Индекс УДК 336.7:004.8
Код ГРНТИ 06.73.55
DOI: 10.22204/2410-4639-2026-129-01-113-126


Е.В. САМОХОДКИН,
А.А. САМОХОДКИНА*

Доверие к системам искусственного интеллекта как условие цифрового суверенитета в финансовом секторе

В данной работе исследуется доверие к системам искусственного интеллекта в финансовом секторе как условие цифрового суверенитета. Цель исследования состоит в теоретико-аналитическом обосновании того, каким образом доверие формируется на индивидуально-поведенческом и институциональном уровнях и при каких условиях оно становится фактором суверенно ориентированного развития финансовой цифровой среды. Методологическая основа включает социотехнический, институциональный и системный подходы, а также анализ, синтез и сравнительное обобщение научной литературы. Показано, что доверие не сводится ни к пользовательской лояльности, ни к технической эффективности алгоритма; оно возникает на пересечении объяснимости, подотчётности, управления модельным риском, контроля над данными и инфраструктурной автономии. Обосновано, что в условиях непрозрачности и внешних технологических зависимостей доверие приобретает суррогатный характер и не способствует укреплению цифрового суверенитета.

Ключевые слова: искусственный интеллект, доверие к искусственному интеллекту, цифровой суверенитет, финансовый сектор, объяснимый искусственный интеллект, модельный риск, социотехническая инфраструктура

Введение

В последние годы в финансовом секторе фиксируется ускоренная институционализация алгоритмических решений: машинно-обучаемые модели всё чаще встраиваются в кредитование, комплаенс-контуры, антифрод, клиентскую аналитику, инвестиционное консультирование, вследствие чего доверие к данным системам начинает функционировать как

* **Самоходкин Евгений Вячеславович** — аспирант кафедры социологии, психологии управления и истории, ФГБОУ ВО «Государственный университет управления».
E-mail: rodentforme@gmail.com
Самоходкина Алиса Андреевна — аспирант кафедры социологии, психологии управления и истории ФГБОУ ВО «Государственный университет управления».
E-mail: alisaelzon@gmail.com

предпосылка легитимного использования, устойчивого масштабирования, общественного принятия в условиях растущего общественного и научного интереса к искусственному интеллекту (ИИ) [1, 2]. При этом в междисциплинарной литературе доверие к ИИ трактуется как готовность субъектов полагаться на выводы системы, делегируя часть когнитивной работы и принимая связанные риски, что в финансовой сфере приобретает повышенную остроту из-за высокой цены ошибок, асимметрии информации, потенциальных каскадных эффектов на уровне рынков [3]. В российских исследованиях влияние ИИ на финансовый сектор также описывается как амбивалентное: одновременно расширяется инструментальный потенциал управления рисками и усиливается уязвимость к новым классам технологических, правовых, этических неопределённостей, влияющих на доверительный контур взаимодействия «клиент — организация — регулятор» [4, 5].

Данная проблематика встраивается в более широкую рамку цифрового суверенитета, поскольку финансовая инфраструктура, будучи критической для экономической безопасности, оказывается зависимой от данных, вычислительных мощностей, программных стеков, облачных архитектур, поставщиков моделей, режимов трансграничной передачи информации, а значит — от внешних правил, стандартов, юрисдикций, технологических траекторий [6]. В зарубежной политэкономической и правовой аналитике цифровой суверенитет описывается как стремление к восстановлению контроля над цифровым пространством через регуляторные, инфраструктурные и институциональные механизмы, при этом ключевым объектом такого контроля становится именно способность управлять цифровыми зависимостями и их рисковым профилем [7]. В отечественной научной дискуссии цифровой суверенитет, как правило, связывается с контролем над данными, цифровыми технологиями и инфраструктурой, а также с задачей минимизации внешней технологической зависимости, что для финансового сектора означает необходимость согласования инноваций с требованиями безопасности и правовой определённости [8].

Внутри данной рамки доверие к системам ИИ выступает необходимой предпосылкой суверенизации: когда решения модели не поддаются верификации, а причины вывода остаются непрозрачными, контроль подменяется верой в корректность поставщика или в «самоочевидность» результата, что противоречит логике суверенного управления критическими цифровыми активами [9, 10]. Исследования по финансовой объяснимости показывают, что даже при наличии высоких показателей точности дефицит интерпретируемости порождает регуляторное напряжение, снижает организационную готовность к внедрению, осложняет внутренние процедуры аудита и постфактум-разбирательств по спорным кейсам [11, 12]. В прикладной финансовой литературе объяснимый ИИ (XAI, eXplainable Artificial Intelligence — совокупность моделей, процедур и инструментов, с помощью которых обеспечивается интерпретируемое представление результата работы системы) описывается не только как набор методов интерпретации, но и как средство обоснования решения, позволяющее снижать доверительный разрыв между алгоритмической рациональностью и институциональной ответственностью финансовой организации [13]. В российских работах по объяснимости кредитного скоринга акцентируется, что прозрачность модели рассматривается как функциональное требование, поскольку устойчивость и интерпретируемость напрямую связаны с возможностью обоснования кредитных решений и снижением конфликтности взаимодействия с клиентом [14].

Даже при достижении приемлемой объяснимости сохраняется слой управленческой проблематики, связанный с управлением модельным риском: машинно-обучаемые методы, будучи более гибкими и часто более точными, создают новые

формы неопределённости для валидации, мониторинга, сопоставления качества модели и связанных с риском издержек, что в финансовых организациях трансформирует традиционные контуры управления модельным риском (model risk management) [15, 16]. В систематических обзорах подчёркивается, что расширение применения машинного обучения в управлении банковскими рисками сопровождается ростом сложности моделей, повышением требований к данным, появлением уязвимостей, связанных с переносимостью результатов и корректностью обобщения, вследствие чего доверие становится зависимым от зрелости процедур контроля жизненного цикла модели [17]. Дополнительно в исследованиях финансового риск-менеджмента указывается, что практики алгоритмического аудита и объяснимости рассматриваются как инфраструктурные компоненты «доверенной» финансовой аналитики, а не как факультативные меры коммуникации [18]. В российской академической периодике риски внедрения ИИ в банковском секторе также связываются с недостаточной прозрачностью решений и с организационными ограничениями контроля, что усиливает разрыв между технологическим внедрением и институциональным доверием [19].

