Доверие к ИИ-агентам: Российский бизнес оценивает ИИ

Технологии Доверия

Исследование ТеДо «Российский бизнес оценивает ИИ», декабрь 2025

Цель исследования:

Оценка уровня доверия и готовности бизнеса к взаимодействию с ИИ-агентами

Формирование **подхода** для российских компаний, позволяющего **безопасно и эффективно интегрировать ИИ-технологии** в свои процессы.

Участники опроса:
1000 + 1 руководителей компаний и ИТ-директоров*, включая 270 представителей среднего и крупного бизнеса

Участники глубинных интервью:
Крупные игроки финансового рынка

Задачи:

1. Выявить текущий уровень ИИ в основных видах деятельности и внутренних процессах
2. Оценить уровень доверия при передаче процесса принятия решений на сторону ИИ
3. Оценить готовность бизнеса к взаимодействию с ИИ-агентами клиентов, ИИ-агентов партеров и ИИ-сотрудниками
4. Проверить гипотезу, что доверие к ИИ-агентам является значимым барьером, замедляющим внедрение ИИ в бизнесе

Исследование по Оценке уровня доверия и готовности бизнеса к взаимодействию с ИИ-агентами проводилось совместно с аналитическим центром НАФИ

Выступление представляет результаты исследования

*приоритетные отрасли компаний: банковская, информационные технологии, металлургическая и горнодобывающая, нефтегазовая отрасль, розничная торговля, транспорт и логистика, электроэнергетика и ЖКХ, агропромышленный комплекс

Технологии Доверия

ИИ-агенты: дань моде или новый стандарт рынка?

ИИ в научной среде определяется через **рациональных агентов на протяжении последних 70 лет.** Повышенное внимание к теме ИИ-агентов в 2025 году в российской и международной бизнес-среде связано **с ростом автономности решений** на базе ИИ и с переходом к платформенному подходу.

Волны развития ИИ

Участник глубинного интервью

Революция, которая сейчас происходит в ИИ-агентах, в **большей степени связана с изменениями в архитектуре и сервисах**, чем со значимым прорывом в агентности.

Чтобы стать настоящим ИИ-агентом, система должна **уметь самостоятельно обучаться, взаимодействовать со средой, делать выводы и находить свои ошибки** в реальном времени.

То есть ключевой критерий для ИИ-агента – это **автономность**. На данный момент он не достигнут в полной мере.

Источники: Кай-фу Ли Сверхдержавы искусственного интеллекта; S. Russell, P. Norving Artificial Intelligence, a Morden Approach; Глубинные интервью ТеДо с представителями технологичных Банков и компаний, проведенные в октябре-декабре 2025

Технологии Доверия

Эволюционный подход к определению ИИ-агентов

ИИ-агент – это программа или система, способная самостоятельно выполнять задачи, принимать решения и взаимодействовать с окружающей средой на основе искусственного интеллекта.

Эволюция агентных систем

Источник: Сбер, Разработка и применение мультиагентных систем в корпоративной среде

Отсутствие надежного планирования у ИИ-агентов подтверждается результатами бенчмарков (PlanBench): при решении задач **сложного логического планирования** качество результата LLM **снижается на 50%-95%.**

На **горизонте 2-3 лет** ожидается реализация ключевых задач для повышения автономности работы ИИ-агента:

- Внедрение детальных подходов в части обучения с подкреплением
- Создание фундаментальных агентов
- Разработка мультиагентных архитектур

Источники: Stanford University Human-Centered Artificial Intelligence (HAI), The 2025 AI Index Report
Subbarao Kambhampati, PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change, *Альянс в сфере ИИ, Горизонты искусственного интеллекта: Какими будут технологии ИИ через 10 лет*

Технологии Доверия

ИИ-агенты: проблемы и риски

Сдерживающим фактором массового внедрения современных ИИ-агентов является **уровень безопасности и доверия**, которые объясняются следующими вызовами:

