

Научная статья

УДК 336.71 + 004.8

JEL: G21 + D82

<https://doi.org/10.18184/2079-4665.2026.17.3.432-447>

Агентская теория в условиях использования ИИ-агентов в финансовом посредничестве

Зверькова Татьяна Николаевна¹¹ Оренбургский государственный университет имени В. А. Бондаренко; Оренбург, Россия¹ tnzverkova@mail.ru, <https://orcid.org/0000-0002-6540-6154>

Аннотация

Цель исследования заключается в теоретическом и прикладном обосновании трансформации агентских отношений в финансовом посредничестве в условиях применения ИИ-агентов.

Методы. В исследовании использованы теоретический, сравнительный и структурно-аналитический подходы, позволившие сопоставить положения агентской теории с внедрением ИИ в финансовом посредничестве. Классические положения агентской теории последовательно соотнесены с условиями передачи функций алгоритмическим системам.

Результаты работы. Установлено, что при внедрении ИИ агентские издержки не исчезают, но меняется источник их возникновения. Выявлено, что в алгоритмической среде остаточные потери уже не могут объясняться только расхождением интересов между принципалом и агентом, поскольку поведение системы определяется параметрами обучения, качеством данных, выбранной целью и возможностью последующей верификации результата. Предложена схема, позволяющая различать источники остаточных потерь и наблюдать, как они формируются на последовательных стадиях, на примере кредитного скоринга. Показано, что положение доверителя в таких условиях меняется, поскольку прошлые действия банка, закрепленные в обучающем массиве, сами становятся источником будущих отклонений.

Выводы. Проведенное исследование показывает, что применение ИИ-агентов требует уточнения агентской теории применительно к алгоритмическому исполнителю, поскольку источник издержек смещается от личного интереса к условиям обучения, составу данных, пригодности ранее обученной системы и возможности объяснить действие программы. Сопоставление согласованной и роевой организации ИИ-агентов позволяет установить, что различие между ними заключается не в составе агентских потерь, а в способе их возникновения, накопления, распределения между цифровыми исполнителями и последующего объяснения. Результаты работы могут быть использованы при дальнейшем исследовании ИИ-агентов в финансовом посредничестве, а также при разработке процедур внедрения алгоритмов в банковской практике.

Ключевые слова: агентская теория, искусственный интеллект, ИИ-агенты, агентские издержки, остаточные потери, финансовое посредничество

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Для цитирования: Зверькова Т. Н. Агентская теория в условиях использования ИИ-агентов в финансовом посредничестве // МИР (Модернизация. Инновации. Развитие). 2026. Т. 17. № 2. С. 432–447

EDN: <https://elibrary.ru/gxtwhj>, <https://doi.org/10.18184/2079-4665.2026.17.3.432-447>

© Зверькова Т. Н., 2026



Контент доступен под лицензией Creative Commons Attribution 4.0 License.
The content is available under Creative Commons Attribution 4.0 License.

Agency theory in the context of using AI agents in financial intermediation

Tatyana N. Zverkova¹

¹ Orenburg State University named after V. A. Bondarenko; Orenburg, Russia

¹ tnzverkova@mail.ru, <https://orcid.org/0000-0002-6540-6154>

Abstract

Purpose: to provide a theoretical and applied justification for the transformation of agency relationships in financial intermediation in the context of the AI agents use.

Methods: the study utilized theoretical, comparative, and structural-analytical approaches, enabling a comparison of the agency theory tenets with the integration of AI in financial intermediation. Classic tenets of the agency theory are consistently correlated with the conditions for the transfer of functions to algorithmic systems.

Results: it was established that with the AI integration, agency costs do not disappear, but their source changes. It was found that in algorithmic environment, residual losses can no longer be explained solely by the divergence of interests between the principal and the agent, as system behaviour is determined by training parameters, data quality, the chosen goal, and the possibility of subsequent verification of the result. A framework is proposed that allows distinguishing the sources of residual losses and observing how they develop at successive stages, using credit scoring as an example. It is also shown that the principal's position in such conditions changes, as the bank's past actions, embedded in the training dataset, themselves become a source of future deviations.

Conclusions and Relevance: the study demonstrates that the use of AI agents requires a refinement of the agency theory as it applies to algorithmic agents, since the source of costs shifts from the personal interest to the training conditions, data composition, suitability of the previously trained system, and the ability to explain the program's actions. A comparison of coordinated and swarm-based AI agent organizations reveals that the difference between them lies not in the composition of agency losses, but in the way they arise, accumulate, distribute among digital agents, and are subsequently explained. The results of this study can be used in further research on AI agents in financial intermediation, as well as in the development of procedures for implementing algorithms in banking practices.

Keywords: agency theory, artificial intelligence, AI agents, agency costs, residual losses, financial intermediation

Conflict of Interest. The author declares that there is no Conflict of Interest.

For citation: Zverkova T. N. Agency theory in the context of using AI agents in financial intermediation. *MIR (Modernization. Innovation. Research)*. 2026; 17(3):432–447. (In Russ.)

EDN: <https://elibrary.ru/gxtwhj>. <https://doi.org/10.18184/2079-4665.2026.17.3.432-447>

© Zverkova T. N., 2026

Введение

Традиционная агентская теория связывала неблагоприятные последствия передачи функций с поведением исполнителя, который в условиях неполной информации преследует собственную выгоду и потому не всегда действует в полном соответствии с интересами принципала. Такая логика долгое время не вызывала сомнений, поскольку в центре анализа находился человек, чьи действия можно было соотносить с личными стимулами, степенью добросовестности и возможностью контроля. Однако в финансовом посредничестве все большая часть операций, связанных с оценкой, отбором, обработкой информации и подготовкой выводов, выполняется обучаемыми алгоритмами. Использование ИИ-агентов требует пересмотра исходного допущения о том, что исполнитель обладает

собственным интересом и реагирует на стимулы. Когда исполнителем выступает программа, не обладающая намерением, собственным интересом и восприимчивостью к стимулам, меняется не только состав участников агентских отношений, но и объяснение причин возникающих потерь.

В банках и платформенных компаниях ИИ-агенты все чаще применяются в тех процессах, где требуется быстрый поиск, сопоставление сведений и подготовка выводов. Их использование позволяет расширять масштаб обработки без сопоставимого увеличения участия человека. В результате роль специалиста постепенно смещается от непосредственного исполнения к постановке задачи, разбору спорных случаев и корректировке параметров. Одновременно появляются трансакционные потери агентского типа, которые нельзя объяснить личной

мотивацией исполнителя или привычной логикой контроля. Усиливается информационная асимметрия между заказчиком и цифровой системой, возрастает значение неопределенности и непрозрачности, а моральный риск и остаточные потери нуждаются в уточнении теоретического обоснования.

Дальнейшее развитие ИИ-агентов связано с переходом от разрозненного применения к совместной работе в распределенной конфигурации. В такой среде ошибка, возникшая в одном звене, уже не остается локальной, поскольку она передается далее, влияет на последующие суждения и, по мере согласования между несколькими цифровыми исполнителями, увеличивает общий объем потерь. Именно поэтому агентские издержки при использовании ИИ-агентов нельзя рассматривать только применительно к одной программе. Они формируются в ходе взаимодействия нескольких цифровых исполнителей, что требует пересмотра представлений о природе агента, происхождении потерь и способах их выявления.

Необходимость объяснить такие потери определяет цель исследования, которая заключается в теоретическом и прикладном обосновании трансформации агентских отношений в финансовом посредничестве при использовании ИИ-агентов.

В соответствии с поставленной целью в статье решаются следующие задачи:

- выявить и систематизировать границы применимости агентской теории при передаче отдельных операций ИИ-агентам;
- показать, как в алгоритмической среде меняется происхождение агентских издержек, если поведение системы определяется параметрами обучения, свойствами исходной информации и заданной целью;
- определить условия, при которых положения агентской теории могут использоваться для анализа агентских отношений с участием ИИ-агентов.