В итоге проблематика доверия к системам ИИ в финансовом секторе может быть описана как многослойная: техническая объяснимость, модельный риск, справедливость и недискриминация, организационная подотчётность, управляемость данных, суверенность инфраструктуры — данные пласты взаимно усиливают друг друга, формируя эффект критического ограничения, при котором провал в одном измерении обесценивает усилия в других. Следовательно, в рамках данной работы проблематика темы задаётся не вопросом «доверять или не доверять» ИИ как технологии, а вопросом о том, каким образом доверие конструируется как проверяемое, институционально закреплённое свойство социотехнической системы, обеспечивая цифровой суверенитет финансового сектора в условиях ускоренной алгоритмизации.

Методология исследования

Методология исследования носит междисциплинарный характер и основана на сочетании социотехнического, институционального и системного подходов, что позволяет рассматривать доверие к системам ИИ в финансовом секторе одновременно как поведенческий, организационный и инфраструктурный феномен. В качестве основных методов используются анализ и синтез научной литературы, сравнительный анализ российских и зарубежных исследований, а также концептуальное обобщение, направленное на выявление взаимосвязи между объяснимостью, подотчётностью, управлением модельным риском, контролем над данными и цифровым суверенитетом.

Объектом исследования выступают общественные отношения, организационные практики и инфраструктурные процессы, складывающиеся в ходе внедрения и использования систем ИИ в финансовом секторе в условиях цифровой трансформации. Предметом исследования являются механизмы формирования доверия к системам ИИ как к индивидуально воспринимаемому и институционально обеспечиваемому свойству социотехнической инфраструктуры финансового сектора, а также связь данных механизмов с обеспечением цифрового суверенитета.

Цель исследования состоит в теоретико-аналитическом обосновании того, каким образом доверие к системам ИИ в финансовом секторе формируется на поведенческом и институциональном уровнях и при каких условиях данное доверие становится фактором укрепления цифрового суверенитета.

Ограничения исследования могут быть сформулированы следующим образом:

1. Данная работа носит теоретико-аналитический характер и не включает самостоятельную эмпирическую проверку выдвигаемых положений на материале опросов, интервью, кейс-стади

либо экспериментальных данных. Вследствие данного обстоятельства выводы ориентированы прежде всего на концептуальное обобщение и интерпретацию научной литературы.

2. Исследование ограничено финансовым сектором, поскольку именно в данной сфере доверие к ИИ приобретает особую значимость вследствие высокой цены ошибки, чувствительности данных, регуляторной плотности и системной значимости цифровой инфраструктуры. Соответственно, полученные выводы не претендуют на автоматическое распространение на иные отрасли без дополнительной адаптации.
3. Источниковая база ограничена академическими статьями, научными публикациями и исследовательскими работами, тогда как отраслевые отчёты, консалтинговые обзоры, корпоративные документы и иные непроверенные по академическим стандартам материалы в исследовательскую рамку не включаются. Данное ограничение повышает научную строгость, однако одновременно сужает спектр прикладных наблюдений.

Методологическая рамка данной работы выстроена таким образом, чтобы обеспечить согласованное рассмотрение доверия к ИИ в финансовом секторе сразу в нескольких взаимосвязанных плоскостях — поведенческой, институциональной, инфраструктурной и суверенной.

Доверие как индивидуально-поведенческая готовность к принятию решений, сгенерированных искусственным интеллектом

Доверие на данном уровне целесообразно интерпретировать как готовность субъекта допустить алгоритм в контур собственного выбора, признать его рекомендации релевантными и, в предельном случае, скорректировать под них финансовое поведение. Данная трактовка подтверждается банковской эмпирикой: в исследовании принятия ИИ в банковских сервисах отношение к технологии оказалось наиболее сильным непосредственным предиктором намерения использования ($\beta = 0.574$), тогда как воспринимаемое доверие сохранило самостоятельное положительное влияние на поведенческое намерение ($\beta = 0.135$) [20]. В исследованиях robo-advisory (цифровой сервис, который с опорой на алгоритмы и/или модели машинного обучения формирует персонализированные финансовые рекомендации) сходная логика воспроизводится на иной выборке: начальный уровень доверия формируется не только ожидаемой функциональной полезностью, но и предрасположенностью к доверию, воспринимаемой способностью системы оправдать ожидания, социальным влиянием, организационной поддержкой и ценностным восприятием сервиса, тогда как цифровая неграмотность и сложность финтех-решений способны ослаблять доверительный импульс уже на входе [21]. Отсюда следует, что индивидуальная готовность принимать решения, сгенерированные ИИ, обусловливается не одной лишь точностью алгоритма; она возникает в точке пересечения технологических характеристик, пользовательской компетентности и институционального контекста, в котором цифровой выбор воспринимается как допустимый и безопасный.

При этом доверие в финансовой среде носит неоднородный характер, поскольку включает когнитивный, аффективный и социальный компоненты, по-разному влияющие на согласие с машинной рекомендацией. В исследовании сектора BFSI (Banking, Financial Services and Insurance — банковский, финансовый и страховой сектор) было показано, что намерение использовать ИИ как средство поддержки решения определяется аффективным и когнитивным доверием, тогда как делегирование решения цифровому агенту требует дополнительно социального доверия, то есть признания допустимости самой формы алгоритмического посредничества [22]. Экспериментальные исследования финансового консультирования демонстрируют

ещё более существенную деталь: пользователи чаще следуют совету романтического партнёра, чем автономного ИИ-советника, а связь между источником совета и готовностью ему следовать опосредуется одновременно когнитивным и аффективным доверием, причём последнее действует сильнее; вместе с тем антропоморфизация robo-advisor и гибридная схема «человек с ИИ-поддержкой» заметно снижают выраженность AI aversion (алгоритмическая неприязнь / избегание ИИ — это наблюдаемая в экспериментах и полевых исследованиях тенденция реже следовать рекомендациям алгоритма или отказываться от делегирования, особенно после обнаружения ошибки или при наличии риска, даже когда объективно алгоритм в среднем не хуже человека) [23]. Аналитически данное обстоятельство принципиально важно для темы цифрового суверенитета: если готовность принять ИИ-решение возрастает за счёт реляционного оформления интерфейса и эмоционального смягчения восприятия, то доверие начинает опираться не столько на проверяемость и контроль, сколько на поведенческую архитектуру убеждения, а значит, может быть институционально хрупким и манипулятивно уязвимым.

