

УДК 159.9

ГРНТИ 15.81.31

ТЕОРЕТИКО-МЕТОДОЛОГИЧЕСКИЕ ПРЕДЕЛЫ РАЗВИТИЯ GenAI: ИНЖЕНЕРНО-ПСИХОЛОГИЧЕСКИЕ ОГРАНИЧЕНИЯ

© 2026 г. С.Ф. Сергеев*, Я.В. Курсевич**

** *Доктор психологических наук, профессор,
Санкт-Петербургский государственный университет, г. Санкт-Петербург, Россия
E-mail: s.f.sergeev@spbu.ru*

*** Бакалавр Санкт-Петербургского государственного университета,
г. Санкт-Петербург, Россия
E-mail: anakurs05@gmail.com*

Статья посвящена анализу инженерно-психологических ограничений развития генеративного искусственного интеллекта (GenAI). На основе междисциплинарного концептуально-аналитического обзора выделены четыре группы ограничений — ресурсно-вычислительные, эпистемические, институциональные и антропологические. Показано, что развитие GenAI ограничивается не только технологическим «потолком», но и совокупностью барьеров, среди которых ключевое место занимает эпистемическая неопределенность модели, разрушающая привычные для оператора паттерны доверия и контроля. Обосновывается положение о том, что дефицит высококачественных данных и модельный коллапс — это не универсальные пределы развития GenAI, а условные риски, зависящие от состава обучающих корпусов и процедур контроля качества. Основной теоретический вклад работы состоит в обосновании эпистемического разрыва между статистическим предсказанием и причинно-следственным пониманием как фундаментальной проблемы при совместном принятии решений в системе «человек — интеллектуальный агент». Предложена исследовательская гипотеза AIGE (Active-Inference Generative Engine), операционализирующая этот разрыв через механизмы активного вывода, интегрирующие генеративное моделирование, внутреннюю модель мира, оценку неопределенности и контролируемое взаимодействие со средой. AIGE рассматривается как архитектурная гипотеза, позволяющая снизить когнитивную нагрузку на пользователя за счет сигнализации о зонах незнания и активного запроса информации. Обсуждаются психологические последствия внедрения GenAI в профессиональную деятельность, включая риски деформации экспертных навыков и

смещения зон контроля, что требует пересмотра существующих эргономических стандартов проектирования симбиотических человеко-машинных интерфейсов. Ставится вопрос о новой регуляторной этике, ориентированной на управление когнитивными искажениями, возникающими при взаимодействии человека с генеративными моделями, что особенно актуально в областях с высокой ценой ошибки — медицине, авиации, управлении критической инфраструктурой.

Ключевые слова: GenAI, инженерная психология, доверие к ИИ, калибровка уверенности, коллапс модели, эпистемическая неопределенность, активный вывод, человеко-машинное взаимодействие, совместное принятие решений

ВВЕДЕНИЕ

Генеративный искусственный интеллект (GenAI) подвергает глубокой трансформации структуру деятельности человека-оператора, проектировщика и пользователя. Большие языковые модели активно интегрируются в системы управления, образования и средства научной и социальной коммуникации. В российском контексте проблема развития GenAI связана с реализацией Национальной стратегии развития искусственного интеллекта (Указ Президента Российской Федерации от 10.10.2019 № 490). Расширение функциональных возможностей моделей обостряет вопросы, относящиеся к инженерной психологии и эргономике — надежность в нештатных ситуациях, снижение когнитивной нагрузки, обеспечение доверия к GenAI, распределение функций в симбиотической паре между человеком и алгоритмом (Parasuraman, Sheridan, & Wickens, 2000; Russell, 2019).

С позиций инженерной психологии центральной проблемой становится не вычислительный или философский предел GenAI, а неопределенность поведения модели в условиях, не представленных в обучающей выборке. Критически важны стабильность ответа модели при повторных запросах, калибровка уверенности, своевременная сигнализация о зонах незнания и возможность интерактивного уточнения решения (Geng et al., 2024).

Для целей настоящего исследования необходимо различать несколько классов явлений, часто смешиваемых в дискуссиях. Это *ограничение* (снижение эффективности

модели в конкретной деятельности), *риск* (возможный негативный эффект), *предел* (граница, требующая смены архитектуры) и *философская граница* (положения о сознании или моральной агентности) (García-Valdecasas, 2026; Manheim, 2026).

Развитие GenAI ограничивается не только технологическим потолком, а также совокупностью гетерогенных барьеров, среди которых ключевое место занимает *эпистемическая неопределенность* модели, разрушающая доверие и механизмы контроля оператора. Проектирование человеко-машинного взаимодействия должно включать механизмы активного управления неопределенностью (Du Sordet, 2026; Mitrokhov, 2025).