Вызовы

- **Отсутствие строгой формальной теории** (невозможность гарантировать безопасность без формализации устойчивости)
- **Выравнивание целей** (предотвращение генерации вредоносных или неэтичных результатов)
- **Объяснимость и прозрачность** (непонимание внутренней логики работы «черного ящика»)
- **Безопасная разработка и эксплуатация (MLSecOps)** (уязвимости в коде, данных и модели; новые угрозы для агентов)
- **Противодействие злоупотреблениям** (использование ИИ для взлома, дипфейков, дезинформации)

Источник: AI Alliance Network, Горизонты искусственного интеллекта: Какими будут технологии ИИ через 10 лет

Ключевые исследовательские задачи

- **Разработка методов снижения рисков** от вредоносных данных и исследование фундаментальных математических границ **для гарантий безопасности.**
- Создание методов постфактум-объяснения сложных моделей, а также разработка **требований и протоколов тестирования систем на объяснимость.**
- Создание **инфраструктуры безопасной разработки**
- Разработка методов **обнаружения и защиты от дипфейков**, включая технологии цифровых водяных знаков.

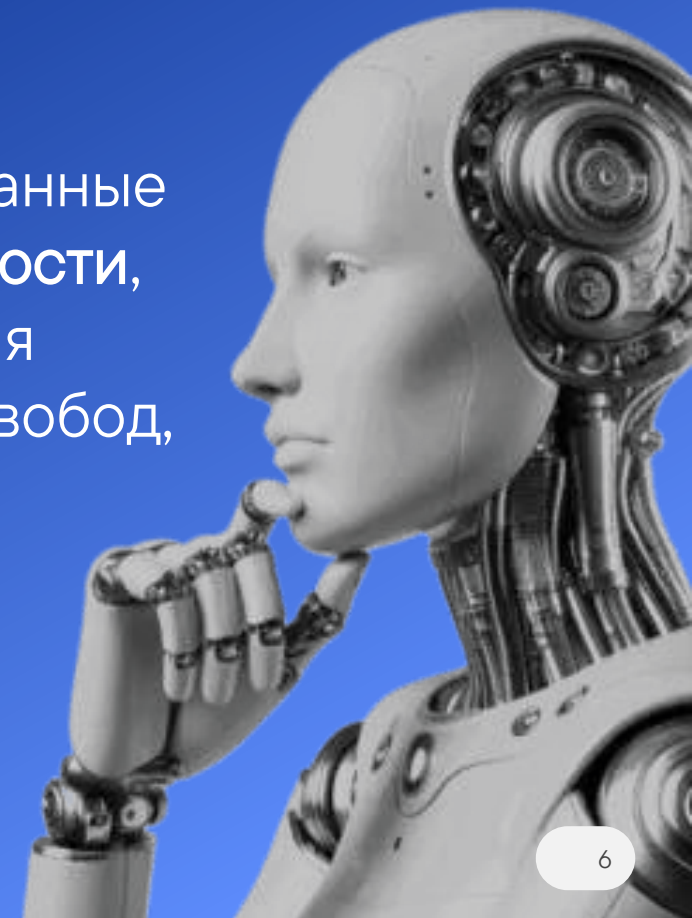
Потенциальный эффект

- Переход от эмпирической безопасности к верифицируемой, создавая агентов с **доказуемыми границами безопасного поведения**
- Автономные агенты будут действовать в рамках заданных этических и социальных норм
- Визуализация принятия агентом решения
- **Защита от отравления наборов данных** и атак на модели, инструменты поиска уязвимостей в сторонних библиотека
- Идентификация контента, созданного или модифицированного агентами, и **защита каналов коммуникации агентов** от подделки информации

Что значит доверять ИИ-агенту?

Доверенные технологии искусственного интеллекта –

технологии, отвечающие **стандартам безопасности**, разработанные с учетом **принципов объективности, недискриминации, этичности**, исключающие при их использовании возможность причинения вреда человеку и нарушения его основополагающих прав и свобод, нанесения **ущерба интересам общества и государства**.



