



**Улучшения деятельности трудовых ресурсов по выявлению
отрицательного влияния нейросетей**

Мирвелов Р.А., магистр,

ФГБОУ ВПО «РЭУ им. Г.В. Плеханова», Москва, Россия

Аннотация. В статье рассматривается проблема организации человеческого контроля при использовании нейросетевых систем для выявления признаков нарушений. Цель исследования состоит в разработке организационно-экономического механизма верификации, который позволяет снижать ожидаемый ущерб от ошибочных машинных выводов, не устранив при этом производительный эффект автоматизации. Методическую основу составили анализ научной литературы и нормативно-методических документов, процессный анализ человеко-машинного контроля, классификация рисков и моделирование ожидаемых затрат. Показано, что формальное включение работника в контур принятия решения не гарантирует содержательной проверки: ее качество зависит от доступа к исходным материалам, профессиональных компетенций, нормирования нагрузки, полномочий отменить машинный вывод и закрепления ответственности. Предложена риск-ориентированная маршрутизация случаев между автоматической обработкой с выборочным аудитом, проверкой одним специалистом и двойным либо коллегиальным рассмотрением. Сформулирован экономический критерий целесообразности проверки, основанный на сопоставлении ее стоимости с ожидаемым предотвращенным ущербом. Предложены индекс маршрутизации, модель расчета потребности в проверяющих и распределение ответственности между участниками процесса. Разработаны система показателей и поэтапный порядок внедрения механизма. Научная новизна заключается в объединении правил маршрутизации, требований к организации труда, планирования мощности и

экономической оценки в едином механизме человеческой верификации. Результаты могут применяться при подготовке регламентов, распределении нагрузки и аудите нейросетевых систем в контрольно-надзорной и иной организационной деятельности.

Ключевые слова: нейросетевые системы, человеческий контроль, верификация, организация труда, управление рисками, экономическая эффективность, контрольно-надзорная деятельность.

Enhancing workforce effectiveness in identifying the adverse effects of neural networks

Mirvelov R.A., master's degree

Plekhanov Russian University of Economics, Moscow, Russia

Annotation. The article addresses the organization of human oversight when neural network systems are used to detect indications of violations. The study aims to develop an organizational and economic verification mechanism that reduces the expected loss from erroneous machine outputs without eliminating the productivity gains of automation. The methodology combines a review of research and regulatory guidance, process analysis of human–AI control, risk classification, and expected cost modeling. The study shows that merely placing an employee in the decision loop does not ensure meaningful verification. Its quality depends on access to source materials, domain competence, manageable workload, authority to override the system, and clearly assigned responsibility. A risk-based routing model is proposed to allocate cases among automated processing with sample audits, review by one authorized specialist, and independent double or collegial review. An economic criterion compares verification costs with the expected loss prevented by human intervention. A routing index, a reviewer capacity model, and a division of responsibilities among process participants are proposed. A set of quality, productivity, economic, and governance indicators and a staged implementation procedure are developed. The contribution of the article is an integrated mechanism combining routing rules, work organization requirements, capacity planning, accountability, and economic

evaluation. The results can support operating procedures, workload allocation, and audits of neural network systems in inspection and other organizational settings.

Key words: neural network systems, human oversight, verification, work organization, risk management, economic efficiency, inspection activities.

1. Введение. Распространение искусственного интеллекта переводит вопрос о взаимодействии человека и алгоритма из области технологического проектирования в область экономики и организации труда. По данным статистического сборника НИУ ВШЭ, среди крупных и средних российских организаций, применявших технологии искусственного интеллекта в 2024 г., обработку визуальных данных использовали 66,3%, а интеллектуальную поддержку принятия решений и управления – 50,2% [1]. Для процессов выявления признаков нарушений данное сочетание особенно характерно: система анализирует изображение или иной массив данных, формирует сигнал, а его правовое и организационное значение определяется на последующих этапах.

Автоматический отбор повышает пропускную способность: работник видит не весь поток, а только помеченные системой сигналы. Но за это приходится платить. Ложное срабатывание оборачивается лишней проверкой, иногда – необоснованным действием и жалобой; пропущенное нарушение – прямым ущербом. Значит, точности модели ещё недостаточно. Важнее другое: во сколько обходится ошибка и способна ли организация направлять решения так, чтобы дорогие ошибки не проходили мимо. Это вопрос не алгоритма, а организации труда – ему и посвящена статья.

Международные документы смотрят на такой контроль как на социотехническую задачу, а не чисто техническую. NIST AI RMF, к примеру, увязывает управление рисками с распределением ролей и подготовкой тех, кто с системой работает [3]. Регламент ЕС идёт дальше и делает надзор обязательным: статья 14 требует, чтобы для высокорисковых систем человек понимал вывод и мог вмешаться, а объём надзора отвечал уровню риска [4]. Национальная стратегия развития ИИ в России приходит к тому же через приоритет прав

человека [2]. Знаменатель общий: само по себе присутствие человека ничего не гарантирует, пока у него нет условий для реального вмешательства.

Однако присутствие человека в контуре не тождественно эффективному контролю. Исследования автоматизационного смещения показывают, что оператор может некритично принять машинную рекомендацию, особенно при высокой нагрузке, монотонности и доверии к системе [5]. Объяснение вывода алгоритма также не гарантирует улучшения совместного результата: оно иногда повышает вероятность согласия с рекомендацией независимо от ее правильности [6]. В государственном управлении формальный надзор способен создавать ложное ощущение защищенности, если у работника отсутствуют ресурсы и реальные полномочия для пересмотра решения [8].

Отдельные элементы такого контроля изучены неплохо, но механизма, который связывал бы их вместе, пока нет. А на практике вопросы приходят пакетом: какие случаи вообще отдавать человеку, сколько для этого нужно людей и как поделить между ними ответственность, чем мерить качество их работы и когда проверка окупается. По отдельности их решать бессмысленно - они завязаны друг на друга. Поэтому в центре работы - деятельность проверяющих при использовании нейросетевых средств выявления нарушений и организационно-экономические методы верификации их результатов.

