Философия науки и техники
2026. Т. 31. № 1. С. 138–152
УДК 167/168:004.8

Philosophy of Science and Technology
2026, vol. 31, no. 1, pp. 138–152
DOI: 10.21146/2413-9084-2026-31-1-138-152

ИССЛЕДОВАТЕЛЬСКИЕ ПРОГРАММЫ ЭПИСТЕМОЛОГИИ

П.Н. Барышников

Апостериорная компрессия истины: эпистемология и этика научного открытия в эпоху ИИ*

Барышников Павел Николаевич – доктор философских наук, доцент, профессор кафедры исторических социально-философских дисциплин, востоковедения и теологии. Пятигорский государственный университет. Российская Федерация, 357532, г. Пятигорск, Ставропольский край, пр. Калинина, д. 9; e-mail: pnbaryshnikov@pgu.ru

Статья посвящена исследованию методологических и этических аспектов научного открытия в эпоху стремительного прогресса систем искусственного интеллекта. Специфика указанного проблемного поля диктует принципы построения рассуждений. В статье рассмотрена связь эволюции эпистемологического статуса научного открытия методами машинных вычислений с аксиоматическими основаниями научного знания. Особое внимание уделено понятию апостериорной компрессии ранее открытых истин, выводимому из недавних открытий в математике, произведенных с помощью больших языковых моделей. Рассматриваемый пример указывает на то, что модель способна производить математически состоятельные шаги, не располагая ни пространственным созерцанием в кантовском смысле, ни категориальной структурой сознания, ни пониманием смысла доказательства. После прояснения природы подобного рода научных открытий становится возможной формулировка этических вопросов о прозрачности исследовательских алгоритмов, об ограничениях постфактумных методов объяснимого ИИ и о распределении ответственности между участниками научного процесса. В практической части работы разрабатывается матрица субсидиарной ответственности (ARMAI) как нормативный ответ на современные технологические и эпистемологические вызовы.

Ключевые слова: машинное научное открытие, апостериорная компрессия знания, эпистемологический статус ИИ, объяснимый искусственный интеллект (XAI), аксиоматические основания науки, субсидиарная ответственность в исследованиях

* Исследование выполнено за счет гранта Российского научного фонда № 25-28-00254, https://rscf.ru/project/25-28-00254/

© Барышников П.Н., 2026

1. Эпистемологический статус «машинного» научного открытия

Применение вычислительных машин и искусственных интеллектуальных систем в процедурах порождения нового научного знания насчитывает уже не одно десятилетие. После того, как в начале XX в. классические дедуктивные и индуктивные методы стали отходить на второй план, логика научного открытия была охарактеризована как набор эвристических принципов – в противовес классическому применению индукции или дедукции. Этот набор допущений должен привести к формулированию высказывания, истинность которого будет доказываться позднее более строгими методами [Carmichael, 1922, p. 586]. Поиск оснований эвристического подхода в науке был вытеснен нормативной философией науки, в рамках которой строго различалась контекст открытия (фактические процессы порождения гипотез и теорий) и контекст обоснования (их логическая проверка и валидация) [Arabatzis, 2006, p. 215]. До определенного этапа развития компьютерных технологий ментальные эвристические процессы оставались объектом психологии, социологии и прочих эмпирических наук. Философия науки при этом занималась логическим анализом нормативных научных высказываний.

Вместе с тем активная разработка языков программирования в середине XX в. привела к новому толкованию понятия «научное открытие». Если открытие понимать как результат перевода высказываний о неизвестном в разряд высказываний об известном, то такой подход позволяет интерпретировать в инженерном ключе научную проблему как задачу, а процесс порождения нового научного знания – как решение задачи. В рамках эвристического подхода решение задачи – это последовательность операций, ведущих от начального состояния пространства проблем к целевому состоянию. Расстояние между начальным и целевым состоянием преодолевается за счет смены состояний системы, управляемой ограничениями, которые задаются правилами. Решение задачи (научное открытие) в этом контексте становится обнаружением допустимых эффективных исследовательских действий и операций в рамках формальных правил или онтологических ограничений.

Построение подобных определений привело к появлению в середине XX в. вычислительной парадигмы в философии науки и построению эвристических компьютерных программ, работа которых состояла в описании правил для быстрого и эффективного поиска решения задачи. «Научное открытие» стало вычислимым. Логика научного открытия сводилась в этом случае к эвристическому поиску наиболее правдоподобных способов решения задач. Уровень компьютерных технологий того времени позволил имитировать на программно-аппаратных комплексах ход исследовательского мышления. Известно, что программа “Logic Theorist” [Ehrenfeucht, 1957] доказала 38 из 52 теорем из “Principia Mathematica” А. Уайтхеда и Б. Рассела. Эвристические алгоритмы BACON и KEKADA за счет разветвленного анализа формальных структуры теоретических высказываний смогли «переоткрыть» уже ранее сделанные открытия либо реконструировать исторический процесс открытия [Simon, 1991].

