufficient components for general intelligence (Laird, 2012; Langley, 2006). Such studies would inform both AI development and psychological theory.

Second, the integration of embodiment deserves attention. Current implementations typically operate on abstract symbolic representations with limited sensorimotor grounding. However, developmental psychology suggests that abstract reasoning emerges from concrete physical experience, with bodily interaction shaping conceptual structures. Robotic implementations of Machine Psychology systems that learn through physical manipulation and navigation could reveal whether embodiment proves essential or merely helpful for developing genuine understanding.

Third, social learning and communication merit investigation. Human intelligence develops within social contexts where knowledge is transmitted culturally rather than rediscovered individually (Sharma et al., 2023). Multi-agent systems implementing Machine Psychology principles could explore how artificial agents might learn from each other, develop shared conceptual frameworks, and coordinate behavior—all critical for AGI systems intended to interact with humans and other AI systems in complex environments.

Fourth, the relationship between Machine Psychology approaches and contemporary neural network methods requires clarification. Rather than viewing these as competing paradigms, hybrid architectures that combine symbolic reasoning with distributed representation learning might leverage the strengths of both approaches (Rumelhart et al., 1986). Neural networks excel at pattern recognition and perceptual processing, while symbolic systems provide interpretable reasoning and planning. Identifying optimal integration strategies represents a promising research direction.

Finally, evaluation metrics and benchmarks specific to Machine Psychology approaches require development. Existing AI benchmarks emphasize performance on narrowly defined tasks, often measuring pattern recognition rather than genuine understanding or flexible reasoning (Chollet, 2019). Benchmarks that assess transfer learning, contextual adaptation, and the ability to acquire new relational frames through minimal training would better evaluate progress toward AGI as conceived within the Machine Psychology framework.

CONCLUSION

Machine Psychology represents a promising paradigm for advancing toward Artificial General Intelligence through systematic integration of psychological learning principles with computational reasoning systems. By grounding AI development in established theories of operant conditioning and Relational Frame Theory (Skinner, 1938; Hayes et al., 2001), this approach provides a principled pathway toward flexible, adaptive artificial cognition that resembles natural intelligence in important respects. The bidirectional learning model—where psychological principles inform AI architecture while computational implementations refine psychological theories—creates a virtuous cycle accelerating progress in both fields (Johansson, 2024a).

Recent implementations using NARS demonstrate the viability of translating psychological principles into computational systems capable of adaptive learning and derived relational reasoning (Johansson, 2024b, 2025; Hammer & Lofthouse, 2020). These proof-of-concept studies, while limited in scope, suggest that fundamental aspects of human intelligence can be captured computationally through attention to underlying mechanisms rather than surface-level imitation of behavioral outputs. The ability of these systems to generalize learned relationships to novel stimuli and adapt flexibly to changing contingencies represents qualitative progress toward genuine artificial intelligence.

The relevance of Machine Psychology extends beyond academic interest to practical concerns about AI deployment in diverse global contexts. As nations across Central Asia and beyond invest heavily in artificial intelligence infrastructure (Sayfiddinov, 2024; Nurgazina et al., 2025), understanding how to develop systems that adapt to local cultural, linguistic, and economic conditions becomes critical. The principle-based approach outlined here offers advantages in resource efficiency and interpretability that may prove particularly valuable in emerging AI ecosystems.

Significant challenges remain before human-level artificial general intelligence becomes reality. Questions about the completeness of current psychological theories, the computational tractability of proposed mechanisms, and the necessity of additional components all require empirical investigation (Morris et al., 2024; Lundervold, 2025). However, Machine Psychology provides a coherent framework for organizing this research agenda and measuring progress toward the ultimate goal of creating artificial systems that match or exceed human cognitive capabilities across domains.

The potential societal implications of successful AGI development—both opportunities and risks—underscore the importance of pursuing this research thoughtfully and responsibly (Elyoseph et al., 2024). Machine Psychology’s emphasis on interpretable mechanisms and principled design may facilitate oversight and governance compared to opaque neural network approaches. As we progress toward increasingly capable AI systems, maintaining human agency and ensuring beneficial alignment with human values becomes paramount. The integration of psychological understanding into AI development provides tools for addressing these challenges while advancing the science of intelligence itself.

References

1. Cassidy, S., Roche, B., Colbert, D., Stewart, I., & Grey, I. M. (2016). A relational frame skills training intervention to increase general intelligence and scholastic aptitude. Learning and Individual Differences, 47, 222-235.
2. Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.
3. Dillon, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences, 27(7), 597-600.
4. Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2024). An ethical perspective on the democratization of mental health with generative artificial intelligence. JMIR Mental Health, 11, e58011.
5. Goertzel, B., Lian, R., Arel, I., de Garis, H., & Chen, S. (2010). A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures. Neurocomputing, 74(1-3), 30-49.
6. Hammer, P., & Lofthouse, T. (2020). OpenNARS for Applications: Architecture and control. In International Conference on Artificial General Intelligence (pp. 193-204). Springer.
7. Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). Relational Frame Theory: A Post-Skinnerian Account of Human Language and Cognition. New York: Kluwer Academic/Plenum Publishers.