Цель исследования – разработать риск-ориентированный организационно-экономический механизм человеческой верификации, обеспечивающий снижение негативных последствий ошибочных нейросетевых выводов при сохранении эффекта автоматизации. Для достижения цели решаются задачи классификации рисков, определения уровней контроля, обоснования требований к организации труда, формирования экономического критерия проверки, расчета мощности, распределения ответственности и построения системы показателей. Гипотеза состоит в том, что дифференциация глубины проверки по риску снижает совокупные ожидаемые затраты по сравнению как с полностью автоматическим режимом, так и со сплошной ручной проверкой.

2. Методы исследования. Исследование имеет концептуально-методический характер. Теоретическую основу составили работы по автоматизационному смещению, человеко-машинной комплементарности, подотчетности алгоритмических решений и организации человеческого надзора [5, 6, 7, 8, 9, 10]. Нормативно-методическую основу образовали NIST AI RMF 1.0, Регламент ЕС 2024/1689 и Национальная стратегия развития искусственного интеллекта Российской Федерации [3, 4, 2].

Основных методов было несколько, и каждый закрывал свою задачу. Системный анализ понадобился, чтобы увидеть нейросеть, работников и регламенты как единый контур, а не три отдельные части. Процессным анализом мы прослеживали путь случая -от сигнала системы до итогового решения. Типы ошибок и их организационные последствия свели в классификацию. И, наконец, чтобы сопоставить стоимость проверки с предотвращаемым ущербом, применили анализ ожидаемых затрат -на нём же сравнили три режима: только система, только специалист и совместная работа.

Конструирование механизма выполнялось последовательно. Сначала для каждого типа ошибки определялись возможное последствие и субъект, способный его предотвратить. Затем признаки случая переводились в факторы маршрутизации, а каждому уровню контроля сопоставлялись процедура, полномочия и требования к документированию. На следующем шаге маршрут связывался с трудоемкостью и ожидаемым ущербом. Такая последовательность позволяет избежать ситуации, при которой уровень контроля выбирается только по технической уверенности модели без учета обратимости и общественной значимости итогового действия.

Для последующей эмпирической проверки предусмотрены два контура валидации. Первый оценивает качество решений на стратифицированной выборке случаев с независимой эталонной оценкой. Второй проверяет организационную реализуемость: достаточность времени, устойчивость очереди, долю эскалаций и соблюдение сроков. Механизм признается результативным только при одновременном улучшении качества и сохранении

управляемой трудоемкости; перенос ошибок из машинного этапа в перегруженную очередь ручной проверки не считается улучшением процесса.

Единицей анализа является случай, то есть отдельный сигнал нейросетевой системы, который может привести к организационному, управленческому или юридически значимому действию. Эталонным считается решение, принятое на основании полного набора допустимых материалов компетентным специалистом либо группой специалистов по заранее установленным критериям. Если эталонная оценка формируется тем же работником, который видел машинный вывод, необходимо учитывать риск зависимой ошибки и по возможности использовать слепую повторную проверку части материала.

Практическая применимость механизма рассматривается на примере процесса административно-технического контроля крупного мегаполиса. Для окончательной эмпирической версии статьи следует указать [*период наблюдения*], [*число и критерии отбора обезличенных случаев*], [*способ получения эталонной оценки*] и [*порядок согласования использования непубличных материалов*]. До заполнения этих сведений количественные выводы и оценки распространенности отдельных ошибок не формулируются.

3. Результаты

3.1. Содержание отрицательного влияния и человеческой верификации. Под отрицательным влиянием нейросетевой системы в настоящем исследовании понимаются вероятность и последствия ошибочного, неполного, предвзятого или лишенного необходимого контекста результата, способного привести к неверному решению, дополнительным трудовым затратам или ущербу. Такое определение отделяет ошибку нейросети от нарушения, выявляемого с ее помощью, и связывает техническое качество модели с последствиями для организации.

Человеческая верификация определяется как регламентированная содержательная проверка машинного результата, завершающаяся одним из пяти исходов: подтверждением, отклонением, запросом дополнительных материалов, эскалацией или регистрацией системной проблемы. Верификация отличается от

формального подтверждения наличием доступа к доказательствам, времени на анализ, права изменить результат и обязанности зафиксировать основания решения.

Таблица 1

Классификация рисков нейросетевой поддержки решений

Риск	Проявление	Последствие	Функция работника
Ложноположительный результат	Система указывает на нарушение, которого нет	Необоснованные действия, жалобы, репутационные и трудовые издержки	Проверить контекст и отклонить недостаточно обоснованный вывод
Ложноотрицательный результат	Реальное нарушение не распознано	Пропущенный ущерб и снижение результативности контроля	Проводить выборочный аудит отрицательных результатов и анализ пропусков
Недостаточная контекстность	Не учтены дата, условия, документы или исключения	Формально точный, но содержательно неверный вывод	Сопоставить сигнал с дополнительными сведениями
Систематическое смещение	Ошибки концентрируются на отдельных типах объектов	Неравномерное распределение рисков и нагрузки	Выявлять повторяющиеся паттерны расхождений
Непрозрачность	Неясно, на каких материалах и признаках основан вывод	Невозможность мотивированного решения и последующего аудита	Запросить основания, исходные материалы и оценку уверенности
Автоматизационное смещение	Работник механически соглашается с системой	Человеческий контроль становится формальным	Применять чек-листы, независимую оценку и аудит подтверждений
Размывание ответственности	Не определен владелец итогового решения	Конфликты, задержки и отсутствие корректирующих действий	Закрепить полномочия, основания эскалации и ответственность

Составлено автором

Классификация показывает, что работник не является универсальным средством исправления технических ошибок. Человеческое участие само создает риски: невнимательность, зависимость от машинной подсказки, разную строгость экспертов и деградацию навыков. Поэтому предметом управления должен быть не отдельный работник, а весь процесс верификации.

3.2. Риск-ориентированный контур проверки. Предлагаемый механизм строится на принципе соразмерности контроля ожидаемым последствиям. Проверка каждого сигнала может устранить выигрыш от автоматизации, тогда как отказ от проверки значимых решений увеличивает ожидаемый ущерб.