Доверие к ИИ в финансах не является статическим состоянием; оно развивается во времени, переживает фазы усиления, шока и частичного восстановления. В динамическом исследовании взаимодействия человека с ИИ-финансовым консультантом было установлено, что пользователи первоначально демонстрируют algorithmic appreciation (алгоритмическое предпочтение — эмпирически фиксируемая тенденция пользователей придавать рекомендациям/оценкам алгоритма больший вес, чем сопоставимым рекомендациям человека), однако единичная ошибка вызывает существенное падение доверия ($\eta^2 = 0.141$), после чего качественно выстроенные объяснения способны не только восстановить, но и усилить доверительный уровень по сравнению с исходной точкой; дополнительно было показано, что финансовая грамотность усиливает чувствительность к ошибке и одновременно повышает способность к восстановлению доверия после объяснения [24]. Следовательно, индивидуально-поведенческое доверие не тождественно максимальной готовности делегировать выбор машине; напротив, оно ближе к калиброванной готовности, в которой сохраняется способность к критической дистанции, а данная способность непосредственно соотносится с автономией субъекта — одной из фундаментальных предпосылок цифрового суверенитета.

При этом авторы исходят из того, что методологически недостаточно рассматривать доверие к системам ИИ лишь как следствие эмоциональной приемлемости интерфейса либо пользовательской привычки к взаимодействию с цифровым сервисом. Подобная редукция чрезмерно упрощает сам предмет анализа, поскольку в финансовом секторе доверие формируется не на уровне поверхностного впечатления от удобства, а в пространстве более сложной оценки — насколько алгоритмическое решение является понятным, насколько оно поддаётся интерпретации, в какой мере сохраняется возможность соотнести его логику с характером конкретной финансовой ситуации. Именно поэтому доверие к ИИ должно рассматриваться не как эмоциональная реакция на технологическую новизну, а как результат осмысленного соотнесения функциональной эффективности, прозрачности, безопасности и пределов допустимой алгоритмической автономии.

В данной связи особую значимость приобретает вопрос объяснимости. Если пользователь сталкивается с решением, логика которого для него принципиально закрыта, доверие оказывается лишённым когнитивного основания и начинает опираться либо на вынужденное принятие, либо на репутационный капитал организации, либо на инерцию цифрового поведения. Однако подобная конфигурация является внутренне неустойчивой, поскольку в ситуации финансового риска одного символического авторитета недостаточно. Для

того чтобы готовность принять решение, сгенерированное ИИ, приобрела устойчивый характер, требуется не абстрактная прозрачность как декларация, а содержательная объяснимость, позволяющая понять, почему система пришла именно к данному выводу, где пролегают границы применимости рекомендации, каким образом учитывается персональный риск-профиль и в какой степени пользователь сохраняет возможность влиять на дальнейшую траекторию решения.

Вместе с тем и сама объяснимость не должна мыслиться упрощённо. Формальное раскрытие участия ИИ в принятии решения ещё не гарантирует усиления доверия. Напротив, подобное раскрытие может производить обратный эффект, если оно не сопровождается понятной для пользователя интерпретацией смысла алгоритмической обработки. Иначе говоря, доверие подрывается не столько самим фактом участия ИИ, сколько отсутствием ясности относительно его роли, компетенции и ограничений. Отсюда вытекает важное аналитическое следствие: для индивидуально-поведенческой готовности принципиально значима такая форма раскрытия, при которой снимается неопределённость относительно оснований решения, условий использования данных и возможностей оспаривания результата.

Данное обстоятельство позволяет перейти к более широкому пониманию доверия как многосоставной конструкции. В финансовом секторе пользователь фактически имеет дело не с изолированным алгоритмом, а с комплексной социотехнической конфигурацией, включающей данные, модель, интерфейс, процедуры верификации, нормативные ограничения, внутренний контроль, механизмы ответственности. Поэтому доверие к ИИ не исчерпывается убеждённостью в математической корректности модели. Оно включает доверие к качеству исходных данных, к корректности их обработки, к устойчивости алгоритма за пределами обучающей выборки, к способности организации выявлять ошибки, к наличию институциональных фильтров, не допускающих произвольной цифровой власти над пользователем. В данном смысле объяснимость выступает существенным, но не единственным элементом доверительной архитектуры.

Логика дальнейшего рассуждения подводит к необходимости учёта человекоориентированного измерения. Индивидуальная готовность принять решение, сгенерированное ИИ, возникает, как представляется, не в условиях полной замены субъекта машиной, а там, где сохраняется представление о контролируемости цифрового инструмента. Когда алгоритм встроен в контур, допускающий проверку, корректировку, человеческое вмешательство и институциональную ответственность, доверие приобретает более рациональную и, следовательно, более устойчивую форму. Если же система воспринимается как автономный и недоступный пересмотру механизм, пользовательская готовность к принятию её рекомендаций становится либо вынужденной, либо ситуативной, либо основанной на неосведомлённости. Для финансовой сферы, где цена ошибки может быть значительной, подобная зависимость имеет принципиальное значение.

