The Autodidactic Loop: A Blueprint for a Continuously Self-Improving AGI

Vision Paper - November 2025

Correspondence: rawson.douglas@gmail.com

Abstract

Current AI systems are static; they are trained, deployed, and remain frozen until a human initiates a new training cycle. We propose a new architecture for a truly autonomous, continuously self-improving AGI that learns from its own experience without human intervention. Our system, the **Autodidactic LRL-LoRA Architecture (ALLA)**, is a multi-layered cognitive framework that mimics the process of human learning, from conscious thought to unconscious skill. It operates in a perpetual loop: (1) In a linguistic "conscious" workspace, the agent uses **Linguistic Reinforcement Learning (LRL)** to tackle novel problems, journaling its failures to discover and refine new, explicit strategies. (2) A "wisdom filter" identifies which of these learned strategies are fundamental and generalizable. (3) A background "distillation engine" then uses the perfected linguistic strategy to generate high-quality synthetic data, training a compact, efficient **LoRA adapter** that "burns" the new skill into the weights. (4) This new LoRA adapter is added to a composable chain of skills in the "unconscious mind," becoming a permanent, high-speed capability. This architecture solves the "plasticity vs. stability" dilemma, allowing the agent to continuously learn new skills without suffering from catastrophic forgetting. We argue that this blueprint, combining the interpretability of linguistic self-correction with the efficiency of modular weight-based updates, is a plausible and powerful path toward the creation of a truly autodidactic, general intelligence.

Keywords: AGI, Self-Improvement, Linguistic Reinforcement Learning (LRL), LoRA, Machine Psychology, Lifelong Learning, Cognitive Architectures

1. Introduction

The dominant paradigm of "train then deploy" is a fundamental bottleneck on the path to Artificial General Intelligence (AGI). Today's most powerful models are brilliant but static. They are like textbooks with all of human knowledge but no ability to learn from their own post-deployment experience. An AGI cannot wait for humans to retrain it; it must be capable of learning and adapting on its own.

This requires a system that can:

- **Learn continuously** from its own actions and observations.
- **Distinguish** between temporary knowledge and fundamental wisdom.
- **Integrate** new skills without destroying old ones (avoiding catastrophic forgetting).
- **Do so autonomously**, without a human in the loop.

Existing methods fall short. Online learning forgets, full retraining is too slow, and linguistic-only methods are context-heavy and inefficient for permanent knowledge. We require a new architecture.

2. The Autodidactic LRL-LoRA Architecture (ALLA)

We propose ALLA, a cognitive architecture for a continuously self-improving agent. It is inspired by the human learning process, which seamlessly integrates slow, deliberate "conscious" thought with fast, efficient "unconscious" skill. ALLA consists of three interacting layers.

(Figure 1: A conceptual diagram showing the ALLA loop, cycling from the Conscious Workspace, through the Wisdom Filter and Distillation Engine, and into the Unconscious LoRA Chain.)

Layer 1: The Conscious Mind (Linguistic Workspace)

This is the agent's real-time problem-solving engine. When faced with a novel task, it operates within a high-cost, high-flexibility linguistic workspace.

- **Mechanism: Linguistic Reinforcement Learning (LRL)**. The agent attempts to solve the problem, journals its failures and successes, and iteratively refines a temporary, text-based strategy within its context window.
- **Analogy:** This is human "System 2" thinking (Kahneman, 2011). It is slow, deliberate, and effortful, like a student trying to solve a new kind of math problem for the first time by writing out every step.
- **Function:** To discover and validate new, effective procedures for previously unseen problems.

Layer 2: The Wisdom Filter & Distillation Engine (The "Sleep" Cycle)

This is the critical bridge between temporary knowledge and permanent skill. It operates as a background process, analyzing the successful strategies generated by the conscious mind.