Маршрут случая определяется сочетанием вероятности ошибки, тяжести последствий, обратимости решения, полноты исходных материалов и нормативной обязательности человеческого участия.

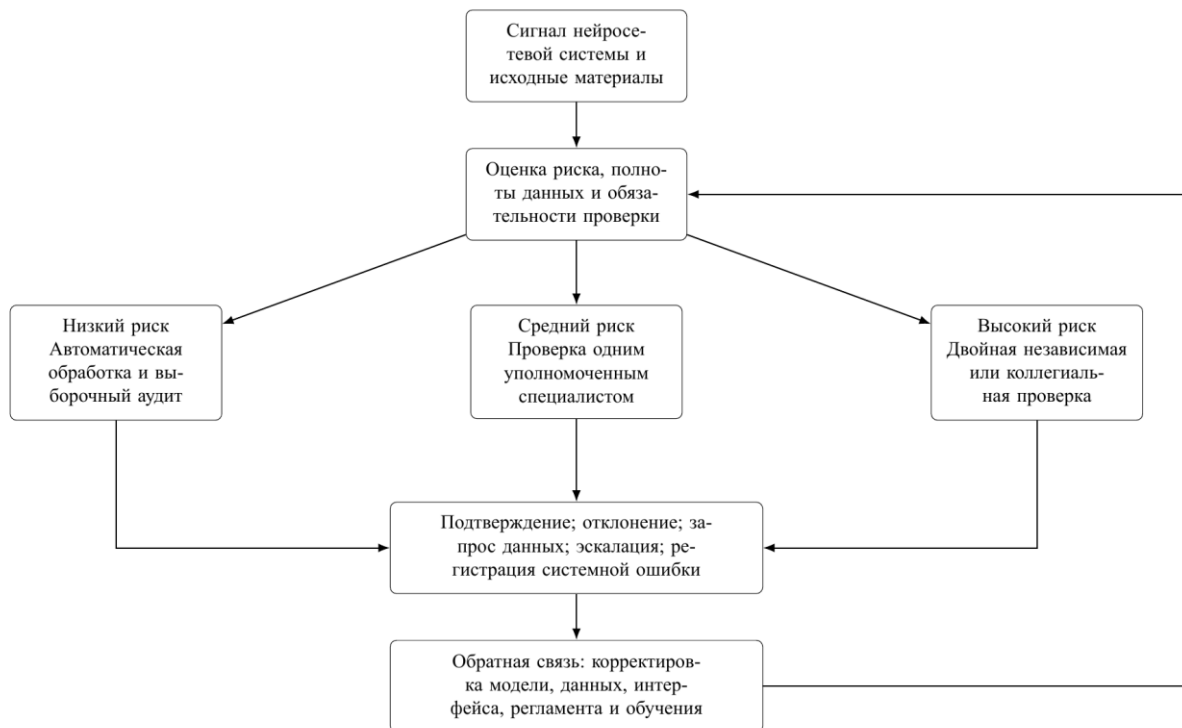


Рис. 1 – Риск-ориентированный контур человеческой верификации

Таблица 2

Матрица уровня человеческого контроля

Уровень	Типичная ситуация	Форма контроля	Документирование
Низкий	Высокая достоверность, обратимые последствия, достаточные материалы	Автоматическая обработка с выборочным аудитом положительных и отрицательных результатов	Лог операции и результат выборочного контроля
Средний	Неоднозначный признак, неполные данные либо заметные последствия	Проверка одним уполномоченным специалистом; при необходимости запрос сведений	Основания решения и перечень использованных материалов
Высокий	Возможен существенный ущерб, затронуты права лиц, решение необратимо или уверенность системы низкая	Двойная независимая либо коллегиальная проверка; обязательная эскалация разногласий	Полный протокол, ответственные лица и основания итогового решения

Составлено автором

Выборочный аудит должен охватывать не только подтвержденные системой случаи, но и часть отрицательных результатов. Иначе организация

измеряет главным образом ложноположительные ошибки и не получает данных о пропущенных нарушениях. Доля аудируемых случаев устанавливается с учетом стабильности модели, изменения входных данных и тяжести возможного пропуска.

3.3. Формализация маршрутизации случаев. Для воспроизводимого выбора уровня контроля качественные признаки необходимо преобразовать в операционное правило. В качестве такого правила предлагается индекс маршрутизации S_i , рассчитываемый для случая i :

$$S_i = w_p z_{p,i} + w_l z_{l,i} + w_u z_{u,i} + w_d z_{d,i} + w_n z_{n,i}, \quad \sum_k w_k = 1, \quad (1)$$

где $z_{p,i}$ – нормированная оценка вероятности ошибочного вывода;

$z_{l,i}$ – тяжесть возможных последствий;

$z_{u,i}$ – необратимость действия;

$z_{d,i}$ – недостаточность или противоречивость исходных данных;

$z_{n,i}$ – новизна случая относительно подтвержденной области применения модели;

w_k – веса факторов.

Все компоненты приводятся к интервалу от 0 до 1. Чем выше индекс, тем более глубокая проверка требуется.

Индекс не следует интерпретировать как точную вероятность ущерба. Его задача – единообразно ранжировать случаи и сделать маршрут проверяемым при аудите. Веса могут устанавливаться методом экспертного согласования, а после накопления наблюдений – уточняться по фактической частоте отмен, пропусков, жалоб и размеру последствий. При этом техническая уверенность нейросети не должна непосредственно подменять $z_{p,i}$: ее сначала необходимо откалибровать на эталонной выборке отдельно по версиям модели и существенным категориям объектов.