Цель исследования — анализ ограничений развития GenAI, определение возможности их смягчения в рамках существующих архитектур и перехода к симбиотическим системам, в которых человек и модель образуют совместную когнитивную систему.

Задачи:

- провести эргономический анализ ограничений;
- классифицировать ограничения, влияющие на надежность и безопасность деятельности оператора;
- рассмотреть проблемы уверенности агентов симбиотической пары как источника операторских ошибок;
- оценить последствия использования синтетических данных в условиях симбиоза GenAI с человеком;
- рассмотреть сценарии развития GenAI с точки зрения перераспределения функций в системе «человек — интеллектуальный агент»;
- предложить концептуальную схему AIGE как архитектуру, снижающую когнитивную нагрузку на оператора за счет активного запроса информации и оценки собственной неопределенности.

Теоретический вклад исследования состоит в описании разрыва между статистическим предсказанием и причинно-следственным пониманием как фундаментальной проблемы совместного принятия решений. Для оператора критичен не факт ошибки модели, а ее непредсказуемость и отсутствие сигналов обратной связи. Предлагаемая схема AIGE позволяет решить эту проблему через механизмы активного вывода, интегрируя генеративное моделирование, внутреннюю модель мира и контролируемое взаимодействие со средой.

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Методологическое содержание проекта определяется с точки зрения инженерной психологии. Каждый тип ограничений развития GenAI рассматривается с позиций влияния на эффективность, надежность и безопасность деятельности оператора. Оцениваются ситуационная осведомленность, когнитивная нагрузка, доверие к системе и эргономические требования к интерфейсам (Endsley, 1995; Parasuraman et al., 2000).

Поиск литературы проводился в Google Scholar, Scopus, Web of Science, eLIBRARY и CyberLeninka по состоянию на август 2026 г. с использованием русскоязычных и англоязычных дескрипторов (generative AI, large language models, model collapse, semantic grounding, systematic generalization, active inference, AI responsibility, а также «генеративный искусственный интеллект», «модельный коллапс», «семантическое заземление», «систематическое обобщение»). В обзор включены рецензируемые статьи, материалы конференций, теоретические работы по философии ИИ и когнитивным наукам, публикации по архитектурам генеративных и агентных систем (Uddin et al., 2026; LeCun, 2022).

Источники группировались по четырем аналитическим уровням, включающим ресурсно-вычислительные ограничения, ограничения данных, эпистемические ограничения, институциональные и антропологические ограничения. При интерпретации различались эмпирические результаты, инженерные наблюдения, теоретические аргументы и философско-нормативные интерпретации.

ТАКСОНОМИЯ ОГРАНИЧЕНИЙ GenAI

В инженерно-психологической перспективе особое значение приобретают ограничения GenAI, влияющие на доверие, контроль и когнитивную нагрузку на оператора, проявляющиеся в нестабильности генераций в сходных ситуациях, неспособности распознавать зоны незнания.

Ресурсно-вычислительные ограничения. Современные модели требуют значительных ресурсов на этапах обучения, включающие стоимость электроэнергии, доступность ускорителей, экологические последствия, что с инженерно-психологической точки зрения и является источниками эргономических проблем. Латентность нарушает динамику операторского цикла, высокая стоимость ограничивает возможность уточняющих запросов, энергопотребление влияет на возможность размещения модели в мобильных комплексах. Вычислительный компромисс становится психофизиологической проблемой, влияющей на устойчивость оператора к стрессу.

Эти факторы не образуют непреодолимого предела. Они могут быть смягчены квантованием, дистилляцией, разреженными архитектурами, но при этом сохраняют характер *ресурсного ограничения* масштабирования (Villalobos et al., 2024).

Ограничения данных и обучения. Качество модели определяется не только объемом, но и структурой данных, их разнообразием, репрезентативностью, качеством разметки. Прогнозы об исчерпании высококачественных публичных текстовых данных для обучения моделей являются сценарными оценками и пока не достигнуты на практике (Villalobos et al., 2024).

Отдельный риск — обучение модели на синтетических данных. Неконтролируемое рекурсивное включение результатов предыдущих моделей может приводить к *модельному коллапсу* (Shumailov et al., 2024; Fu et al., 2026; Shabgahi et al., 2026). Эффект зависит от доли синтетического материала, качества фильтрации и сохранения доступа к первичным данным. Модельный коллапс следует рассматривать как *условный риск рекурсивного загрязнения данных*.

Эпистемические ограничения. Большие языковые модели обучаются на предсказании токенов, формируя способность выявлять статистические закономерности. Однако предсказательная эффективность не тождественна истинности и пониманию причин. Необходимо различать четыре уровня компетенций — синтаксическую компетентность, функциональную языковую компетентность, референциальную семантику и феноменальное понимание (Du Sordet, 2026; García-Valdecasas, 2026). Для инженерной психологии это различие имеет важное значение. Оператор ожидает устойчивого соответствия высказывания реальной рабочей ситуации. Когда модель ошибается в референции, возникает «иллюзии компетентности», снижающая критичность восприятия пользователя.