Обзор литературы и исследований

Исследования в области FinTech, ИИ и алгоритмического делегирования в финансовом посредничестве

В научных работах по финансовому посредничеству вопросы использования ИИ обычно рассматриваются в связи с автоматизацией процессов обработки информации, ускорением аналитических расчетов и ростом числа самостоятельных действий, выполняемых алгоритмическими агентами. Ранние исследования финансовых технологий были сосредоточены преимущественно на раскрытии их сущности и общих тенденциях развития

[1, 2]. Более поздние работы сместили внимание к измерению влияния цифровых моделей на посредническую деятельность и изменению банковского сектора под воздействием новых технологий [3, 4].

Группировка исследований по искусственному интеллекту (далее – ИИ), выполненная с применением библиометрического и контентного анализа, показывает, что он рассматривается преимущественно как средство повышения точности прогнозирования, выявления рисков и автоматизации операций [5]. Та же направленность наблюдается и в работах, посвященных его практическому использованию, где основное внимание уделяется качеству прогноза, его зависимости от свойств исходных данных и вопросам получения сопоставимых оценок при изменении принципов исследования [6, 7].

В работах по машинному обучению (далее – МО) описание обычно строится вокруг процедур подготовки исходной информации, построения прогнозов, оценки возможных потерь и последующего ограничения их масштаба [8]. Такая логика фиксирует порядок действий, но не затрагивает вопрос о том, как меняется характер асимметрии при использовании алгоритмов. В статье [9] обосновывается, что отказ от заранее заданных правил в пользу обучения на массивах переносит источник ущерба с поведения исполнителя на сам процесс формирования алгоритмической оценки. Эмпирические наблюдения подтверждают, что асимметрия информации в этом случае не исчезает, а возникает в иной форме, поскольку она связана с невозможностью полностью проследить ход расчетов.

Схожая логика прослеживается и в исследованиях обоснования действий автономных программ. Алгоритмы рассматриваются в них как особый тип агентов, доверие к которым не может строиться только на двустороннем отношении между пользователем и кодом [10]. Высокая сложность расчетов приводит к тому, что даже доступ к внутренним последовательностям действий не устраняет трудности аудита.

Экспериментальные исследования показывают, что ChatGPT способен уменьшать оптимистические искажения аналитиков при прогнозировании показателей [11], а также формировать инвестиционные рекомендации, статистически связанные с будущей доходностью [12]. Такие свидетельства усиливают аргументы в пользу передачи части полномочий алгоритмам. Одновременно в работах [13, 14] обосновывается, что применение непрозрачных систем в задачах с высокой ценой ошибки сужает возможности последующего контроля, а объяснение уже полученного результата не позволяет установить источник отклонения.

Отдельным направлением является литература по аудиту алгоритмов. В таких работах подчеркивается, что надежный контроль над ИИ требует сочетания технических и институциональных мер, заимствованных из методик финансового аудита и инженерного тестирования [15, 16].

Таким образом, существующие исследования фиксируют устойчивый разрыв между изучением последствий цифровизации кредитных организаций и теоретическим объяснением агентских отношений с участием алгоритмов. Финансово-экономическая литература подробно рассматривает влияние FinTech на посредничество, риск, сетевую связанность и состав финансовых услуг [17–19]. В то же время, работы, посвященные МО и использованию кодов для оценки заявок, сопоставления признаков и классификации объектов анализа, сосредоточены главным образом на доверии, качестве данных, смещениях и возможностях последующей верификации. При этом такие затруднения обычно рассматриваются как технические или организационные и не соотносятся с категориями агентской теории. В результате, изменения в отношениях между финансовым посредником, клиентом и цифровым исполнителем рассматриваются отдельно от допуска алгоритмов к оценке клиентов, сопоставлению данных и подготовке заключений, хотя на практике эти процессы уже давно переплелись.

Между тем, использование ИИ принципиально меняет условия делегирования прав принятия решений программам. Поведение алгоритма определяется не личным выбором исполнителя, а параметрами обучения, свойствами данных, целью, способом настройки и возможностью последующей верификации результата. Это означает, что источник отклонения уже нельзя объяснять только расхождением интересов между доверителем и агентом. При встраивании ИИ-агентов в банковские продукты и услуги предмет анализа смещается от мотивов исполнителя к критерию обучения, использованным данным и настройке программы.

Поэтому сохраняется задача соединить два направления анализа в едином теоретическом подходе. С одной стороны, требуется учесть последствия цифровизации финансового посредничества. С другой стороны, необходимо показать, каким образом положения агентской теории могут использоваться для анализа ситуаций, где отрицательный эффект возникает не из поведения исполнителя, а из обучающей базы, логики настройки и ограниченной возможности последующей проверки. Именно эта задача определяет предмет настоящего исследования.

Теоретические постулаты агентской теории при использовании ИИ-агентов

В трактовке M. Jensen и W. Meckling [20] агентские отношения строятся на несовпадении интересов принципала и исполнителя, что порождает затраты на наблюдение, принятие обязательств и остаточные потери. Когда исполнителем становится алгоритм, механизм расхождения интересов теряет объяснительную силу, поскольку у программы отсутствует собственная выгода. Однако из этого не следует исчезновение транзакционных потерь агентского типа. Передача функции обучаемой программе не устраняет отрицательный итог, а меняет причину его появления. Остаточный ущерб сохраняется, но возникает уже не из конфликта интересов, а из несовпадения между целью кода и тем результатом, которого ожидает принципал.

Именно здесь проявляется проблема алгоритмического делегирования. Система стремится улучшить показатель, заданный при обучении, например, точность прогноза дефолта или долю выявленных подозрительных операций. Однако он не всегда совпадает с тем, ради чего банк использует алгоритм, поскольку требуется не только высокий процент верных расчетов, но и обоснованная оценка риска, корректный учет признаков клиента и возможность объяснить автоматизированный вывод. Поэтому программа может успешно пройти проверку на тестовой выборке и при этом дать нежелательный для банка результат не из-за сбоя, а вследствие строгого следования заданному показателю.

Схожее смещение наблюдается и в понимании контроля, поскольку проверка уже не может сводиться к наблюдению за исполнителем и сопоставлению его поведения с интересом доверителя. В агентской схеме он строился вокруг поведения человека, которому передавалась часть полномочий, и чьи действия можно было соотнести с личной выгодой и условиями договора. При допуске программы к самостоятельной оценке такой подход теряет прежний смысл, поскольку наблюдение за ходом вычисления само по себе не раскрывает происхождение автоматизированного вывода. Поэтому проверка переносится на постановку цели, состав обучающей выборки, обоснование параметров до запуска системы и устойчивость ее работы при последующем применении. Меняется и содержание обязательств, которые уже нельзя свести к заверениям исполнителя о добросовестности. Они приобретают форму требований к ведению записей, сохранению последовательности изменений, фиксации исправлений и объяснению спорного действия.

Дальнейшее развитие этой логики позволяет уточнить, почему положения B. Holmstrom [21], J.

Stiglitz [22], О. Hart [23], О. Williamson [24] и Н. Simon [25] сохраняют объяснительную ценность, но требуют адаптации к алгоритмической среде.

У В. Holmstrom [21] агентская проблема раскрывается через невозможность полного наблюдения за действиями агента. Принципалу доступен только итог, тогда как усилия агента остаются вне видимости, вследствие чего отрицательные последствия не могут быть полностью устранены даже при добросовестном поведении. В алгоритмической среде это получает иную форму, поскольку программа не скрывает сведения намеренно, но часто не представляет ход расчетов в виде, пригодном для подтверждения. Возникает ситуация, в которой доверитель располагает полной технической фиксацией работы алгоритма, но все же не может восстановить последовательность принятия решения. Тем самым появляется новый тип информационных расходов, не сводимый к обычному рассогласованию сведений.