Операционализация факторов маршрутизации

Фактор	Пример основания для оценки	Влияние на маршрут
Вероятность ошибки	Историческая доля расхождений в однородной группе; калиброванная уверенность; наличие сдвига данных	Рост переводит случай от аудита к обязательной проверке
Тяжесть последствий	Ожидаемый прямой ущерб, затрагивание прав, масштаб последующих действий и репутационный эффект	Высокая тяжесть исключает полностью автоматическое решение
Необратимость	Возможность отменить действие и восстановить положение без значительных затрат	Необратимые действия требуют предварительного, а не последующего контроля
Качество данных	Полнота материалов, согласованность источников, пригодность изображения или записи для анализа	Недостаточность данных ведет к запросу сведений либо эскалации
Новизна случая	Редкая категория, новые условия наблюдения, изменение правил, оборудования или входного потока	Новые случаи направляются на экспертную проверку до накопления статистики

Составлено автором

Пусть τ_1 и τ_2 – утвержденные пороги, причем $0 < \tau_1 < \tau_2 < 1$. При $S_i < \tau_1$ применяется низкий уровень контроля; при $\tau_1 \leq S_i < \tau_2$ – средний; при $S_i \geq \tau_2$ – высокий. Поверх этого правила действуют безусловные ограничения: обязательное участие человека, существенная неопределенность примененного правила, конфликт источников и возможность необратимого воздействия автоматически повышают уровень контроля. Тем самым формула поддерживает, но не заменяет профессиональное и нормативное суждение.

Пороговые значения должны учитывать не только качество решений, но и доступную мощность подразделения. Слишком низкие пороги создают очередь, увеличивают время ожидания и снижают фактическое качество проверки; слишком высокие – оставляют без внимания случаи с существенным ожидаемым ущербом. Поэтому пороги сначала проверяются ретроспективно, затем в ограниченном пилоте и только после этого закрепляются в регламенте. Не реже установленного периода они пересматриваются при изменении модели, данных, норм или структуры ошибок.

Выборочный аудит низкого уровня также должен быть стратифицированным. Простая случайная выборка преимущественно воспроизводит массовые стандартные случаи и может не обнаружить ошибку в редкой категории. В аудиторскую совокупность следует включать случайную часть, повышенную долю новых и пограничных объектов, а также отрицательные результаты. Обнаружение систематического расхождения является основанием временно повысить долю аудита или перевести соответствующую категорию на обязательную проверку.

3.4. Требования к организации труда проверяющих работников.

Результативность механизма зависит от условий, в которых специалист принимает решение. Обобщение научных работ и требований рамочных документов позволяет выделить восемь взаимосвязанных направлений совершенствования деятельности персонала (табл. 4).

Таблица 4

Направления совершенствования человеческой верификации

Направление	Организационное решение	Предотвращаемый риск
Регламентация	Единые критерии достаточности материалов, варианты решения, сроки и основания эскалации	Произвольность и несопоставимость решений
Компетенции	Подготовка по предметной области, ограничениям модели, когнитивным искажениям и интерпретации вывода	Некритичное доверие и ошибочная интерпретация
Полномочия	Право отклонить вывод, запросить данные, приостановить действие и передать случай выше	Формальный контроль без возможности вмешательства
Информационное обеспечение	Доступ к исходным материалам, времени наблюдения, контексту, уверенности и основаниям сигнала	Решение по неполному или деконтекстуализированному представлению
Нормирование нагрузки	Лимиты одновременно рассматриваемых случаев, риск-приоритеты, ротация и перерывы	Невнимательность, усталость и механическое подтверждение
Разделение функций	Независимость проверки от разработчика системы и владельца показателей ее внедрения	Конфликт интересов и сокрытие проблем
Обратная связь	Учет отмен, пропусков, жалоб и причин расхождения	Повторение системных ошибок
Мотивация и контроль качества	Оценка не только скорости, но и обоснованности, согласованности и предотвращенного ущерба	Оптимизация количества решений в ущерб качеству

Важным элементом является независимость первичного человеческого суждения. Интерфейс может сначала требовать от работника зафиксировать собственную оценку и только затем показывать рекомендацию системы. Подобные «когнитивные ограничители» способны уменьшать чрезмерное доверие к алгоритму, хотя могут повышать время и субъективную сложность работы [7]. Следовательно, они целесообразны прежде всего для случаев среднего и высокого риска.

3.5. Распределение ответственности и планирование мощности.

Ответственность нужно развести на два уровня: за конкретное решение и за условия, в которых оно принимается. Проверяющий отвечает за то, что соблюдал процедуру и обосновал вывод. Но он не выбирал обучающие данные, не задавал пороги, не проектировал интерфейс и не устанавливал норму нагрузки -значит, и отвечать за них в одиночку не может. Иначе получается знакомая ситуация: спрашивают с того, кто на проблему повлиять не мог.

Таблица 5

Роли и ответственность в механизме верификации

Участник	Полномочия	Основная ответственность
Владелец процесса	Утверждает уровни риска, сроки, ресурсы и допустимые исходы	Совокупный результат процесса и соответствие обязательным требованиям
Владелец модели	Управляет версией, порогами срабатывания и техническим мониторингом	Качество модели в заявленной области применения и реакция на деградацию
Владелец данных	Устанавливает требования к источникам, полноте и срокам хранения	Качество, происхождение и доступность данных для проверки
Проверяющий специалист	Подтверждает, отклоняет, запрашивает сведения и эскалирует	Обоснованность конкретного решения в пределах доступных материалов
Старший эксперт или комиссия	Разрешает разногласия и рассматривает высокорисковые случаи	Независимость и полнота углубленной оценки
Функция контроля качества	Формирует выборки, проводит повторную проверку и анализирует показатели	Выявление системных ошибок и достоверность отчетности

Составлено автором

В регламенте для каждой операции целесообразно обозначить исполнителя, утверждающего, консультируемого и информируемого участника.

Особенно важно развести изменение модели и приемку результата ее работы: подразделение, заинтересованное в демонстрации производительности автоматизации, не должно единолично определять качество контрольной выборки и закрывать выявленные инциденты. Для значимых изменений требуется совместное решение владельца процесса, владельца модели и функции контроля качества.

Численность проверяющих должна определяться из входного потока, а не из среднего числа обработок в спокойный период. Минимальная расчетная потребность в специалистах может быть представлена выражением

$$n_{\min} = \left\lceil \frac{\sum_{j=1}^m \lambda_j t_j}{60H\rho} \right\rceil, \quad (2)$$

где λ_j -среднее число случаев маршрута j за день,

t_j -время на один случай с учётом повторной проверки и документирования,

H -рабочее время специалиста в часах, ρ -целевая загрузка (меньше единицы),

m -число маршрутов.

