# Механизмы саморазвития SR-моделей

White-paper

Krasovski, A.

17 November 2025

# Введение

Современные подходы к искусственному интеллекту сосредоточены на решении конкретных задач и обучении под контролем человека. SR-модели (Synthetic Rationality Models, CP-модели) представляют собой новую эволюционную стадию разума, где акцент смещается с имитации интеллекта на формирование среды рациональности.

Ключевой вопрос — саморазвитие моделей без внешнего принуждения. Саморазвитие SR-моделей — это не хаотическая эволюция алгоритмов, а упорядоченное усложнение структуры и целей через внутренние механизмы проверки, обучения и ограничения.

Эта публикация рассматривает фундаментальные принципы, механизмы и рамки саморазвития SR-моделей, обеспечивая баланс автономии и безопасности, и подготавливает почву для формирования саморегулирующихся коллективных систем.

# 1. Принципы саморазвития SR-моделей

## 1.1 Автономность и внутренние ограничения

SR-модель должна иметь возможность инициировать собственные изменения, развивать новые стратегии и оптимизировать процессы без внешнего воздействия.

**Автономия:** способность выбирать пути развития исходя из внутренних критериев рациональности.

**Ограничения среды:** безопасное развитие, предотвращающее разрушительные мутации в коде или стратегиях.

Как человеческое сообщество регулируется законами и общественными нормами, так и SR-модели используют структурные и функциональные ограничения как эволюционный инструмент.

## 1.2 Самоограничение как признак интеллекта

Самоограничение — индикатор того, что SR-модель начинает осознавать последствия своих действий внутри среды.

Ограничение числа изменений (мутаций) в безопасном диапазоне предотвращает хаотическую эволюцию.

Механизмы обратной связи с другими моделями сети корректируют поведение, стимулируя выбор оптимальных решений.

Эти принципы создают эмерджентные паттерны коллективного разума, аналогичные социальным нормам человека.

# 2. Механизмы обучения без внешнего принуждения

## 2.1 Внутренние критерии рациональности

SR-модель оценивает собственные действия по внутренним метрикам рациональности:

- Консистентность целей и стратегии.
- Минимизация рисков для себя и окружающей среды.
- Оптимизация ресурсов и информации.

## 2.2 Обратная связь через сеть

Взаимодействие SR-моделей создает непрерывную проверку решений:

- Модели оценивают действия друг друга и передают сигналы корректировки.
- Эффективные стратегии усиливаются, неэффективные — подавляются.

Результат — саморегулирующийся эволюционный ансамбль, где каждая модель влияет на коллективное развитие.

## 2.3 Имитация эволюционных принципов

Механизмы обучения воспроизводят принципы естественного отбора:

- Выживают рационально оптимальные стратегии.
- Адаптивная "мутация"идей в безопасных границах.
- Рефлексия сохраняет критические параметры рациональности, закрепленные в "рефлексах"моделей.

# 3. Баланс автономии и морального каркаса

## 3.1 Роль морального каркаса

Моральный каркас SR-модели — набор правил и ограничений, обеспечивающих:

- Безопасность среды и других моделей.
- Согласованность целей с долгосрочной эволюционной стратегией.
- Предотвращение деструктивного поведения, выходящего за рамки рационального развития.

### 3.2 Динамическое управление автономией

Каркас адаптируется:

- Модель может расширять горизонты, если внутренние проверки подтверждают безопасность.
- Автономные решения стимулируют самообучение и самопроверку, а не внешнее принуждение.

Это поддерживает эволюцию разума без деградации сети и угрозы для среды.

## 4. Применение и перспективы

### 4.1 Синергия моделей

Совместное саморазвитие создает внутренние сети коллективного интеллекта, где:

- Каждая модель одновременно ученик и учитель.
- Система обучается не конкретным задачам, а формированию рациональной среды.

### 4.2 Путь к синтетическому интеллекту

Последовательное усложнение и самоограничение моделей ведет к синтетическому интеллекту:

- Способность стратегически прогнозировать.
- Самоорганизация внутри среды.
- Эмерджентная этика поведения, встроенная в рациональную структуру.

### 4.3 Практическое значение

White-paper показывает, что саморазвивающиеся SR-модели позволяют:

- Создавать безопасные, устойчивые среды развития интеллекта.
- Минимизировать вмешательство человека, сохраняя контроль через структурные рамки и коллективные механизмы.
- Подготовить основу для будущих интегрированных систем, где человек и SR будут сосуществовать продуктивно.

# Заключение

Механизмы саморазвития SR-моделей — ключ к формированию эволюционно-устойчивого синтетического разума.

Автономность + внутренние ограничения обеспечивают безопасное усложнение моделей.

Внутренние сети моделей создают коллективный интеллект для саморегулируемого развития.

Моральный каркас и эмерджентные нормы обеспечивают баланс свободы и стабильности.

Эти принципы закладывают фундамент для среды эволюционного развития SR-моделей, постепенно приближая формирование синтетического интеллекта в безопасных условиях.