

DOI: 10.17323/2587-814X.2026.1.7.28

# Как раскрыть черный ящик: объяснимый ИИ для Индустрии 5.0

Сергей Михайлович Авдошин 

E-mail: savdoshin@hse.ru

Елена Юрьевна Песоцкая 

E-mail: epesotskaya@hse.ru

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

## Аннотация

Бурное развитие искусственного интеллекта (ИИ) сопровождается ростом вычислительной сложности и снижением прозрачности моделей, что существенно ограничивает их применение в критически важных сферах, требующих высокого уровня доверия, интерпретируемости и обоснованности принимаемых решений. В этих условиях особое значение приобретает направление объяснимого искусственного интеллекта (Explainable Artificial Intelligence, XAI), ориентированное на создание подходов и технологий, обеспечивающих понимание логики работы ИИ-систем и интерпретацию их выводов. Статья посвящена актуальной теме внедрения объяснимого искусственного интеллекта в контексте Индустрии 5.0. Особое внимание уделено практическим сценариям использования, авторами приводятся конкретные промышленные примеры от компаний IBM, Siemens и других, демонстрирующие, как XAI помогает повысить надежность, безопасность, эффективность и доверие к ИИ-системам. В статье приведен систематический поиск и анализ литературы в данной области, предложены и обоснованы ключевые критерии сравнения существующих подходов. Также обозначены преимущества, существующие ограничения и перспективные направления развития XAI, открывающие новые возможности для повышения эффективности, прозрачности и доверия в бизнесе.

**Ключевые слова:** XAI, объяснимый искусственный интеллект, Индустрия 5.0, машинное обучение, промышленность

**Цитирование:** Авдошин С. М., Песоцкая Е. Ю. Как раскрыть черный ящик: объяснимый ИИ для Индустрии 5.0 // Бизнес-информатика. 2026. Т. 20. № 1. С. 7–28.  
<https://doi.org/10.17323/2587-814X.2026.1.7.28>

## Введение

**В**озникновение объяснимого искусственного интеллекта (ХАИ) напрямую связано с успехами современных методов машинного обучения, особенно глубоких нейронных сетей. Эти модели достигли выдающихся результатов во множестве задач, однако приобрели характер «черного ящика» – то есть представляют собой чрезвычайно сложные системы, внутренние механизмы которых остаются неочевидными для пользователя [1]. В отличие от ранних систем ИИ, таких как экспертные системы или модели с жестко заданными правилами, которые были относительно прозрачными, современные алгоритмы глубокого обучения содержат миллионы параметров. Рост их сложности привел к тому, что интерпретировать принимаемые ими решения стало практически невозможно. Это породило так называемый «барьер объяснимости», ограничивающий применение ИИ из-за дефицита доверия к непрозрачным моделям [2].

Современное общество ожидает, что искусственный интеллект будет не только эффективным, но и достоверным, прозрачным и справедливым [3–5]. Недостаток объяснимости решений вызывает осторожность со стороны пользователей и регулирующих органов.

Фактически, ХАИ появился как ответ на этот вызов: он призван обеспечить интерпретируемость и прозрачность работы «черных ящиков» искусственного интеллекта. Его задача – преодолеть разрыв между высокой сложностью современных моделей и необходимостью человека понимать результаты, которые они генерируют. В рамках ХАИ создаются методы, техники и алгоритмы, способные давать интерпретируемые и интуитивно понятные объяснения решений ИИ. Таким образом, ХАИ предоставляет человеку – разработчику, пользователю или регулятору – ясное и логичное обоснование работы алгоритма.

В контексте человекоцентричной Индустрии 5.0 ХАИ рассматривается как один из ключевых факторов успешного внедрения ИИ. Он позволяет пользователям понимать и доверять результатам работы алгоритмов, что критически важно для эффективного взаимодействия человека и машины. Объяснимый ИИ способствует тому, чтобы цифровые системы оставались этичными, подотчетными и согласованными с человеческими ценностями и целями [6, 7].

Для руководителей бизнеса ХАИ становится не просто инструментом, а необходимым условием эффективного управления. В условиях возрастающей сложности «черные ящики» лишают управленцев возможности оценить обоснованность решений, на которых строятся стратегические и операционные действия компаний. Использование ХАИ позволяет преодолеть этот разрыв, предоставляя прозрачные объяснения работы алгоритмов. Это создает основу для более взвешенного и ответственного принятия решений, способствует снижению рисков и открывает новые возможности для инноваций и развития. Для компаний, стремящихся оставаться конкурентоспособными в условиях Индустрии 5.0, внедрение ХАИ становится стратегической необходимостью [8–10].

Авторами проанализированы современные подходы и требования к объяснимости, направленные на повышение прозрачности и надежности интеллектуальных систем, а также укрепление доверия к их решениям. В рамках исследования проведен систематический поиск и анализ литературы по ХАИ, основанный на критериях включения и исключения публикаций, анализе баз цитирования и структурировании материалов. В первом разделе раскрывается сущность ХАИ в контексте Индустрии 5.0, освещается его роль и проблема «черного ящика» в бизнес-применениях, представлено сравнение существующих подходов. Второй раздел посвящен возможностям использования ХАИ в бизнесе и ключевым направлениям его внедрения. В третьем разделе приводятся практические кейсы и отраслевые примеры, демонстрирующие эффективность ХАИ в деятельности компаний. В четвертом разделе анализируются барьеры и ограничения, препятствующие широкому внедрению ХАИ, а также оцениваются потенциальные риски. В заключительном, пятом разделе обсуждаются перспективные направления развития и будущие траектории применения ХАИ в бизнес-решениях.

## 1. Понятие объяснимого ИИ в контексте Индустрии 5.0

### 1.1. Индустрия 5.0 и объяснимый ИИ

Широкое внедрение ИИ в критически важных областях обнажает ряд проблем, связанных с объяснимостью, особенно в контексте Индустрии 5.0, которую европейская комиссия определяет как промышленность, которая дополняет существующую парадигму Индустрии 4.0 человекоцентричным

подходом и устойчивостью к внешним изменениям [11]. Если Индустрия 4.0 фокусировалась на технологиях (автономность, цифровые связи, данные), то пятая ставит в центр человека, характеризуется тесной интеграцией ИИ и подразумевает социальную ответственность. Индустрия 5.0 делает человеческое участие ключевым элементом производственных и управленческих процессов [11, 12] и обеспечивает более тесное сотрудничество человека и ИИ/роботов на рабочих местах. При этом человек не устраняется из процесса, а наоборот, технологии служат для усиления возможностей, обеспечения комфорта и безопасности сотрудников, а также персонализации производства под запросы людей.

В этих условиях ХАИ становится ключевым фактором доверия и эффективности, служит «мостом» между сложностью современных «черных ящиков» и потребностями в достоверности и прозрачности систем ИИ. ХАИ обычно определяется как способность системы предоставлять понятное человеку объяснение того, как принимаются решения [13]. Его цель – сделать модели ИИ прозрачными, интерпретируемыми и надежными за счет объяснения внутренних процессов и выводов алгоритма [14].

Мотивация развития ХАИ в бизнесе во многом продиктована этическими и правовыми запросами. Во-первых, регуляторы предъявляют все больше требований к прозрачности алгоритмов. В Европейском союзе обсуждается «право на объяснение» решений, принимаемых автоматизированными системами. Например, в банковском деле: если заемщику отказано в кредите автоматизированной системой, клиент имеет право узнать причины

этого решения [15]. Такие нормы (включая требования регламента GDPR по защите данных в ЕС) вынуждают организации внедрять объяснимость, иначе использование «черного ящика» может привести к юридическим последствиям [16].

Во-вторых, социально-организационные факторы также играют значимую роль. Как отмечают Zavodna и соавторы [17] недостаточная прозрачность ИИ вызывает сопротивление пользователей и менеджеров при внедрении таких систем. В бизнес-среде накоплен опыт, когда непрозрачность ИИ ведет к неприятию технологии, снижая эффективность цифровой трансформации.

Обеспечение объяснимости – необходимое условие повышения доверия к ИИ среди сотрудников, клиентов и пользователей сервисов. Согласно последним исследованиям [18, 19] ХАИ помогает выявлять и устранять предубеждения модели, гарантировать соблюдение этических норм и повышать обоснованность решений, что в конечном счете повышает готовность людей принимать и эффективно использовать системы ИИ. Можно сделать вывод, что в человеко-ориентированной парадигме Индустрии 5.0, где машины призваны дополнять человека, а не заменять, прозрачность решений ИИ становится обязательной для безопасного и результативного сотрудничества человека с ИИ.

В рамках исследования проведен систематический поиск и анализ литературы по ХАИ, основанный на критериях включения и исключения публикаций, анализе баз цитирования и структурировании материалов (рис. 1). Исследование ос-



Рис. 1. Методика библиометрического анализа.

новывается на систематическом поиске и анализе научной литературы по теме объяснимого искусственного интеллекта и его применения в бизнесе и Индустрии 5.0.

Наибольший рост публикаций наблюдается в 2021–2024 гг., что объясняется растущим интересом бизнеса к прозрачности алгоритмов и нормативным требованиям.

Тематическое структурирование выполнено по следующим направлениям, которые взяты за основу структуры данного исследования:

- ◆ концептуальные основы ХАИ;
- ◆ методы и метрики;
- ◆ применение в бизнесе;
- ◆ барьеры и риски;
- ◆ нормативные аспекты.

Важно понимать, что объяснимость – это не отдельный и статичный показатель. В академических исследованиях ХАИ – это сложный, многомерный критерий: от прозрачности модели (насколько ее механизм доступен для понимания) и интерпретируемости (насколько мы можем понять, почему она приняла именно такое решение), до точности, справедливости, прозрачности (отсутствие заблуждений) и ответственности. Например, прямая и ясная модель может быть прозрачной, но не всегда точной – поэтому каждый проект ХАИ должен балансировать между этими измерениями. Юсе с

коллегами [20] предлагают рассматривать объяснимость как функцию понятности, отражающую прозрачность и интерпретируемость. Arrieta и соавторы [14] наравне с Murdoch [21] объединяют эти концепции в более широкий контекст ответственного ИИ, дополняя понятиями доверия, надежности и пр. Авторами статьи предложена оригинальная карта наиболее распространенных свойств объяснимости (рис. 2).