Именно на данном уровне проблематика доверия начинает непосредственно сопрягаться с цифровым суверенитетом. Поведенческое доверие пользователя не возникает в нормативном вакууме — оно является производным от того, каким образом организован контроль над данными, инфраструктурой и цифровыми посредниками. Если данные циркулируют по непрозрачным маршрутам, если вычислительная логика решения вынесена за пределы доступного контроля, если критически значимые компоненты финансовой цифровой среды зависят от внешних поставщиков и закрытых технологических стеков, доверие теряет суверенное основание. В таком случае пользователь может демонстрировать внешнюю готовность следовать алгоритмическому выводу, однако данная готовность будет опираться не на институционально обеспеченную

уверенность, а на структурную асимметрию — на невозможность проверить, понять и оспорить решение.

Следовательно, в контексте цифрового суверенитета доверие должно интерпретироваться как переживаемое пользователем выражение более глубокой архитектурной упорядоченности. Безопасность, понятность, оспоримость, контролируемость, ограниченность алгоритмической автономии — данные свойства приобретают не только прикладное, но и суверенное значение. Чем выше способность финансовой системы гарантировать субъекту доступное понимание принципов работы ИИ, защиту данных, процедурную справедливость и наличие реальных каналов пересмотра решения, тем в большей степени индивидуально-поведенческое доверие перестаёт быть психологической случайностью и превращается в социально и институционально обоснованную готовность к использованию алгоритмических инструментов.

При данном подходе становится очевидным, что доверие к решениям, сгенерированным ИИ, не может быть понято как спонтанная лояльность цифровой инновации. Оно складывается в результате сложной калибровки между воспринимаемой компетентностью системы, качеством объяснений, режимом обращения с данными, степенью человеческого контроля и субъективной способностью пользователя удерживать автономную позицию в ситуации алгоритмического воздействия. Чем в большей мере пользователь понимает, кто контролирует данные, каким образом проверяется модель, где проходят границы автоматизации и каким способом может быть оспорен результат, тем выше вероятность формирования доверия, совместимого с логикой цифрового суверенитета.

Таким образом, индивидуально-поведенческая готовность к принятию решений, сгенерированных ИИ, должна рассматриваться не как частный поведенческий эффект цифровизации, а как стратегически значимый индикатор состояния финансовой цифровой среды. Там, где доверие опирается на содержательную объяснимость, институциональную подотчётность и контролируемость инфраструктуры, оно способствует укреплению суверенной модели развития. Там же, где оно формируется на основе непрозрачности, интерфейсного удобства и скрытых цифровых зависимостей, возникает лишь видимость доверия, лишённая устойчивого нормативного и технологического основания.

Доверие как институционально обеспеченное свойство социотехнической инфраструктуры

В рамках анализа доверия к системам ИИ в финансовом секторе продуктивным представляется смещение оптики с индивидуального восприятия на институциональную среду, поскольку в высокорисковых цифровых контурах доверие формируется не столько в сознании пользователя, сколько в архитектуре проверки, подотчётности и управляемости алгоритмического решения. В российской академической дискуссии данная логика выражена предельно отчётливо: доверие к ИИ связывается с качеством знаний и данных, границами применимости модели, процедурами валидации и верификации, объяснимостью вывода и ролью человека в принятии окончательного решения, вследствие чего доверенным может считаться лишь такой ИИ, который является подотчётным, надёжным и объяснимым [25].

Данная постановка особенно значима для финансового сектора, где алгоритмическая система включается не в нейтральную информационную среду, а в инфраструктуру распределения доступа к капиталу, оценки кредитоспособности, обнаружения мошенничества и управления рисками. Поэтому институциональное доверие не может выводиться из одной лишь предиктивной эффективности: обзор объяснимого ИИ в финансовой сфере показывает, что в финансовой сфере доверительность решений непосредственно увязывается с прозрачностью, способностью интерпретировать ошибки, подот-

чётностью и корректным выбором объяснительных инструментов, причём сами объяснения подлежат оценке по критериям правдоподобия и добросовестности, иначе возрастает риск ложной уверенности в корректности модели [11]. В той же логике показывается, что внедрение машинного обучения в кредитный скоринг действительно повышает прогностическую результативность, однако одновременно производит новые формы модельного риска, прежде всего в плоскости надзорной валидации, технологической прозрачности и проверяемости модели, а неопределённость относительно того, каким образом данные риски будут оцениваться регулятором, уже сама по себе способна тормозить инновацию [16]. Из данного обстоятельства следует принципиальный вывод: доверие к ИИ в финансах возникает там, где организация способна доказуемо показать, почему ему можно доверять, где проходят пределы его применимости и каким образом устраняются институциональные последствия ошибки.

Вместе с тем доверие как свойство социотехнической инфраструктуры не исчерпывается внутренними процедурами банка либо финтех-платформы, поскольку в высокорисковых контурах оно носит распределённый характер. В исследовании, посвящённом управлению высокорисковыми системами ИИ, выполненном на материале здравоохранения и кредитного скоринга, подчёркивается, что кредитный скоринг представляет собой сеть взаимодействующих акторов — финансовых организаций, кредитных бюро, разработчиков, регуляторов и пользователей, вследствие чего формирование доверия не может быть локализовано в одной точке; на уровне поставщика решения оно требует тестирования, валидации, постоянного мониторинга, прозрачности и механизмов оспаривания решений, на институциональном уровне — доверия к регулирующим органам и установленным ими стандартам, на общественном уровне — уверенности в справедливости цифрового порядка в целом [26]. Более того, в данной работе специально подчёркивается, что пользователи должны быть способны адекватно оценивать степень надёжности и допустимые пределы доверия к ИИ-системам, иначе формируются режимы избыточного доверия или недостаточного доверия, в равной мере деструктивные для высокорисковых сфер. Следовательно, институционально обеспеченное доверие предполагает не просто наличие регуляции, а наличие такой регуляции и такой организационной практики, которые делают доверительные свойства системы наблюдаемыми, проверяемыми и поддающимися калибровке для всех участников инфраструктуры.