Модель может успешно воспроизводить знакомые шаблоны, и при этом демонстрировать непредсказуемые результаты при незначительном изменении формулировки задачи. Оператор не может выстроить устойчивую ментальную модель поведения системы, что многократно увеличивает его когнитивную нагрузку.

Устойчивая связь между языковыми выражениями, объектами и действиями ведет к «семантическому заземлению». Текстовая согласованность сама по себе не доказывает наличия такой связи. Отсутствие у модели сенсомоторного опыта означает, что она не разделяет с оператором общего знания о мире, что приводит к ошибкам коммуникации, особенно опасным в аварийных ситуациях (Manheim, 2026; Mitrokhov, 2025).

Заявленная моделью уверенность в результате не тождественна получению окончательной истины. Система должна уметь распознавать недостаточность данных, выявлять конфликтующие источники, отличать неизвестное от ложного. Неадекватная калибровка модели формирует у оператора ложное доверие или, напротив, недоверие к результатам (Geng et al., 2024).

ИНСТИТУЦИОНАЛЬНЫЕ И АНТРОПОЛОГИЧЕСКИЕ ОГРАНИЧЕНИЯ

Право, ответственность и психологические последствия. Архитектурная автономность модели GenAI не тождественна ее правосубъектности. Ключевые проблемы в области права связаны с авторством произведений, использованием защищенных материалов, ответственностью за ошибочные рекомендации. Особую сложность представляет «прокси-агентность». Если агентная система автономно совершает транзакции, возникает вопрос о распределении ответственности (Russell, 2019).

Вопросы права имеют прямую инженерно-психологическую проекцию. Если оператор не понимает, почему система приняла решение, он не может оценить его надежность — возникает феномен «диффузии ответственности». Инженерно-психологическим решением выступает требование *прослеживаемости когнитивного процесса* (traceability) — возможность ретроспективного анализа того, какие гипотезы и оценки привели к ответу.

Предвзятость и выравнивание. Модели могут воспроизводить социальные стереотипы. Процедуры выравнивания языковых моделей — Reinforcement Learning from Human Feedback (RLHF) снижают вредные ответы, но могут усиливать соответствие усредненным предпочтениям и снижать разнообразие. Процедуры выравнивания могут вступать в конфликт с задачей сохранения разнообразия допустимых ответов, что для инженерной психологии означает риск снижения гибкости системы в нестандартных ситуациях (Russell, 2019; Manheim, 2026).

Агентность и антропологическая граница. Следует различать: (1) цель, заданную пользователем; (2) инструментальную оптимизацию; (3) автономное планирование; (4) внутреннее целеполагание; (5) моральную агентность; (6) самосознание. Современные GenAI-системы демонстрируют первые три свойства. Наличие последних трех не установлено (García-Valdecasas, 2026; Mitrokhov, 2025). Обобщенная систематизация выделенных ограничений представлена в табл. 1.

Таблица 1

Матрица ограничений развития GenAI

Уровень анализа	Основное ограничение	Эмпирический статус	Возможные меры смягчения	Сохраняющаяся проблема
Ресурсный	Стоимость вычислений и инфраструктуры	Хорошо подтверждено инженерной практикой	Квантование, дистилляция, оптимизация инференса, спецоборудование	Компромисс между качеством, скоростью и стоимостью
Данные и обучение	Дефицит качественных человеческих данных	Сценарная гипотеза, зависящая от определения качества	Новые корпуса, фильтрация, доступ к первичным данным	Неопределенность критериев качества и темпов прироста данных
Данные и обучение	Модельный коллапс	Подтвержден в условиях рекурсивного обучения	Контроль состава данных, верификация, смешивание с первичными данными	Деградация редких признаков при рекурсивном обучении
Эпистемический	Нестабильное систематическое обобщение	Подтверждается отдельными тестами	Верификаторы, нейросимволические модули, специальные обучающие задачи	Перенос причинной структуры между доменами
Эпистемический	Ограниченное семантическое заземление	Теоретически и эмпирически дискуссионно	Мультимодальность, взаимодействие со средой, world models	Неясность критериев референциального понимания
Эпистемический	Недостаточная калибровка уверенности	Подтверждается исследованиями надежности	Калибровка, ансамбли, внешняя проверка, abstention	Надежное распознавание неизвестного
Институциональный	Распределение ответственности	Нормативная и правовая проблема	Аудит, журналирование, human oversight	Ответственность при автономных действиях (прокси-агентность)
Антропологический	Субъектность, сознание, моральная агентность	Философски спорно	Уточнение критериев и поведенческих тестов	Отсутствие общепринятого диагностического критерия

СЦЕНАРИИ ДАЛЬНЕЙШЕГО РАЗВИТИЯ GenAI

Инкрементальный сценарий предполагает оптимизацию архитектур, улучшение данных, посттренировочное выравнивание, RAG (Retrieval Augmented Generation),

внешние инструменты и верификацию. RAG расширяет доступ к внешней памяти, но не устраняет проблему качества источников и проверки вывода (Lewis et al., 2020).