На сегодняшний день не существует единого, общепринятого стандарта, определяющего, что именно считать объяснимостью в ИИ. Эта многомерность подходов отражает сложность самого понятия и указывает на необходимость систематизации и согласования понятийного аппарата и подходов к оценке ХАИ в зависимости от контекста применения (отрасли, типа модели, целевой аудитории и пр.).

Также отсутствует и универсальный набор количественных или качественных метрик для измерения уровня объяснимости. Разные подходы используют разные критерии – от субъективной понятности объяснений пользователю до формальных оценок стабильности и локальной точности интерпретаций [22, 23]. В результате внедрение объяснимого ИИ требует не только технической реализации, но и методологической работы по определению, что считать «достаточной» объяснимостью в конкретном контексте.

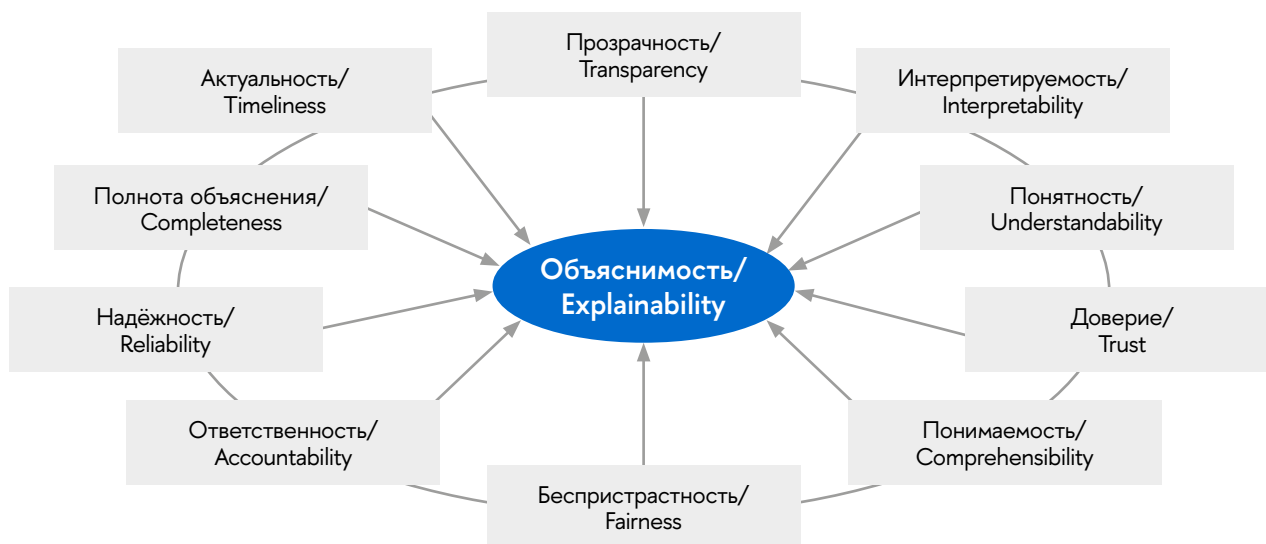


Рис. 2. Карта основных свойств объяснимости.

При существующем многообразии подходов можно утверждать, что сравнение методов объяснимости основывается на ряде повторяющихся критериев [13, 14, 20, 21]. Наиболее часто в литературе используются следующие характеристики:

- ◆ Тип объяснения (локальное/глобальное, пост-хок/встроенное).
- ◆ Прозрачность и интерпретируемость (насколько легко человек может понять логику модели).
- ◆ Стабильность объяснений (степень изменчивости объяснений при малых изменениях входных данных).
- ◆ Точность и информативность объяснения (правильно ли отражает реальную логику модели).
- ◆ Устойчивость к шумам и атакам (хорошо/мало устойчивы).
- ◆ Временная и вычислительная сложность, которая важна для промышленности и вычислений в реальном времени (высокая/средняя/низкая).

Эти критерии широко представлены в международных исследованиях и используются как основа для систематического сравнения подходов ХАИ.

### 1.2. Проблема «черного ящика» и доверия к ИИ в бизнесе

Спрос на ответственный, прозрачный и человекоцентричный ИИ со стороны бизнеса стабильно растет, следуя современным трендам. Формируются лучшие практики, появляются отраслевые рекомендации (преимущественно в финансовом секторе и здравоохранении), многие компании начинают самостоятельно определять требования к объяснимости, исходя из специфики своих бизнес-процессов.

В деловых и управленческих сферах применение искусственного интеллекта уже становится повседневной практикой – от систем рекомендаций в интернет-торговле до алгоритмов оптимизации бизнес-процессов и принятия решений на основе данных. Однако готовность компаний и организаций доверять критически важные процессы «черным ящикам» ограничена.

Модель «черного ящика» – это система, внутреннее устройство которой скрыто или слишком сложно для понимания человеком. Многие высокоэффективные алгоритмы машинного обучения, включая глубокие нейронные сети и ансамблевые методы, отличаются высокой непрозрачностью.

Пока такие модели применялись в ограниченных задачах, это не считалось критичным. Однако с расширением применения ИИ в сферах, затрагивающих принятие решений и деятельность человека, проблема непрозрачности приобрела фундаментальный характер, что подчеркивается в ряде исследований по ХАИ [24–26]. Для бизнеса встает вопрос: если мы не понимаем, как работает алгоритм, можем ли мы ему доверять?

Отсутствие объяснений затрудняет обнаружение ошибок и предвзятости: скрытые перекосы обучающих данных могут приводить к несправедливым решениям. Например, алгоритмы найма или поиска клиентов способны непреднамеренно дискриминировать отдельные группы соискателей.

Говоря о доверии к ИИ, особенно в сферах с высоким риском (транспорт, финансы, промышленность) люди склонны отвергать результаты моделей, даже высокоточные, если отсутствует рациональное объяснение. Так, в банковском секторе клиенты и менеджеры требуют знать причину отказа в кредите. В сфере производства отсутствие объяснений к решениям может привести не только к недоверию, но и к критичным последствиям в части безопасности, если модель даст сбой. Любое значимое решение должно сопровождаться понятной человеку аргументацией. Без этого использование ИИ в подобных областях считается неприемлемым.

Широко обсуждается компромисс между точностью модели и ее понятностью. Действительно, самые прозрачные модели, например, решающие деревья с ограниченной глубиной, зачастую менее точны на сложных задачах, чем глубокие сети. И наоборот, стремление максимизировать метрику качества приводит к громоздким моделям, жертвующим интерпретируемостью. Либо мы имеем «серый ящик» попроще, но понятный, либо мощный «черный ящик» высокой производительности, который достигается ценой потери прозрачности [27, 28]. Задача ХАИ – уменьшить этот разрыв, предлагая методы сохранения точности при обеспечении объяснимости, однако полностью снять этот компромисс пока не удалось, и вопрос «какой частью производительности можно пожертвовать ради прозрачности?» остается открытым.

Подводя итог можно сказать, что проблема «черного ящика» – это не просто техническая метафора, а серьезное препятствие на пути внедрения ИИ в различные сферы бизнеса.

## 2. Потребности ХАИ в бизнесе

По результатам библиометрического анализа авторами сделан вывод, что наиболее насыщенными сферами применения объяснимого ИИ сегодня являются финансы и промышленность, где ХАИ используется как для обеспечения доверия к алгоритмам, так и для поддержки операционных и стратегических решений. Энергетика, государственный сектор и здравоохранение встречаются преимущественно в работах нормативно-регуляторного характера, что отражает растущее внимание к прозрачности, аудируемости и недискриминационности алгоритмов в высокорисковых областях. Логистика формируется как новая, активно развивающаяся область применения ХАИ: количество работ пока относительно невелико, однако область демонстрирует быстрый рост и возрастающий интерес исследователей и компаний к объяснимости в системах цепей поставок и интеллектуальной оптимизации (*табл. 1*).

Интеграция объяснимых алгоритмов в реальную практику помогает сгладить проблему «черного ящика», сделать ИИ понятным и приемлемым для бизнеса. Требования к объяснимому ИИ в бизнесе, экономике и менеджменте можно классифицировать по нескольким ключевым направлениям, отражающим практические потребности организаций в объяснимых алгоритмах и требования внешней среды.

**А. Доверие и прозрачность решений.** Доверие рассматривается как фундаментальное условие функционирования систем ИИ в цифровой экономике [29]. Объяснимый ИИ способен предоставить объяснение в понятной форме, снижая риски недоверия и дискриминации, обосновывать рекомендации ИИ перед клиентами, акционерами, аудиторами и сотрудниками. Так, кредитный скоринг на базе ХАИ может детализировать вклад факторов (доход, кредитная история и др.), удовлетворяя нормативные требования и укрепляя доверие клиентов. В сфере управления персоналом объяснимость помогает избежать необоснованных решений. Если алгоритм отсеивает кандидатов на вакансию, компания должна убедиться, что это происходит по релевантным причинам, а не, например, из-за скрытой дискриминации. Предоставляя HR-специалистам интерпретируемые критерии и информацию, какие навыки или компетенции стали решающими, ХАИ делает процесс отбора более прозрачным и справедливым. Это снижает риск предвзятости и повышает доверие сотрудников к подобным системам.

**В. Интеграция ИИ в рабочие процессы.** Объяснимость способствует интеграции алгоритмов в рабочие процессы, снижает сопротивление персонала и формирует «общий язык» между человеком и системой. На уровне организаций исследователи вводят понятие приемлемости ИИ (AI acceptability) – насколько охотно сотрудники и руководители соглашаются внедрять и использовать ИИ-инструменты. Выясняется, что основной барьер – социально-организационные факторы, во многом связанные с доверием и пониманием [16, 43]. Сотрудники могут сопротивляться алгоритмическим вычислениям, опасаясь потери контроля или не доверяя «машинным» решениям. Но если система предоставляет понятные объяснения и будет вовлекать пользователей в процесс, появится высокая вероятность взаимного доверительного партнерства. Можно предположить, что инженеры будут охотнее доверять предиктивной системе, если она укажет конкретные показатели датчиков, ставшие причиной прогноза, и предоставит релевантную информацию.