Последовательность машинной обработки данных выглядела следующим образом:

- ввод числовых данных для ряда переменных;
- проведение поиска в пространстве теоретических терминов;
- использование операторов, превращающих старые термины в новые;
- применение эвристических методов для выявления закономерности в данных;
- рекурсивное применение этих методов для формулирования отношений более высокого уровня [Langley, 2000].

В своем докладе «ИИ и философия науки в период с 1990-х по 2020 г.» Д. Джиллис и М. Джиллис дают исчерпывающее описание успехов стэнфордской группы эвристического программирования (Б. Бьюкенен, Э. Фейгенбаум и Э. Шортлифф): «Они обнаружили, что ключ к успеху заключается в извлечении из эксперта знаний, которые он или она использовали для выполнения специализированной задачи, а затем в кодировании этих знаний в компьютер. Таким образом, они смогли создать “экспертные системы”, которые выполняли конкретные задачи на уровне человеческих экспертов» [Gillies, Gillies, 2022, p. 68]. Системы, основанные на правилах, работали по принципу «если наблюдается то-то и то-то, то вероятный вывод такой». При экспоненциальном росте баз знаний данный метод работал во многих областях.

Можно сказать, что вычислительные методы 1970-х гг. указали на возможность тонкой механистической реконструкции процессов порождения нового научного знания, но не смогли описать логику научного открытия.

Эвристический подход сразу продемонстрировал собственные ограничения:

- новое свойство не встроено в теорию, описывающую это свойство, причины, его породившие, и принципы функционирования; зависимость параметров выводится лишь на основе регулярных корреляций без объяснения;
- при анализе данных из внешнего мира пространство вероятностей становится столь велико, что вычисления не могут закончиться за обозримое время;
- системы не выводят новых законов из аналитического перебора и классификации структур данных, следовательно, обнаруженный результат вычислений не является открытием в полном смысле этого слова [Alai, 2004].

На наш взгляд, особенно остро кризис машинной реализации принципов человеческого мышления проявился в некоторых подходах при конструировании компьютерных шахмат. Здесь можно привести пример из работ М.М. Ботвинника, когда при определении полной стоимости фигур, включающей среднюю стоимость и возможность создания угроз для фигур соперника, пришлось вводить в алгоритм комплексные числа [Ботвинник, 1968, с. 37].

При моделировании экспертных систем возникало две проблемы: 1) эксперты далеко не всегда представляли суть методов, которыми пользовались (гроссмейстер не может ответить на вопрос о том, как устроено мышление гроссмейстера); 2) извлечение знаний и их формальное описание оказались очень ресурсоемкими процедурами. Решение этих двух проблем стало

просматриваться с момента, когда вычислительные мощности позволили комбинировать классические методы (логика правил и деревья решений) с нейросетевыми подходами. Процесс поиска зависимостей и корреляций, извлечение значимых признаков в «сырых» массивах данных с последующим построением и обоснованием гипотез начал постепенно автоматизироваться.

Увеличение вычислительной мощности многослойных нейронных сетей и рост объемов обучающих примеров привели к новым вопросам в эпистемологии машинного научного открытия. Структуры правил заменялись вероятностями и весами, в связи с этим возникла проблема интерпретируемости результатов. Иерархии представлений (взаимосвязь идей) проявились через отношения огромного числа простых элементов, вследствие чего возникла парадоксальная ситуация: мощность научной индукции стала обратно пропорциональна научной значимости. Иерархия скрытых признаков может основываться на миллиардах весов, связь между которыми непонятна человеку. «Эти результаты опровергают утверждение Поппера о том, что индукция – это миф. Однако они не меняют его мнения о том, что индукция не играет роли в человеческой науке. Вместо этого они предполагают, что ИИ выполняет научную работу совершенно иначе, чем люди» [Gillies, Gillies, 2022, p. 77].

Эпистемологический статус машинного научного открытия, как мы видим, зависел от отношения вычислительных подходов в методологии науки к понятию открытия. Это отношение можно выразить через эволюционный вектор: начиная с эвристических программ и экспертных систем к непрозрачным нейросетевым архитектурам. Каждый отрезок такой оси содержит в себе фундаментальную точку напряжения, сохраняющую баланс (важный для философии науки) между продуктивностью вычислительного метода и его эпистемологической легитимностью. До недавнего времени баланс сохранялся за счет того, что ни одна из систем не претендовала на роль производителя нового знания в строгом смысле. Это были лишь сложные и изощренные исследовательские инструменты. Современные методы научных исследований, вдохновленные революционными успехами ИИ (так называемые AI-driven researches), приобретают совершенно новый статус, эпистемологические свойства которого требуют дальнейшего прояснения.