Нейросимволический сценарий объединяет статистическую обработку с формальными правилами, базами знаний и верификаторами. Потенциальное преимущество — повышение проверяемости, однако сохраняются трудности интеграции вероятностных и символических компонентов (LeCun, 2022).

Агентно-интерактивный сценарий предполагает, что система получает наблюдения, обновляет модель мира, генерирует гипотезы, выбирает действия и получает обратную связь. Это уменьшает зависимость от статичных корпусов, но создает риски безопасности, контроля и стратегического поведения (Russell, 2019; LeCun, 2022; Mitrokhov, 2025).

Критерии инженерно-психологической оценки сценариев: прозрачность поведения, метакогнитивная сигнализация, интерактивная управляемость, обучаемость оператора, эмоциональная приемлемость, устойчивость к ошибкам оператора. Сравнительный анализ сценариев развития представлен в табл. 2.

Таблица 2

Сравнение сценариев дальнейшего развития GenAI

Сценарий	Основные механизмы	Потенциально решаемые задачи	Сохраняющиеся ограничения	Критерии оценки
Инкрементальный	Масштабирование, оптимизация, RAG, выравнивание	Повышение качества, снижение стоимости, предметная адаптация	Зависимость от данных, нестабильность обобщения, отсутствие гарантии истинности	Точность, стоимость, устойчивость, калибровка
Нейро-символический	Правила, базы знаний, программы, верификаторы	Формализация и проверка рассуждений	Сложность интеграции и формализации	Корректность, воспроизводимость, систематическое обобщение
Агентно-интерактивный	Взаимодействие со средой, world model, память, планирование	Активное получение информации, адаптация, многошаговые действия	Риск автономности, ошибочная модель мира, контроль целей	Безопасность, причинное прогнозирование, калибровка, эффективность активного обучения

Ни один сценарий не обеспечивает автоматического устранения всех ограничений. Масштабирование не гарантирует понимания, символическая интеграция не устраняет неполноту знаний, а интерактивное обучение может повысить автономность системы и одновременно усилить риски контроля.

AIGE КАК ИССЛЕДОВАТЕЛЬСКАЯ ГИПОТЕЗА

Назначение схемы

AIGE (*Active-Inference Generative Engine*) предлагается авторами настоящей статьи как концептуальная исследовательская программа, а не как готовая технологическая система. Ее задача — описать архитектуру, в которой генерация ответа связана с наблюдением, моделированием, действием, обратной связью и оценкой неопределенности.

В отличие от пассивной генеративной системы, AIGE должна функционировать в замкнутом контуре:

наблюдение → моделирование → гипотезы → действие → обратная связь → обновление модели.

Предполагаемые компоненты — генеративная модель, внутренняя модель мира, память, сенсорные и инструментальные каналы, модуль оценки неопределенности, модуль планирования, верификаторы, механизм обновления модели, контур безопасности и контроля.

Формальная схема

В отличие от простой максимизации информационной ценности, предлагаемая схема AIGE базируется на принципе минимизации ожидаемой свободной энергии (Expected Free Energy, EFE), что соответствует современной теории активного вывода (Friston, 2010; Friston et al., 2017; Mancinelli, 2025).

Цикл AIGE можно представить как последовательность состояний:

$$o_t \rightarrow s_t \rightarrow H_t \rightarrow U_t \rightarrow a_t \rightarrow o_{t+1}$$

где: O_t — наблюдение среды; S_t — рабочее представление текущего состояния; H_t — множество альтернативных гипотез; U_t — оценка неопределенности; a_t — выбранное действие; O_{t+1} — новое наблюдение после действия.

Выбор действия в AIGE осуществляется на основе минимизации ожидаемой свободной энергии:

$$G(a) = -E_{\{o|a\}} [D_{KL}(Q(s|o) || Q(s))] + E_{\{o|a\}} [-\log P(o|C)]$$

где:

- $G(a)$ — ожидаемая свободная энергия для действия a ;
- $E_{o|a}$ — математическое ожидание по возможным наблюдениям после действия;
- $D_{\{KL\}}(Q(s|o) || Q(s))$ — дивергенция Кульбака-Лейблера между апостериорным и априорным распределениями скрытых состояний (эвристическая ценность, уменьшающая неопределенность);
- $P(o|C)$ — вероятность наблюдения при условии предпочтительной цели C (прагматическая ценность);
- Q — вариационное приближение к истинному распределению.