Источник: Национальная стратегия развития искусственного интеллекта на период до 2030 года

Результаты исследования «Российский бизнес оценивает ИИ»

Технологии Доверия

Применение ИИ в российском бизнесе носит фрагментарный характер и сосредоточено в области клиентского сервиса и HR. Причина – отсутствие инфраструктуры

Технологии Доверия

1

Большинство кейсов применения генеративного ИИ на российском рынке связано с внедрением чат-ботов и голосовых помощников. Для прочих задач ИИ используется как co-pilot для **поиска информации** и **генерации текста**.

Фрагментарный характер применения связан с **отсутствием ИТ-инфраструктуры** и **системных платформенных решений**, необходимых для масштабного внедрения ИИ агентов повсеместно в широком круге процессов в российском бизнесе.

Участник глубинного интервью

> ИИ-агенты хороши в тех задачах, где мы можем принять ошибки, которые они совершают. К таким задачам в первую очередь относятся поиск и анализ информации. **ИИ-агенты подходят для сотрудника бэк-офиса как co-pilot.**

Участник глубинного интервью

> На практике бывает сложно провести грань между инвестициями в ИИ и инвестициями в инфраструктуру, которая позволяет разрабатывать ИИ. **Инвестиции в определенные области ИТ** будут оказывать **значимый эффект на внедрение ИИ уже в 2026 году.**

1.1. В каких сферах деятельности компании используют ИИ?*, % от всех опрошенных

1.2. Для каких именно задач используют ИИ?*, в % от всех опрошенных

**Многовариантный вопрос, сумма ответов может быть не равна 100%.*

Бизнес не готов доверять ИИ-агентам выполнение задач без постоянного контроля

2 Несмотря на декларируемое доверие к системам ИИ и к ИИ-агентам, бизнес склонен **оставлять за собой контроль.**

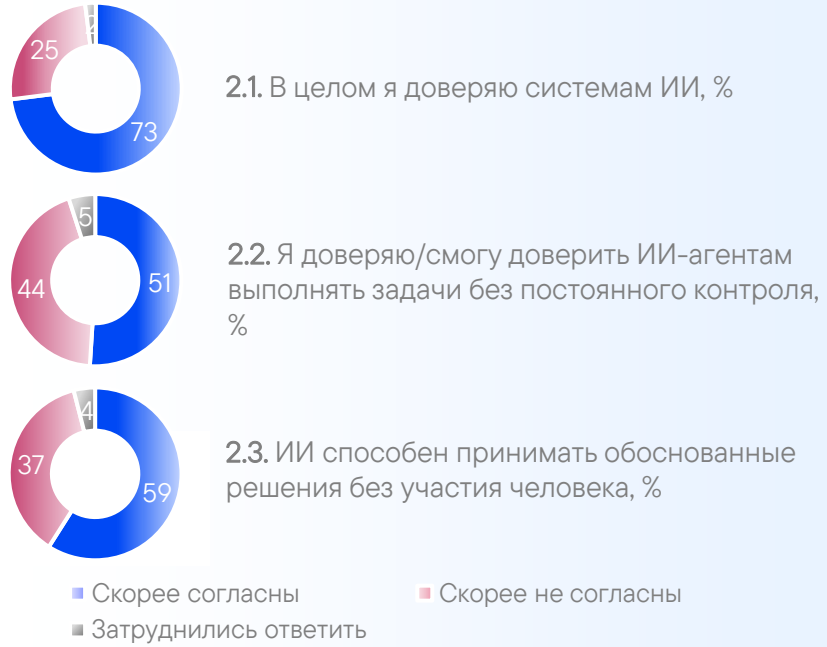
Причина – в текущем уровнем развития моделей, технологий и процессов:

- отсутствие платформенного подхода к разработке и внедрению ИИ и контрольной среды
- недостаточная точность LLM-моделей, склонность к галлюцинациям
- существенная зависимость результата работы ИИ-агента от качества данных и процессов. Отсутствие в компаниях доверенных источников данных и процессного подхода к управлению бизнесом
- накапливание ошибок ИИ-агентами при реализации многошаговых процессов и прочие фундаментальные ограничения ИИ-агентов.