Ж. Stiglitz связывает информационную асимметрию с неравным доступом к знанию, когда агент располагает информацией, недоступной принципалу, и может использовать это преимущество в собственных интересах [22]. В случае ИИ-агентов такая интерпретация требует уточнения, поскольку затруднение возникает не из намеренного сокрытия сведений, а из недоступности внутренней логики для содержательной проверки. Проблема состоит в том, что основание полученного итога не представлено в явной форме и потому не может быть полностью восстановлено ни заказчиком, ни разработчиком алгоритма.

Концепция неполных контрактов О. Hart [23] объясняет, почему соглашение не может заранее охватить все состояния, возникающие в ходе исполнения, и потому оставляет агенту пространство для усмотрения. При передаче части полномочий алгоритмам к этому добавляется иной по природе источник неопределенности, связанный с поведением уже обученной системы за пределами знакомой ей выборки. Такое затруднение проявляется не в момент заключения соглашения, а после ввода кода в работу, когда его действия начинают влиять на последующие операции. Повторное согласование не устраняет возникшую проблему, поскольку к моменту ее выявления алгоритм обычно уже успевает произвести множество итераций с нарастающими последствиями. Поэтому распределения остаточных правомочий, предложенного О. Hart, для таких случаев недостаточно. Наряду с этим, требуется заранее закрепить технические обязанности по постоянному наблюдению за работой программы, подтверждению ее действий и своевременному внесению исправлений.

Институциональный подход О. Williamson [24] позволяет рассматривать отношения не через личные мотивы участников, а через устойчивые правила, процедуры согласования и способы ограничения оппортунизма при длительном взаимодействии. Поэтому его положения имеют значение там, где часть действий передается ИИ. В такой ситуации контроль через санкции уже не дает нужного результата, поскольку наказание обращено к субъекту, способному соотнести последствия со своим поведением и изменить его. Алгоритм же не обладает ни собственным интересом, ни способностью воспринимать санкцию как побуждение к исправлению. Поэтому верификация должна строиться не вокруг последующего взыскания, а на основе заранее установленных ограничений к созданию и использованию системы. Эти условия охватывают состав и качество исходной информации, а также тестирование до ввода кода в работу и после его внедрения [26]. Кроме того, необходимы обязательная фиксация действий программы и всех ее изменений, а также точное разграничение обязанностей между разработчиком, владельцем данных и тем, кто обеспечивает работу вычислительной среды.

Н. Simon [25] рассматривал ограниченную рациональность как свойство агента, действующего при неполной информации и недостаточных познавательных возможностях. Человек не способен охватить весь доступный объем сведений и выбрать наилучший вариант из всех возможных, поэтому обычно останавливается на выборе, который признается приемлемым при существующих ограничениях. Переносить это понятие на алгоритм некорректно, поскольку у него нет рациональности в том смысле, который вкладывал Н. Simon. Точнее говорить о заранее заданных пределах, при которых автоматизированная система не ищет лучший вариант как мыслящий субъект, а работает в границах, заложенных при обучении и настройке. Полученный итог определяется качеством данных, способом обучения, выбранным критерием ошибки и доступным объемом вычислений. Поэтому отклонение от наилучшего варианта возникает из-за исходных технических и математических допущений. По этой причине такие аномалии нельзя устранить ни волевым усилием, ни правильно выстроенными стимулами, поскольку их причина находится не в мотивации, а в самом способе построения кода.

Проведенное сопоставление показывает, что привычное понимание агента описывает лишь сам факт делегирования задачи и наличие отрицательных последствий, возникающих в ходе ее выполнения. Этого уже недостаточно, чтобы объяснить их происхождение в цифровой системе. При использовании ИИ ущерб связан не с интересами исполнителя, не

с его скрытым поведением и не с его усмотрением, а с постановкой цели, качеством данных, обучением, возможностью обоснования правильности вывода и последующим использованием результата [27]. Поэтому агентская проблема здесь сохраняется, но ее источник смещается из мотивации исполнителя в область написания и применения кода.

Материалы и методы

Исследование имеет теоретико-прикладной характер и направлено на объяснение агентских издержек при передаче отдельных операций обучаемым цифровым системам. Для решения поставленных задач использовано сочетание теоретического, сравнительного и прикладного анализа. Такой подход позволяет соотнести положения агентской теории с применением ИИ в практике кредитных организаций и показать, как меняется происхождение отрицательных последствий, когда исполнение переходит от работника к ИИ-агенту.

Теоретическая часть построена на последовательном сопоставлении идей М. Jensen, W. Meckling [20], В. Holmstrom [21], J. Stiglitz [22], О. Hart [23], О. Williamson [24] и Н. Simon [25] с ситуацией, в которой часть работы выполняет обучаемая цифровая система. Это позволяет установить, какие положения агентской теории сохраняют объяснительную силу, а какие уже не охватывают изменившийся характер отношений и потому нуждаются в пересмотре. Сравнительный анализ используется для выявления различий в происхождении трансакционных потерь агентского типа, способе наблюдения, содержании обязательств, положении принципала и природе остаточного неблагоприятного эффекта при таком изменении роли исполнителя.

В прикладной части используется аналитическое выделение внутренних факторов остаточных потерь. Совокупный ущерб раскладывается на причины, связанные с ограниченной пригодностью ранее обученной модели, исторической смещенностью данных, расхождением между целью обучения и результатом, а также с невозможностью объяснить полученный вывод. Метод логического вывода применяется для установления причинной связи между устройством модели, базой данных, выбранной целью обучения, наблюдаемыми последствиями и перераспределением контрольных функций внутри банка.

Результаты исследования

Изменение содержания агентских издержек и декомпозиция остаточных потерь

Классическая версия агентской теории в интерпретации М. Jensen и W. Meckling [20] исходит из того, что совокупные агентские издержки склады-

ваются из расходов на контроль, принятие обязательств и остаточных потерь:

$$C = M + B + R, \quad (1)$$

где M – затраты принципала на контроль действий агента; B – затраты агента на принятие обязательств перед принципалом; R – остаточные потери, сохраняющиеся после согласования условий между сторонами.

Формула (1) имеет две предпосылки:

- агент обладает интересами, не совпадающими со стимулами принципала, поэтому после передачи ему права действовать возникает вероятность отклонения результата от поставленной цели;
- агент в пределах переданных ему полномочий самостоятельно определяет последовательность действий и тем самым влияет на итог порученной ему операции.

Применительно к алгоритму эти положения уже не действуют в прежнем виде. У цифрового исполнителя нет личной выгоды, поэтому он не может рассматриваться как носитель интереса, противопоставленного принципалу. Передача ему отдельных полномочий по обработке, отбору и подготовке решений не устраняет потери, поскольку отрицательные последствия продолжают возникать, но содержание слагаемых меняется. По этой причине для алгоритмической среды целесообразно не отказываться от классической формулы полностью, а сохранить ее общую логику. Акцент при этом переносится с мотивации исполнителя на свойства цифровой системы, порядок ее применения и причины возникающих отклонений.

Совокупные агентские издержки при таком участии цифрового исполнителя могут быть представлены следующим образом:

$$C_{AI} = M_{AI} + B_{AI} + R_{AI}, \quad (2)$$

где M_{AI} – расходы на контроль алгоритма; B_{AI} – расходы на техническое подтверждение надежности и работоспособности алгоритмической системы; R_{AI} – остаточные потери, сохраняющиеся после тестирования и введения обязательных правил использования.