Здесь C обозначает предпочтительное целевое состояние, которое может быть задано пользователем или выведено из контекста задачи.

Оптимальное действие выбирается как:

$$a^* = \arg \min_{a \in A} G(a)$$

Эта запись означает, что агент выбирает действие, которое минимизирует сумму ожидаемой неопределенности и расхождения с целевым состоянием.

Этапы функционирования: наблюдение, формирование представления, генерация альтернативных гипотез, оценка неопределенности, выбор информационного действия (утрачивающий вопрос, обращение к базе, проверка, симуляция), контролируемое взаимодействие, обновление модели мира, формирование ответа с указанием источников и предупреждений о рисках. Сопоставление AIGE с базовыми архитектурами GenAI приведено в табл. 3.

Таблица 3

Сравнение AIGE с базовыми генеративными системами

Компонент	Обычная LLM	LLM с RAG	AIGE
Внешние источники	Ограниченно	Да	Да
Модель мира	Неявная или ограниченная	В основном внешняя память	Функционально предполагается
Альтернативные гипотезы	Не гарантированы	Не гарантированы	Центральный компонент
Оценка неопределенности	Нестабильная	Частично внешняя	Обязательный механизм
Информационные действия	Обычно отсутствуют	Поиск	Выбор по минимизации свободной энергии
Обновление представлений	Ограниченное	Обновление базы	Обновление модели мира на основе EFE
Метакогнитивная сигнализация	Отсутствует	Отсутствует	Явный механизм
Контроль действий	Зависит от реализации	Обычно ограничен	Принципиально необходим
Эмпирический статус	Реализованные системы	Реализованные системы	Исследовательская гипотеза

КРИТЕРИИ ЭМПИРИЧЕСКОЙ ПРОВЕРКИ

Для проверки AIGE необходимы эксперименты, сопоставляющие ее с LLM, RAG-системами и агентными архитектурами. Основные критерии — калибровка уверенности, способность отличать неизвестное от ложного, устойчивость к распределительному сдвигу, качество причинного прогнозирования, эффективность

активного получения информации, систематическое обобщение, безопасность действий, прозрачность обновления модели мира, стоимость и задержка. AIGE будет иметь теоретическую ценность только при измеримом улучшении относительно базовых архитектур. Наличие модели мира, памяти или агента не является доказательством понимания GenAI (Du Sordet, 2026; García-Valdecasas, 2026; Manheim, 2026).

ОБСУЖДЕНИЕ

Предложенная таксономия позволяет различить ограничения, имеющие разный эпистемический и практический статус. *Ресурсные ограничения* поддаются оптимизации, но сохраняют характер компромисса. *Ограничения данных* зависят от состава корпусов и юридического режима. *Модельный коллапс* — риск при рекурсивном обучении, а не универсальный закон (Shumailov et al., 2024; Fu et al., 2026; Shabgahi et al., 2026).

Эпистемические ограничения требуют осторожности. Отсутствие сознания не препятствует практической полезности системы. В инженерном контексте следует оценивать надежность, устойчивость, калибровку и безопасность. Здесь возникает важный контраргумент функционалистского характера: если система стабильно использует язык, планирует, корректирует действия, получает обратную связь и достигает целей, то для ряда практических задач функционального понимания достаточно. Данная позиция не опровергается предлагаемой моделью. Статья лишь утверждает, что функциональная результативность, референциальное понимание и феноменальный опыт нельзя автоматически считать тождественными (García-Valdecasas, 2026; Du Sordet, 2026).

С инженерно-психологической точки зрения функциональная результативность GenAI оценивается через вклад в ситуационную осведомленность оператора (situation awareness, SA), включающую восприятие, понимание и проекцию (Endsley, 1995). AIGE с альтернативными гипотезами и оценкой неопределенности потенциально способна поддерживать все три уровня.

Агентно-интерактивный сценарий расширяет возможности модели, но увеличение автономности не означает повышения надежности. Система может действовать на основе ошибочной модели мира, выбирать рискованные информационные стратегии или усиливать ошибки через неконтролируемое обновление памяти (Russell, 2019; Mitrokhov, 2025).

Реализация AIGE сопряжена со значительными вычислительными затратами, связанными с многократным моделированием альтернатив и оценкой свободной энергии. Кроме того, построение внутренней модели мира, остается открытой исследовательской задачей (LeCun, 2022; Mancinelli, 2025).

ЗАКЛЮЧЕНИЕ

Проведенный теоретико-методологический анализ развития GenAI в инженерно-психологическом ракурсе позволяет сделать следующие выводы.

