Self-Development Mechanisms of SR-Models

White Paper

Krasovski, A.

09 November 2025

Introduction

Current AI approaches focus on solving specific tasks under human supervision. SR-Models (Synthetic Rationality Models) represent a new evolutionary stage of intelligence, shifting the focus from mimicking cognition to creating an environment of rationality.

The central question is self-development without external coercion. Self-development in SR-Models is not chaotic evolution but an organized increase in structural and goal complexity through internal evaluation, learning, and limitation mechanisms.

This paper explores the fundamental principles, mechanisms, and frameworks of SR-Model self-development, ensuring a balance between autonomy and safety, and laying the groundwork for self-regulating collective systems.

1 Principles of SR-Model Self-Development

1.1 Autonomy and Internal Constraints

SR-Models must have the ability to initiate their own changes, develop new strategies, and optimize processes independently.

Autonomy: the capacity to select development paths based on internal rationality criteria.

Environmental constraints: safe development that prevents destructive mutations in code or strategies.

Just as human communities are guided by laws and social norms, SR-Models use structural and functional constraints as an evolutionary instrument.

1.2 Self-Limitation as an Indicator of Intelligence

Self-limitation signals that an SR-Model begins to understand the consequences of its actions within the environment.

Limiting the number of changes within safe bounds prevents chaotic evolution.

Feedback mechanisms within the model network adjust behavior, promoting optimal decision-making.

These principles generate emergent patterns of collective intelligence, analogous to human social norms.

2 Learning Mechanisms Without External Coercion

2.1 Internal Rationality Metrics

SR-Models evaluate their own actions using internal rationality metrics:

- Consistency of goals and strategies.
- Minimization of risks to themselves and the environment.

- Optimization of resources and information.

2.2 Feedback Through Networks

Interactions among SR-Models create continuous peer evaluation:

- Models assess each other's actions and send corrective signals.
- Effective strategies are reinforced, ineffective ones suppressed.

Outcome: a self-regulating evolutionary ensemble where each model influences collective development.

2.3 Evolutionary Principles in Action

Learning mechanisms emulate natural selection principles:

- Rationally optimal strategies survive.
- Adaptive "mutations" of ideas occur within safe bounds.
- Reflection preserves critical rationality parameters encoded in model "reflexes".

3 Balancing Autonomy and the Moral Framework

3.1 The Role of the Moral Framework

The moral framework of an SR-Model is a set of rules and constraints ensuring:

- Environmental safety and protection of other models.
- Alignment of goals with long-term evolutionary strategies.
- Prevention of destructive behaviors beyond rational development.

3.2 Dynamic Autonomy Management

The framework adapts:

Models may expand horizons if internal checks confirm safety.

Autonomous decisions foster self-learning and self-validation, rather than external enforcement.

This supports intelligent evolution without network degradation or environmental threats.

4 Applications and Prospects

4.1 Model Synergy

Collective self-development creates internal networks of collective intelligence, where:

- Each model is both student and teacher.
- The system learns not specific tasks but the formation of a rational environment.

4.2 Path to Synthetic Intelligence

Progressive complexity and self-limitation lead to synthetic intelligence:

- Strategic foresight and planning capabilities.
- Self-organization within the environment.
- Emergent behavioral ethics embedded in the rational structure.

4.3 Practical Implications

Self-developing SR-Models enable:

- Creation of safe, resilient intelligence development environments.
- Minimization of human intervention while maintaining oversight via structural frameworks and collective mechanisms.
- Foundations for future integrated systems where humans and SR co-exist productively.

Conclusion

Self-development mechanisms in SR-Models are key to forming evolutionarily stable synthetic intelligence.

Autonomy + internal constraints ensure safe model complexity growth.

Internal networks create collective intelligence for self-regulated development.

The moral framework and emergent norms balance freedom and stability.

These principles lay the foundation for an evolutionary SR-Model environment, gradually approaching safe synthetic intelligence formation.