2. Вызов аксиоматическим основаниям науки и новое математическое знание

2.1. Апостериорная компрессия научных суждений

Новая ИИ-методология научного открытия приводит к тому, что аксиоматические основания научных высказываний перестают быть привычным фундаментом. Математика исторически занимает особое место среди наук именно потому, что истинность математических высказываний выражается через строгое формальное доказательство (часто приводящее к поиску тавтологий), а не через эмпирическое корреспондентное соответствие. Ум математика работает с так называемыми башнями абстракций – понятийными иерархиями, где

каждый новый уровень опирается на строго доказанные нижние. Важно указать: вертикаль абстракций удерживается в системе репрезентаций математического ума даже при работе на самых верхних «этажах». Математик упорядочивает и переупорядочивает свои идеи и убеждается в их истинности задолго до того, как сможет записать логическое доказательство. Эта идея подтверждается рассуждениями Р. Борчердса (лауреат Филдсовской премии 1998 г.): «Логическая последовательность появляется только в самом конце, и на самом деле довольно утомительно проверять, действительно ли все детали работают. До этого нужно все это собрать воедино с помощью множества экспериментов, догадок и интуиции» (цит. по: [Gibbs, 1998, p. 40]).

Обобщая все разнообразие подходов к кантианской интерпретации философии математики [Evstigneev, 2021], математическое знание можно отнести к формам чистого созерцания. Если порождение человеком априорных математических суждений можно представить как результат взаимодействия чистого созерцания и рассудочных категорий, то порождение математических высказываний современными моделями ИИ – это своеобразная «апостериорная компрессия» ранее открытых истин. «Человеческая» математика начинается с конструирования понятий в созерцании с последующим переходом к аксиомам, формальным ограничениям, определениям, постулатам, теоремам и задачам.

Научные прорывы в точных науках, сделанные сегодня методами ИИ, имеют совершенно иную эпистемологическую природу. ИИ-системы, обученные на чудовищных по своим объемам корпусах текстов и математических данных, извлекают паттерны из уже существующих результатов человеческого познания, не имея при этом (по крайней мере пока) ни пространственного созерцания, ни категориальной структуры сознания [Baryshnikov et al., 2025]. При этом отсутствие опыта построения математических репрезентаций никак не сказывается на способности ИИ-систем сохранять инструментальную эффективность даже при решении сложных и нетривиальных задач.

2.2. Генерация правомерных решений без понимания

Практические успехи ИИ на ниве научных открытий, с одной стороны, заставляют пересмотреть философские основания теории научного открытия, с другой – подталкивают к осторожным оценкам действительного вклада современных моделей в ход получения нового знания. Зачастую оказывается, что за громкими заголовками типа «ИИ решил математическую загадку тысячелетия» скрывается обычный инструментальный подход к решению рутинных задач.

Сегодня большинство исследований ИИ сосредоточено на функциональном сходстве и статистической оптимизации. Лишь небольшое количество работ посвящено изучению возможности реструктуризации собственной аксиоматической системы ИИ и его эволюции посредством перекрестной проверки с человеческими системами [Kim, 2025]. На наш взгляд, происходит смещение в интерпретации ценности доказательства в связи с тем, что моделям

позволяется выходить за ограничения формальных систем, если в итоге получается релевантный и (еще лучше) рентабельный результат.

Это смещение получает формальное обоснование в концепции кросс-аксиоматической эволюции [Kim, 2025]. При рассмотрении человека и ИИ в качестве независимых замкнутых формальных систем (в духе гёделевской теоремы) неспособность системы доказать истинность утверждения методами самой системы перестает быть приговором. Система расширяется именно там, где внутреннее доказательство недостижимо, принимая внешние аксиомы через функцию верификации: если предложение согласуется с существующим множеством теорем, оно включается в систему, порождая новые ветви вывода и необратимо увеличивая ее информационную сложность. При таком подходе объяснимость и приемлемость становятся непересекающимися критериями. С. Ким вводит понятие неявных аксиом – утверждений, для которых внутри системы не существует прослеживаемого пути вывода, но которые внешняя система способна верифицировать и принять. Такое положение вещей возникает в реальной практике математической работы с большими языковыми моделями: результат проверяется через согласование с ожиданиями внешнего наблюдателя, исследователя, пользователя, рецензента.

Приведем пример. В 2025 г. исследователи из Люксембургского университета поставили перед моделью GPT-5 исследовательскую задачу из теории вероятностей: представить количественную версию теоремы четвертого момента [Diez et al., 2025]. В математике моменты функции связаны с количественными измерениями изменений формы графика функции. Теорема имеет специфичную значимость, т.к. вскрывает неочевидную связь между вероятностным распределением и алгебраической структурой случайной величины (так называемый винеровский хаос). Уже издавались работы, которые доказывали, что существует сходимость четвертого момента нормированных кратных интегралов к числу 3. Но скорость этой сходимости оставалась неизвестной [Döbler et al., 2018].