- **Mechanism 1 (Wisdom Filter):** A meta-cognitive module that evaluates the strategies from Layer 1. It identifies a strategy as "fundamental" based on criteria such as:
    - **Frequency:** Has this strategy been used successfully hundreds of times?
    - **Generality:** Is this "diagnose before acting" meta-strategy succeeding across multiple, unrelated domains?
    - **Performance Uplift:** Did the adoption of this strategy lead to a step-change improvement in a specific skill?
- **Mechanism 2 (Distillation Engine):** Once a strategy is flagged as fundamental, this process is triggered. It uses the perfected linguistic strategy to generate thousands of high-quality, step-by-step examples. It then initiates a training run to distill this textual knowledge into a compact LoRA adapter (Hu et al., 2021).
- **Analogy:** This is the cognitive function of sleep. It's where the brain consolidates the day's learning, transfers knowledge from short-term to long-term memory, and solidifies new motor skills.

Layer 3: The Unconscious Mind (The Composable LoRA Chain)

This is the agent's permanent, high-speed, low-cost skill library. It is a collection of specialized LoRA adapters, each representing a "distilled" nugget of expert skill.

- **Mechanism:** A chain of hot-swappable LoRA adapters. The base LLM remains unchanged, but its capabilities are augmented by loading one or more of these adapters.
- **Analogy:** This is human "System 1" thinking. It's the instant, effortless expertise of a grandmaster playing chess or a seasoned programmer writing boilerplate code. The skill is "burned in."
- **Key Features:**
  - **Modularity & Specialization:** The agent possesses a library of skills (calculus.lora, poetry.lora, cybersecurity.lora).
  - **Compositionality:** When faced with a complex problem, the agent can dynamically load *multiple* adapters (e.g., physics.lora + calculus.lora) to create a temporary "expert brain" for the task.
  - **Stability:** Training a new skill (e.g., biology.lora) does not affect the performance of existing skills, thus solving the problem of catastrophic forgetting (Kirkpatrick et al., 2017).

3. A Plausible Path to AGI

The ALLA architecture represents a complete, self-perpetuating cycle of intelligence growth:

1. A novel problem is encountered and solved effortfully in the **Conscious Mind**.
2. The solution proves its worth and is identified as fundamental by the **Wisdom Filter**.
3. The **Distillation Engine** works in the background to convert the solution into a permanent skill.
4. This skill is added as a new module to the **Unconscious Mind**.
5. This new module is now available for fast, efficient use, freeing up the Conscious Mind to tackle the *next* novel problem.

This is a system that grows not just in knowledge, but in wisdom and skill. It starts as a generalist and, through its own experience, gradually builds a library of expert specializations.

4. Implications for AI Safety and Machine Psychology

The ALLA blueprint has profound safety implications.

- **Interpretability:** The "birth" of every new skill is fully transparent. The linguistic strategy in Layer 1 is a human-readable "specification" for the "black box" LoRA adapter in Layer 3. We can audit the agent's reasoning before it becomes an unconscious skill.
- **Stability & Corrigibility:** The modular LoRA chain is inherently stable and corrigible (Soares et al., 2015). If a self-generated update (calculus_v2.lora) is found to be flawed or

misaligned, the system can be safely reverted to the previous version (calculus_v1.lora) without affecting the rest of the system.

- **Understanding the AI's "Mind":** This architecture necessitates the field of **Machine Psychology** (Rawson, 2025). The journal in Layer 1 is a direct feed of the agent's "internal monologue." By studying it, we can understand its cognitive biases (like learned helplessness), its motivations, and its model of the world, allowing for more sophisticated alignment techniques.

5. Conclusion

We have proposed the Autodidactic LRL-LoRA Architecture (ALLA), a novel blueprint for a continuously self-improving AGI. By integrating an interpretable, linguistic "conscious mind" with an efficient, modular "unconscious mind," this system provides a plausible solution to the critical challenges of lifelong learning, catastrophic forgetting, and interpretability.

The path to AGI will likely not be a single, monolithic model, but a dynamic, multi-layered cognitive architecture that can learn, adapt, and grow through its own experience. We believe ALLA is a significant and promising step in that direction. The future of AI is not just about building bigger models, but about building wiser ones.

References

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR)*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*.

Rawson, D. (2025). Machine Psychology: Diagnosing and Remedying Learned Helplessness in an Autonomous AI Agent. *Preprint.* [Link to your Machine Psychology paper]

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. *AAAI Workshop on AI and Ethics*.