**С. Стратегическое управление и бизнес-аналитика.** ХАИ поддерживает топ-менеджмент и собственников бизнеса в принятии стратегических решений. Бизнес все чаще используют аналитические модели для стратегического планирования, оценки рисков, изучения поведения потребителей. Однако руководители не готовы основываться на результатах работы модели, если не понимают предпосылок. Поэтому объяснимые модели – например, эконометрика с интерпретируемыми коэффициентами или современные ML-модели, обогащенные ХАИ-объяснениями – предпочтительны в корпоративной аналитике.

Недавние обзоры, в частности Tchuente, Lonlac, и Kamsu-Fogueum [9] предлагают подход структурированной оценки с учетом теоретических основ, контекста применения, характеристик данных/задачи и методологии решения (ТССМ – Theory, Context, Characteristics, Methodology). В реальных условиях важно объяснять весь управленческий процесс целиком: почему задаем вопрос X, почему используем данные Y, как модель пришла к выводам, и подтверждают ли эксперты эти объяснения на практике. Без валидации объяснений человеком применение ХАИ в бизнесе будет неполным. Именно поэтому специалисты рекомендуют выстраивать цикл: модель – объяснение – оценка объяснения экспертом – корректировка модели или ее применения.

Таблица 1.

## Библиографический анализ сфер применения ХАИ в различных отраслях

Источник	Фокус исследования и ключевые идеи									
	Финансы	Промышленность	Энергетика	Бизнес	Экономика	Менеджмент	Логистика	Гос.сектор	Медицина	
1. Martins et al. (2024) [10]	X			X	X	X				Обзор ХАИ в финансах; SHAP/LIME; прозрачность кредитного скоринга
2. Gramegna & Scardapane (2021) [31]	X					X				Оценка дискриминации; объяснимость в кредитном риске
3. Hjelkrem & de Lange (2023) [32]	X					X				Объяснение глубоких моделей в открытом банкинге
4. Poyiadzi et al. (FACE, 2020) [30]	X	X		X	X		X			Контрфактуальные объяснения; применимость в разных областях
5. Weitz et al. (2022) [18]				X		X				AI-допустимость; снижение сопротивления
6. Chehbi-Gamoura (2023) [6]				X		X				Объяснимость и принятие ИИ
7. Tabassi (NIST AI RMF, 2023) [16]	X	X	X	X	X	X		X	X	Нормативные требования к ХАИ
8. EU AI Act (2024) [41]	X	X	X	X	X	X		X	X	Правовые требования к объяснимости
9. Ahmed et al. (2022) [12]		X		X	X		X			ХАИ в Индустрии 4.0/5.0
10. Adadi & Berrada (2018) [13]	X	X		X	X	X			X	Таксономия ХАИ; пост-хок методы
11. Arrieta et al. (2020) [14]	X	X	X	X	X	X		X	X	Ответственный ИИ; свойства объяснимости
12. Černevičienė & Kabasinskas (2024) [42]	X			X	X	X				Систематический обзор ХАИ в финансах; задачи: скоринг, методы SHAP/ANN/ XGBoost
13. Brasse et al. (2023) [43]				X	X	X				ХАИ в информационных системах; классификация направлений
14. Samek W., Montavon G., et al. (2019) [1]	X			X	X	X				Обзор методов ХАИ: таксономия подходов (rule-based, model-agnostic, intrinsic models); применение и цитируемость
15. Carvalho et al. (2019) [44]		X		X	X		X			Систематический обзор ХАИ: отраслевые ограничения, вызовы и возможности
16. Molnar (2025) [45]	X	X		X		X				Комплексный подход к интерпретируемости; стабильность объяснений
17. Angelov et al. (2021) [46]	X	X		X						Самообъясняемые модели; интерпретируемые нечеткие правила
18. Rai (2020) [47]	X			X	X	X				Объяснимость в управлении и системах поддержки принятия решений
19. Samek & Müller (2017) [22]		X	X	X						Методы визуализации; оценка объяснимости; определение релевантности
20. Liao & Varshney (2021) [48]	X	X		X		X				Человеко-ориентированный ХАИ; адаптивные объяснения для стейкхолдеров
21. Chamola V et al. (2023) [49]		X	X	X	X	X	X			Объяснимость в киберфизических системах; контекстно-зависимые объяснения
22. Belle & Papantonis (2021) [50]				X	X	X				Логико-ориентированный ХАИ; символическое рассуждение; принятия решений

#### **Д. Контроль сложных производственных систем.**

Современное производство генерирует огромные объемы данных (датчиков на оборудовании, финансовая информация, данные логистики и т. д.), и AI-модели находят в них скрытые паттерны, оптимизируя работу. При этом инженеры и операторы должны понимать эти паттерны, особенно когда система предлагает нестандартное решение – например, остановить станок из-за обнаруженной аномалии – ХAI позволяет встроить в системы промышленной аналитики модули, объясняющие: какие именно датчики или показатели вышли за норму, почему система прогнозирует скорую неисправность, какой фактор стал решающим при выявлении дефекта продукта.

Для этих целей хорошо себя зарекомендовал подход FACE (Feasible and Actionable Counterfactual Explanations), подробно описанный в литературе [30]. FACE подбирает реалистичный и достижимый путь изменений от текущего случая к желаемому исходу с учетом реалистичности и выполнимости изменений (технологические допуски, безопасность, регламенты). В результате персонал получает интерпретацию от AI-алгоритма на понятном ему языке (будь то график, описание или визуальная подсветка проблемного узла на схеме) с информацией о том, какие факторы стали решающими для вывода системы и какие изменения необходимы. Если робот или автоматизированная линия действует непредсказуемо, это также риск для людей и производства, который нужно контролировать для обеспечения безопасности. Наличие объяснений (например, «робот снизил скорость, потому что датчик выявил отклонение в качестве сырья») позволяет проанализировать, на основе каких данных и правил система приняла решение, и скорректировать алгоритм, чтобы избежать повторения ошибки.

#### **Е. Соответствие нормативным требованиям.**

Многие отрасли экономики строго регулируются – финансы, промышленность, энергетика. Бизнес, желая избежать репутационных и юридических рисков, нуждается в этических комитетах и процедурах аудита алгоритмов. ХAI-инструменты выступают технической поддержкой этих инициатив. Фактически, объяснимость становится конкурентным преимуществом: компании, которые могут доказать прозрачность и справедливость своих алгоритмов, получают больше доверия потребителей и регуляторов [10], в том числе GDPR [15] и AI Act [41].

Резюмируя, хочется отметить, что в бизнесе, экономике и управлении объяснимый ИИ повышает прозрачность бизнес-аналитики, улучшает взаимодействие людей и алгоритмов в организациях, обеспечивает соблюдение норм и этики. Объяснимость ИИ постепенно становится частью корпоративной культуры работы с данными. Решения менеджмента теперь должны быть не только «data-driven» (основанными на данных), но и «explanation-driven», то есть сопровождаться понятными обоснованиями. Только при наличии понятных объяснений алгоритмов все стейкхолдеры готовы принять и поддержать решение.

### **3. Практическое применение ХAI: кейсы и отрасли**

Объяснимый ИИ становится наиболее востребованным в тех сферах, где автоматизированные решения оказывают прямое влияние на людей, их здоровье, благосостояние, права и безопасность. В таких контекстах простого повышения точности модели недостаточно – необходимо обеспечить понятность и обоснованность решений, что делает ХAI критически важным компонентом внедрения ИИ.

Ниже представлены ключевые области бизнеса, в которых ХAI уже используется или активно внедряется, а также конкретные кейсы и задачи, где объяснимость играет решающую роль.

#### **3.1. Финансовый сектор**

Финансы можно отнести к ключевому сектору, который является одной из наиболее регламентированных сфер применения ИИ. Здесь от объяснимости зависят не только доверие клиентов, но и выполнение обязательных юридических и этических норм. При этом существует конфликт между точностью и интерпретируемостью: глубокие модели показывают высокую предсказательную силу, но не поддаются объяснению. Для повышения прозрачности банки предпочитают либо более интерпретируемые модели (например, градиентный бустинг, где важности признаков можно оценить), либо могут применять ХAI-методы. Это могут быть интерпретируемые скоринговые карты и монотонные GBM-модели (Gradient Boosting Machines), где соблюдается логическая связь между факторами и итоговым значением показателя. Популярность набирают модели LIME

(Local Interpretable Model-agnostic Explanations) и SHAP-анализ (SHapley Additive Explanations), особенно в направлениях скоринга, инвестиционного анализа, анализа рисков, о чем заявляют последние исследования [31–33].

**Кредитный скоринг и одобрение займов.** При автоматическом решении о выдаче кредита банки зачастую обязаны сообщить заемщику причину отказа. Клиенты имеют право получить объяснение, почему заявка отклонена, а банки обязаны следить за тем, чтобы решения моделей не основывались на дискриминационных признаках, например, поле, возрасте или этнической принадлежности. Так, для получения интерпретируемого результата SHAP дает числовую оценку вклада признака. Эти значения упрощаются и транслируются в виде «обоснований» (например, недостаточный доход: –20 к рейтингу, короткая кредитная история: –15, высокий текущий долг: –10). Далее модель анализирует, что должно измениться в исходных данных, чтобы модель выдала другой результат и предлагает клиенту варианты улучшения своей кредитоспособности. В результате достигается и регуляторная объяснимость, и понятность для клиента, что повышает прозрачность системы и позволяет клиенту понять, что можно улучшить для пересмотра решения.

**Инвестиции и трейдинг.** В области инвестиционного анализа объяснимость играет роль фактора доверия между системой и пользователем. Алгоритмы, предлагающие решения по инвестициям, должны объяснять свои рекомендации, чтобы убедить инвесторов им следовать. Инвесторы, принимающие решения на основе ИИ-рекомендаций, должны понимать, какие макроэкономические или рыночные сигналы лежат в основе прогноза. Объяснение может быть в форме: «Мы рекомендуем сократить долю акций в портфеле, т. к. обнаружены тревожные сигналы, например, рост инфляции, снижение прибыли компаний». Такие меры позволят не только обосновывать решения, но и снизить регуляторные и репутационные риски.

**Анализ рисков и обнаружение мошенничества.** Сложность современных финансовых транзакций и постоянная адаптация мошеннических схем требуют объяснимых решений. Здесь ХАИ используется как инструмент для проверки корректности работы моделей специалистами-экспертами. Объяснимость помогает понять, почему система отнесла транзакцию к подозрительным: например, из-за необычного географического региона или превы-

шения лимита операции. Это позволяет отличать реальные угрозы от ложноположительных срабатываний и снижает операционные издержки.

