УДК 004.891.3

Проблема доверия к «черному ящику» в современных медицинских системах искусственного интеллекта

Власова Вита Владимировна vlasova.vita@gmail.com

МГТУ им. Н.Э. Баумана, Москва, Россия

Аннотация. *Рассмотрена проблема доверия к медицинским системам искусственного интеллекта, функционирующим как «черный ящик» и выдающим клинически значимые рекомендации при ограниченной интерпретируемости внутренних механизмов вывода. Источники недоверия включают непрозрачность причинно-следственных связей, риск смещения* (bias) *и снижение качества, а также «разрывы ответственности» между разработчиком, медицинской организацией и врачом-пользователем. Предложена рамка доверия, объединяющая проверяемость, объяснимость* (XAI-*артефакты*), *управляемость* (*человеческий контроль и оспаривание*) *и доказательность* (*клиническая валидация и эксплуатационное сопровождение*). *Показано, что объяснения должны быть встроены в жизненный цикл медицинского ИИ — от требований к данным и документации до регуляторного соответствия и процедур внедрения. Сделан вывод о необходимости перехода от «доверия к модели» к «доверию к социотехнической системе», включающей модель, данные, интерфейс, клинический контекст и распределение ответственности.*

Ключевые слова: *медицинский искусственный интеллект, черный ящик, доверие, объяснимый ИИ, ответственность, клиническая валидация, человеческий контроль*

The problem of trust in the black box in modern medical artificial intelligence systems

Vlasova Vita Vladimirovna vlasova.vita@gmail.com

BMSTU, Moscow, Russia

Abstract. This paper addresses the trust challenge in medical artificial intelligence (AI) systems that operate as "black boxes", producing clinically relevant recommendations with limited transparency of internal decision logic. Drivers of distrust include opacity of reasoning, bias and performance degradation under dataset shift, and "responsibility gaps" between developers, healthcare institutions, and clinicians. A trust framework is proposed combining verifiability (auditability and traceability), explainability (XAI artifacts), controllability (human oversight, override and contestability), and evidential support (clinical validation and post-market monitoring). The paper argues that explanations alone do not guarantee trust; they must be integrated into the medical AI lifecycle through data and documentation requirements, regulatory compliance, and deployment procedures. Trust should therefore be targeted at the socio-technical system rather than the model alone.

Keywords: medical Artificial Intelligence, black box, trust, explainable AI, accountability, clinical validation, human oversight

Введение. Активное внедрение систем искусственного интеллекта (ИИ) в клиническую практику охватывает интерпретацию медицинских изображений, анализ сигналов, обработку электронных медицинских карт (EHR), прогнозирование исходов и поддержку выбора тактики лечения [1]. Во многих из этих задач лидирующие результаты достигаются за счет глубоких нейронных сетей, которые характеризуются высокой эффективностью, но низкой интерпретируемостью для пользователя [2].

Непрозрачность алгоритма в медицине приобретает особую значимость по причине высокой цены ошибки и необходимости клинического обоснования решений. Врач в реальном процессе лечения действует в условиях неполноты данных, конкурирующих гипотез и ограничений по времени, и потому нуждается не только в рекомендации, но и в возможности критически оценить ее уместность, сопоставить с клиническими руководствами и индивидуальными особенностями пациента [3].

Дополнительной проблемой является различие между «лабораторной достоверностью» и «достоверностью в эксплуатации». Модели могут показывать высокие показатели на ретроспективных наборах данных, но сталкиваться с дрейфом данных при переносе в другую клинику, при смене оборудования, протоколов обследования, схем кодирования или демографического состава пациентов. В этих условиях доверие не может быть статическим, оно требует поддержания через мониторинг качества, контроль смещений и управляемость обновлений.

На уровне институциональной практики доверие связано с распределением ответственности. Если врач несет ответственность за конечное решение, но не получает достаточной информации о данных обучения, ограничениях применимости и логике вывода, возникает «разрыв ответственности», который снижает готовность использовать систему. Международные рекомендации по этике и управлению ИИ в здравоохранении подчеркивают необходимость прозрачности и подотчетности на уровне всей социотехнической системы, а не только на уровне модели [4].

Цель работы — систематизировать причины недоверия к «черным ящикам» в медицинских ИИ-системах и предложить практическую рамку мер, повышающих доверие без снижения клинической полезности.

Методы и материалы; результаты. Работа выполнена в формате аналитического обзора с нормативно-ориентированным компонентом. Были изучены публикации по теме объяснимого ИИ (XAI) в здравоохранении, описывающие методы объяснений и практики их оценки, работы по теме «разрыва ответственности» и проблематике подотчетности при использовании непрозрачных алгоритмов [5].