73% в целом доверяют системам ИИ

44% считают, что ИИ-агентам нужен постоянный контроль

2.1-2.3. Доверие и контроль ИИ, в % от всех опрошенных



Готовность к взаимодействию с ИИ-агентами во внешнем контуре ниже из-за юридических рисков

3 Бизнес склонен больше доверять ИИ-агентам, выполняющим задачи вместо сотрудников, но не партнеров по бизнесу или ИИ-агентов, заключающих сделки от имени клиентов.

3.1-3.4. Готовность взаимодействия с ИИ-агентами, в % от всех опрошенных

- **Высокий уровень** доверия при взаимодействии с **ИИ-сотрудниками** объясняется относительно **распространенным применением ИИ** при решении **ad-hoc задач** в режиме co-pilot.
- В случае работы с конфиденциальной информацией уровень ответственности повышается, поэтому снижается доверие.
- Доверие при взаимодействии с ИИ-агентами во внешнем контуре (**ИИ-клиенты, ИИ-партнеры**) **на порядок ниже**, по причине наличия юридических барьеров в части **незакрепленного правового статуса ИИ-агентов.**

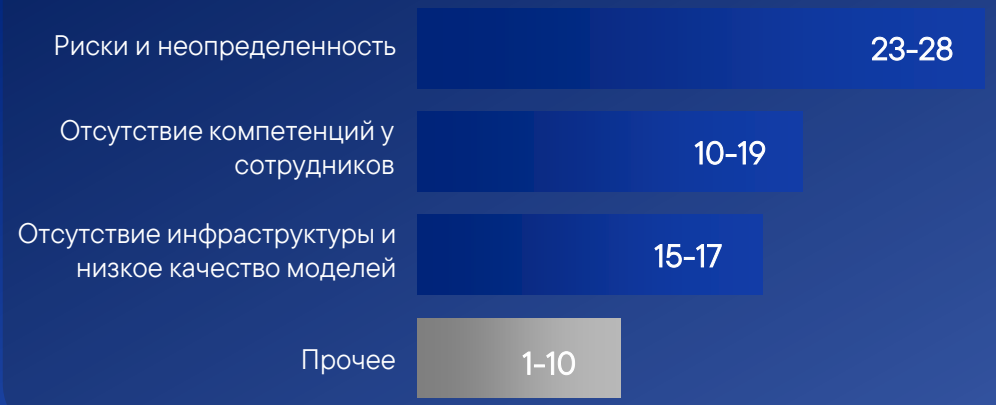
Отсутствие инфраструктуры для разработки – важный фактор снижения доверия к ИИ-агентам

4

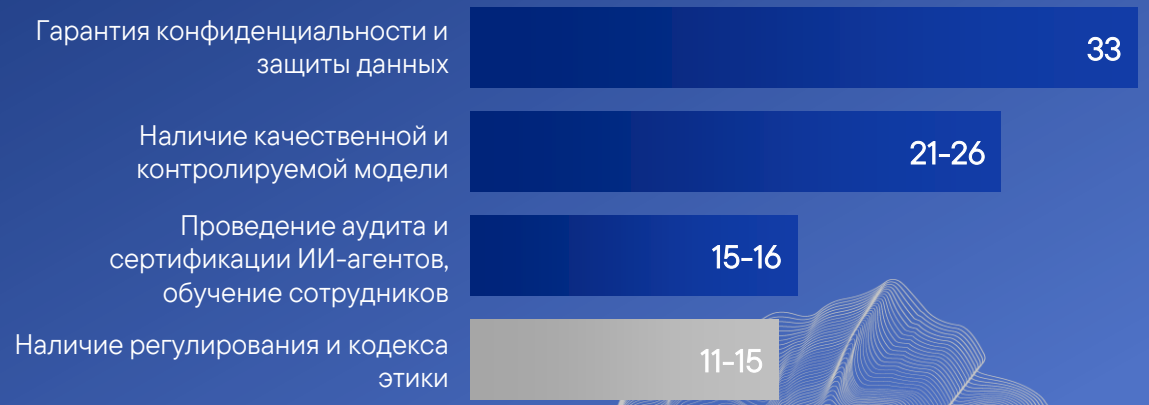
Отсутствие платформенных решений по разработке ИИ приводит к повышению существенности барьеров развития ИИ-агентов в части обеспечения безопасности, надежности и качества моделей.
Бизнес не может доверить конфиденциальные данные иностранным LLM моделям, а российские модели пока не внедрены на уровне платформенного решения.