С позиций цифрового суверенитета данная проблематика углубляется, поскольку институциональное доверие к ИИ оказывается неотделимым от контроля над данными, цифровой инфраструктурой и внешними технологическими зависимостями. В российской научной литературе цифровой суверенитет определяется через способность государства контролировать собственные цифровые данные, информацию, технологии и инфраструктуру, обеспечивая безопасность, конфиденциальность, киберустойчивость и относительную независимость от внешних поставщиков, причём в структуру цифрового пространства прямо включаются данные, информационные системы, сети, инфраструктура, нормы регулирования и механизмы управления [8]. В результате доверие к ИИ-системе в финансовом секторе не может быть признано институционально зрелым, если критически важные контуры хранения, обработки и интерпретации данных вынесены в непрозрачные внешние юрисдикции, а внутренний субъект управления не обладает полнотой контроля над жизненным циклом модели.

Зарубежные исследования суверенитета финансовых данных подтверждают данную логику уже на уровне систематического обзора. В работе, выполненной по протоколу PRISMA на материале 52 исследований, показано, что страны и регио-

ны существенно расходятся в подходах к регулированию ИИ и обеспечению суверенитета финансовых данных; при этом из 13 выявленных методов обеспечения суверенитета данных лишь 5 прямо адресуют вопросы безопасности, хотя именно безопасность признаётся наиболее значимым компонентом. Одновременно подчёркивается необходимость формирования устойчивой и согласованной нормативно-правовой рамки, способной совместить доверие, защиту частной жизни и автономию данных в финансовом секторе [27]. Там же фиксируется, что возрастающая опора финансовых систем на цифровые процессы усиливает риски, связанные с защитой частной жизни, регуляторной согласованностью, алгоритмическими смещениями, отравлением данных и состязательными атаками, вследствие чего без надлежащего надзора и защитных рамок ИИ в финансах способен подрывать доверие, справедливость и соблюдение законодательства о суверенитете данных. В аналитическом плане это означает следующее: цифровой суверенитет не является внешним политико-правовым фоном для доверия к ИИ; напротив, он выступает его инфраструктурной предпосылкой, поскольку именно контроль над данными и платформами делает возможными подлинную проверку модели, эффективный аудит и реальную подотчётность.

Тем самым доверие как институционально обеспеченное свойство социотехнической инфраструктуры должно пониматься как результат согласованного действия нескольких взаимосвязанных контуров: верифицируемости модели, прозрачности её поведения, распределения ответственности между участниками, наличия процедур обжалования, устойчивости цифровой инфраструктуры и суверенного контроля над данными. В логике цифрового суверенитета доверие приобретает качество стратегического критерия: если финансовая организация и регулирующая среда способны не только использовать ИИ, но и управлять его рисками, контролировать инфраструктурные зависимости, локализовать ответственность и обеспечивать автономию данных, тогда доверие становится институциональным ресурсом суверенного развития; если же алгоритмическая система встроена в непрозрачные, внешне управляемые и слабо оспоримые контуры, тогда речь идёт не о доверии в строгом смысле, а о вынужденной зависимости, маскируемой под цифровую эффективность.

Заключение

Проведённое исследование позволило установить, что доверие к системам ИИ в финансовом секторе не может быть сведено ни к субъективной расположенности пользователя к цифровым сервисам, ни к формальному соответствию алгоритма заданным техническим параметрам. В рамках данной работы было показано, что доверие представляет собой многослойную и институционально опосредованную конструкцию, формирующуюся на пересечении индивидуального восприятия, организационных механизмов контроля, нормативной архитектуры, качества данных и инфраструктурной автономии. Вследствие данного обстоятельства доверие следует интерпретировать не как внешнее сопровождение цифровой трансформации, а как одно из её базовых условий, от которого зависит возможность легитимного, устойчивого и суверенно ориентированного внедрения ИИ в финансовые процессы.

В ходе анализа было выявлено, что на индивидуально-поведенческом уровне доверие выражается в готовности субъекта принять алгоритмически сгенерированное решение в контур собственного выбора, однако данная готовность не является простой реакцией на удобство интерфейса либо на технологическую новизну. Напротив, она определяется более сложной конфигурацией факторов, включающей воспринимаемую компетентность системы, содержательную объяснимость, понятность границ применимости модели, защищённость данных, возможность сохранения

человеческого контроля и процедурную оспоримость результата. Тем самым было обосновано, что доверие к ИИ в финансовой сфере должно рассматриваться как калиброванная готовность к взаимодействию с алгоритмической системой, а не как безусловное делегирование ей решения. Данная калибровка особенно значима в условиях цифрового суверенитета, поскольку именно она позволяет сохранить субъектную автономию в ситуации возрастающей алгоритмизации финансового поведения.

Одновременно было показано, что ограничение анализа только поведенческим измерением доверия является методологически недостаточным. Доверие в высокорисковых цифровых контурах возникает и воспроизводится в границах социотехнической инфраструктуры, где решающую роль начинают играть процедуры валидации и верификации моделей, качество управления модельным риском, прозрачность логики вывода, распределение ответственности между участниками, механизмы надзора, обжалования и аудита. При данном подходе доверие предстаёт как институционально обеспеченное свойство системы, а его наличие определяется не декларацией надёжности, а способностью финансовой организации и регулирующей среды доказуемо подтверждать корректность, подотчётность и управляемость алгоритмического решения. Из данного вывода следует, что доверие не может быть признано устойчивым, если оно не поддержано зрелыми организационными и правовыми механизмами контроля.

Особое значение в рамках проведённого исследования приобрела связка доверия и цифрового суверенитета. Было установлено, что доверие к ИИ в финансовом секторе утрачивает устойчивое основание в тех случаях, когда критически значимые контуры хранения, обработки и интерпретации данных оказываются вынесенными в непрозрачные внешние юрисдикции, когда жизненный цикл модели не поддаётся полноценному внутреннему контролю, а инфраструктурные зависимости маскируются под технологическую эффективность. Следовательно, цифровой суверенитет должен рассматриваться не как внешний политико-правовой фон, а как инфраструктурная предпосылка доверия. Именно контроль над данными, вычислительными ресурсами, цифровыми платформами, регуляторными рамками и процедурами ответственности создаёт условия, при которых доверие становится не иллюзией надёжности, а подтверждаемым свойством финансовой цифровой среды.