После передачи части полномочий программе контроль, обязательства и остаточные потери сохраняются, однако их содержание меняется. Контроль в данном случае относится уже не к наблюдению за действиями работника в ходе исполнения поручения, а к аудиту качества исходной информации, корректности поставленной цели, устойчивости работы и допустимости использования получаемых выводов в конкретной операции.

Обязательства также получают другой смысл, поскольку речь идет уже не о личной добросовестности агента, а о требованиях к фиксации изменений, происхождению данных и возможности проверить работу системы до ее применения.

Остаточные потери больше не объясняются расхождением интересов между принципалом и исполнителем, поскольку их первопричина переносится в свойства самой программы, обучающую выборку и ограничения последующей обоснованности вывода. Тем самым формула сохраняет связь с подходом M. Jensen и W. Meckling [20], но переносит объяснение потерь с поведения агента на устройство программы и порядок ее применения.

M_{AI} отражает расходы, связанные не с наблюдением за действиями исполнителя, а с аудитом того, может ли программа использоваться в конкретной процедуре. На практике к ним относятся ревизия целевой функции и состава обучающей выборки, испытание системы на нетипичных массивах, а также наблюдение за ее устойчивостью после ввода в эксплуатацию. Тем самым, внимание переносится с поведения агента на этапы создания, настройки и использования цифрового контура.

Расходы B_{AI} также получают иное содержание. Речь идет уже не о заверениях агента и не о его деловой репутации, а о закрепленных технических обязанностях, включающих сохранение версий, регистрацию изменений, раскрытие порядка обучения, возможность доказать полученный вывод и возврат к ранее подтвержденной версии кода.

Наибольшее изменение относится к R_{AI} , поскольку именно здесь проявляется источник потерь, который уже нельзя свести к мотивации исполнителя. Для раскрытия содержания остаточных потерь при передаче отдельных полномочий алгоритму предлагается следующая формула¹:

$$R_{AI} = C_{arch} + C_{data} + C_{obj} + C_{опас}, \quad (3)$$

где C_{arch} – потери, возникающие из-за неполного охвата будущих состояний; C_{data} – потери, порождаемые искажениями и историческим смещением обучающей выборки; C_{obj} – расхождение между критерием работы алгоритма и целью принципала; $C_{опас}$ – потери, связанные с невозможностью объяснить конкретное действие даже при наличии полной технической документации.

Потери типа C_{arch} связаны с тем, что поведение уже обученной модели нельзя надежно предсказать в

ситуациях, которых не было в исходном массиве. Для банков это означает возможность регулярных сбоев после изменения обстоятельств, появления новых категорий заемщиков или структуры операций. Потери в таком случае появляются не в момент исполнения договора, а раньше, когда формируются обучающая выборка, параметры настройки и пределы применения программы.

C_{arch} не являются следствием ошибки разработчика или недобросовестности поставщика. Речь идет об ограничении, встроенном в сам способ обучения модели, поскольку любая система обобщает закономерности прошлого и переносит их на будущие ситуации, которые могут уже не совпадать с прежним опытом. Из этого следует, что уменьшение C_{arch} требует не только более тщательного обучения, но всего установления границ применения и постоянного наблюдения за ее устойчивостью при практическом использовании.

Смещение данных C_{data} отражает ситуацию, в которой прошлые действия и последствия включаются в основу обучения, а затем возвращаются в виде устойчивого набора новых отказов, оценок и отборов. В этом случае принципал перестает быть только стороной, которая несет потери от работы цифрового исполнителя, поскольку он сам становится первопричиной тех данных, которые определяют последующее поведение программы. Поэтому работа с этим источником требует не только технической очистки массива, но и выявления тех предпосылок будущего ущерба, которые были созданы самим банком.

Расхождение целевой функции C_{obj} показывает, что алгоритм может улучшать заданный ему показатель и одновременно ухудшать итог, которого в действительности добивается кредитная организация. Скоринговая модель способна сокращать долю невозвратов путем чрезмерного отсека платеспособных заемщиков, а антифрод может повышать долю выявленных подозрительных операций ценой блокировки добросовестных клиентов. В обоих случаях система работает корректно в пределах поставленной перед ней цели, однако сама задача оказывается неточно согласованной с интересом принципала. По этой причине C_{obj} отличается от остальных показателей тем, что его появление в наибольшей степени связано с действиями самого принципала. Именно он определяет, какой показатель считать приоритетным и какое соотношение риска, доходности и отказов

¹ Прим. Автора: Данная формула используется для разграничения различных источников потерь и не предполагает их точного арифметического суммирования. В реальности такие источники могут пересекаться и усиливать друг друга. Например, непрозрачность затрудняет своевременное выявление выхода модели за пределы прежних правил обучения, поэтому при эмпирической проверке потребовалось бы учитывать их совместное действие. В настоящей работе формула применяется как средство разграничения причин ущерба, а не как способ его количественного расчета.

признавать допустимым. Следовательно, потери в данном случае возникают не из-за сбоя кода, а по причине неточной постановки цели, на которую он был ориентирован.

Непрозрачность вывода $C_{опас}$ описывает разрыв между технической фиксацией работы системы и содержательным объяснением основания индивидуального кредитного заключения. Банк может располагать журналами операций, входными признаками и промежуточными вычислениями, но при этом не имеет возможности объяснить причину отказа по отдельной заявке. Поэтому непрозрачность следует рассматривать как самостоятельный источник остаточных потерь, а не как побочное свойство сложного кода.

Принципиальное значение имеет то, что каждая причина потерь имеет индивидуальное происхождение, поэтому устранение одной из них не ведет к исчезновению остальных. Это положение требует отдельного обоснования, поскольку перечисленные основания действительно могут восприниматься как связанные между собой.

Различие между C_{arch} и C_{data} состоит в том, что первый показатель связан с ограниченной пригодностью уже обученной модели при изменении внешних обстоятельств, тогда как второй определяется свойствами данных, на которых она была обучена. Модель может быть обучена на репрезентативных и несмещенных данных, и при этом давать значительные отклонения после изменения состава заемщиков или структуры заявок. И наоборот, исторически смещенный массив способен воспроизводить прежние искажения даже при относительно устойчивой внешней среде.

Различие между C_{data} и C_{obj} заключается в том, что исправление записей не устраняет расхождение между показателем, который улучшает алгоритм, и результатом, которого добивается банк. Скоринг может быть обучен на репрезентативном массиве, но по-прежнему повышать точность прогноза дефолта вместо улучшения доходности кредитного портфеля. В таком случае первоисточник потерь находится не в параметрах, а в способе постановки задачи.

Различие между C_{obj} и $C_{опас}$ состоит в том, что правильное задание цели обучения само по себе не делает итоговый вывод релевантным. Модель может быть ориентирована непосредственно на рост прибыли банка и, вместе с тем, оставаться такой, что конкретный отказ невозможно объяснить клиенту, службе внутреннего контроля или надзорному органу. Следовательно, корректность цели и объяснимость автоматизированного суждения относятся к разным источникам остаточных потерь.

Особого рассмотрения требует соотношение C_{arch} и $C_{опас'}$ поскольку при изменении внешних обстоятельств непрозрачность алгоритма затрудняет своевременное установление причин ухудшения качества работы. Если расчетная схема позволяет изучить основания полученной оценки, то банк способен раньше выявить, связано ли ухудшение с выходом за пределы прежних правил обучения. Когда основание такого вывода остается неясным, обнаружение причины отклонений запаздывает, а потери успевают накопиться в большем объеме. Из этого следует, что данные следствия не только различаются по происхождению, но и способны усиливать действие друг друга.

Верификация предложенной схемы на примере кредитного скоринга

Прикладная ценность предложенной схемы определяется тем, насколько точно она позволяет установить момент формирования будущего ущерба и раскрыть причины его появления. Поэтому далее рассматривается не набор разрозненных операций, а единый цикл, в котором прослеживается движение от постановки цели до проверки и последующей корректировки. Такой способ изложения позволяет показать, как разные виды агентских потерь возникают на отдельных стадиях применения ИИ-агентов и затем соединяются в совокупный ущерб.