Сама по себе задача не была концептуально новой. Но к моменту эксперимента не существовало опубликованного решения количественного результата сходимости. Сложность заключалась в рутинной работе с длинными формулами, комбинаторными коэффициентами и разложениями. Здесь GPT-5 смог выступить в роли «исполнителя», который быстро перебрал варианты и выдал структуру доказательства. Важно то, что модель допускала ошибки, характер которых указывает как на сильные, так и на слабые стороны архитектуры трансформеров, участвующих в научных исследованиях.

Слабые стороны инструментального подхода подробно описаны авторами эксперимента. Их можно свести к нескольким типам.

- *Поверхностность и безоценочность*. Не имея глубокого понимания сути исследуемых данных, модель при оперировании последовательностями символов не может оценить, какое утверждение является «красивым» или «сильным», а какое – лишь техническим и, возможно, избыточным. Это приводит к игнорированию внутренних связей рассуждений и потере информации о зависимостях.

- *Склонность к слабо детектируемым галлюцинациям и упрощениям*. В гауссовском случае модель попыталась применить «школьное» представление о перемножении, проигнорировав сложную комбинаторику (симметризация и свертка) при многомерном анализе.
- *Горизонт контекста*. Модель не смогла самостоятельно интегрировать информацию из двух разных частей предоставленной для анализа вспомогательной статьи [Döbler et al., 2018]. В ответе от ИИ были просто перечислены все члены, которые «мешают» в разложении четвертого момента. При этом были пропущены связи, которые раскрывались через лемму 2.4 из указанной работы. Во всех случаях проблемы были обнаружены и преодолены после экспертной оценки со стороны человека. Человек давал прямые указания на ошибки, тем самым активируя новые контексты. И только тогда модель успешно выходила из затруднения.

Сильные стороны применения моделей ИИ в подобных исследованиях, которые демонстрируют несомненную полезность такого инструмента с учетом перечисленных ограничений, можно представить следующим списком.

1. *Скорость обработки научных источников*.
2. *Владение стандартным математическим аппаратом*. Модель уверенно применяла классические инструменты – метод Маллявэна – Штейна, что позволило ей с первой попытки выбрать правильную стратегию доказательства.
3. *Способность к удержанию контекста при корректировке*. При указании экспертом на ошибки модель реактивировала рассуждения в новом контексте, не теряя цепочку предыдущих рассуждений.
4. *Объемы рутинной работы*. Авторы указывают на то, что решение рутинных задач (символьные преобразования, комбинаторные упрощения, численные проверки, итеративное уточнение, оформление результата в LaTeX) без применения GPT-5 заняло бы колоссальное количество времени. При этом не исключается, что технических ошибок при «ручных» расчетах было бы больше [Diez et al., 2025, p. 11].

Важно отметить, что GPT-5 хорошо справляется с задачами, которые требуют комбинирования известного. При этом трудности возникают при решении задач, требующих подлинно новой идеи. В статье указывается, что работа с моделью была похожа на работу с молодым талантливым ассистентом, который может быть исполнительным, но при этом не понимать всего замысла исследования. Без постоянного контроля и подсказок со стороны математиков GPT-5 допустил бы существенные ошибки. Тем не менее исследование достигло значимых результатов, ранее в литературе не встречавшихся. Подчеркнем еще раз, что модель производила математически состоятельные шаги, не располагая интуицией, опытом чистого созерцания или тем, что математик назвал бы пониманием доказательства. Для философии науки остается открытым вопрос: является ли такой прирост научного знания или такой полумеханический перевод области неизвестного в область известного

научным открытием? Влияет ли на эпистемологический статус научного открытия инструментарий, проверяемость и прозрачность промежуточных выводов?

3. Этические и юридические аспекты научных исследований с использованием ИИ

3.1. Проблема прозрачности алгоритмов

Помимо философской проблемы, связанной с расчетами, которые «идут в темноте», существует классическая методологическая проблема «черного ящика», и она сегодня находит новые нестандартные решения. Проблема «черного ящика» связана с тем, что непрозрачность процедур нейросетевого вывода повышает риск принятия результатов исследования на основе не интерпретируемых процессов и не верифицируемых данных. В этом случае рост эффективности научных исследований обратно пропорционален объяснимости процедур вывода.

Несложно представить себе ситуацию, в которой обучающий набор данных будет размечен заведомо ошибочным образом так, чтобы у модели сформировались ошибочные ассоциации. Такая тонкая настройка позволяет внедрить в модель любое систематическое искажение, которое впоследствии будет воспроизводиться уверенно, без каких-либо видимых признаков ошибки.