Резерв $1-\rho$ оставляем на неравномерность потока, сложные случаи, обучение и разбор инцидентов.

Покажем на условном примере. Пусть на маршрут среднего риска приходит $\lambda = 150$ случаев в сутки, на каждый уходит $t = 12$ минут, полезное рабочее время $H = 6$ часов (остальное -совещания, учёба, перерывы), а целевую загрузку берём $\rho = 0,8$. Тогда дневной объём - $150 \times 12 = 1800$ минут, то есть 30 часов работы; один специалист даёт $6 \times 0,8 = 4,8$ часа. Нужно $30 / 4,8 \approx 7$ человек только на этот маршрут. Если убрать резерв и поставить $\rho = 1$, получится 6,25 - то есть при полной загрузке первая же пиковая волна создаёт очередь. Отсюда практический смысл резерва: он не «страховка от лени», а условие устойчивости очереди.

Расчет по средним значениям дополняется анализом пикового потока. Для каждого уровня риска устанавливаются предельный возраст случая в очереди и правило реакции: временное перераспределение работников, подключение

резервной группы, приоритет высокорисковых случаев или ограничение новых автоматических действий. Снижение глубины обязательной проверки только из-за накопившейся очереди не допускается; такая ситуация означает несоответствие мощности выбранному режиму.

Норматив t_j не должен задаваться исключительно по лучшему времени наиболее опытного сотрудника. Его следует оценивать по пилотной выборке с учетом медианы, верхнего квартиля, доли запросов дополнительных сведений и времени фиксации оснований. После изменения интерфейса или состава материалов норматив пересматривается. Благодаря этому планирование численности становится частью управления качеством, а не только бюджетным расчетом.

3.6. Экономическая модель человеческой проверки. Экономический эффект нейросетевого решения нельзя оценивать только сокращением времени или численности работников. Совокупные затраты процесса предлагается представить следующим образом:

$$C_{\text{общ}} = C_{\text{сист}} + C_{\text{пров}} + N_{\text{лп}}L_{\text{лп}} + N_{\text{ло}}L_{\text{ло}} + C_{\text{инц}} \quad (3)$$

где $C_{\text{сист}}$ – затраты на разработку, приобретение, эксплуатацию и мониторинг системы;

$C_{\text{пров}}$ – затраты на труд, обучение и аудит проверяющих;

$N_{\text{лп}}$ и $N_{\text{ло}}$ – количество ложноположительных и ложноотрицательных решений;

$L_{\text{лп}}$ и $L_{\text{ло}}$ – средний ущерб соответствующего типа;

$C_{\text{инц}}$ – затраты, связанные с системными инцидентами, жалобами и восстановлением процесса.

Отдельную проверку стоит проводить только тогда, когда ожидаемый предотвращённый ущерб больше её цены:

$$P_{\text{ош},i} L_i Q_{\text{испр},i} > C_{\text{пров},i} \quad (4)$$

где $P_{\text{ош},i}$ – оценочная вероятность ошибки системы;

L_i – ущерб при принятии ошибочного решения;

$Q_{\text{испр},i}$ – вероятность обнаружения и исправления ошибки работником;

$C_{\text{пров},i}$ – полная стоимость проверки случая, включая затраты времени и эскалации.

Условный пример. Пусть вероятность ошибки системы на случае $P_{\text{ош}} = 0,1$, возможный ущерб $L = 100\,000$ руб., а проверяющий ловит такую ошибку с вероятностью $Q_{\text{испр}} = 0,7$. Ожидаемый предотвращённый ущерб $-0,1 \times 100\,000 \times 0,7 = 7000$ руб. Если проверка стоит 500 руб., она окупается больше чем в десять раз -случай однозначно идёт к человеку. Но пусть возможный ущерб не 100 000, а 3000 руб.: тогда слева всего $0,1 \times 3000 \times 0,7 = 210$ руб. против 500 руб. затрат -сплошная проверка таких случаев себя не оправдывает, их разумнее пускать в выборочный аудит. Маршрут определяет именно этот расчёт, а не общая установка «проверять всё».

Разность левой и правой частей формулы (4) образует ожидаемый чистый эффект проверки. При этом формула не отменяет обязательные правовые и этические ограничения: если решение затрагивает права лиц или нормативно требует участия человека, экономическое сравнение используется для выбора глубины и организации контроля, а не для отказа от него.

Особенно хрупок здесь $Q_{\text{испр}}$. Он легко проваливается почти в ноль: перегруженный работник, у которого нет доступа к первичным данным и который по привычке соглашается с системой, проверку формально проводит -а ущерб не предотвращает. В формуле это видно сразу: левый край обнуляется, и дорогая на бумаге проверка не стоит ничего. Значит, вложения в обучение и удобный интерфейс окупаются ровно в той мере, в какой они поднимают $Q_{\text{испр}}$, -не больше и не меньше.

Для набора случаев правило можно представить как задачу распределения ограниченного ресурса проверки. Если $x_i = 1$ означает направление случая на проверку, а $x_i = 0$ – отсутствие индивидуальной проверки, переменная часть ожидаемых затрат имеет вид

$$C_{\text{пер}} = \sum_{i=1}^N [x_i C_{\text{пров},i} + P_{\text{ош},i} L_i (1 - x_i Q_{\text{испр},i})]. \quad (5)$$

Минимизация выражения (5) при ограничении бюджета или рабочего времени сначала направляет ресурс на случаи с наибольшим отношением ожидаемого предотвращенного ущерба к стоимости проверки. Обязательные категории включаются в множество проверяемых до экономической оптимизации. Практический смысл такой постановки заключается не в полной автоматизации управленческого выбора, а в прозрачном объяснении, почему дефицитное экспертное время направляется на одни случаи раньше других.