Наиболее подходящим примером для такого анализа выступает кредитный скоринг, поскольку в нем одновременно проявляются выбор цели, свойства обучающего массива, пределы пригодности ранее обученной модели и трудность пояснения конкретного отказа.

В табл. 1 скоринг представлен как последовательность этапов, на каждом из которых может возникнуть самостоятельная причина будущих потерь. Такая логика показывает, что отрицательные последствия складываются постепенно, а не появляются одномоментно в одной точке. Уже при постановке цели может быть выбран показатель, удобный для обучения, но не совпадающий с тем эффектом, которого банк добивается в кредитовании. При подготовке обучающего массива в него могут попасть прежние правила отказа, накопленные искажения отбора и ограничения, сложившиеся в прошлой практике. Во время предварительного тестирования могут остаться незамеченными пределы пригодности скоринговой модели в новых обстоятельствах, а также невозможность содержательно объяснить отдельный отказ. После начала применения эти причины переходят в упущенный доход, ухудшение качества кредитного портфеля, рост числа спорных отказов и претензии.

Представленная схема показывает, что в скоринге будущие остаточные потери возникают не в одной

Таблица 1

Кредитный скоринг как пример возникновения агентских издержек в условиях передачи части решений алгоритмической системе

Table 1

Credit scoring as an example of the formation of agency costs in the context of transferring some decisions to an algorithmic system

Этап	Содержание	Преобладающий вид издержек (источник потерь)	Накопление потерь и признаки выявления
1	2	3	4
Постановка цели	Определяется назначение скоринга и какой итог считается приемлемым	$M_{AI}, B_{AI}, \text{далее возможно } R_{AI} (C_{obj})$	Формирование будущего ущерба начинается уже на стадии выбора целевого показателя, если программа ориентируется на точность прогноза, а не на итог. Такое расхождение проявляется в сочетании точности прогноза дефолта с низкой доходностью выдачи, снижением доли одобренных платежеспособных заемщиков и сокращением прибыльных клиентов
Подготовка обучающего массива	Загружаются исторические данные о выданных кредитах, отказах, просрочках и характеристиках заемщиков	$M_{AI}, B_{AI}, \text{далее возможно } R_{AI} (C_{data})$	Закрепление будущих потерь происходит через структуру обучающего массива, где сохраняются прежние правила отказа, исторические искажения отбора и отсутствуют сведения о клиентах, не получивших кредит ранее. Это выявляется по воспроизведению старой структуры отказов, ухудшению работы с новыми группами заемщиков и снижению устойчивости при изменении состава заявок
Обучение и настройка модели	Выявляются зависимости между признаками заемщика и ожидаемым результатом	$B_{AI}, \text{далее возможно } R_{AI} (C_{data}, C_{obj})$	На этой стадии закрепляются статистические связи, которые выглядят убедительно на проверочной выборке, но могут сохранить историческое смещение или узость выборочной цели. О проблеме говорит сочетание высоких проверочных показателей с последующим расхождением между скоринговым баллом и качеством новых выдач
Испытание перед вводом в работу	Оценивается, можно ли использовать программу	$M_{AI}, B_{AI} (C_{data}, C_{obj}, C_{grac} \text{ могут быть не выявлены полностью})$	Скрытое накопление потерь продолжается при слабой предварительной экспертизе, когда не выявлены пределы применения системы, расхождение между целью и экономической отдачей, а также трудность объяснения отдельного отказа. Это обнаруживается через отсутствие стрессовых испытаний, ограниченность тестирования на новых сегментах, неясные границы применимости и слабую объяснимость отказов
Рассмотрение конкретной заявки	Рассчитывается скоринговый балл, на основании которого заявка одобряется либо отклоняется	$R_{AI} (C_{data}, C_{obj}, C_{grac})$	На этой стадии скрытые отклонения переходят в наблюдаемую форму, поскольку ошибка в оценке уже влияет либо на упущенный доход, либо на последующее ухудшение качества кредитного портфеля. Это проявляется в росте числа спорных отказов, снижении доли надежных клиентов среди одобренных и появлении необъяснимых различий между схожими заявками
Работа после изменения внешних условий	Экономическая ситуация, платежное поведение клиентов или структура заявок меняются по сравнению с прошлым периодом	$M_{AI}, \text{далее } R_{AI} (C_{arch})$	После изменения внешних условий потери связаны с тем, что программа должна опираться на прежние зависимости, тогда как структура уже изменилась. Это фиксируется по снижению прогнозной точности на новых выдачах, росту ошибок первого и второго рода и ухудшению качества кредитного портфеля
Проверка отказа клиентом, службой внутреннего контроля или надзорным органом	Поясняется, почему по конкретной заявке был вынесен отказ	$R_{AI}, \text{а также дополнительные } M_{AI} \text{ и } B_{AI} (C_{grac})$	Потери начинают складываться уже не только в кредитном портфеле, но и в виде споров, дополнительных сверок и внутренней нагрузки, если основание заключения нельзя ясно раскрыть. На это указывают рост числа обращений клиентов, трудности внутреннего суждения, невозможность подготовить содержательный ответ и увеличение времени на спорные случаи

Окончание таблицы 1
End of Table 1

Этап	Содержание	Преобладающий вид издержек (источник потерь)	Накопление потерь и признаки выявления
1 Мониторинг послед- ствий и корректировка	2 Отслеживается качество применения скоринга и определяет участок, требующий исправления	3 M_{AI}, B_{AI} (зависит от того, какой источник потерь выявлен)	4 Сохранение отрицательной динамики после корректировки указывает на неверное опре- деление источника потерь, когда изменения вносятся не в тот участок, где заложена при- чина отклонения. Такая ситуация проявляется в повторении прежних ошибок после перена- стройки и отсутствии улучшения качества заключения, несмотря на внесенные изменения

Разработано автором
Developed by the author

точке, а последовательно, по мере перехода от выбора цели к подготовке данных, диагностике, принятию решения и последую- щему наблюдению за его итогами. Поэтому одно и то же неблаго- приятное заключение может иметь различное происхождение, а его объяснение требует разграничения источников потерь внутри самой процедуры кредитования.

Роевая организация ИИ-агентов в кредитном скоринге

В предшествующем изложении алгоритмическое распределение задач рассматривалось как последовательный процесс, в кото- ром заключение скоринга проходит проверку, согласование и последующую корректировку. Далее необходимо сопоставить этот порядок с конфигурацией, в которой итог складывается из ло- кальных взаимодействий множества ИИ-агентов без единого со- гласующего узла (рой). В ней один и тот же источник остаточных потерь уже не проходит через заранее установленную последова- тельность стадий, а распределяется между параллельно действу- ющими участками обработки кредитной заявки. По этой причине переход к рою ИИ-агентов следует рассматривать не как простое усложнение вычислений, а как изменение последовательности формирования результата.

Применительно к скорингу рой можно описать как совокупность ИИ-агентов, каждый из которых оценивает заявку на основе на- бора данных, признаков или критерия. Ни один из агентов не по- лучает исключительного права на окончательный вывод. Вместо этого они обмениваются частичными сигналами, уточняют свои оценки под влиянием соседних агентов и постепенно приходят к распределенному согласию, из которого возникает одобрение либо отказ. В такой форме организации итоговое суждение не задается в одной точке, а складывается как общий результат множества локальных взаимодействий. Для целей настоящего ис- следования существенно не то, следует ли считать такую модель заведомо лучшей по сравнению с многоагентной схемой с со- гласованием, а то, как именно в ней ведут себя источники оста- точных потерь (табл. 2).