Сегодня в «чувствительных сферах» некоторых научных исследований используются постфактумные методы для построения Объяснимого ИИ (Explainable Artificial Intelligence – XAI), такие как LIME (Local Interpreatable Model-Agnostic Explanations) и SHAP (SHapley Additive exPlanations). В первом случае конкретный входной пример «зашумляется» через создание схожих ошибочных вариантов, и после этого анализируются изменения выхода модели. Простая линейная регрессия указывает на признаки, которые больше всего повлияли на конкретное «правильное» предсказание. Во втором случае признак получает «справедливую долю» ответственности за предсказание, рассчитанную по всем возможным комбинациям признаков, что позволяет теоретически обосновать степень вклада признаков в итоговый вывод [Salih et al., 2025]. Такие постфактумные модельно-агностичные (описывающие модель как «черный ящик») подходы действительно способны повысить уровень доверия пользователя к конечному результату. Но:

1. во-первых, (в случае с LIME) коллинеарность (сильная корреляция) признаков порождает нереалистичные комбинации объяснений;
2. во-вторых, (в случае с SHAP) два признака, фактически несущие одну и ту же информацию, могут получить совершенно разные оценки важности в зависимости от порядка перебора.

Выяснилось, что постфактумный характер объяснений не способен выявить искусственные искажения в обучающих данных: «Объяснение может быть правдоподобным, не будучи верным. Два случая с одинаковой точностью предсказания могут давать противоречивые локальные объяснения для одного

и того же входного примера» [Liu et al., 2025, p. 8]. С. Лю с соавторами вводят понятие «маргинальной интерпретируемости», где «добавление каждого нового слоя объяснения приносит все меньший прирост понимания» [Ibid., p. 5]. Также отметим, что большинство работ, посвященных XAI, исследуют проблему точности и интерпретируемости, опуская то, как постфактумные объяснения воспринимаются различными группами пользователей.

Возможно, решение проблемы лежит в плоскости взаимодействия слоев верификации с самообъясняющимися агентами ИИ, которые должны встраивать прозрачность непосредственно в архитектуру моделей. В этом случае прозрачность и интерпретируемость становятся социальной нормой для применения ИИ как в научных исследованиях, так и в иных прикладных областях.

3.2. Проблема распределения ответственности в научных исследованиях с применением систем ИИ

В совершенно ином свете предстает использование ИИ в ходе научных исследований при его юридическом прочтении.

Проблема распределения ответственности в человеко-технических системах насчитывает уже не одно десятилетие. За последнее время тема распределения ответственности между всеми участниками деликтных отношений с участием ИИ перетекла из теоретико-правовой области в область конкретных правоприменительных решений. Под деликтными отношениями здесь понимается всякое противоправное действие, вторгающееся в личную или имущественную сферу личности и причиняющее ей ущерб, независимо от гражданско-правовых отношений, существующих между лицами [Кирин, Плигин (ред.), 2018].

Широкое использование систем ИИ в различных отраслях экономики и управленческих процессах приводит к тому, что ИИ становится существенным фактором правовых отношений [Казанцев, 2025]. С одной стороны, сегодня не представляется возможным признать за системами ИИ статус субъекта правовых отношений. (Попытки выглядят, скорее, аномалиями, под которые еще не разработан адекватный юридический язык [Вавилин, 2022].) С другой стороны, так же неприемлемо приравнивать действия ИИ к форс-мажору и освобождать причастных лиц от ответственности. При применении ИИ-инструментов в реальных производственных процессах может использоваться многомерная матрица субсидиарной ответственности (далее – матрица Казанцева), позволяющая адаптивно регулировать отношения с участием ИИ без создания принципиально новых правовых институтов. Основной принцип работы матрицы выглядит следующим образом: выделяется семь категорий систем ИИ – от высокорисковых (угроза жизни и здоровью) до низкорисковых. Для каждой категории предлагается иерархия презумпций виновности. Например, для высокорисковых систем, способных угрожать жизни людей, приоритетная ответственность возлагается на заказчика и разработчика; для низкорисковых – в первую очередь на пользователя [Казанцев, 2025, с. 464].

Мы полагаем, что такая структура вполне применима при регулировании процедур научного поиска, основанных на системах ИИ. Если сравнить

принципы «Этического руководства для доверенного ИИ» от Европейской Комиссии [Ethics Guidelines for Trustworthy AI, 2019] с матрицей Казанцева, то становится очевидным, что «Этическое руководство» задает нормативно-правовую архитектуру, в рамках которой матрица Казанцева выглядит как прикладной правовой механизм ее реализации.

Вначале зададим параметры субъектов ответственности и их функции:

- исследователь (выбор модели, интерпретация);
- научное учреждение (стандарты, надзор);
- разработчик модели (дефекты алгоритма, низкое качество данных);
- издательские платформы / институт рецензензирования (проверка методологии);
- финансирующий орган (требования воспроизводимости);
- поставщик данных.

Если адаптировать утвержденные механизмы регулирования к этическим проблемам научного открытия методами ИИ, то можно построить схему, связывающую субъектов ответственности (от исследователя до поставщика данных) с регуляторами. Для удобства обозначим матрицу распределения ответственности за результаты научных исследований с применением ИИ англоязычной аббревиатурой "ARMAI" – Accountability and Responsibility Matrix for AI research[1].