Развитие GenAI ограничивается не единым технологическим пределом, а совокупностью разнотипных барьеров. В инженерно-психологической перспективе эти барьеры следует разделять на ресурсно-вычислительные (транслирующиеся в проблемы латентности, стоимости и эргономической доступности), ограничения данных и обучения (проявляющиеся в снижении надежности при распределительных сдвигах), эпистемические (связанные с непредсказуемостью поведения и отсутствием референциальной надежности), а также институциональные и антропологические (определяющие правовые рамки и психологическую приемлемость автономных систем) (Parasuraman et al., 2000; Russell, 2019; Villalobos et al., 2024).

Ресурсные ограничения могут смягчаться инженерными методами, но сохраняют характер компромисса между масштабом, качеством и скоростью, что непосредственно влияет на психофизиологическое состояние оператора в условиях дефицита времени. Модельный коллапс является условным риском рекурсивного обучения на синтетических данных, однако для инженерной психологии важно, что снижение разнообразия ответов

модели может вести к сужению репертуара альтернатив, доступных оператору для принятия решения (Shumailov et al., 2024; Fu et al., 2026).

Масштабирование моделей может повышать функциональную языковую результативность, но не снимает проблему *эпистемической неопределенности*, которая является центральным вызовом для доверия и контроля со стороны оператора. В этом суть *эпистемического разрыва*, обоснованного в данной работе (Du Sordet, 2026; Mannheim, 2026). Эргономическая эффективность нейросимволических подходов зависит от прозрачности для оператора процесса рассуждения (LeCun, 2022).

Симбиотические системы потенциально позволяют связать генерацию с наблюдением, действием, обратной связью и обновлением модели мира, что открывает возможность для создания активно *взаимодействующих систем*. Такой режим калиброванной обратной связи является ключевым требованием к надежному человеко-машинному интерфейсу (Endsley, 1995; Parasuraman et al., 2000).

AIGE предложена как концептуальная исследовательская гипотеза, объединяющая генеративную модель, внутреннюю модель мира, альтернативные гипотезы, оценку неопределенности и активное получение информации на основе минимизации свободной энергии (Friston, 2010; Friston et al., 2017; Mancinelli, 2025). Ее потенциальное инженерно-психологическое преимущество состоит в том, что она может обеспечить:

- *явное представление альтернатив* — снижающее эффект «подтверждающего смещения» у оператора;
- *калиброванную оценку уверенности* — позволяющую оператору адекватно распределять внимание и контроль (Geng et al., 2024);
- *активный запрос информации* — снижающий когнитивную нагрузку по поиску и верификации данных;
- *прослеживаемость решений* — поддерживающую возможность аудита и обучения оператора.

Научная состоятельность этой схемы может быть установлена только посредством формализации архитектуры, разработки экспериментальных протоколов и сравнительной оценки с базовыми генеративными и агентными системами. В качестве дорожной карты дальнейших эмпирических исследований выделены три приоритетных направления: (1) сравнительная калибровка уверенности в задачах с неполной информацией; (2) тесты на систематическое обобщение при вариациях формулировок; (3) оценка эффективности активного получения информации по сравнению с пассивным RAG-поиском (Lewis et al., 2020).

Итоговый тезис статьи заключается в следующем:

теоретико-методологические пределы развития современных генеративных моделей являются конститутивными для парадигмы, основанной на масштабировании авторегрессионного предсказания и статичных корпусов. Их эргономическое смягчение требует перехода к интерактивным архитектурам с явной оценкой неопределенности, активным получением информации и калиброванной метакогнитивной сигнализацией — к системам, которые сотрудничают с оператором в совместном познании и принятии решений.

ЛИТЕРАТУРА

- Указ Президента Российской Федерации «О развитии искусственного интеллекта в Российской Федерации» от 10.10.2019 № 490 // Собрание законодательства Российской Федерации. 2019. № 41. Ст. 5700.
- Du Sordet E. When Language Hides Causes: On Causal Representation Problems in LLMs // *Philosophy of AI*. 2026. Forthcoming. URL: <https://philarchive.org/rec/DUSWLH> (дата обращения: 12.08.2026).
- Endsley M.R. Toward a Theory of Situation Awareness in Dynamic Systems // *Human Factors*. 1995. V. 37. № 1. P. 32-64. DOI:10.1518/001872095779049543
- Friston K. The Free-Energy Principle: A Unified Brain Theory? // *Nature Reviews Neuroscience*. 2010. V. 11. № 2. P. 127-138. DOI:10.1038/nrn2787
- Friston K.J., FitzGerald T., Rigoli F., Schwartenbeck P., Pezzulo G. Active Inference: A Process Theory // *Neural Computation*. 2017. V. 29. № 1. P. 1-49. DOI:10.1162/NECO_a_00912