Сопоставление формы с согласованием и роевой самооргани- зацией показывает, что переход к распределенному взаимодей- ствию не устраняет остаточные потери, а меняет последователь- ность их возникновения и дальнейшего выявления. В системе с согласованием отрицательное последствие легче связать с опре- деленной стадией операции. В рое тот же итог оказывается более распределенным, а его происхождение труднее отнести к одному шагу или одному участнику.

Практический смысл такого сопоставления состоит в том, что банк получает возможность выбирать не только между двумя варианта- ми обработки заявки, но и между различными способами локали- зации и объяснения будущих потерь. Для скоринга из этого вытека- ет несколько последствий:

- потери, связанные с данными, возникают там, где множество агентов работает с неодинаковыми фрагментами обучающего массива и по-разному воспроизводит ранее сложившиеся иска- жения отбора;
- ущерб, обусловленный ограниченной пригодностью обученной системы, проявляется в тех случаях, когда часть агентов про- должает опираться на прошлые зависимости после изменения

Таблица 2

Роевая организация ИИ-агентов в кредитном скоринге и перераспределение остаточных потерь

Table 2

Swarm organization of AI agents in credit scoring and redistribution of residual losses

Свойство	Проявление	Ослабление источников потерь	Возникающий риск
Отсутствие единого согласующего агента	Одобрение или отказ формируются не в одной точке, а через совокупность частичных проверок	$C_{\text{итог}}$ ослабление возможно, поскольку сбой одного агента не превращается автоматически в обязательное правило для остальных	Утрата общего критерия, при котором совокупность локальных реакций не приводит к наилучшему для банка результату
Локальные взаимодействия между агентами	Каждый агент передает соседним агентам не готовый итог, а частичный сигнал, степень уверенности или скорректированную оценку	$C_{\text{итог}}$ снижение возможно, если объяснение отказа собирается из локально понятных оснований, зафиксированных на отдельных участках проверки	Расхождение между группами агентов, когда разные части роя дают несовместимые основания для дальнейшего отбора
Распределение данных между агентами	Агенты работают с разными наборами заемщиков, наборами признаков или временными срезами	$C_{\text{итог}}$ снижение вероятно в тех случаях, когда историческое искажение воспроизводится разными агентами не одинаково	Если смещение присутствует во всем массиве равномерно, рой не устраняет его, а лишь переносит в распределенную форму
Различие критериев оценки	Одни агенты ориентируются на вероятность дефолта, другие на доходность сделки, третьи на устойчивость пояснения	$C_{\text{итог}}$ сдерживание достижимо, поскольку наличие конкурирующих критериев не позволяет одному показателю полностью заменить цель банка	Ошибочное равновесие, при котором достигнутое согласие роя не совпадает с целью банка
Самоорганизация итогового вывода	Итог возникает по мере накопления сигналов, а не утверждается отдельным координатором	Все 4 источника могут частично усиливать или компенсировать друг друга, если рой не позволяет одному дефектному сигналу определить позицию всей системы	Трудности определения происхождения отклонения, поскольку становится невозможно точно установить, какой агент и на каком шаге превратил локальное отклонение в окончательный отказ
Отсутствие обязательного человеческого подтверждения	Сотрудник банка не участвует в каждом определении, а наблюдает лишь за общим поведением роя или вмешивается позже	В отдельных случаях сокращается задержка при рассмотрении заявки и быстрее обнаруживаются массовые отклонения в статистике выдач	Ослабление оперативного вмешательства в отдельный ошибочный отказ, поскольку корректировка переносится с уровня заявки на уровень последующего наблюдения

Разработано автором
Developed by the author

внешних обстоятельств, состава заемщиков или структуры заявок;

- расхождение между формальным показателем и целью банка особенно наглядно проявляется в тех случаях, когда согласованное действие нескольких агентов строится вокруг критерия, не совпадающего с интересом кредитора;
- непрозрачность также получает особое выражение, поскольку итоговый отказ возникает не из одного вывода, а из совокупности сигналов, вследствие чего его основание труднее соотнести с одним определенным источником.

Сравнение двух способов наделения программы полномочиями позволяет рассматривать не просто разные варианты обработки заявки, а различные вариации возникновения и распределения потерь. В тех случаях, когда для банка на первом месте стоит возможность последовательно установить место возникновения отклонения и дать понятное объяснение по конкретной заявке, более приемлемой остается согласованная процедура. Когда приоритет смещается к распределению риска ошибки между множеством агентов, интерес приобретает роевая самоорганизация.

Новизна данного сопоставления состоит не в описании роя как более сложной формы ИИ, а в применении единой схемы остаточных потерь к двум способам организации работы ИИ-агентов. Это дает возможность показать, что переход от согласования к самоорганизации меняет не набор источников потерь, а последовательность их перераспределения, накопления и частичного устранения.

Обсуждение

Сравнение схемы с согласованием и роевой самоорганизации позволяет точнее показать пределы предложенного подхода и оценить, как он объясняет распределение остаточных потерь. В первом случае потери возникают внутри заранее заданной последовательности процедур, поэтому их легче связать с конкретным участком процесса. Во втором случае последствие распределяется между параллельно действующими агентами и уже не сводится к отдельному выводу, разовой проверке или одному источнику ошибки. Следовательно, сопоставляемые варианты различаются не перечнем потерь, а местом их возникновения, способом передачи между участниками, скоростью последующего обнаружения и корректировки.

Переход от иерархического согласования к роевой самоорганизации не устраняет расхождение между целью банка и действиями программ, которым передается часть обработки. Потери возникают уже не у одного исполнителя, а в связях между ИИ-агентами, когда один алгоритм использует сведения и промежуточные расчеты другого. В согласованной схеме

каждое суждение проходит через установленную последовательность проверки, поэтому ошибку легче соотнести с определенным действием или массой сведений. В итоге складывается через обмен сигналами между несколькими агентами, каждый из которых работает со своей частью данных и собственным правилом оценки. Поэтому последствие нельзя уверенно отнести к одной программе или одному звену обработки, поскольку оно возникает через передачу сигналов, взаимную корректировку и последующее согласование. Такой порядок усложняет изучение потерь, так как требуется учитывать не только качество отдельного алгоритма, но и переход ошибки между агентами, ее усиление при обмене и превращение локального сбоя в общий итог.

Выбор между системой согласования и роевой самоорганизацией нельзя сводить только к скорости вычислений или удобству настройки программы. Он зависит от того, какой источник остаточных потерь банк считает для себя наиболее чувствительным. Когда на первый план выходят возможность объяснить отказ, установить причину заключения и точно распределить обязанности внутри операции, предпочтение чаще получает схема согласования. Если больший вес придается распределению ошибки между многими агентами и способности быстрее приспосабливаться к меняющимся обстоятельствам, интерес может представлять и роевая форма. В обоих случаях предложенный подход позволяет рассматривать такой выбор как сопоставление разных способов возникновения, накопления и частичного снижения остаточных потерь.

Выводы

Проведенное исследование показывает, что перенос принятия решений в алгоритмическую среду не устраняет агентскую проблему, а меняет причину ее возникновения. Когда суждение формируется ИИ-агентами, остаточный вывод уже нельзя объяснить конфликтом интересов между доверителем и агентом. У цифрового исполнителя отсутствуют выгода, воля и восприимчивость к стимулам. В этих ситуациях агентские потери сохраняются, однако их происхождение смещается в сферу постановки цели обучения, отбора и исторических искажений данных, ограниченной пригодности алгоритма к новым правилам и непрозрачности получаемых выводов.

Отдельного внимания заслуживает изменение положения принципала и обоснованности работы ИИ-агента. В предложенной схеме банк уже нельзя рассматривать только как сторону, которая несет последствия отказов, одобрений и изменения условий. Его прежняя практика, закрепленная в обучающей базе, влияет на последующую обработку клиентских запросов и тем самым участвует

в формировании остаточных потерь. Непрозрачность перестает быть только техническим свойством вычислительного процесса, так как невозможность обосновать отказ или одобрение заявки перед клиентом, внутренней проверкой и надзорным органом сама становится самостоятельным источником таких потерь.