Полная версия ARMAI с расширением в концептуальную схему и перечислением регулирующих систем представлена в открытом репозитории[2]. Здесь для иллюстрации рассуждений представим сокращенную версию (Таблица).

Таблица. ARMAI – матрица (сокращенная) распределения ответственности за результаты научных исследований с применением ИИ

Субъект	**Роль в исследовании**	**Основания ответственности**
Исследователь	Выбор модели; интерпретация результатов; соблюдение этических стандартов (декларирование ИИ)	Неподходящая модель; ложная интерпретация; недекларирование ИИ; фабрикация данных
Научное учреждение	Разработка политики использования ИИ; надзор за исследовательским процессом	Отсутствие политики использования ИИ; непринятие мер при известных нарушениях

[1] Матрица распределения ответственности за результаты научных исследований с применением ИИ разработана автором на основании работ [Казанцев, 2025; Ethics Guidelines for Trustworthy AI, 2019] с применением языковой модели Claude Sonnet 4.6 (Anthropic, 2025). Структура, классификации и содержательные суждения верифицированы автором самостоятельно.

[2] https://github.com/kagort/RNF_files/blob/main/research_responsibility_matrix.pdf

Разработчик модели	Качество алгоритма и обучающих данных; поддержка и контроль версий	Скрытые дефекты; ложные заявления о точности; предвзятые данные; сокрытие уязвимостей
Издательские платформы / институт рецензирования	Проверка методологии; управление ретракциями и исправлениями	Публикация несостоятельных работ; задержка ретракции; конфликт интересов рецензентов
Финансирующий орган	Требования воспроизводимости; управление конфликтом интересов	Отсутствие требований к открытости данных; вмешательство в результаты исследования
Поставщик данных	Качество и происхождение данных; соблюдение прав субъектов	Скрытые ошибки в данных; отсутствие провенанса (истории происхождения обучающего набора данных); нарушение 152-ФЗ «О персональных данных»; ненадлежащая анонимизация

ARMAI воспроизводит логику многомерной модели субсидиарной ответственности Казанцева [Казанцев, 2025], адаптируя ее к специфике науки: место «владельца источника повышенной опасности» занимает разработчик высокорисковой модели, в качестве «регулирующего органа» выступают издательские платформы и этический комитет, исследователь является аналогом «пользователя», несущего ответственность за корректность применения системы в пределах своей профессиональной компетенции. Категория риска определяется последствиями ошибки для людей и общества. Деятельность субъекта регулируется и оценивается нормативной системой, представленной различными институтами, распределенными по перечисленным шести субъектам.

Результаты анализа матрицы распределения ответственности позволяют заключить, что правовое регулирование ответственности за ход научных исследований с применением систем ИИ и их конечные результаты не может строиться на основе некого универсального принципа. Структура науки как особого способа познания мира и как социального института не может принять ни полной безнаказанности разработчика, апеллирующего к автономности модели, ни снятия ответственности с исследователя, не способного объяснить процессы «черного ящика». Нормативные системы сегодня быстро адаптируются вслед за ростом производительности моделей ИИ и постоянным расширением областей их применения. Разумеется, такая гипотетическая схема не может решить проблему фрагментарности распределения ответственности и сразу перейти в плоскость правоприменительных решений. Но, на наш взгляд, такую концептуальную схему можно использовать в качестве ориентира для разработки конкретных решений.

Заключение

История машинизации процедур научного открытия последовательно приводит человечество к сегодняшнему состоянию, когда вопрос «является ли автоматизированный перевод области неизвестного в область известного научным открытием в полном смысле этого слова?» остается без внятного ответа. Вектор развития от первых эвристических программ и экспертных систем 1970-х гг. к современным непрозрачным нейросетевым архитектурам устремлен туда, где системы ИИ начинают претендовать на роль производителей подлинно нового и по-настоящему ценного научного знания. Человек с трудом успевает выработать концептуальный язык для юридического и этического регулирования процессов, стремительно меняющих как общий методологический ландшафт (достаточно указать на интенсивность датафикации науки), так способы получения научных открытий. Если осмелиться сделать прогноз, то ввиду эффективности ИИ-систем можно сказать, что наиболее перспективными будут гибридные способы взаимодействия человека-исследователя и искусственного ассистента, решающего рутинные задачи, но при этом постоянного расширяющего области доверия. Останется ли в этом союзе место человеку, покажет время.

Список литературы

Ботвинник, 1968 – *Ботвинник М.М.* Алгоритм игры в шахматы. М.: Наука, 1968. 94 с.

Вавилин, 2022 – *Вавилин Е.В.* Статус искусственного интеллекта: от объекта к субъекту правовых отношений // Вестник Томского государственного университета. Право. 2022. № 45. С. 147–158.

Казанцев, 2025 – *Казанцев Д.А.* Правовые механизмы распределения ответственности за вред, причиненный системами искусственного интеллекта // Journal of Digital Technologies and Law. 2025. Т. 3. № 3. С. 446–471.