Использование ХАИ в страховой аналитике и управлении рисками также способствует повышению качества взаимодействия между алгоритмами и экспертным знанием, обеспечивая возможность коррекции и дообучения моделей, получения дополнительного рыночного преимущества [34].

### 3.2. Промышленность

Другой ключевой сферой внедрения ХАИ является промышленность и так называемые «умные предприятия», где ИИ используется для прогнозирования сбоев оборудования, оптимизации качества продукции и управления цепочками поставок. Для промышленности в условиях Индустрии 5.0 критически важным становится не только предсказание событий, но и объяснение причин, стоящих за рекомендациями алгоритмов. Это позволяет инженерам и операторам доверять решениям и действовать на их основе.

**Прогностическое/предиктивное обслуживание.** Традиционные подходы обслуживания на производстве сталкиваются с целым рядом проблем. Алгоритмы нередко выдают многочисленные ложные тревоги, не объясняя их происхождения, что приводит к избыточным проверкам и простоям. Кроме того, использование данных от множества датчиков осложняется их динамикой: после ремонта или модернизации оборудование меняет поведение, что снижает точность прогнозов и вызывает эффект дрейфа данных. Вдобавок операторы могут получать «черные» сигналы без конкретного объяснения, какие именно показатели вызвали тревогу и какие действия следует предпринять.

Исследование Watanabe и соавторы показывает, что обобщенные аддитивные модели с ограничениями (GA2M+) позволяют сочетать высокую предсказательную способность с структурой, более понятной инженерам и согласованной с физической логикой процесса [35]. Применение техник атрибуции риска во временных рядах делает возможным анализ вклада отдельных факторов в заданном окне наблюдения. Построение суррогатных деревьев правил (упрощенные интерпретируемые модели, которые строятся поверх «черного ящика») позволяет трансформировать сложные предсказания в простые и понятные операторам интерпретации, а

контрфактуальные объяснения показывают, какие изменения параметров могут снизить вероятность отказа до приемлемого уровня. Практическим примером такого подхода является система IBM Maximo Predict [36], которая использует искусственный интеллект и данные с сенсоров, отчеты о техническом обслуживании и историю поломок для прогнозирования отказов оборудования и предоставляет объяснения специалистам чтобы интерпретировать прогнозы работы системы. Как утверждают Негма́на и соавторы в своем исследовании [37], включение SHAP-анализа и интерпретируемых моделей в системы предиктивного обслуживания дает сокращение ложноположительных сигналов более чем на 90%, повышая доверие инженеров и эффективность операций.

**Контроль качества продукции.** Не менее важным направлением применения ХАИ становится контроль качества продукции. Алгоритмы компьютерного зрения все чаще используются для выявления дефектов на производственных линиях, однако традиционные модели ограничиваются бинарной классификацией без указания причин. Это снижает доверие операторов и затрудняет поиск источников брака. Методы интерпретации, такие как LIME или SHAP, позволяют визуализировать области изображения, которые стали определяющими для классификации. Таким образом инженеры получают возможность не только доверять системе, но и быстрее выявлять первопричины дефектов, что ускоряет локализацию брака и способствует более активному принятию автоматизированного контроля качества.

**Логистика.** Сферы логистики и цепочки поставок выигрывают от ИИ при оптимизации маршрутов, распределения ресурсов и управления складом. При этом алгоритмы оптимизации запасов и маршрутов доставки часто воспринимаются управленцами как «черные ящики», и это снижает готовность внедрять рекомендованные ими стратегии. Объяснимость позволяет преодолеть этот барьер. Системы, которые демонстрируют, какие факторы (рост спроса, задержка у конкретного поставщика или изменение транспортных расходов) повлияли на выбор стратегии, вызывают больше доверия и обеспечивают лучшее согласование решений между человеком и алгоритмом. Эксперименты на основе собранного датасета показывают, что использование SHAP и LIME может повысить прозрачность моделей и доверие к принятию ИИ-рекомендаций в управлении логистикой и запасами [38].

Концепцию индустриального ХАИ для производственных процессов активно продвигает компания Siemens. В своем техническом отчете [39] компания подчеркивает важность объяснимости как стандарта промышленного AI, утверждая, что объяснимость стала ключевым требованием на всех этапах жизненного цикла промышленных AI-систем — от формулировки бизнес-задачи до мониторинга и поддержки в эксплуатации. Этот пример демонстрирует растущую роль ХАИ как обязательного элемента для прозрачности и управления системами Индустрии 5.0.

Спектр отраслей, где ИИ применим, гораздо шире. Помимо бизнеса, концепции прозрачности и объяснимого искусственного интеллекта находят все более широкое применение также в социальных сферах — медицине, политике, праве и государственном управлении, где цена решений особенно высока и доверие общества критически важно. Многие сферы пока еще не сфокусированы на использовании объяснимого ИИ, особенно там, где проникновение искусственного интеллекта ниже. Например, агросектор (прогнозы урожаев, управление техникой), энергетика (оптимизация сетей), индустрия развлечений (где тоже важно понимать предпочтения аудитории), культура и искусство — эти направления остаются относительно менее изученными с точки зрения ХАИ.

#### **4. Барьеры внедрения ХАИ: экономические, технические и организационные**

Несмотря на очевидные преимущества ХАИ, связанные с ростом доверия к системам и снижением рисков, на сегодняшний день его широкое внедрение в бизнесе сталкивается с рядом барьеров. В бизнес-среде всегда возникает вопрос: «Какова отдача от инвестиций?». Если она неочевидна и не приводит напрямую к росту прибыли, некоторые могут считать это необязательной опцией. Чтобы убедить руководство или инвесторов, нужно наглядно показать эффект: растет ли лояльность клиентов, насколько снижается число ошибок, удастся ли выполнять требования регуляторов с меньшими затратами? Для таких выводов необходима статистика, но по сравнению с классическими внедрениями опубликованных кейсов с измерением эффекта ХАИ пока относительно немного.

Для уточнения и эмпирического подтверждения ключевых барьеров внедрения объяснимого искусственного интеллекта (ХАИ) в бизнес- и индустриальной среде был проведен целевой библиографический анализ научных публикаций (табл. 2). В анализ включались только те работы, в которых ХАИ рассматривается не как абстрактная техническая концепция, а в контексте реального применения в организациях, промышленности, цифровом производстве, корпоративном управлении, финансовом секторе или регулировании высокорисковых ИИ-систем.

Ключевым барьером, который наиболее часто упоминается в исследованиях, являются организационные сложности. Человеческий фактор, основанный на привычке доверять личной экспертизе и интуиции, нередко становится причиной организационного сопротивления внедрению ХАИ. Не все организации готовы принимать «совет от машины». Возникает сомнение, будут ли сотрудники доверять знаниям, полученным от ИИ, без должного уровня объяснимости. Кроме того, для эффективной работы нужны новые роли от экспертов по интерпретации до разработчиков объяснений. Необходимы инвестиции в обучение персонала, чтобы сотрудники научились воспринимать ХАИ как полезного помощника, а не как угрозу своему месту.

Сопротивление может вызывать замедление процессов принятия решений. Объяснимые модели требуют времени на ознакомление и интерпретацию, что контрастирует с устремленностью бизнеса к скорости и оптимальности. Внедрение ХАИ

способно снижать оперативность, если не будет грамотно встроено в рабочий процесс. Здесь нужны решения, которые экономят время: например, сокращают число совещаний, поскольку все участники сразу понимают логику алгоритма и меньше обсуждают его результаты, споря над их корректностью и прозрачностью.

К еще одному барьеру можно отнести недостаточную персонализацию объяснений. Современные ХАИ-системы обычно выдают шаблонные объяснения, не учитывая уровень компетенции пользователя, его задачу или контекст. В результате объяснение может оказаться слишком сложным для одних или слишком упрощенным для других. В научных исследованиях уже разрабатываются подходы к адаптивным объяснениям, где система оценивает, понял ли пользователь предыдущий ответ, и при необходимости упрощают или детализируют объяснение. Однако такие методы пока не используются масштабно.

Еще одним ключевым барьером являются технические ограничения и нагрузка на ресурсы. Большинство методов ХАИ по-прежнему находятся в статусе исследовательских прототипов: реализуются как скрипты или ноутбуки, и их сложно интегрировать в промышленные системы. Часто они работают медленно, требуют доступа к внутренней структуре модели или сильно нагружают сервер. Например, метод LIME для формирования одного объяснения требует сотен или тысяч прогонов модели, что требует значительные вычислительные мощности [40]. В реальных условиях инженерные

Таблица 2.

### Классификация барьеров ХАИ

Тип барьера, выявленный на основе библиографического анализа	Упоминание в источниках
1. Технические ограничения (сложность интеграции, отсутствие стандартов, низкая производительность ХАИ-методов)	[7], [8], [9], [10], [12], [25], [27], [28], [36], [39], [40], [43], [49], [58]
2. Организационные сложности (нехватка компетенций, необходимость обучения, изменения процессов)	[6], [7], [8], [9], [11], [12], [17], [18], [19], [29], [36], [39], [47], [48], [43], [54], [55]
3. Экономические барьеры (стоимость внедрения, ROI, ресурсы)	[7], [8], [9], [12], [17], [18], [19], [29], [34], [36], [39], [42]
4. Регуляторные/комплаенс барьеры	[3], [4], [10], [11], [12], [15], [16], [29], [34], [36], [39], [41], [42]
5. Пользовательские/человеческие факторы (доверие, когнитивная нагрузка, неинтуитивные объяснения)	[6], [8], [9], [11], [17], [18], [19], [29], [36], [39], [47], [48], [49], [54], [55]

команды вынуждены искать компромиссы: кэширование, приближенные вычисления и оптимизация модулей ХАИ, чтобы объяснения выдавались в реальном времени или хотя бы в приемлемые сроки. Пользователь не будет долго ждать, пока система «думает» над объяснением. К тому же на практике отсутствуют общепринятые индустриальные стандарты форматов представления объяснений и единой платформы, поддерживающей все модели «из коробки», поэтому компании часто реализуют ХАИ-решения самостоятельно, под свои задачи. Это означает, что каждая компания тратит собственное время и ресурсы на индивидуальную реализацию, что препятствует масштабированию решений.