Дополнительным фактором к повышению доверия к ИИ становится обучение сотрудников, проведение сертификации ИИ-агентов.

4.1. Барьеры при внедрении ИИ*, в % от всех опрошенных



4.2. Факторы повышения доверия к ИИ-агентам*, в % от всех опрошенных



**Многовариантный вопрос, сумма ответов может быть не равна 100%.*

Ключевые наблюдения

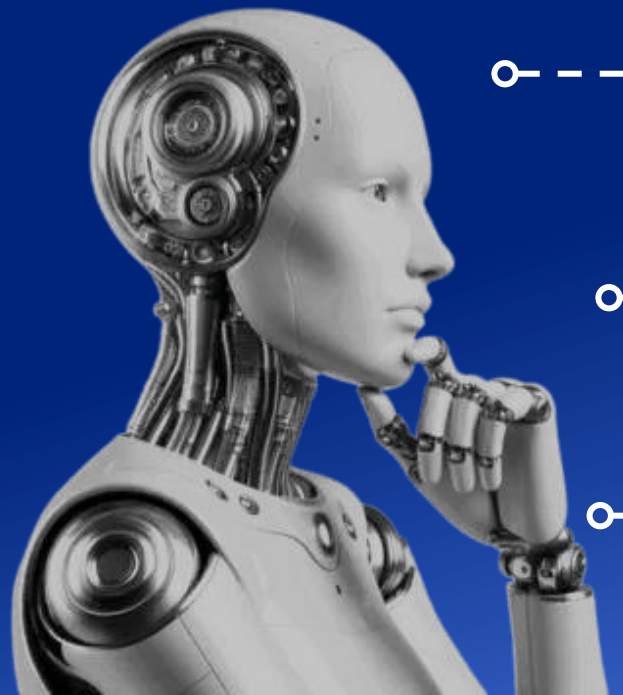
Технологии Доверия

01 **Создание инфраструктуры для разработки** технологий ИИ становится **ключевым драйвером роста доверия** российского бизнеса к ИИ-агентам, влияющее также на безопасность и качество используемых моделей

02 Основные барьеры при внедрении ИИ-агентов касаются **рисков безопасности данных**, неопределенного **правового статуса** ИИ-агентов, отсутствия **контрольной среды** и высоких требований к **качеству фундаментальных LLM моделей**

03 Для масштабного внедрения ИИ-агентов в бизнес необходимо совершенствование **методов процессного управления** и **управления данными**

Рекомендации по дальнейшим шагам



внедрение платформенной инфраструктуры для работы с ИИ агентами, включая обеспечение требований информационной безопасности и создание контрольной среды вокруг ИИ-агентов

тестирование качества ИИ-агентов в конкретных бизнес-процессах и выставление требований к уровню обоснованности принимаемых решений со стороны бизнеса для последующего роста автономности

Повышение качества процессного управления и внедрение методов управления данными, включая создание базы метаданных

Наталья Стрекаль

Директор практики нефинансовых рисков и данных для финансового сектора

+7 (905) 272-76-44
natalia.strekal@tedo.ru

Технологии Доверия