Научная новизна исследования состоит в применении агентской теории к ИИ-агентам не как вспомогательным программам, а как исполнителям, участие которых меняет причины возникновения агентских издержек. Показано, что при использовании ИИ последствия связаны не с оппортунистическим поведением агента, а с условиями обучения, составом исходных данных, пределами пригодности ранее настроенной системы и трудностью содержательного объяснения конкретного действия программы. На этой основе уточнено содержание остаточных потерь и предложено их разделение на 4 группы: неполный охват будущих состояний обученной системой,

искажения и историческое смещение обучающего массива, расхождение между критерием работы алгоритма и целью принципа, а также невозможность объяснить конкретное действие даже при наличии полной технической документации. Отдельно рассмотрена роевая организация ИИ-агентов, при которой неблагоприятные последствия не исчезают, а распределяются между несколькими программами, действующими без единого согласующего узла. Такой подход позволяет перейти от объяснения агентских издержек через конфликт интересов доверителя и исполнителя к изучению их возникновения при алгоритмическом исполнении, включая коллективные формы работы ИИ-агентов.

Предложенный подход имеет значение для разработчиков алгоритмов и нуждается в дальнейшем эмпирическом исследовании. Перспективным направлением остается анализ трансформации агентских издержек в гибридных формах совместного принятия решений человеком и ИИ-агентами.

Список источников

1. Котляров И.Д. Финтех: сущность и модели реализации // ЭКО. 2018. № 12(534). С. 23–39. EDN: <https://elibrary.ru/yoqgjn>
2. Артеменко Д.А., Зенченко С.В. Цифровые технологии в финансовой сфере: эволюция и основные тренды развития в России и за рубежом // Финансы: теория и практика. 2021. Т. 25. № 3. С. 90–101. EDN: <https://elibrary.ru/rixjhx>. <https://doi.org/10.26794/2587-5671-2021-25-3-90-101>
3. Криничанский К.В., Зеленева Е.С. Финтех-сектор в контексте финансового развития и проблем его измерения // Финансы: теория и практика. 2024. Т. 28. № 5. С. 121–132. EDN: <https://elibrary.ru/yqhgoh>. <https://doi.org/10.26794/2587-5671-2024-28-5-121-132>
4. Тавбулатова З.К., Батукаева Л.С.Э. Финтех-трансформация банковского сектора Российской Федерации в условиях технологического суверенитета // Финансы и кредит. 2025. Т. 31. № 10. С. 46–65. EDN: <https://elibrary.ru/eztyen>. <https://doi.org/10.24891/eztyen>
5. Bahoo S., Cucculelli M., Goga X., Mondolo J. Artificial intelligence in finance: A comprehensive review through bibliometric and content analysis // SN Business & Economics. 2024. Vol. 4. P. 23. <https://doi.org/10.1007/s43546-023-00618-x>
6. Ali H., Zafar M.B., Aysan A.F. Generative AI in finance: Replicability, methodological contingencies, and future research directions // Finance Research Letters. 2025. Vol. 86. Part F. P. 108797. <https://doi.org/10.1016/j.frl.2025.108797>
7. Ali H., Aysan A.F. Navigating the future of finance: The transformative role of generative AI // Journal of Central Banking Law and Institutions. 2026. Vol. 5. Iss. 1. P. 79–124. <https://doi.org/10.21098/jcli.v5i1.431>
8. Nayak K., Shivaram S.B., Nayak S.K. Machine learning framework for algorithmic trading // Computer Sciences and Mathematics Forum. 2025. Vol. 12. Iss. 1. P. 12. <https://doi.org/10.3390/cmsf2025012012>
9. Borch C. Machine learning, knowledge risk, and principal-agent problems in automated trading // Technology in Society. 2022. Vol. 68. P. 101852. <https://doi.org/10.1016/j.techsoc.2021.101852>
10. Dowding K., Taylor B.R. Algorithmic decision-making, agency costs, and institution-based trust // Philosophy and Technology. 2024. Vol. 37. P. 68. <https://doi.org/10.1007/s13347-024-00757-5>
11. Li X., Feng H., Yang H., Huang J. Can ChatGPT reduce human financial analysts' optimistic biases? // Economic and Political Studies. 2024. Vol. 12. Iss. 1. P. 20–33. <https://doi.org/10.1080/20954816.2023.2276965>

12. *Pelster M., Val J.* Can ChatGPT assist in picking stocks? // *Finance Research Letters*. 2024. Vol. 59. P. 104786. <https://doi.org/10.1016/j.frl.2023.104786>
13. *Rudin C.* Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // *Nature Machine Intelligence*. 2019. Vol. 1. P. 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
14. *Kordzadeh N., Ghasemaghaei M.* Algorithmic bias: Review, synthesis, and future research directions // *European Journal of Information Systems*. 2022. Vol. 31. Iss. 3. P. 388–409. <https://doi.org/10.1080/0960085X.2021.1927212>
15. *Mökander J.* Auditing of AI: Legal, ethical and technical approaches // *Digital Society*. 2023. Vol. 2. P. 49. <https://doi.org/10.1007/s44206-023-00074-y>
16. *Metaxa D., Park J.S., Robertson R.E., Karahalios K., Wilson C., Hancock J., Sandvig C.* Auditing algorithms: Understanding algorithmic systems from the outside in // *Foundations and Trends in Human–Computer Interaction*. 2021. Vol. 14. Iss. 4. P. 272–344. <https://doi.org/10.1561/11000000083>
17. *Uddin H., Barai M.K.* Fintech adoption and bank risk, efficiency and stability: Evidence from panel data of selected Asian economies // *FinTech*. 2026. Vol. 5. Iss. 1. P. 14. <https://doi.org/10.3390/fintech5010014>
18. *Cevik S.* The dark side of the moon? Fintech and financial stability // *International Review of Economics*. 2024. Vol. 71. P. 421–433. <https://doi.org/10.1007/s12232-024-00449-8>
19. *An J., Rau R.* Finance, technology and disruption // *The European Journal of Finance*. 2021. Vol. 27. Iss. 4-5. P. 334–345. <https://doi.org/10.1080/1351847X.2019.1703024>
20. *Jensen M.C., Meckling W.H.* Theory of the firm: Managerial behavior, agency costs and ownership structure // *Journal of Financial Economics*. 1976. Vol. 3. Iss. 4. P. 305–360. [https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X)
21. *Holmstrom B.* Moral hazard and observability // *The Bell Journal of Economics*. 1979. Vol. 10. Iss. 1. P. 74–91. URL: <https://people.duke.edu/~qc2/BA532/1979%20Rand%20Holmstrom.pdf> (дата обращения: 06.02.2026)
22. *Stiglitz J.E.* Incentives and risk-sharing in sharecropping // *The Review of Economic Studies*. 1974. Vol. 41. Iss. 2. P. 219–255. URL: <https://www.sci-hub.ru/10.2307/2296714> (дата обращения: 06.02.2026)
23. *Hart O.* *Firms, contracts, and financial structure*. Oxford: Oxford University Press, 1995. 248 p. URL: <http://www.library.fu.ru/files/Hart-Firms.pdf> (дата обращения: 06.02.2026)
24. *Williamson O.E.* The theory of the firm as governance structure: From choice to contract // *Journal of Economic Perspectives*. 2002. Vol. 16. Iss. 3. P. 171–195. <https://doi.org/10.1257/089533002760278776>
25. *Simon H.A.* *The sciences of the artificial*. 3rd ed. Cambridge (Massachusetts): MIT Press, 1996. 231 p. URL: https://monoskop.org/images/9/9c/Simon_Herbert_A_The_Sciences_of_the_Artificial_3rd_ed.pdf (дата обращения: 06.02.2026)
26. *Mei H., Bodog S.A., Badulescu D.* Artificial Intelligence adoption in sustainable banking services: The critical role of technological literacy // *Sustainability*. 2024. Vol. 16. Iss. 20. P. 8934. <https://doi.org/10.3390/su16208934>
27. *Manta O., Vasile V., Rusu E.* Banking transformation through fintech and the integration of Artificial Intelligence in payments // *FinTech*. 2025. Vol. 4. Iss. 2. P. 13. <https://doi.org/10.3390/fintech4020013>

Статья поступила в редакцию 24.04.2026; одобрена после рецензирования 03.06.2026; принята к публикации 17.06.2026

Об авторе:

Зверькова Татьяна Николаевна, кандидат экономических наук, доцент, доцент кафедры банковского дела и страхования; SPIN-код: 3385-9234

Автор прочитал и одобрил окончательный вариант рукописи.