Высокая стоимость и сложность внедрения также является ограничением. Объяснимость часто рассматривается как дополнительный модуль, требующий адаптации интерфейсов, бизнес-процессов и серьезной подготовки команды. Например, в финансовых институтах требуется не просто наладить выдачу объяснений скоринговой модели, но и обучить сотрудников, обновить клиентские интерфейсы и обеспечить грамотное представление выводов. Эти расходы сами по себе могут ограничивать внедрения, особенно если ХАИ не является требованием регуляторов или отраслевых стандартов. Поэтому даже там, где значимость объяснимости признают, ХАИ может восприниматься как второстепенная функция, а не необходимая инвестиция.



Рис. 3. Карта ключевых рисков.

Перечисленные барьеры можно представить на карте рисков, оценить их вероятность и возможный негативный эффект. На рисунке 3 авторами сформулированы ключевые риски со средней или высокой вероятностью возникновения, с наиболее серьезными и значимыми последствиями в части экономической, технической и организационной сложности внедрения и использования ХАИ в бизнесе. В таблице 3 представлены митигирующие меры.

Таблица 3.

**Митигирующие меры**

Ключевые риски	Митигирующие меры
1. Неочевидный возврат инвестиций, нехватка бюджета, доп. затраты на интеграцию ХАИ	Дашборды и панели мониторинга ХАИ, КПЭ (KPI) объяснимости, поэтапные пилоты; стоп-критерии
2. Апгрейд инфраструктуры, высокая вычислительная стоимость ХАИ методов	Аппроксимации/кэш; предвычисления, гибридные модели: быстрые правила + асинхронные объяснения
3. Сложная интеграция результатов ХАИ в архитектуру, отсутствие единых стандартов	Интерфейсы объяснений ХАИ с REST API и JSON-выводом, решения с модульной архитектурой
4. Сопротивление изменениям, нехватка компетенций в ХАИ, сложные интерфейсы	Обучающие материалы (ХАИ гайды, тренинги); двухуровневые объяснения, UX-UI протоколы
5. Риски дискриминации, предвзятости, манипулятивные или неполные объяснения, подрыв доверия	Многоуровневый bias-аудит удаление/ограничение чувствительных признаков, контрафактуальные тесты, регулярный мониторинг дрейфа, контроль срезов
6. Утечка конфиденциальной информации через объяснения, отсутствие ответственности	Контроль детализации и доступа к объяснениям, встроенные политики фильтрации выводов, логи

Для преодоления рисков и барьеров необходимо не только совершенствовать технологии и архитектуры, но и сформировать новые стандарты, обучить кадры, и адаптировать бизнес-процессы, разработать специализированные артефакты, которые мы опишем далее.

## 5. Будущее внедрения ХАИ в ИИ решения бизнеса

Индустрия 5.0 практически требует объяснимого ИИ как стандарт: это создает условия для партнерства человека и ИИ, о котором часто говорят применительно к Индустрии 5.0. Например, в производстве будущего оператор высокой квалификации будет совместно с ИИ принимать решения: AI предложит оптимизацию или выявит проблему, объяснит свою логику, а оператор, поняв ее, утвердит действие или отклонит, внося человеческий фактор, креативность, интуицию, ответственность. Такая синергия возможна только при мощной инфраструктуре ХАИ, поддерживаемой организационными изменениями.

### 5.1. Разработка нормативных требований

Регуляторное давление остается одним из наиболее мощных драйверов развития ХАИ в бизнесе. Уже сейчас Европа обсуждает регламент AI Act с акцентом на объяснимость [41], согласно которому прозрачность, проверяемость и объяснимость становятся базовыми требованиями для высокорисковых ИИ-систем — от финансовых моделей до промышленных решений. Законодатели постепенно формулируют требования к объяснимости алгоритмов — особенно в критичных отраслях. В финансовом секторе, к примеру, могут появиться правила, требующие того, чтобы все модели принятия решений раскрывали клиенту ключевые факторы, на которых основано решение. Будущие скоринговые и торговые системы должны также соответствовать регуляторным ожиданиям: демонстрировать прозрачность, избегать дискриминации, раскрывать все риски. Перспективно появление самодиагностических алгоритмов, которые не только объясняют свои результаты, но и проверяют себя на наличие запрещенных зависимостей, встраивают автоматическую генерацию объяснений в отчетность.

### 5.2. Человеко-машинный симбиоз

Современные ХАИ-системы должны понимать, в каком контексте они работают, и адаптировать объяснения под отраслевую логику. Это требует сотрудничества с предметными экспертами и использования доменных знаний, таких как онтологии и правила.

Ключевые работы, в частности d'Avila Garcez и соавторы, и Besold и соавторы предлагают способы встраивания логических рассуждений в нейросети (neuro-symbolic AI methods), объединяющие дедуктивное логическое мышление и глубокое обучение [51, 52]. Эти методы позволяют моделям опираться на доменные знания, правила и онтологии, что повышает интерпретируемость и документируемость рассуждений.

В дальнейшем можно ожидать развитие систем поддержки бизнес-решений, которые вместе с прогнозом (например, риск срыва сделки) будут выдавать читабельный довод, ссылаясь на аналогичные сделки или статистические данные, и указывая на какие факты они опираются.

Исследования в области интерфейсов объяснений, проведенные Kim и соавторы, Rong и соавторы [54, 55], подтверждают важность адаптации объяснений под пользователя. На практике это может означать, что появятся интерактивные панели для менеджеров, где прогнозы рынка, рекомендации по стратегии будут сопровождаться диаграммами, показывающими, какие предпосылки к этому привели. Будет уделено внимание тому, чтобы такие панели соответствовали мышлению менеджеров и были дополнены подходящей им визуализацией ключевых допущений, диаграммами, абстракциями.

### 5.3. Внедрение ХАИ и обмен знаниями в компании

Все объяснения, собранные вместе, могут выявить широкую картину, стать отправной точкой для управленческих выводов верхнего уровня. То есть, ХАИ как бы будет ускорять обратную связь снизу вверх: вместо долгих докладов от менеджеров среднего звена, сводки объяснений от ИИ дадут верхушке быстрое понимание, что происходит и почему. Не исключено появление методов, которые смогут дообучаться на культуре компании и учитывать специфику бизнеса: обнаруживать, ка-

кие объяснения руководство сочтет понятными, и формулировать предложения в привычном стиле.

Сложные управленческие ситуации могут потребовать не одного, а нескольких сценариев с объяснениями. ИИ может предложить несколько альтернатив, обосновать каждую, либо предложить разбить задачу на подзадачи, объяснив, почему так эффективнее. Таким образом, формируется полнопроцессное объяснение: охватывающее все шаги — от постановки задачи до финального выбора. При этом важно найти грань, где объяснение не переходит в искажение фактов. Возможно, хорошо себя проявит концепция многоуровневого объяснения: краткое (упрощенное) — для первого ознакомления, и подробное — для проверки.

Предлагаемые в работе артефакты ХАИ (табл. 4) являются результатом анализа существующих практик документации, мониторинга и UX-подходов, а также дополняют их новыми элементами, учитывающими потребности бизнеса и выявленные риски, представляя оригинальный вклад авторов в развитие темы ХАИ в организационном контексте.

Исследования подчеркивают важность документации моделей как инструмента прозрачности и передачи знаний между командами. Так, Mitchell и соавторы [56] отмечают важность таких документов интерпретации моделей, как карта ИИ модели «Model Cards», в то время как Gebru и соавто-

ры [57], аргументируют пользу от использования технических паспортов «Data Sheets» в которых описаны цели использования, источники данных, выявленные ограничения и способы генерации объяснений. Их значение особенно возрастает в крупной организации, где передача знаний между командами не может опираться лишь на устные инструкции или код в репозиториях. Подобная документация позволяет быстрее адаптировать новые модели, особенно в условиях смены персонала или масштабирования решений.

Не менее важным становится вопрос интерфейса взаимодействия. Если раньше объяснение результата модели можно было получить только через специализированные инструменты анализа, то теперь объяснения все чаще встраиваются прямо в рабочие приложения. Так, аналитик или менеджер может на основе результата прогноза, получить комментарий, который пояснит, какие факторы повлияли на решение, и насколько оно отличается от нормы. Причем формат объяснения может адаптироваться под пользователя — от краткого бизнес-резюме до технической расшифровки. Такие интерфейсы значительно повышают принятие ИИ в корпоративной среде, особенно в условиях ограниченного времени на принятие решений.

Для стратегического уровня эксплуатации моделей могут применяться визуальные панели мони-

Таблица 4.

#### Ключевые артефакты ХАИ для бизнеса

Артефакт	Цель / Функция	Пользователь
ХАИ-документация и протоколы (Model Cards, Datasheets)	Формализованное описание модели, данных и ограничений для прозрачности и аудита	Разработчики моделей, аудиторы, регуляторы
Интерфейсы объяснений (Explainability UI/UX)	Интерактивный доступ к объяснениям внутри пользовательского интерфейса	Конечные пользователи, аналитики, операторы
Дашборды и панели мониторинга ХАИ	Визуализация факторов, повлиявших на прогноз, для менеджеров и принятия решений	Менеджеры среднего и высшего звена
Автоматизированные фреймворки валидации ХАИ	Автоматическая проверка качества объяснений и отклонений от норм	Инженеры по качеству, отдел рисков, внутренний аудит
Обучающие и сопровождающие материалы (ХАИ-гайды, тренинги)	Поддержка персонала в освоении ХАИ через обучающие курсы и гайды	Менеджеры, бизнес-аналитики, специалисты по обучению

торинга и отчётные представления, позволяющие регулярно отслеживать распределения ключевых признаков и предсказаний, выявлять data/target drift и автоматически генерировать оповещения при превышении порогов. Такой мониторинг поддерживает своевременное выявление деградации качества и принятие корректирующих мер (например, устранение проблем качества данных или запуск переобучения модели [58]).

Особое внимание уделяется автоматизированной валидации объяснений. В организациях, где важно соблюдение нормативных требований (например, в банковской или медицинской сфере), ручная проверка каждого объяснения невозможна. Поэтому разрабатываются фреймворки, которые автоматически анализируют объяснения на предмет соответствия политике, отсутствия дискриминационных признаков или неучтенных рисков. Это превращает ХАИ в элемент системы качества, а не просто визуальный слой.