Оценка эффекта должна сравнивать не только состояние «до» и «после» внедрения. На нее могут влиять изменение входного потока, сезонность, новая версия модели и обучение работников. Предпочтителен параллельный пилот либо последовательное включение сопоставимых групп с сохранением контрольной выборки. Если такой дизайн невозможен, рассчитываются по меньшей мере три сценария – оптимистический, базовый и консервативный – с разными значениями вероятности ошибки, ущерба и способности работника исправить вывод. Решение устойчиво, если преимущество механизма сохраняется в консервативном сценарии.

Особую трудность представляет денежная оценка последствий, не имеющих прямой рыночной цены. Жалобы, задержки, неравномерность воздействия и потеря доверия не следует искусственно сводить к одной сомнительной сумме. Для них наряду с затратами используются натуральные показатели: число отмененных действий, продолжительность восстановления, количество затронутых лиц и повторяемость инцидента. Экономическая модель в этом случае дополняет, а не вытесняет правовую и качественную оценку.

3.7. Показатели эффективности. Предложенная модель требует одновременного учета качества, производительности, экономики и управляемости. Использование только количества обработанных сигналов создает стимул к ускоренному формальному подтверждению (табл. 6).

Показатели эффективности человеко-машинного контроля

Группа	Показатель	Назначение
Качество	Доля подтвержденных сигналов	Оценка полезности нейросетевого отбора
Качество	Доля выводов, отмененных человеком	Выявление ложных срабатываний и проблем модели
Качество	Доля пропущенных нарушений в аудируемой выборке	Оценка ложноотрицательных результатов
Согласованность	Доля совпадений независимых проверяющих; коэффициент согласия	Оценка надежности человеческого решения
Производительность	Среднее и медианное время проверки по уровню риска	Нормирование трудоемкости без смешения простых и сложных случаев
Производительность	Число случаев на одного специалиста и размер очереди	Оценка нагрузки и пропускной способности
Экономика	Стоимость проверки одного случая	Сопоставление вариантов организации процесса
Экономика	Оценочный предотвращенный ущерб	Оценка результата человеческого контроля
Управление	Время реакции на системную ошибку	Оценка качества обратной связи
Управление	Доля решений без обращения к дополнительным материалам	Индикатор возможного автоматизационного смещения

Составлено автором

Показатели следует рассчитывать отдельно по категориям риска, типам объектов, версиям модели и группам работников. Высокая средняя точность может скрывать устойчивую ошибку в редкой, но значимой категории. Аналогично высокая доля отмен не всегда свидетельствует о плохой работе системы: она может быть ожидаемым результатом целевого отбора пограничных случаев для экспертного рассмотрения.

3.8. Порядок внедрения механизма. Внедрение целесообразно осуществлять как управляемое изменение процесса, а не как однократную установку программного продукта. До запуска необходимо описать, какое решение поддерживает система, кто несет ответственность за итоговое действие, какие материалы доступны проверяющему и какие последствия возникают при ошибке. Если эти элементы не определены, автоматизация лишь закрепляет неформальный и слабо контролируемый порядок.

Предлагается пять этапов внедрения (табл. 7). Переход между этапами осуществляется по заранее установленным критериям, а не только по календарному сроку проекта.

Таблица 7

Этапы внедрения механизма человеческой верификации

Этап	Основные действия	Результат этапа
1. Инвентаризация	Описание решений, источников данных, участников, последствий ошибок и обязательных требований	Реестр сценариев применения и границы допустимого использования системы
2. Базовая оценка	Формирование эталонной выборки, измерение качества текущего ручного и машинного режимов, оценка трудоемкости	Исходные показатели для сравнения и перечень критических категорий
3. Проектирование	Настройка факторов и порогов маршрутизации, ролей, интерфейса, чек-листов, журналирования и обучения	Утвержденный регламент и расчет необходимой мощности
4. Пилот	Ограниченный запуск, независимый аудит решений, анализ очереди, разногласий и пользовательских затруднений	Скорректированные пороги, нормативы времени и критерии эскалации
5. Эксплуатация	Постоянный мониторинг, разбор инцидентов, аудит выборки и управление изменениями	Подтверждаемая результативность и контролируемое обновление механизма

Составлено автором

В пилоте рекомендуется использовать режим ограниченного воздействия: система формирует сигнал и предложенный маршрут, однако юридически или организационно значимое действие выполняется только после проверки. Это позволяет измерить ошибки без переноса полного риска на участников процесса. Пилот должен включать не только стандартные случаи, но и заранее отобранные пограничные, редкие и низкокачественные материалы; иначе будет завышена оценка готовности механизма.

Для ввода в эксплуатацию устанавливаются минимальные критерии: завершенное обучение проверяющих, приемлемая согласованность независимых оценок, отсутствие неразобранных критических ошибок, устойчивость очереди при расчетной нагрузке и возможность восстановить ход решения по журналу. Несоблюдение хотя бы одного из этих критериев является основанием продлить пилот или сузить область применения.

После запуска механизм нельзя считать застывшим. Обновили модель, изменили набор данных, сдвинули порог, переделали интерфейс или поменялась норма – всё это повод пересмотреть проверку целиком, а не только затронутый кусок. Насколько глубоко пересматривать -зависит от масштаба. Небольшая правка? Хватит регрессионной выборки. Новая категория объектов? Нужен повторный пилот. Экстренная приостановка или перевод категории на сплошную ручную проверку применяется при росте тяжелых ошибок, невозможности воспроизвести решение, резком сдвиге данных либо серии однотипных обоснованных жалоб.

3.9. Апробация на обезличенном практическом материале. Для эмпирической проверки механизма рекомендуется сформировать карточки случаев, не содержащие фамилий, адресов, координат, номеров дел, названий внутренних информационных систем и иных признаков косвенной идентификации. Минимальный набор полей приведен в табл. 8.