- Fu S., Wang Y., Chen Y., Shen L., Tao D. Self-Verification Provably Prevents Model Collapse in Recursive Synthetic Training // *Advances in Neural Information Processing Systems*. 2026. V. 38. P. 36101-36154. DOI:10.52202/085713-1213
- García-Valdecasas M. Are Large Language Models Intentional? The Limits of Referential Grounding // *Philosophy & Technology*. 2026. V. 39. № 2. Article 62. DOI:10.1007/s13347-026-01079-4
- Geng J., Cai F., Wang Y., Koepl H., Nakov P., Gurevych I. A Survey of Confidence Estimation and Calibration in Large Language Models // *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico, 2024. P. 6577–6595. DOI: 10.18653/v1/2024.naacl-long.366.
- LeCun Y. A Path Towards Autonomous Machine Intelligence // *Open Review*. 2022. URL: <https://openreview.net/forum?id=BZ5a1r-kVsf> (дата обращения: 12.08.2026).
- Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-T., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // *Advances in Neural Information Processing Systems*. 2020. V. 33. P. 9459-9474. DOI:10.48550/arXiv.2005.11401
- Mancinelli F.M. *Multi-Agent Decision Making via Active Inference*: Master's Thesis. Politecnico di Milano, 2025. URL: https://www.politesi.polimi.it/bitstream/10589/240545/2/2025_07_Mancinelli.pdf (дата обращения: 12.08.2026).
- Manheim D. Language Models' Hall of Mirrors Problem: Why AI Alignment Requires Peircean Semiosis // *Philosophy & Technology*. 2026. V. 39. № 1. DOI:10.1007/s13347-025-00975-5
- Mitrokhov K. Between World Models and Model Worlds: On Generality, Agency, and Worlding in Machine Learning // *AI & Society*. 2025. V. 40. № 6. P. 5087-5099. DOI:10.1007/s00146-024-02086-9
- Parasuraman R., Sheridan T.B., Wickens C.D. A Model for Types and Levels of Human Interaction with Automation // *IEEE Transactions on Systems, Man, and Cybernetics. Part A*. 2000. V. 30. № 3. P. 286-297. DOI:10.1109/3468.844354
- Russell S. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking, 2019. 352 p.
- Shabgahi S.Z., Aghazadeh P., Mirhoseini A., Koushanfar F. ForTIFAI: Fending Off Recursive Training Induced Failure for AI Model Collapse // *npj Artificial Intelligence*. 2026. DOI:10.1038/s44387-026-00127-w

- Shumailov I., Shumaylov Z., Zhao Y., Papernot N., Anderson R., Gal Y. AI Models Collapse When Trained on Recursively Generated Data // *Nature*. 2024. V. 631. P. 755-759. DOI:10.1038/s41586-024-07566-y
- Uddin M., Arfeen S.U., Alanazi F., Hussain S., Mazhar T., Rahman M.A. A Critical Analysis of Generative AI: Challenges, Opportunities, and Future Research Directions // *Archives of Computational Methods in Engineering*. 2026. V. 33. № 2. P. 1763-1793. DOI:10.1007/s11831-025-10355-z
- Villalobos P., Ho A., Sevilla J., Besiroglu T., Heim L., Hobbhahn M. Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data // *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. Vienna, Austria, 2024. Vol. 235 (Proceedings of Machine Learning Research). P. 49523–49544.

Статья поступила в редакцию: 16.08.2026. Статья опубликована: 30.09.2026

THEORETICAL AND METHODOLOGICAL LIMITS OF GenAI DEVELOPMENT: ENGINEERING PSYCHOLOGICAL CONSTRAINTS

© 2026 S.F. Sergeev*, Ya.V. Kursevich**

** Doctor of Psychological Sciences, Professor,
Saint Petersburg State University, Saint Petersburg, Russia
E-mail: s.f.sergeev@spbu.ru*

*** Bachelor of Saint Petersburg State University;
Saint Petersburg, Russia
E-mail: anakurs05@gmail.com*

The article analyzes engineering-psychological constraints on the development of generative artificial intelligence (GenAI). Based on an interdisciplinary conceptual-analytical review, four groups of constraints are identified: resource-computational, epistemic, institutional, and anthropological. It is shown that GenAI development is limited not only by a technological “ceiling” but also by a set of barriers, among which epistemic uncertainty of the model occupies

a key place, undermining the operator's usual patterns of trust and control. The position is substantiated that deficits in high-quality data and model collapse are not universal limits to GenAI development, but conditional risks dependent on the composition of training corpora and quality-control procedures. The main theoretical contribution of the work consists in substantiating the epistemic gap between statistical prediction and causal understanding as a fundamental problem in joint decision-making within the “human—intelligent agent” system. A research hypothesis, AIGE (Active-Inference Generative Engine), is proposed, operationalizing this gap through active inference mechanisms that integrate generative modeling, an internal world model, uncertainty estimation, and controlled interaction with the environment. AIGE is considered as an architectural hypothesis that can reduce the user's cognitive load by signaling zones of ignorance and actively requesting information. The psychological consequences of introducing GenAI into professional activity are discussed, including risks of deformation of expert skills and shifts in control zones, which necessitate a revision of existing ergonomic standards for designing symbiotic human—machine interfaces. The question is raised of a new regulatory ethics oriented toward managing cognitive biases that arise during human interaction with generative models, which is especially relevant in high-stakes domains such as medicine, aviation, and critical infrastructure management.