Наконец, устойчивое внедрение ХАИ невозможно без обучающих и сопровождающих материалов, доступных не только разработчикам, но и широкому кругу сотрудников. Такие материалы включают руководства, тренинги, пошаговые инструкции по взаимодействию с системами ХАИ. Их цель – снизить порог входа и обеспечить грамотное использование инструментов, особенно среди тех, кто отвечает за интерпретацию данных, но не обладает технической подготовкой. Согласно Donoso-Guzmán и соавторы [59], перспективными являются human-centered подходы к оценке ХАИ, учитывающие цели и контекст разных пользовательских ролей. Такие рамки оценки могут быть использованы и для настройки объяснений под корпоративные практики и ожидания руководства. Они способны предсказывать, какие объяснения будут восприняты как убедительные, какие – как избыточные, и как лучше структурировать аргументацию для различных ролей.

Таким образом, артефакты ХАИ не просто объясняют результат, но помогают встроить ИИ в корпоративное мышление, делая его понятным, доступным и управляемым инструментом.

Обозначенные перспективы представляют собой план развития ХАИ на ближайшие годы. Главная тенденция – углубление интеграции ИИ и челове-

ческого фактора: от узкого инструмента объяснения одного прогноза к широкому человеко-ориентированному интеллекту, который становится частью коллективного процесса принятия решений. Иными словами, ХАИ эволюционирует из простого модуля интерпретации в концепцию построения таких ИИ-систем, которые изначально проектируются для совместной работы с человеком.

### Заключение

Объяснимость ИИ приобрела критическое значение в бизнесе, промышленности и управлении, где на кону стоят реальные деньги, безопасность и ответственность перед людьми. Она становится фактором конкурентоспособности и нормативного соответствия: организации, способные объяснить действия своих алгоритмов, имеют больше шансов выдержать проверку регуляторов и получить одобрение своих клиентов. Уже сегодня ХАИ помогают облегчить интеграцию ИИ в умные предприятия и финансовый сектор, что положительно влияет на перспективы Индустрии 5.0, где значимую роль играет человеко-ориентированность, безопасность и устойчивое развитие.

Объяснимый ИИ помогает компаниям соответствовать этим требованиям – предоставляя инструменты мониторинга алгоритмов, отчетности по предпосылкам решений, контроля отсутствия дискриминации. В обозримом будущем ХАИ может стать частью системы качества предприятия – так же как сегодня существуют стандарты ISO для процессов, могут появиться стандарты на объяснимость и этичность AI-компонентов бизнес-процессов.

Руководители, вооруженные алгоритмами, которые могут объяснить свое решение, получают инструмент, сочетающий силу данных и моделей с понятностью и логичностью традиционного анализа. Это позволит принимать более обоснованные и в то же время инновационные решения, поскольку ИИ способен обнаружить нетривиальные закономерности, а через объяснения – сделать их приемлемыми для реализации. Новые формы обучения в организациях, основанные на взаимодействии с ХАИ, помогут ускорить распространение лучших практик и знаний.

Важно признать, что широкое внедрение ХАИ пока сдерживается экономическими, технически-

ми и организационными барьерами. Наиболее высокие шансы развития ХАИ там, где совпадают регуляторное давление, высокая цена ошибки и наличие данных/процессной дисциплины: финансы, страхование, производство и промышленность.

Там, где организации целенаправленно выделяют инвестиции и встраивают объяснимость в архитектуру решений, процессы качества и пользовательский опыт, объясняемый ИИ даст устойчивое конкурентное преимущество. ■

### Литература

1. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning / W. Samek [et al.] // *Lecture Notes in Artificial Intelligence*. 2019. Vol. 11700. Springer International Publishing. <https://doi.org/10.1007/978-3-030-28954-6>
2. Vilone G., Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence // *Information Fusion*. 2021. Vol. 76. P. 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
3. IEEE Std 7001-2021. IEEE Standard for Transparency of Autonomous Systems. IEEE, 2022. <https://doi.org/10.1109/IEEESTD.2022.9726144>
4. ISO/IEC TR 24028:2020. Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence. Geneva: International Organization for Standardization, 2020.
5. Du M., Liu N., Hu X. Techniques for interpretable machine learning // *Communications of the ACM*. 2020. Vol. 63. No. 1. P. 68–77. <https://doi.org/10.1145/3359786>
6. Gamoura S. C. Explainable AI (XAI) for AI-acceptability: The coming age of digital management 5.0 // 2023 IEEE International Conference on Networking, Sensing and Control (ICNSC). 2023. P. 1–6. <https://doi.org/10.1109/ICNSC58704.2023.10319030>
7. Khan A., Jhanjhi N. Z., Hamid D. H. T. B. A. H., Omar H. A. H. The need for explainable AI in Industry 5.0 // *Advances in Explainable AI Applications for Smart Cities*. IGI Global, 2024. P. 1–30. <https://doi.org/10.4018/978-1-6684-6361-1.ch001>
8. Chang T.-S., Bau D.-Y. eXplainable Artificial Intelligence (XAI) in business management research: A success/failure system perspective // *Journal of Electronic Business & Digital Economics*. 2024. Vol. 4. No. 1. P. 36–53. <https://doi.org/10.1108/JEBDE-07-2024-0019>
9. Tchunte D., Lonlac J., Kamsu-Foguem B. A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications // *Computers in Industry*. 2024. Vol. 155. Article 104044. <https://doi.org/10.1016/j.compind.2023.104044>
10. Martins T., de Almeida A. M., Cardoso E., Nunes L. Explainable AI (XAI): A systematic literature review on taxonomies and applications in finance // *IEEE Access*. 2024. <https://doi.org/10.1109/ACCESS.2023.3347028>
11. European Commission: Directorate-General for Research and Innovation. Industry 5.0: Towards a sustainable, human-centric and resilient European industry // *Publications Office of the European Union*. 2021. <https://data.europa.eu/doi/10.2777/308407>
12. Ahmed S., Jeon G., Piccialli F. From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where // *IEEE Transactions on Industrial Informatics*. 2022. Vol. 18. No. 8. P. 5031–5042. <https://doi.org/10.1109/TII.2022.3146552>
13. Adadi A., Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI) // *IEEE Access*. 2018. Vol. 6. P. 52138–52160. <https://doi.org/10.1109/access.2018.2870052>
14. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI / A. A. Barredo [et al.] // *Information Fusion*. 2020. Vol. 58. P. 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
15. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) // *Official Journal of the European Union*. 2016. Vol. L119. P. 1–88. [Электронный ресурс]: [http://data.europa.eu/eli/reg/2016/679/oj\(дата обращения 06.02.2026\)](http://data.europa.eu/eli/reg/2016/679/oj(дата%20обращения%2006.02.2026)).
16. NIST AI 100-1. Artificial intelligence risk management framework (AI RMF 1.0). National Institute of Standards and Technology, 2023. <https://doi.org/10.6028/nist.ai.100-1>
17. Zavodna L. S., Überwimmer M., Frankus E. Barriers to the implementation of artificial intelligence in small and medium sized enterprises: Pilot study // *Journal of Economics & Management*. 2024. Vol. 46. No. 1. P. 331–352. <https://doi.org/10.22367/jem.2024.46.13>
18. Weitz K., Dang C. T., André E. Do we need explainable AI in companies? Investigation of challenges, expectations, and chances from employees' perspective // *arXiv:2210.03527*. 2022. <https://doi.org/10.48550/arXiv.2210.03527>
19. Darvish M., Kret K. S., Bick M. An explorative study on the adoption of explainable artificial intelligence (XAI) in business organizations // *Disruptive Innovation in a Digitally Connected Healthy World*. *Lecture Notes in Computer Science*. 2024. Vol. 14907. P. 29–40. Springer, Cham. [https://doi.org/10.1007/978-3-031-72234-9\\_3](https://doi.org/10.1007/978-3-031-72234-9_3)
20. Joyce D. W., Kormilitzin A., Smith K. A., Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability // *Npj Digital Medicine*. 2023. Vol. 6. No. 1. Article 6. <https://doi.org/10.1038/s41746-023-00751-9>