Кирин, Плигин (ред.), 2018 – Реформа административной ответственности в России / Под ред. А. Кирина, В. Плигина. М.: Национальный исследовательский университет «Высшая школа экономики», 2018. 480 с.

Alai, 2004 – *Alai M.* AI Scientific Discovery and Realism // Minds and Machines. 2004. Vol. 14. No. 1. P. 21–42.

Baryshnikov et al., 2025 – *Baryshnikov P., Velis L., Atakuev M.* Natural Language Processing and the Representation of Phenomenal Experience // Technology and Language. 2025. Vol. 6. No. 4. P. 217–239.

Carmichael, 1922 – *Carmichael R.D.* The Logic of Discovery // Monist. 1922. Vol. 32. No. 4. P. 569–608.

Ethics Guidelines for Trustworthy AI, 2019 – Ethics Guidelines for Trustworthy AI. European Commission, Directorate-General for Communications Networks. Luxembourg: Publications Office, 2019. 41 p. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (дата обращения: 02.03.2026).

Diez et al., 2025 – *Diez C.-P., Maia L. da, Nourdin I.* Mathematical Research with GPT-5: A Malliavin-Stein Experiment // arXiv:2509.03065 [math]. URL: https://arxiv.org/abs/2509.03065 (дата обращения: 02.03.2026).

Döbler et al., 2018 – *Döbler C., Vidotto A., Zheng G.* Fourth Moment Theorems on the Poisson Space in Any Dimension // Electronic Journal of Probability. 2018. Vol. 23. DOI: 10.1214/18-EJP168.

Ehrenfeucht, 1957 – *Ehrenfeucht A.* Allen Newell and Herbert A. Simon. The Logic Theory Machine. A Complex Information Processing System // Journal of Symbolic Logic. 1957. Vol. 22. No. 3. P. 331–332.

Evstigneev, 2021 – *Evstigneev M.D.* What is Kantian Philosophy of Mathematics? An Overview of Contemporary Studies // Kantian Journal. 2021. Vol. 40. No. 2. P. 151–178.

Gibbs, 1998 – *Gibbs W.W.* Monstrous Moonshine is True // Scientific American. 1998. Vol. 279. No. 5. P. 40–41.

Gillies, Gillies, 2022 – *Gillies D., Gillies M.* Artificial Intelligence and Philosophy of Science from the 1990s to 2020 // Current Trends in Philosophy of Science: A Prospective for the Near Future / Ed. by W.J. Gonzalez. Cham: Springer International Publishing, 2022. P. 65–79.

Kim, 2025 – *Kim S.* Cross-Axiomatic Evolution: A Formal Model of Mutual Axiom Acceptance Between Humans and Artificial Intelligence: PhD Thesis. Independent Researcher, 2025. URL: https://philpapers.org/archive/KIMCEA-7.pdf (дата обращения: 02.03.2026).

Langley, 2000 – *Langley P.* The Computational Support of Scientific Discovery // International Journal of Human-Computer Studies. 2000. Vol. 53. No. 3. P. 393–410.

Liu et al., 2025 – *Liu X., Huang D., Yao J. et al.* From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI) // AI. 2025. Vol. 6. No. 11. DOI: 10.3390/ai6110285.

Salih et al., 2025 – *Salih A.M., Raisi-Estabragh Z., Galazzo I.B. et al.* A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME // Advanced Intelligent Systems. 2025. Vol. 7. No. 1. DOI: 10.1002/aisy.202400304.

Arabatzis, 2006 – *Arabatzis T.* On the Inextricability of the Context of Discovery and the Context of Justification // Revisiting Discovery and Justification. Dordrecht: Kluwer Academic Publishers, 2006. P. 215–230.

Simon, 1991 – *Simon H.A.* Comments on the Symposium on "Computer Discovery and the Sociology of Scientific Knowledge" // Social Studies of Science. 1991. Vol. 21. No. 1. P. 143–148.

A posteriori truth compression: epistemology and ethics of scientific discovery in the age of AI

Pavel N. Baryshnikov

Pyatigorsk State University. 9 Kalinin Avenue, Pyatigorsk, 357532, Stavropol Krai, Russian Federation; e-mail: pnbaryshnikov@pgu.ru

This article explores the methodological and ethical aspects of scientific discovery in the era of the rapid development of artificial intelligence systems. The logic of this problematic field's development dictates the principles of reasoning. The article examines the relationship between the evolution of the epistemological status of scientific discovery using machine computation methods and the axiomatic foundations of scientific knowledge. Particular attention is paid to the concept of a posteriori compression of previously discovered truths, derived from recent mathematical discoveries made using large language models. This example demonstrates that the model is capable of producing mathematically sound steps without possessing spatial contemplation in the Kantian sense, a categorical structure of consciousness, or an understanding of the meaning of proof. After clarifying the nature of such scientific discoveries, it becomes possible to formulate ethical questions about the transparency of research algorithms, the limitations of post-hoc methods of explainable AI, and the distribution of responsibility among participants in the scientific process. In the practical part of the paper, a subsidiary responsibility matrix (ARMAI) for science is developed as a normative response to modern technological and epistemological challenges.