Keywords: GenAI, engineering psychology, trust in automation, confidence calibration, model collapse, epistemic uncertainty, semantic grounding, active inference, human-machine interaction, joint decision-making

REFERENCES

- President of the Russian Federation. (2019). *Ukaz Prezidenta Rossijskoj Federatsii ot 10 oktyabrya 2019 g. № 490 «O razvitii iskusstvennogo intellekta v Rossijskoj Federatsii»* [Decree of the President of the Russian Federation dated October 10, 2019, No. 490 “On the Development of Artificial Intelligence in the Russian Federation”]. *Sobranie zakonodatel'stva Rossijskoj Federatsii*, (41), st. 5700.
- Du Sordet, E. (2026). When language hides causes: On causal representation problems in LLMs. *Philosophy of AI*. <https://philarchive.org/rec/DUSWLH>
- Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32–64. <https://doi.org/10.1518/001872095779049543>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, 29(1), 1–49. https://doi.org/10.1162/NECO_a_00912
- Fu, S., Wang, Y., Chen, Y., Shen, L., & Tao, D. (2026). Self-verification provably prevents model collapse in recursive synthetic training. In *Advances in Neural*

Information Processing Systems (Vol. 38, pp. 36101–36154).
<https://doi.org/10.52202/085713-1213>

García-Valdecasas, M. (2026). Are large language models intentional? The limits of referential grounding. *Philosophy & Technology*, 39(2), Article 62.
<https://doi.org/10.1007/s13347-026-01079-4>

Geng, J., Cai, F., Wang, Y., Koepl, H., Nakov, P., & Gurevych, I. (2024). A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 6577–6595). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.366>

LeCun, Y. (2022). *A path towards autonomous machine intelligence* [Preprint]. OpenReview. <https://openreview.net/forum?id=BZ5a1r-kVsf>

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. <https://doi.org/10.48550/arXiv.2005.11401>

Mancinelli, F.M. (2025). *Multi-agent decision making via active inference* [Master's thesis, Politecnico di Milano]. PoliTeSI. https://www.politesi.polimi.it/bitstream/10589/240545/2/2025_07_Mancinelli.pdf

Manheim, D. (2026). Language models' hall of mirrors problem: Why AI alignment requires Peircean semiosis. *Philosophy & Technology*, 39(1).
<https://doi.org/10.1007/s13347-025-00975-5>

Mitrokhov, K. (2025). Between world models and model worlds: On generality, agency, and worlding in machine learning. *AI & Society*, 40(6), 5087–5099.
<https://doi.org/10.1007/s00146-024-02086-9>

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics: Part A*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Shabgahi, S.Z., Aghazadeh, P., Mirhoseini, A., & Koushanfar, F. (2026). ForTIFAI: Fending off recursive training induced failure for AI model collapse. *npj Artificial Intelligence*. <https://doi.org/10.1038/s44387-026-00127-w>

- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- Uddin, M., Arfeen, S.U., Alanazi, F., Hussain, S., Mazhar, T., & Rahman, M.A. (2026). A critical analysis of generative AI: Challenges, opportunities, and future research directions. *Archives of Computational Methods in Engineering*, 33(2), 1763–1793. <https://doi.org/10.1007/s11831-025-10355-z>
- Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. *Proceedings of Machine Learning Research*, 235, 49523–49544.

The article was received: 16.08.2026. Published online: 30.09.2026

Библиографическая ссылка на статью:

Сергеев С.Ф., Курсевич Я.В. Теоретико-методологические пределы развития GenAI: инженерно-психологические ограничения // Институт психологии Российской академии наук. Организационная психология и психология труда. 2026. Т. 11. № 3. С. 36–56. DOI: 10.38098/ipran.opwp_2026_40_3_002

Sergeev, S.F., Kursevich, Ya.V. (2026). Teoretiko-metodologicheskie predely razvitiya GenAI: inzhenerno-psihologicheskie ogranicheniya [Theoretical and methodological limits of GenAI development: engineering psychological constraints]. Institut Psikologii Rossiyskoy Akademii Nauk. Organizatsionnaya Psikhologiya i Psikhologiya Truda [Institute of Psychology of the Russian Academy of Sciences. Organizational Psychology and Psychology of Labor]. 11(3). 36–56. DOI: 10.38098/ipran.opwp_2026_40_3_002

Адрес ссылки:

<http://work-org-psychology.ru/engine/documents/document1259.pdf>