References

1. Kotliarov I.D. Fintech: Nature and models of implementation. *ECO*. 2018; (12(534)):23–39. EDN: <https://elibrary.ru/yoqgjn> (In Russ.)
2. Artemenko D.A., Zenchenko S.V. Digital technologies in the financial sector: Evolution and major development trends in Russia and abroad. *Finance: Theory and Practice*. 2021; 25(3):90–101. EDN: <https://elibrary.ru/rixjhx>. <https://doi.org/10.26794/2587-5671-2021-25-3-90-101> (In Russ.)
3. Krinichansky K.V., Zeleneva E.S. Fintech sector in the context of financial development and problems of its measurement. *Finance: Theory and Practice*. 2024; 28(5):121–132. EDN: <https://elibrary.ru/yqhgox>. <https://doi.org/10.26794/2587-5671-2024-28-5-121-132> (In Russ.)
4. Tavbulatova Z.K., Batukaeva L.S.E. Fintech transformation of the Russian banking sector in the context of technological sovereignty. *Finance and Credit*. 2025; 31(10):46–65. EDN: <https://elibrary.ru/eztyen>. <https://doi.org/10.24891/eztyen> (In Russ.)
5. Bahoo S., Cucculelli M., Goga X., Mondolo J. Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. *SN Business & Economics*. 2024; 4:23. <https://doi.org/10.1007/s43546-023-00618-x> (In Eng.)
6. Ali H., Zafar M.B., Aysan A.F. Generative AI in finance: Replicability, methodological contingencies, and future research directions. *Finance Research Letters*. 2025; 86(F):108797. <https://doi.org/10.1016/j.frl.2025.108797> (In Eng.)
7. Ali H., Aysan A.F. Navigating the future of finance: The transformative role of generative AI. *Journal of Central Banking Law and Institutions*. 2026; 5(1):79–124. <https://doi.org/10.21098/jcli.v5i1.431> (In Eng.)
8. Nayak K., Shivaram S.B., Nayak S.K. Machine learning framework for algorithmic trading. *Computer Sciences and Mathematics Forum*. 2025; 12(1):12. <https://doi.org/10.3390/cmsf2025012012> (In Eng.)
9. Borch C. Machine learning, knowledge risk, and principal-agent problems in automated trading. *Technology in Society*. 2022; 68:101852. <https://doi.org/10.1016/j.techsoc.2021.101852> (In Eng.)
10. Dowding K., Taylor B.R. Algorithmic decision-making, agency costs, and institution-based trust. *Philosophy and Technology*. 2024; 37:68. <https://doi.org/10.1007/s13347-024-00757-5> (In Eng.)
11. Li X., Feng H., Yang H., Huang J. Can ChatGPT reduce human financial analysts' optimistic biases? *Economic and Political Studies*. 2024; 12(1):20–33. <https://doi.org/10.1080/20954816.2023.2276965> (In Eng.)
12. Pelster M., Val J. Can ChatGPT assist in picking stocks? *Finance Research Letters*. 2024; 59:104786. <https://doi.org/10.1016/j.frl.2023.104786> (In Eng.)
13. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019; 1:206–215. <https://doi.org/10.1038/s42256-019-0048-x> (In Eng.)
14. Kordzadeh N., Ghasemaghaei M. Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*. 2022; 31(3):388–409. <https://doi.org/10.1080/0960085X.2021.1927212> (In Eng.)
15. Mökander J. Auditing of AI: Legal, ethical and technical approaches. *Digital Society*. 2023; 2:49. <https://doi.org/10.1007/s44206-023-00074-y> (In Eng.)
16. Metaxa D., Park J.S., Robertson R.E., Karahalios K., Wilson C., Hancock J., Sandvig C. Auditing algorithms: understanding algorithmic systems from the outside in. *Foundations and Trends in Human–Computer Interaction*. 2021; 14(4):272–344. <https://doi.org/10.1561/11000000083> (In Eng.)
17. Uddin H., Barai M.K. Fintech adoption and bank risk, efficiency and stability: Evidence from panel data of selected Asian economies. *FinTech*. 2026; 5(1):14. <https://doi.org/10.3390/fintech5010014> (In Eng.)
18. Cevik S. The dark side of the moon? Fintech and financial stability. *International Review of Economics*. 2024; 71:421–433. <https://doi.org/10.1007/s12232-024-00449-8> (In Eng.)
19. An J., Rau R. Finance, technology and disruption. *The European Journal of Finance*. 2021; 27(4-5):334–345. <https://doi.org/10.1080/1351847X.2019.1703024> (In Eng.)

20. Jensen M.C., Meckling W.H. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*. 1976; 3(4):305–360. [https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X) (In Eng.)
21. Holmstrom B. Moral hazard and observability. *The Bell Journal of Economics*. 1979; 10(1):74–91. URL: <https://people.duke.edu/~qc2/BA532/1979%20Rand%20Holmstrom.pdf> (accessed: 06.02.2026) (In Eng.)
22. Stiglitz J.E. Incentives and risk-sharing in sharecropping. *The Review of Economic Studies*. 1974; 41(2):219–255. URL: <https://www.sci-hub.ru/10.2307/2296714> (accessed: 06.02.2026) (In Eng.)
23. Hart O. Firms, contracts, and financial structure. Oxford: Oxford University Press, 1995. 248 p. URL: <http://www.library.fu.ru/files/Hart-Firms.pdf> (accessed: 06.02.2026) (In Eng.)
24. Williamson O.E. The theory of the firm as governance structure: From choice to contract. *Journal of Economic Perspectives*. 2002; 16(3):171–195. <https://doi.org/10.1257/089533002760278776> (In Eng.)
25. Simon H.A. The sciences of the artificial. 3rd ed. Cambridge (Massachusetts): MIT Press, 1996. 231 p. URL: https://monoskop.org/images/9/9c/Simon_Herbert_A_The_Sciences_of_the_Artificial_3rd_ed.pdf (accessed: 06.02.2026) (In Eng.)
26. Mei H., Bodog S.A., Badulescu D. Artificial Intelligence adoption in sustainable banking services: The critical role of technological literacy. *Sustainability*. 2024; 16(20):8934. <https://doi.org/10.3390/su16208934> (In Eng.)
27. Manta O., Vasile V., Rusu E. Banking transformation through fintech and the integration of Artificial Intelligence in payments. *FinTech*. 2025; 4(2):13. <https://doi.org/10.3390/fintech4020013> (In Eng.)

The article was submitted 24.04.2026; approved after reviewing 03.06.2026; accepted for publication 17.06.2026

About the author:

Tatyana N. Zverkova, Candidate of Economic Sciences, Associate Professor, Associate Professor of the Department of Banking and Insurance; SPIN: 3385-9234

The author has read and approved the final version of the manuscript.