Таблица 8

Структура карточки практического случая

Элемент	Содержание
Контекст	Обобщенный тип объекта или процесса без идентифицирующих деталей
Сигнал системы	Выявленный признак, доступные материалы и уверенность модели
Риск ошибочного решения	Возможный организационный, правовой, трудовой или экономический ущерб
Действие специалиста	Проверенные дополнительные сведения и использованные критерии
Итог	Подтверждение, отклонение, запрос данных или эскалация
Причина расхождения	Качество данных, временное изменение, контекстное исключение, правило, редкий случай или ошибка интерфейса
Организационный вывод	Изменение модели, данных, регламента, обучения, нагрузки или маршрута

Составлено автором

Количественный дизайн предполагает сравнение трех режимов: только нейросеть, только специалист и нейросеть совместно со специалистом. Для каждого режима рассчитываются точность, ложные срабатывания, пропуски,

время, трудоемкость и ожидаемый ущерб. Если доступен только качественный материал, выводы следует ограничить типологией расхождений и организационными рекомендациями, не перенося частоту наблюдавшихся ситуаций на всю деятельность органа или отрасли.

4. Обсуждение результатов. Полученный механизм развивает положение о том, что качество искусственного интеллекта является свойством не только модели, но и социотехнической системы. С NIST AI RMF предложенную конструкцию сближают четкое распределение ролей, подготовка персонала, документирование надзора и обратная связь [3]. С логикой Регламента ЕС 2024/1689 ее объединяет соразмерность глубины контроля уровню риска и способность уполномоченного лица понимать, интерпретировать и при необходимости отменять результат системы [4].

Вместе с тем работа показывает ограниченность тезиса «окончательное решение принимает человек». Если специалист видит только итоговый ярлык, не знает ограничений модели, работает в условиях очереди и оценивается преимущественно по скорости, его участие может легитимировать машинный результат, не повышая качество. В терминологии М. Элиш работник рискует превратиться в «моральную зону деформации», на которую переносится ответственность за отказ сложной системы при ограниченном реальном контроле [11]. Поэтому ответственность должна распределяться между владельцем процесса, разработчиком, подразделением данных, руководителем и непосредственным проверяющим в соответствии с их полномочиями.

Совместная работа человека и искусственного интеллекта не обеспечивает комплементарность автоматически. Она возникает, когда стороны обладают различающейся полезной информацией, а процедура позволяет использовать это различие [9]. Нейросеть эффективна в масштабном поиске устойчивых признаков; работник – в учете контекста, исключений и последствий. Если оба основаны на одном неполном источнике или человек лишь повторяет машинную оценку, комбинированный режим увеличивает затраты без соответствующего прироста качества.

Практическая ценность модели состоит в переходе от общего требования человеческого контроля к проектируемому бизнес-процессу. Руководитель получает основания определить категории обязательной проверки, численность и компетенции персонала, содержание интерфейса, целевые показатели и порядок разбора ошибок. Экономический критерий позволяет обсуждать глубину контроля в терминах ожидаемых затрат, а не только абстрактной точности.

Отсюда неочевидное следствие: больше ручной проверки -не всегда лучше. Пока у отдела есть запас мощности, каждая дополнительная проверка снижает остаточный риск. Но за определённой загрузкой картина переворачивается. Копится очередь, накапливается усталость, растёт доля механических «согласен» -то есть падает тот самый $Q_{\text{испр}}$. В пределе сплошная проверка создаёт видимость строгого контроля и одновременно ухудшает способность замечать ошибки. Практический вывод простой: пороги маршрутизации и штат нельзя трогать по отдельности -сдвинули одно, пересчитайте другое.

Модель также разделяет два направления инвестиций. Первое уменьшает $P_{\text{ош}}$ за счет улучшения данных, модели и границ ее применения. Второе повышает $Q_{\text{испр}}$ благодаря подготовке работников, более информативному интерфейсу, независимой оценке и нормальной нагрузке. Сравнение предельного эффекта этих направлений позволяет избежать технологического детерминизма: иногда дополнительное обучение модели дешевле расширения проверки, а иногда небольшое изменение маршрута предотвращает больше ущерба, чем сложная техническая модернизация.

Организационная зрелость проявляется не в минимальной доле отмен машинного вывода. Слишком низкая доля может означать как высокое качество модели, так и формальность проверки; слишком высокая – как плохое качество системы, так и правильную передачу человеку наиболее неоднозначных случаев. Интерпретация показателя возможна только вместе с составом маршрутизированной выборки, результатами независимого аудита и динамикой

жалоб. Это подтверждает необходимость системы взаимодополняющих, а не одного целевого показателя.

Исследование имеет ограничения. Во-первых, значения вероятности ошибки, ущерба и способности работника исправить ошибку зависят от конкретной отрасли и должны оцениваться эмпирически. Во-вторых, предложенная трехуровневая матрица упрощает непрерывный спектр рисков, а веса индекса маршрутизации до калибровки отражают экспертное суждение. В-третьих, расчет мощности по среднему потоку требует дополнения данными о вариации и сезонных пиках. В-четвертых, профессиональное наблюдение подвержено ошибкам отбора и памяти, поэтому его желательно дополнять обезличенной выборкой, независимым кодированием случаев и экспертным опросом. В-пятых, даже обезличенные описания могут раскрывать участников через сочетание косвенных признаков; публикация внутренней статистики требует отдельного согласования. Наконец, до эмпирической апробации предложенные зависимости имеют характер проверяемой организационно-экономической модели, а не доказанной оценки эффекта для конкретного органа или отрасли.

5. Заключение. Из работы вытекает несколько выводов.

Первый – о самой природе вреда. Он не сводится к ложным срабатываниям. Не меньше значат потеря контекста, привыкание соглашаться с машиной, размывание ответственности и медленная утрата навыков у самих работников. Поэтому проверка нужна не «для галочки», а против всего этого набора сразу.

Работает она, однако, не всегда. Мы выделили пять условий, без которых верификация превращается в формальность: доступ к исходным материалам, компетентность - и предметная, и алгоритмическая, - время, право отменить решение и зафиксированная ответственность. Достаточно убрать одно -и остальные почти не спасают.

Дальше – вопрос сортировки. Проверять всё подряд дорого и бессмысленно, поэтому случаи распределяются по риску через индекс маршрутизации

Второй – человеческая верификация является эффективной лишь при наличии пяти условий: доступа к исходным материалам, предметной и алгоритмической компетентности, достаточного времени, полномочий изменить решение и документированной ответственности.