На первом шаге исследования были систематизированы типы объяснимости, форматы представления объяснений в клиническом интерфейсе и риски некорректной интерпретации (в т. ч. правдоподобные, но вводящие в заблуждение объяснения). На втором шаге выделены сценарии возникновения разрывов ответственности — отсутствие доступа к данным и метаданным,

непрозрачность обновлений, неопределенность ролей между разработчиком, медицинским учреждением и врачом, а также недостаточность процедур расследования инцидентов. На третьем шаге сделан вывод о том, что требование, чтобы клиницист мог самостоятельно оценить обоснованность рекомендации, рассматривается как минимальное условие безопасного внедрения и критерий проектирования документации и интерфейса.

Анализ показал, что недоверие к медицинскому ИИ имеет как когнитивные, так и институциональные основания. На когнитивном уровне ключевым фактором является невозможность для врача проверить релевантность рекомендации применительно к конкретному пациенту. Даже если модель демонстрирует высокую точность, клиницисту требуется понимать, какие признаки и контекстные факторы оказали влияние на вывод, и не выходит ли случай за пределы области применимости.

Существенным практическим источником риска является смещение (bias) и снижение эффективности при дрейфе данных. Нерепрезентативность обучающих выборок, изменение демографического состава пациентов, различия в оборудовании и протоколах приводят к тому, что рекомендации могут быть систематически менее надежными для отдельных групп или клинических сценариев, хотя средняя метрика на тестовом наборе остается высокой.

Отдельно выявлена институциональная проблема «разрывов ответственности». При внедрении «черного ящика» врач может оказаться юридически и профессионально ответственным за последствия применения системы, не имея возможности оценить источники данных, известные ограничения, частоту обновлений и причины конкретного вывода. Это снижает готовность использовать систему и усиливает оборонительную практику (избыточные обследования, отказ от алгоритма).

Сопоставление с регуляторными ожиданиями показывает, что требование независимого пересмотра основания рекомендации предполагает предоставление пользователю:

– информации о входных данных и их качестве;

– описания алгоритмического подхода на уровне, достаточном для клинической оценки;

– результатов клинической валидации, включая ограничения применимости и известные риски;

– механизмов управления.

На основе обобщения сформулирована практическая рамка снижения рисков проблемы «черного ящика», где меры сгруппированы по компонентам доверия:

– доказательность (клиническая достоверность и обобщаемость);

– проверяемость (аудит, трассируемость и воспроизводимость);

– объяснимость (артефакты XAI и их клиническая полезность);

– управляемость (надзор, оспаривание);

– подотчетность (распределение обязанностей);

– устойчивость (контроль смещений и дрейфа).

Компоненты доверия к медицинскому ИИ и меры снижения рисков «черного ящика»

Компонент доверия	Типовой риск	Меры	Доказательства/ артефакты
Проверяемость	Невозможность аудита, воспроизводимости и анализа инцидентов	Трассируемость версий; журналирование входов/выходов; воспроизводимые пайплайны	Audit trails; модельные карты; протоколы валидации
Объяснимость	Непонимание факторов; предвзятость автоматизации; ложные объяснения	XAI-выходы в интерфейсе; обучение пользователей; проверка надежности объяснений	LIME/SHAP/салиентность с ограничениями; чек-листы интерпретации
Управляемость человеком	Нельзя безопасно оспорить/отменить рекомендацию	Экспертный надзор; режим отказа	Протоколы использования; обучение
Доказательность	Нет подтвержденной клинической пользы; плохая обобщаемость	Клиническая валидация; анализ подгрупп; оценка переносимости	Отчеты о валидации; ограничения применимости; метрики
Подотчетность	«Разрыв ответственности» при вреде и обновлениях	Распределение ролей; договорные обязанности; процедуры расследования	Политики обновлений; отчеты об инцидентах
Устойчивость	Снижение качества при дрейфе данных и изменении клинических процессов	Мониторинг дрейфа; контроль смещений; триггеры переоценки модели	Панели мониторинга; пороги тревог; план переобучения
Кибербезопасность	Атаки/подмена данных; утечки	Контроль доступа; тесты безопасности; защита данных	Политики информационной безопасности; журналирование; оценка рисков
Интеграция в процесс	Неправильное использование из-за плохого UX и расхождение в рабочих процессах	Проектирование под клинический сценарий; пилотирование; обратная связь	Протокол внедрения; обучение; анализ ошибок использования

Для синтеза выводов доверие представлено как многокомпонентная конструкция (см. таблицу). Данное представление использовано как «матрица кодирования» для сопоставления рисков и мер на этапах жизненного цикла: данные → разработка → валидация → внедрение → эксплуатация → обновления и эксплуатационное сопровождение.