В результате было обосновано, что доверие к системам ИИ в финансовом секторе имеет двойственную природу. С одной стороны, оно проявляется как индивидуально-поведенческая готовность субъекта признать алгоритмическую рекомендацию допустимой и значимой для собственного выбора. С другой стороны, оно формируется как институционально закреплённое качество социотехнической инфраструктуры, обеспечиваемое прозрачностью, подотчётностью, управляемостью рисков, процедурной справедливостью и суверенным контролем над данными и технологическими зависимостями. Данные измерения не противостоят друг другу, а образуют единую конструкцию: без институциональной обеспеченности поведенческое доверие становится хрупким и манипулятивно уязвимым, тогда как без пользовательской готовности даже наиболее развитая инфраструктура не приобретает полноценной социальной легитимности.

Тем самым основная исследовательская идея получает подтверждение: в условиях ускоренной алгоритмизации финансового сектора доверие к ИИ должно пониматься как проверяемое, институционально сконструированное и нормативно насыщенное свойство цифровой среды, выступающее необходимым условием цифрового суверенитета. Чем выше степень объяснимости, подотчётности, контролируемости и инфраструктурной автономии финансовых ИИ-систем, тем более вероятным становится формирование доверия, совместимого с задачами устойчивого и безопасного разви-

тия. И наоборот, там, где алгоритмические решения встроены в непрозрачные, внешне зависимые и слабо оспоримые контуры, возникает не доверие в строгом смысле, а лишь его функциональный суррогат, не укрепляющий, а размывающий суверенные основания финансовой цифровизации.

Итак, по итогам проведённой работы представляется возможным заключить, что доверие к системам ИИ в финансовом секторе должно рассматриваться как стратегический ресурс, от состояния которого зависит не только эффективность конкретных цифровых решений, но и способность финансовой системы сохранять управляемость, автономию и институциональную устойчивость в условиях технологической трансформации. Именно в данной плоскости доверие перестаёт быть вторичным социально-психологическим эффектом и превращается в одну из центральных категорий анализа цифрового суверенитета.

ЛИТЕРАТУРА

1. Ridzuan N.N., Masri M., Anshari M., Fitriyani N.L., Syafrudin M. AI in the Financial Sector: The Line between Innovation, Regulation and Ethical Responsibility // Information. 2024. Vol. 15, No. 8. Art. 432. DOI: 10.3390/info15080432.
2. Samokhodkin E.V., Elzon A.A. Analysis of the Relationship Between Scientific Interest and Publication Dynamics on Artificial Intelligence in the Russian Federation (2020–2024) // Scientific and Technical Information Processing. 2025. Vol. 52, No. 2. P. 152. DOI: 10.3103/S0147688225700182.
3. Afroogh S., Akbari A., Malone E., Kargar M., Alambeigi H. Trust in AI: Progress, Challenges, and Future Directions // Humanit. Soc. Sci. Commun. 2024. Vol. 11. Art. 1568. DOI: 10.1057/s41599-024-04044-8.
4. Samokhodkin E.V., Elzon A.A. Social Threats of Artificial Intelligence: Classification and Risk Assessment // Scientific and Technical Information Processing. 2025. Vol. 52, No. 3. P. 263. DOI: 10.3103/S0147688225700674.
5. Кузин М. А. Оценка влияния развития искусственного интеллекта на финансовый сектор экономики // Вестник евразийской науки. 2023. Т. 15, № s4. Art. 31FAVN423.
6. Rodriguez Pita P., Perez-Martinez J., Urueña López A. From Concept to Method: A Framework for Measuring Digital Sovereignty // SSRN. 2025. DOI: 10.2139/ssrn.5683701.
7. Farrand B., Farrand Carrapico H. Digital Sovereignty and Taking Back Control: From Regulatory Capitalism to Regulatory Mercantilism in EU Cybersecurity // European Security. 2022. Vol. 31, No. 3. P. 435. DOI: 10.1080/09662839.2022.2102896.
8. Авдийский В.И., Иванов А.В., Царегородцев А.В. Взаимосвязь цифрового суверенитета и цифрового пространства: новые вызовы и перспективы // Вестник евразийской науки. 2024. Т. 16, № s3. Art. 21FAVN324.
9. Balasubramaniam N., Kauppinen M., Rannisto A., Hiekkanen K., Kujala S. Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements // Information and Software Technology. 2023. Vol. 159. Art. 107197. DOI: 10.1016/j.infsof.2023.107197.
10. Timokhovich A.N., Samokhodkin E.V., Elzon A.A. Classification of Epistemological Threats Posed by Artificial Intelligence // Scientific and Technical Information Processing. 2025. Vol. 52, No. 4. P. 318. DOI: 10.3103/S0147688225700741.
11. Yeo W.J., Van Der Heever W., Mao R., Cambria E., Satapathy R., Mengaldo G. A Comprehensive Review on Financial Explainable AI // Artif. Intell. Rev. 2025. Vol. 58. Art. 189. DOI: 10.1007/s10462-024-11077-7.
12. Кантаев Н.К. Перспективы использования методов машинного обучения для оценки кредитных рисков // Вестник евразийской науки. 2025. Т. 17, № s2. Art. 70FAVN225.
13. Weber P., Carl K. V., Hinz O. Applications of Explainable Artificial Intelligence in Finance — a Systematic Review of Finance, Information Systems, and Computer Science Literature // Manag. Rev. Q. 2024. Vol. 74, No. 2. P. 867. DOI: 10.1007/s11301-023-00320-0.
14. Якубин А. Я. Искусственный интеллект в кредитном скоринге: разработка, обучение и применение модели // Молодой учёный. 2025. № 19 (570). С. 33-35.