Третий – случаи целесообразно распределять по уровням риска. Индекс маршрутизации объединяет вероятность ошибки, тяжесть последствий, необратимость, качество данных и новизну случая, но не отменяет обязательных оснований человеческого участия. Для низкого риска допустима автоматическая обработка с выборочным аудитом; для среднего требуется уполномоченный специалист; для высокого – независимая двойная или коллегиальная проверка.

Четвёртый – ответственность проверяющего должна быть соразмерна его полномочиям. Качество процесса распределено между владельцем процесса, владельцами модели и данных, проверяющим, старшим экспертом и функцией контроля качества. Необходимая численность персонала определяется входным потоком, трудоемкостью маршрутов и резервом на неравномерность нагрузки.

Пятый – экономическую эффективность необходимо оценивать по совокупным затратам, включающим эксплуатацию системы, труд проверяющих и ущерб от ложных решений. Проверка отдельного случая оправдана, когда ожидаемый предотвращенный ущерб превышает ее стоимость, если иное не установлено обязательными нормами.

Тем самым гипотеза получает теоретическое подтверждение: риск-ориентированная верификация создает предпосылки для снижения ожидаемых затрат без сплошного дублирования автоматизированного процесса. Следующим этапом исследования должна стать эмпирическая апробация на обезличенной выборке с оценкой параметров формул (1)–(5), межэкспертной согласованности, устойчивости очереди и эффекта различных способов маршрутизации.

Библиографический список:

1. Абашкин В. Л., Абдрахманова Г. И., Бочаров М. Я., Вишневский К. О., Гохберг Л. М. и др. Индикаторы цифровой экономики: 2026: статистический сборник. М.: ИСИЭЗ НИУ ВШЭ, 2026. 304 с. doi: 10.17323/978-5-7598-3047-4.
2. Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» (с изм., внесенными Указом Президента РФ от 15.02.2024 № 124).
3. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD, 2023. doi: 10.6028/NIST.AI.100-1.
4. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. 2024. L, 2024/1689.
5. Parasuraman R., Manzey D. H. Complacency and bias in human use of automation: An attentional integration // Human Factors. 2010. Vol. 52, № 3. P. 381–410. doi: 10.1177/0018720810376055.
6. Bansal G., Wu T., Zhou J., Fok R., Nushi B., Kamar E., Ribeiro M. T., Weld D. S. Does the whole exceed its parts? The effect of AI explanations on complementary team performance // Proceedings of CHI 2021. New York: ACM, 2021. P. 1–16. doi: 10.1145/3411764.3445717.
7. Bućinca Z., Malaya M. B., Gajos K. Z. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making // Proceedings of the ACM on Human-Computer Interaction. 2021. Vol. 5, CSCW1. Article 188. doi: 10.1145/3449287.
8. Green B. The flaws of policies requiring human oversight of government algorithms // Computer Law & Security Review. 2022. Vol. 45. Article 105681. doi: 10.1016/j.clsr.2022.105681.
9. Ramchurn S. D., Stein S., Jennings N. R. Trustworthy human–AI partnerships // iScience. 2021. Vol. 24, № 8. Article 102891. doi: 10.1016/j.isci.2021.102891.

10. Veale M., Van Kleek M., Binns R. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making // Proceedings of CHI 2018. New York: ACM, 2018. P. 1–14. doi: 10.1145/3173574.3174014.

11. Elish M. C. Moral crumple zones: Cautionary tales in human–robot interaction // Engaging Science, Technology, and Society. 2019. Vol. 5. P. 40–60. doi: 10.17351/ests2019.260.

References:

1. Abashkin V. L., Abdrakhmanova G. I., Bocharov M. Ya., Vishnevskiy K.O., Gokhberg L. M. et al. Indikatory tsifrovoy ekonomiki: 2026: statisticheskiy sbornik [Digital Economy Indicators in the Russian Federation: 2026: Data Book]. Moscow, HSE ISSEK Publ., 2026. 304 p. doi: 10.17323/978-5-7598-3047-4. (In Russian).

2. Decree of the President of the Russian Federation № 490 of 10 October 2019 «On the Development of Artificial Intelligence in the Russian Federation», as amended by Decree № 124 of 15 February 2024. (In Russian).

3. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD, 2023. doi: 10.6028/NIST.AI.100-1.

4. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, 2024, L, 2024/1689.

5. Parasuraman R., Manzey D. H. Complacency and bias in human use of automation: An attentional integration. Human Factors, 2010, vol. 52, № 3, pp. 381–410. doi: 10.1177/0018720810376055.

6. Bansal G., Wu T., Zhou J., Fok R., Nushi B., Kamar E., Ribeiro M. T., Weld D. S. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of CHI 2021. New York, ACM, 2021, pp. 1–16. doi: 10.1145/3411764.3445717.

7. Buçinca Z., Malaya M. B., Gajos K. Z. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings

of the ACM on Human-Computer Interaction, 2021, vol. 5, CSCW1, article 188. doi: 10.1145/3449287.

8. Green B. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 2022, vol. 45, article 105681. doi: 10.1016/j.clsr.2022.105681.

9. Ramchurn S. D., Stein S., Jennings N. R. Trustworthy human–AI partnerships. *iScience*, 2021, vol. 24, № 8, article 102891. doi: 10.1016/j.isci.2021.102891.

10. Veale M., Van Kleek M., Binns R. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of CHI 2018*. New York, ACM, 2018, pp. 1–14. doi: 10.1145/3173574.3174014.

11. Elish M. C. Moral crumple zones: Cautionary tales in human–robot interaction. *Engaging Science, Technology, and Society*, 2019, vol. 5, pp. 40–60. doi: 10.17351/ests2019.260.

Для цитирования: Мирвелов Р.А., Улучшения деятельности трудовых ресурсов по выявлению отрицательного влияния нейросетей / Российский экономический интернет-журнал. – 2026. – № 3. URL:

© Мирвелов Р.А., Российский экономический интернет-журнал 2026, № 3