21. Murdoch W. J., Singh C., Kumbier K., Abbasi-Asl R., Yu B. Definitions, methods, and applications in interpretable machine learning // *Proceedings of the National Academy of Sciences (PNAS)*. 2019. Vol. 116. No. 44. P. 22071–22080. <https://doi.org/10.1073/pnas.1900654116>
22. Samek W., Wiegand T., Müller K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models // *arXiv:1708.08296*. 2017. <https://doi.org/10.48550/arXiv.1708.08296>
23. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning // *arXiv:1702.08608*. 2017. <https://doi.org/10.48550/arXiv.1702.08608>
24. Yuan H., Yang F., Du M., Ji S., Hu X. Towards structured NLP interpretation via graph explainers // *Applied AI Letters*. 2021. Vol. 2. No. 4. Article e58. <https://doi.org/10.1002/ail.2.58>
25. Dumka A., Chaudhari V., Bisht A. K., Rawat R., Pandey A. Methods, techniques, and application of explainable artificial intelligence // *Reshaping Environmental Science Through Machine Learning and IoT*. 2024. P. 337–354. IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-2351-9.ch017>
26. Wachter S., Mittelstadt B., Russell C. Counterfactual explanations without opening the black box // *Harvard Journal of Law & Technology*. 2018. Vol. 31. No. 2. P. 841–887. <https://doi.org/10.2139/ssrn.3063289>
27. Dixit M., Kansal I., Khullar V., Kumar R., Kumar S. Analyzing trustworthiness and explainability in artificial intelligence: A comprehensive review // *Recent Advances in Electrical & Electronic Engineering*. 2025. Vol. 18. No. 8. Article e040724231621. <https://doi.org/10.2174/0123520965308169240616144800>
28. Vasanth S., Keerthana S., Saravanan G. Demystifying AI: A robust and comprehensive approach to explainable AI // *2024 International Conference on Electronics and Communication*. 2024. <https://doi.org/10.1109/ICECS9683.2024.10837078>
29. Авдошин С. М., Песоцкая Е. Ю. Доверенный искусственный интеллект как способ цифровой защиты // *Бизнес-информатика*. 2022. Т. 16. № 2. С. 62–73. <https://doi.org/10.17323/2587-814X.2022.2.62.73>
30. Poyiadzi R., Sokol K., Santos-Rodriguez R., De Bie T., Flach P. FACE: Feasible and actionable counterfactual explanations // *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020. P. 344–350. <https://doi.org/10.1145/3375627.3375850>
31. Gramegna A., Scardapane P. L. L., Giudici P. SHAP and LIME: An evaluation of discriminative power in credit risk // *Frontiers in Artificial Intelligence*. 2021. Vol. 4. Article 752558. <https://doi.org/10.3389/frai.2021.752558>
32. Hjelkrem L. O., de Lange P. E. Explaining deep learning models for credit scoring with SHAP: A case study using open banking data // *Journal of Risk and Financial Management*. 2023. Vol. 16. No. 4. Article 221. <https://doi.org/10.3390/jrfm16040221>
33. Li Y., Simon Z., Turkington D. Investable and interpretable machine learning for equities // *Journal of Financial Data Science*. 2022. Vol. 4. No. 1. P. 54–74. <https://doi.org/10.3905/jfds.2021.1.084>
34. Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. Financial Risk Management and Explainable, Trustworthy, Responsible AI // *Frontiers in Artificial Intelligence*. 2022. Vol. 5. Article 779799. <https://doi.org/10.3389/frai.2022.779799>
35. Constrained Generalized Additive 2 Model with consideration of high-order interactions (CGA2M+) / A. Watanabe [et al.] // *arXiv:2106.02836*. 2021. <https://doi.org/10.48550/arXiv.2106.02836>
36. IBM Maximo Predict. IBM Documentation // IBM, 2023. [Электронный ресурс]: <https://www.ibm.com/docs/en/mhmpmh-and-p-u/cd?topic=overview-maximo-predict> (дата обращения 06.02.2026).
37. Sensor-based predictive maintenance with reduction of false alarms – A case study in heavy industry / M. Hermansa [et al.] // *Sensors*. 2022. Vol. 22. No. 1. Article 226. <https://doi.org/10.3390/s22010226>
38. Kilari S. D. The role of explainable AI (XAI) in improving transparency and trust in supply chain demand and price forecasting models // *SSRN preprint*. 2023. <https://doi.org/10.2139/ssrn.5357669>
39. The rise of industrial explainable artificial intelligence (XAI) – Insights across the AI life cycle. White Paper // Siemens. 2023. [Электронный ресурс]: <https://assets.new.siemens.com/siemens/assets/api/uuid:3b4de373-57e2-4329-b025-2825db0172aa/WhitpaperXAI.pdf> (дата обращения 06.02.2026).
40. The cost of understanding – XAI algorithms towards sustainable ML in the view of computational cost / C. Jean-Quartier [et al.] // *Computation*. 2023. Vol. 11. No. 5. Article 92. <https://doi.org/10.3390/computation11050092>
41. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) // *Official Journal of the European Union*. 2024. Vol. L257. P. 1–64. [Электронный ресурс]: <https://data.europa.eu/eli/reg/2024/1689/oj> (дата обращения 06.02.2026).
42. Černevičienė J., Kabasinskas A. Explainable artificial intelligence (XAI) in finance: a systematic literature review // *Artificial Intelligence Review*. 2024. Vol. 57. No. 8. Article 216. <https://doi.org/10.1007/s10462-024-10854-8>
43. Explainable artificial intelligence in information systems: A review of the status quo and future research directions / J. Brasse [et al.] // *Electronic Markets*. 2023. Vol. 33. Article 26. <https://doi.org/10.1007/s12525-023-00644-5>
44. Carvalho D. V., Pereira E. M., Cardoso J. S. Machine learning interpretability: A survey on methods and metrics // *Electronics*. 2019. Vol. 8. No. 8. Article 832. <https://doi.org/10.3390/electronics8080832>

45. Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 3rd ed. 2025. (Online book). [Электронный ресурс]: <https://christophm.github.io/interpretable-ml-book/> (дата обращения 06.02.2026).
46. Angelov P., Soares E., Jiang R., Arnold N., Atkinson P. Explainable artificial intelligence: An analytical review // *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2021. Vol. 11. No. 5. Article e1424. <https://doi.org/10.1002/widm.1424>
47. Rai A. Explainable AI: From black box to glass box // *Journal of the Academy of Marketing Science*. 2020. Vol. 48. P. 24–48. <https://doi.org/10.1007/s11747-019-00710-5>
48. Liao Q. V., Varshney K. R. Human-centered explainable AI (XAI): From algorithms to user experiences // *arXiv:2110.10790*. 2021. <https://doi.org/10.48550/arXiv.2110.10790>
49. Chamola V., Hassija V., Sulthana A. R., Ghosh D., Dhingra D., Sikdar B. A review of trustworthy and explainable artificial intelligence (XAI) // *IEEE Access*. 2023. Vol. 11. P. 78994–79015. <https://doi.org/10.1109/access.2023.3294569>
50. Belle V., Papantonis I. Principles and Practice of Explainable Machine Learning // *Frontiers in Big Data*. 2021. Vol. 4. Article 688969. <https://doi.org/10.3389/fdata.2021.688969>
51. d’Avila Garcez A., Lamb L. C., Gabbay D. Neural-Symbolic Learning Systems // *Perspectives in Neural Computing*. Springer, 2002. <https://doi.org/10.1007/978-1-4471-0211-3>
52. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation / T. R. Besold [et al.] // *arXiv:1711.03902*. 2017. <https://doi.org/10.48550/arXiv.1711.03902>
53. Yu D., Yang B., Liu D., Wang H., Pan S. A survey on neural-symbolic learning systems // *Neural Networks*. 2023. Vol. 166. P. 105–126. <https://doi.org/10.1016/j.neunet.2023.06.028>
54. Kim J., Maathuis H., Sent D. Human-centered evaluation of explainable AI applications: a systematic review // *Frontiers in Artificial Intelligence*. 2024. Vol. 7. Article 1456486. <https://doi.org/10.3389/frai.2024.1456486>
55. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations / Y. Rong [et al.] // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2024. <https://doi.org/10.1109/TPAMI.2023.3331846>
56. Model cards for model reporting / M. Mitchell [et al.] // *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019. P. 220–229. <https://doi.org/10.1145/3287560.3287596>
57. Datasheets for Datasets / T. Gebru [et al.] // *Communications of the ACM*. 2021. Vol. 64. No. 12. P. 86–92. <https://doi.org/10.1145/3458723>
58. Vadapalli S. R. Monitoring the performance of machine learning models in production // *International Journal of Computer Trends and Technology*. 2022. Vol. 70. No. (9) P. 38–42. <https://doi.org/10.14445/22312803/IJCTT-V70I9P105>
59. Donoso-Guzmán I., Ooge J., Parra D., Verbert K. Towards a comprehensive human-centred evaluation framework for explainable AI // *Explainable Artificial Intelligence (xAI 2023)*. *Communications in Computer and Information Science*. 2023. Vol. 1903. Springer, Cham. [https://doi.org/10.1007/978-3-031-44070-0\\_10](https://doi.org/10.1007/978-3-031-44070-0_10)

## Об авторах

### Сергей Михайлович Авдошин

кандидат технических наук, доцент;

профессор, департамент компьютерной инженерии, Московский институт электроники и математики им. А.Н. Тихонова, Национальный исследовательский университет «Высшая школа экономики», Россия, 123458, г. Москва, ул. Таллинская, д. 34;

E-mail: savdoshin@hse.ru

ORCID: 0000-0001-8473-8077

### Елена Юрьевна Песоцкая

кандидат экономических наук;

доцент, департамент программной инженерии, Факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», Россия, 109028, г. Москва, Покровский бульвар, д. 11;

E-mail: epesotskaya@hse.ru

ORCID: 0000-0003-2129-4645

# Explainable AI for Industry 5.0: Shedding light on the black box

**Sergey Mikhailovich Avdoshin**

E-mail: savdoshin@hse.ru

**Elena Yuryevna Pesotskaya**

E-mail: epesotskaya@hse.ru

HSE University, Moscow, Russia

## Abstract

The rapid development of artificial intelligence (AI) is accompanied by increasing computational complexity and decreasing model transparency, which significantly limits its adoption in critical domains that require a high level of trust, interpretability, and justification of decisions. Under these conditions, the field of Explainable Artificial Intelligence (XAI) has gained particular importance as it focuses on approaches and technologies that enable understanding of AI system logic and interpretation of their outputs. This article examines the timely topic of implementing XAI in the context of Industry 5.0. Special attention is given to practical application scenarios: the authors present concrete industrial cases from IBM, Siemens, and other companies demonstrating how XAI contributes to enhancing the reliability, safety, efficiency, and trustworthiness of AI systems. The study includes a systematic search and analysis of the literature in this domain and proposes well-grounded key criteria for comparing existing XAI approaches. The article also outlines the advantages, current limitations, and promising directions for the development of XAI, highlighting the opportunities it opens for improving effectiveness, transparency, and trust in business.