15. Самоходкин Е. В., Эльзон А. А. Подходы к формализации нормативного поведения автономных интеллектуальных агентов // Искусственный интеллект и принятие решений. 2025. № 4. С. 35. DOI: 10.14357/20718594250403.
16. Alonso Robisco A., Carbó Martínez J. M. Measuring the Model Risk-Adjusted Performance of Machine Learning Algorithms in Credit Default Prediction // Financial Innovation. 2022. Vol. 8. Art. 70. DOI: 10.1186/s40854-022-00366-1.
17. Heß V.L., Damásio B. Machine Learning in Banking Risk Management: Mapping a Decade of Evolution // International Journal of Information Management Data Insights. 2025. Vol. 5, No. 1. Art. 100324. DOI: 10.1016/j.jjimei.2025.100324.
18. Fritz-Morgenthal S., Hein B., Papenbrock J. Financial Risk Management and Explainable, Trustworthy, Responsible AI // Front. Artif. Intell. 2022. Vol. 5. Art. 779799. DOI: 10.3389/frai.2022.779799.
19. Гозгешев Э.А. Искусственный интеллект в банковском секторе: возможности и риски // Вестник евразийской науки. 2024. Т. 16, № s6. Art. 49FAVN624.
20. Byambaa O., Yondon C., Rentsen E., Darkhijav B., Rahman M. An Empirical Examination of the Adoption of Artificial Intelligence in Banking Services: the Case of Mongolia // Futur. Bus. J. 2025. Vol. 11, No. 1. Art. 76. DOI: 10.1186/s43093-025-00504-y.
21. Bashir Z., Farooq S., Iqbal M.S., Aamir M. The Role of Trust in Financial Robo-Advisory Adoption: A Case of Young Retail Investors in Pakistan // Sustainable Futures. 2025. Vol. 9. Art. 100538. DOI: 10.1016/j.sftr.2025.100538.
22. Pathak A., Bansal V. AI as Decision Aid or Delegated Agent: The Effects of Trust Dimensions on the Adoption of AI Digital Agents // Computers in Human Behavior: Artificial Humans. 2024. Vol. 2, No. 2. Art. 100094. DOI: 10.1016/j.chbah.2024.100094.
23. Hermann E., Alberhasky M. The Trusted Partner for Financial Decision Making: Romantic Partner or AI? // Computers in Human Behavior. 2026. Vol. 177. Art. 108877. DOI: 10.1016/j.chb.2025.108877.
24. Han J., Ko D. Trust Formation, Error Impact, and Repair in Human-AI Financial Advisory: A Dynamic Behavioral Analysis // Behav. Sci. 2025. Vol. 15, No. 10. Art. 1370. DOI: 10.3390/bs15101370.
25. Кобринский Б.А. Доверие к технологиям искусственного интеллекта // Искусственный интеллект и принятие решений. 2024. № 3. С. 3–17. DOI: 10.14357/20718594240301.
26. Bartsch S., Behn O., Benlian A., Brownsword R. [et al.] Governance of High-Risk AI Systems in Healthcare and Credit Scoring // Business & Information Systems Engineering. 2025. Vol. 67, No. 4. P. 563. DOI: 10.1007/s12599-025-00944-4.
27. Patil A., Mishra B., Chockalingam S., Misra S., Kvalvik P. Securing Financial Systems Through Data Sovereignty: a Systematic Review of Approaches and Regulations // Int. J. Inf. Secur. 2025. Vol. 24. Art. 159. DOI: 10.1007/s10207-025-01074-4.

Trust in Artificial Intelligence Systems as a Condition for Digital Sovereignty in the Financial Sector

Evgeniy Vyacheslavovich Samokhodkin — Postgraduate Student, Department of Sociology, Management Psychology and History, State University of Management.
E-mail: rodentforme@gmail.com

Alisa Andreevna Samokhodkina — Postgraduate Student, Department of Sociology, Management Psychology and History, State University of Management.
E-mail: alisaelzon@gmail.com

This paper examines trust in artificial intelligence systems in the financial sector as a condition of digital sovereignty. The study aims to provide a theoretical and analytical explanation of how trust is formed at the individual-behavioral and institutional levels and under what conditions it becomes a factor of sovereignty-oriented development of the financial digital environ-

ment. The methodological framework combines socio-technical, institutional, and systemic approaches, as well as analysis, synthesis, and comparative generalization of academic literature. It is demonstrated that trust cannot be reduced either to user loyalty or to the technical efficiency of an algorithm; rather, it emerges at the intersection of explainability, accountability, model risk management, data control, and infrastructural autonomy. It is argued that under conditions of opacity and external technological dependencies, trust becomes surrogate in nature and does not contribute to the strengthening of digital sovereignty.

Keywords: artificial intelligence, trust in artificial intelligence, digital sovereignty, financial sector, explainable artificial intelligence, model risk, socio-technical infrastructure