Обсуждение полученных результатов. Полученные результаты согласуются с тем, что объяснимость является необходимым, но недостаточным условием доверия. Во-первых, многие XAI-методы строят апостериорные объяснения, которые могут быть визуально убедительными, но не гарантируют отражения истинных причинных механизмов модели. Это особенно критично в клинических задачах, где ложная уверенность может усилить предвзятость автоматизации — склонность переоценивать рекомендации системы.

Во-вторых, в клинической среде объяснение должно быть функциональным: оно должно помогать врачу (а) выявлять ошибочные рекомендации; (б) понимать границы применимости; (в) соотносить вывод с клиническими руководствами и контекстом пациента. Поэтому ключевой критерий полезности XAI — не визуальная убедительность, а проверяемость и возможность поддержать клиническое рассуждение, включая указание на неопределенность и чувствительность к входным данным.

В-третьих, доверие невозможно сформировать только на уровне интерфейса, если отсутствуют организационные и регуляторные механизмы. Регуляторный подход подчеркивает необходимость того, чтобы пользователь мог самостоятельно пересмотреть основания рекомендаций; это требует документации, журналирования и прозрачных процедур обновления, а также ответственности за мониторинг качества после внедрения [6, 7].

С организационной точки зрения полезно различать ответственность за:

– разработку и сопровождение (включая управление версиями и исправлениями);

– локальную адаптацию и валидацию в медицинской организации;

– клиническое применение и окончательное решение.

Четкое распределение этих ролей снижает «разрыв ответственности» и повышает готовность врачей использовать систему, понимая пределы своей ответственности и инструменты контроля.

Этические рекомендации настаивают, что управление рисками должно учитывать справедливость, отсутствие дискриминации и безопасность на протяжении жизненного цикла. Это означает регулярные проверки смещений, мониторинг дрейфа, анализ инцидентов и корректирующие действия, а также вовлечение заинтересованных сторон (врачей, пациентов, администраторов) в оценку реальной клинической ценности и приемлемости системы.

В итоге доверие целесообразно определять как свойство социотехнической системы — модель является лишь одним элементом, а доверие формируется через доказательность, управляемость и подотчетность, поддержанные

данными и процедурами. Это смещает фокус с вопроса «можно ли объяснить нейросеть?» к вопросу «можно ли безопасно управлять системой в реальной клинической практике?»

Заключение. Проблема доверия к «черному ящику» в медицинских системах ИИ является социотехнической и определяется сочетанием непрозрачности выводов, рисков смещения/дрейфа и институциональных «разрывов ответственности».

Показано, что повышение доверия требует интеграции объяснимости (XAI) с проверяемостью, клинической валидацией и эксплуатационным сопровождением, а также предоставления пользователю информации, достаточной для независимого пересмотра основания рекомендаций и осуществления человеческого контроля.

Практическая рамка повышения доверия должна включать документацию и трассируемость версий, регламенты обновлений и уведомлений, обучение пользователей, мониторинг дрейфа и смещений, процедуры расследования инцидентов и распределение ответственности между разработчиком, медицинским учреждением и врачом.

Список источников

[1] Muhammad D. Unveiling the black box: A systematic review of Explainable Artificial Intelligence methods applied to medical image analysis. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11382209/ (accessed 24.12.2025).

[2] Sadeghi Z. A review of Explainable Artificial Intelligence in healthcare. URL: https://www.sciencedirect.com/science/article/pii/S0045790624002982 (accessed 24.12.2025).

[3] Carr D. Responsibility Gaps and Black Box Healthcare AI. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11003475/ (accessed 24.12.2025).

[4] Ethics and governance of artificial intelligence for health. URL: https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content (accessed 24.12.2025).

[5] U.S. Food and Drug Administration. Clinical Decision Support Software: Final Guidance for Industry and FDA Staff. URL: https://www.fda.gov/media/162345/download (accessed 24.12.2025).

[6] Sidorova E.Y. Quality management of AI-driven clinical decision support systems. International Journal for Quality Research, 2024, vol. 18, no. 4, pp. 939–952. https://doi.org/10.24874/IJQR18.04-01

[7] Malashin I.P., Masich I.S., Tynchenko V.S. et al. Image Text Extraction and Natural Language Processing of Unstructured Data from Medical Reports. Machine Learning and Knowledge Extraction, 2024, vol. 6, no. 2, pp. 1361–1377. https://doi.org/10.3390/make6020064