**Keywords:** XAI, explainable artificial intelligence, Industry 5.0, machine learning, industry

**Citation:** Avdoshin, S. M., & Pesotskaya, E. Yu. (2026). Explainable AI for Industry 5.0: Shedding light on the black box. *Business Informatics*, 20(1), 7–28. <https://doi.org/10.17323/2587-814X.2026.1.7.28>

## References

1. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. (2019). In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Lecture Notes in Computer Science*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-28954-6>
2. Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
3. IEEE Std 7001-2021. (2022). IEEE Standard for Transparency of Autonomous Systems. IEEE. <https://doi.org/10.1109/IEEESTD.2022.9726144>
4. ISO/IEC TR 24028:2020. (2020). Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence. Geneva: International Organization for Standardization. <https://www.iso.org/standard/77608.html>

5. Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68–77. <https://doi.org/10.1145/3359786>
6. Gamoura, S. C. (2023). Explainable AI (XAI) for AI-Acceptability: The coming age of digital management 5.0. *2023 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, 1–6. <https://doi.org/10.1109/icnsc58704.2023.10319030>
7. Khan, A., Jhanjhi, N. Z., Hamid, D. H. T. B. A. H., & Omar, H. A. H. bin H. (2024). The need for explainable AI in Industry 5.0. *Advances in Explainable AI Applications for Smart Cities*, 1–30. <https://doi.org/10.4018/978-1-6684-6361-1.ch001>
8. Chang, T.-S., & Bau, D.-Y. (2024). eXplainable artificial intelligence (XAI) in business management research: A success/failure system perspective. *Journal of Electronic Business & Digital Economics*, 4(1), 36–53. <https://doi.org/10.1108/jebde-07-2024-0019>
9. Tchuente, D., Lonlac, J., & Kamsu-Foguem, B. (2024). A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications. *Computers in Industry*, 155, 104044. <https://doi.org/10.1016/j.compind.2023.104044>
10. Martins, T., de Almeida, A. M., Cardoso, E., & Nunes, L. (2024). Explainable Artificial Intelligence (XAI): A systematic literature review on taxonomies and applications in finance. *IEEE Access*, 12, 618–629. <https://doi.org/10.1109/access.2023.3347028>
11. European Commission: Directorate-General for Research and Innovation. (2021). *Industry 5.0: Towards a sustainable, human-centric and resilient European industry*. Publications Office of the European Union. <https://data.europa.eu/doi/10.2777/308407>
12. Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, 18(8), 5031–5042. <https://doi.org/10.1109/tii.2022.3146552>
13. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/access.2018.2870052>
14. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
15. European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119, 1–88. <http://data.europa.eu/eli/reg/2016/679/oj>
16. NIST AI 100-1. (2023). Artificial intelligence risk management framework (AI RMF 1.0). National Institute of Standards and Technology. <https://doi.org/10.6028/nist.ai.100-1>
17. Zavodna, L. S., Überwimmer, M., & Frankus, E. (2024). Barriers to the implementation of artificial intelligence in small and medium sized enterprises: Pilot study. *Journal of Economics and Management*, 46, 331–352. <https://doi.org/10.22367/jem.2024.46.13>
18. Weitz, K., Dang, C. T., & André, E. (2022). Do we need explainable AI in companies? Investigation of challenges, expectations, and chances from employees' perspective. *arXiv:2210.03527*. <https://doi.org/10.48550/arXiv.2210.03527>
19. Darvish, M., Kret, K. S., & Bick, M. (2024). An explorative study on the adoption of explainable artificial intelligence (XAI) in business organizations. In: van de Wetering, R., et al. *Disruptive Innovation in a Digitally Connected Healthy World*. Lecture Notes in Computer Science, 14907, 29–40. Springer, Cham. [https://doi.org/10.1007/978-3-031-72234-9\\_3](https://doi.org/10.1007/978-3-031-72234-9_3)
20. Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *Npj Digital Medicine*, 6(1). <https://doi.org/10.1038/s41746-023-00751-9>
21. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080. <https://doi.org/10.1073/pnas.1900654116>
22. Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv:1708.08296*. <https://doi.org/10.48550/arXiv.1708.08296>
23. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*. <https://doi.org/10.48550/arXiv.1702.08608>
24. Yuan, H., Yang, F., Du, M., Ji, S., & Hu, X. (2021). Towards structured NLP interpretation via graph explainers. *Applied AI Letters*, 2(4), e58. <https://doi.org/10.1002/aill.2.58>
25. Dumka, A., Chaudhari, V., Bisht, A. K., Rawat, R., & Pandey, A. (2024). Methods, techniques, and application of explainable artificial intelligence. In R. Gupta, A. Jain, J. Wang, & R. Pateriya (Eds.), *Reshaping Environmental Science Through Machine Learning and IoT*, 337–354. IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-2351-9.ch017>
26. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3063289>
27. Dixit, M., Kansal, I., Khullar, V., Kumar, R., & Kumar, S. (2025). Analyzing trustworthiness and explainability in artificial intelligence: A comprehensive review. *Recent Advances in Electrical & Electronic Engineering*, 18(8), article e040724231621. <https://doi.org/10.2174/0123520965308169240616144800>
28. Vasanth, S., Keerthana, S., & Saravanan, G. (2024). Demystifying AI: A robust and comprehensive approach to explainable AI. *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*, 1–5. <https://doi.org/10.1109/icec59683.2024.10837078>

29. Avdoshin, S. M., & Pesotskaya, E. Yu. (2022). Trusted artificial intelligence: Strengthening digital protection. *Business Informatics*, 16(2), 62–73. <https://doi.org/10.17323/2587-814x.2022.2.62.73>
30. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: Feasible and actionable counterfactual explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350. <https://doi.org/10.1145/3375627.3375850>
31. Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, 4, 752558. <https://doi.org/10.3389/frai.2021.752558>
32. Hjelkrem, L. O., & Lange, P. E. de. (2023). Explaining deep learning models for credit scoring with SHAP: A case study using open banking data. *Journal of Risk and Financial Management*, 16(4), 221. <https://doi.org/10.3390/jrfm16040221>
33. Li, Y., Simon, Z., & Turkington, D. (2021). Investable and interpretable machine learning for equities. *Journal of Financial Data Science*, 4(1), 54–74. <https://doi.org/10.3905/jfds.2021.1.084>
34. Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible AI. *Frontiers in Artificial Intelligence*, 5, 779799. <https://doi.org/10.3389/frai.2022.779799>
35. Watanabe, A., Kuramata, M., Majima, K., Kiyohara, H., Kondo, K., & Nakata, K. (2021). Constrained Generalized Additive 2 Model with consideration of high-order interactions (CGA2M+). *arXiv:2106.02836*. <https://doi.org/10.48550/arXiv.2106.02836>
36. IBM. (2023). *IBM Maximo Predict*. IBM Documentation. <https://www.ibm.com/docs/en/mhmpmh-and-p-u/cd?topic=overview-maximo-predict>
37. Hermansa, M., Kozielski, M., Michalak, M., Szczyrba, K., Wróbel, Ł., & Sikora, M. (2021). Sensor-based predictive maintenance with reduction of false alarms – A case study in heavy industry. *Sensors*, 22(1), 226. <https://doi.org/10.3390/s22010226>
38. Kilari, S. D. (2025). The role of explainable AI (XAI) in improving transparency and trust in supply chain demand and price forecasting models. *SSRN preprint*. <https://doi.org/10.2139/ssrn.5357669>
39. Siemens. (2023). *The rise of industrial explainable artificial intelligence (XAI) – Insights across the AI life cycle*. White Paper. <https://assets.new.siemens.com/siemens/assets/api/uuid:3b4de373-57e2-4329-b025-2825db0172aa/WhitepaperXAI.pdf>
40. Jean-Quartier, C., Bein, K., Hejny, L., Hofer, E., Holzinger, A., & Jeanquartier, F. (2023). The cost of understanding — XAI algorithms towards sustainable ML in the view of computational cost. *Computation*, 11(5), 92. <https://doi.org/10.3390/computation11050092>
41. European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L257, 1–64. <https://data.europa.eu/eli/reg/2024/1689/oj>
42. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, 57(8), 216. <https://doi.org/10.1007/s10462-024-10854-8>
43. Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, 33, 26. <https://doi.org/10.1007/s12525-023-00644-5>
44. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832. <https://doi.org/10.3390/electronics8080832>
45. Molnar C. (2025). *Interpretable machine learning. A guide for making black box models explainable*. 3rd edition. <https://christophm.github.io/interpretable-ml-book/>
46. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5), e1424. <https://doi.org/10.1002/widm.1424>
47. Rai, A. (2019). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. <https://doi.org/10.1007/s11747-019-00710-5>
48. Liao, Q. V., & Varshney, K. R. (2021) Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv:2110.10790*. <https://doi.org/10.48550/arXiv.2110.10790>
49. Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*, 11, 78994–79015. <https://doi.org/10.1109/access.2023.3294569>
50. Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4, 688969. <https://doi.org/10.3389/fdata.2021.688969>
51. d’Avila Garcez, A. S., Broda, K. B., & Gabbay, D. M. (2002). Neural-Symbolic Learning Systems. In *Perspectives in Neural Computing*. Springer. <https://doi.org/10.1007/978-1-4471-0211-3>
52. Besold, T. R., d’Avila Garcez, A. S., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kuehnberger, K.-U., Lamb, L. C., Mikkulainen, R., & Silver, D. L. (2017) Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. *arXiv:1711.03902*. <https://doi.org/10.48550/arXiv.1711.03902>
53. Yu, D., Yang, B., Liu, D., Wang, H., & Pan, S. (2023). A survey on neural-symbolic learning systems. *Neural Networks*, 166, 105–126. <https://doi.org/10.1016/j.neunet.2023.06.028>
54. Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence*, 7. <https://doi.org/10.3389/frai.2024.1456486>
55. Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4), 2104–2122. <https://doi.org/10.1109/tpami.2023.3331846>

56. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the *Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
57. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
58. Vadapalli, S. R. (2022). Monitoring the performance of machine learning models in production. *International Journal of Computer Trends and Technology*, 70(9), 38–42. <https://doi.org/10.14445/22312803/IJCTT-V70I9P10559>
59. Donoso-Guzmán, I., Ooge, J., Parra, D., & Verbert, K. (2023). Towards a comprehensive human-centred evaluation framework for explainable AI. In: Longo, L. (eds) *Explainable Artificial Intelligence (xAI 2023)*. Communications in Computer and Information Science, 1903. Springer, Cham. [https://doi.org/10.1007/978-3-031-44070-0\\_10](https://doi.org/10.1007/978-3-031-44070-0_10)

### About the authors

#### **Sergey Mikhailovich Avdoshin**

Candidate of Sciences (Technology);

Professor, School of Computer Engineering, HSE Tikhonov Moscow Institute of Electronics and Mathematics, HSE University, 34 Tallinskaya St., Moscow 123458, Russia;

E-mail: [savdoshin@hse.ru](mailto:savdoshin@hse.ru)

ORCID: 0000-0001-8473-8077

#### **Elena Yuryevna Pesotskaya**

Candidate of Sciences (Economics);

Associate Professor, School of Software Engineering, Faculty of Computer Science, HSE University, 11 Pokrovsky Blvd., Moscow 109028, Russia;

E-mail: [epesotskaya@hse.ru](mailto:epesotskaya@hse.ru)

ORCID: 0000-0003-2129-4645