REFERENCES

1. Ridzuan N.N., Masri M., Anshari M., Fitriyani N.L., Syafrudin M. AI in the Financial Sector: The Line between Innovation, Regulation and Ethical Responsibility // Information. 2024. Vol. 15, No. 8. Art. 432. DOI: 10.3390/info15080432.
2. Samokhodkin E.V., Elzon A.A. Analysis of the Relationship Between Scientific Interest and Publication Dynamics on Artificial Intelligence in the Russian Federation (2020–2024) // Scientific and Technical Information Processing. 2025. Vol. 52, No. 2. P. 152. DOI: 10.3103/S0147688225700182.
3. Afroogh S., Akbari A., Malone E., Kargar M., Alambeigi H. Trust in AI: Progress, Challenges, and Future Directions // Humanit. Soc. Sci. Commun. 2024. Vol. 11. Art. 1568. DOI: 10.1057/s41599-024-04044-8.
4. Samokhodkin E.V., Elzon A.A. Social Threats of Artificial Intelligence: Classification and Risk Assessment // Scientific and Technical Information Processing. 2025. Vol. 52, No. 3. P. 263. DOI: 10.3103/S0147688225700674.
5. Kuzin M.A. Otsenka vliyaniya razvitiya iskusstvennogo intellekta na finansovyy sektor ekonomiki // Vestnik evraziyskoy nauki. 2023. T. 15, № s4. Art. 31FAVN423 (in Russian).
6. Rodriguez Pita P., Perez-Martinez J., Urueña López A. From Concept to Method: A Framework for Measuring Digital Sovereignty // SSRN. 2025. DOI: 10.2139/ssrn.5683701.
7. Farrand B., Farrand Carrapico H. Digital Sovereignty and Taking Back Control: From Regulatory Capitalism to Regulatory Mercantilism in EU Cybersecurity // European Security. 2022. Vol. 31, No. 3. P. 435. DOI: 10.1080/09662839.2022.2102896.
8. Avdiyskiy V.I., Ivanov A.V., Tsaregorodtsev A.V. Vzaimosvyaz tsifrovogo suvereniteta i tsifrovogo prostranstva: novye vyzovy i perspektivy // Vestnik evraziyskoy nauki. 2024. T. 16, № s3. Art. 21FAVN324 (in Russian).
9. Balasubramaniam N., Kauppinen M., Rannisto A., Hiekkanen K., Kujala S. Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements // Information and Software Technology. 2023. Vol. 159. Art. 107197. DOI: 10.1016/j.infsof.2023.107197.
10. Timokhovich A.N., Samokhodkin E.V., Elzon A.A. Classification of Epistemological Threats Posed by Artificial Intelligence // Scientific and Technical Information Processing. 2025. Vol. 52, No. 4. P. 318. DOI: 10.3103/S0147688225700741.
11. Yeo W.J., Van Der Heever W., Mao R., Cambria E., Satapathy R., Mengaldo G. A Comprehensive Review on Financial Explainable AI // Artif. Intell. Rev. 2025. Vol. 58. Art. 189. DOI: 10.1007/s10462-024-11077-7.
12. Kantaev N.K. Perspektivy ispolzovaniya metodov mashinnogo obucheniya dlya otsenki kreditnykh riskov // Vestnik evraziyskoy nauki. 2025. T. 17, № s2. Art. 70FAVN225 (in Russian).
13. Weber P., Carl K.V., Hinz O. Applications of Explainable Artificial Intelligence in Finance — a Systematic Review of Finance, Information Systems, and Computer Science Literature // Manag. Rev. Q. 2024. Vol. 74, No. 2. P. 867. DOI: 10.1007/s11301-023-00320-0.

14. Yakubin A.Ya. Iskusstvennyy intellekt v kreditnom skoringe: razrabotka, obuchenie i primenenie modeli // Molodoy uchenyy. 2025. № 19 (570). S. 33–35 (in Russian).
15. Samokhodkin E.V., Elzon A.A. Podkhody k formalizatsii normativnogo povedeniya avtonomnykh intellektualnykh agentov // Iskusstvennyy intellekt i prinyatie resheniy. 2025. № 4. S. 35. DOI: 10.14357/20718594250403 (in Russian).
16. Alonso Robisco A., Carbó Martínez J.M. Measuring the Model Risk-Adjusted Performance of Machine Learning Algorithms in Credit Default Prediction // Financial Innovation. 2022. Vol. 8. Art. 70. DOI: 10.1186/s40854-022-00366-1.
17. Heß V.L., Damásio B. Machine Learning in Banking Risk Management: Mapping a Decade of Evolution // International Journal of Information Management Data Insights. 2025. Vol. 5, No. 1. Art. 100324. DOI: 10.1016/j.jjimei.2025.100324.
18. Fritz-Morgenthal S., Hein B., Papenbrock J. Financial Risk Management and Explainable, Trustworthy, Responsible AI // Front. Artif. Intell. 2022. Vol. 5. Art. 779799. DOI: 10.3389/frai.2022.779799.
19. Gozgeshev E.A. Iskusstvennyy intellekt v bankovskom sektore: vozmozhnosti i riski // Vestnik evraziyskoy nauki. 2024. T. 16, № s6. Art. 49FAVN624 (in Russian).
20. Byambaa O., Yondon C., Rentsen E., Darkhijav B., Rahman M. An Empirical Examination of the Adoption of Artificial Intelligence in Banking Services: the Case of Mongolia // Futur. Bus. J. 2025. Vol. 11, No. 1. Art. 76. DOI: 10.1186/s43093-025-00504-y.
21. Bashir Z., Farooq S., Iqbal M.S., Aamir M. The Role of Trust in Financial Robo-Advisory Adoption: A Case of Young Retail Investors in Pakistan // Sustainable Futures. 2025. Vol. 9. Art. 100538. DOI: 10.1016/j.sftr.2025.100538.
22. Pathak A., Bansal V. AI as Decision Aid or Delegated Agent: The Effects of Trust Dimensions on the Adoption of AI Digital Agents // Computers in Human Behavior: Artificial Humans. 2024. Vol. 2, No. 2. Art. 100094. DOI: 10.1016/j.chbah.2024.100094.
23. Hermann E., Alberhasky M. The Trusted Partner for Financial Decision Making: Romantic Partner or AI? // Computers in Human Behavior. 2026. Vol. 177. Art. 108877. DOI: 10.1016/j.chb.2025.108877.
24. Han J., Ko D. Trust Formation, Error Impact, and Repair in Human-AI Financial Advisory: A Dynamic Behavioral Analysis // Behav. Sci. 2025. Vol. 15, No. 10. Art. 1370. DOI: 10.3390/bs15101370.
25. Kobrinskiy B.A. Doverie k tekhnologiyam iskusstvennogo intellekta // Iskusstvennyy intellekt i prinyatie resheniy. 2024. № 3. S. 3–17. DOI: 10.14357/20718594240301 (in Russian).
26. Bartsch S., Behn O., Benlian A., Brownsword R. [et al.] Governance of High-Risk AI Systems in Healthcare and Credit Scoring // Business & Information Systems Engineering. 2025. Vol. 67, No. 4. P. 563. DOI: 10.1007/s12599-025-00944-4.
27. Patil A., Mishra B., Chockalingam S., Misra S., Kvalvik P. Securing Financial Systems Through Data Sovereignty: a Systematic Review of Approaches and Regulations // Int. J. Inf. Secur. 2025. Vol. 24. Art. 159. DOI: 10.1007/s10207-025-01074-4.