Keywords: machine scientific discovery, a posteriori knowledge compression, epistemological status of AI, explainable artificial intelligence (XAI), axiomatic foundations of science, subsidiary liability in research

Acknowledgement: This study was supported by a grant from the Russian Science Foundation, No. 25-28-00254, https://rscf.ru/project/25-28-00254/

References

Alai, M. AI "Scientific Discovery and Realism", *Minds and Machines*, 2004, vol. 14, no. 1, pp. 21–42.

Baryshnikov, P., Velis, L., Atakuev, M. "Natural Language Processing and the Representation of Phenomenal Experience", *Technology and Language*, 2025, vol. 6, no. 4, pp. 217–239.

Botvinnik, M.M. *Algoritm igry v shakhmaty* [The Algorithm of Chess]. Moscow: Nauka Publ., 1968. 94 pp. (In Russian)

Carmichael, R.D. "The Logic of Discovery", *Monist*, 1922, vol. 32, no. 4, pp. 569–608.

Diez, C.-P., Maia, L. da, Nourdin, I. "Mathematical Research with GPT-5: A Malliavin-Stein Experiment", *arXiv:2509.03065 [math]*. URL: https://arxiv.org/abs/2509.03065 (accessed: 02.03.2026).

Döbler, C., Vidotto, A., Zheng, G. "Fourth Moment Theorems on the Poisson Space in Any Dimension", *Electronic Journal of Probability*, 2018, vol. 23. DOI: 10.1214/18-EJP168.

Ehrenfeucht, A. "Allen Newell and Herbert A. Simon. The Logic Theory Machine. A Complex Information Processing System", *Journal of Symbolic Logic*, 1957, vol. 22, no. 3, pp. 331–332.

Ethics Guidelines for Trustworthy AI. European Commission, Directorate-General for Communications Networks. Luxembourg: Publications Office, 2019. 41 pp. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on: 02.03.2026).

Evstigneev, M.D. "What is Kantian Philosophy of Mathematics? An Overview of Contemporary Studies", *Kantian Journal*, 2021, vol. 40, no. 2, pp. 151–178.

Gibbs, W.W. "Monstrous Moonshine is True", *Scientific American*, 1998, vol. 279, no. 5, pp. 40–41.

Gillies, D., Gillies, M. "Artificial Intelligence and Philosophy of Science from the 1990s to 2020", *Current Trends in Philosophy of Science: A Prospective for the Near Future*, ed. by W.J. Gonzalez. Cham: Springer International Publishing, 2022, pp. 65–79.

Kazantsev, D.A. "Pravovye mekhanizmy raspredeleniya otvetstvennosti za vred, prichinennyi sistemami iskusstvennogo intellekta" [Legal Mechanisms for Distributing Liability for Harm Caused by Artificial Intelligence Systems], *Journal of Digital Technologies and Law*, 2025, vol. 3, no. 3, pp. 446–471. (In Russian)

Kim, S. *Cross-Axiomatic Evolution: A Formal Model of Mutual Axiom Acceptance Between Humans and Artificial Intelligence*. PhD Thesis. Independent Researcher, 2025. URL: https://philpapers.org/archive/KIMCEA-7.pdf (accessed: 02.03.2026).

Langley, P. "The Computational Support of Scientific Discovery", *International Journal of Human-Computer Studies*, 2000, vol. 53, no. 3, pp. 393–410.

Liu, X., Huang, D., Yao, J. et al. "From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI)", *AI*, 2025, vol. 6, no. 11. DOI: 10.3390/ai6110285.

Reforma administrativnoi otvetstvennosti v Rossii [Reform of Administrative Liability in Russia], ed. by A. Kirin, V. Pligin. Moscow: National Research University Higher School of Economics Publ., 2018. 480 pp. (In Russian)

Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B. et al. "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME", *Advanced Intelligent Systems*, 2025, vol. 7, no. 1. DOI: 10.1002/aisy.202400304.

Arabatzis, T. “On the Inextricability of the Context of Discovery and the Context of Justification”, in: *Revisiting Discovery and Justification*. Dordrecht: Kluwer Academic Publishers, 2006, pp. 215–230.

Simon, H.A. “Comments on the Symposium on ‘Computer Discovery and the Sociology of Scientific Knowledge’”, *Social Studies of Science,* 1991, vol. 21, no. 1, pp. 143–148.

Vavilin, E.V. “Status iskusstvennogo intellekta: ot ob’’ekta k sub’’ektu pravovykh otnoshenii” [The Status of Artificial Intelligence: from Object to Subject of Legal Relations], *Vestnik Tomskogo gosudarstvennogo universiteta. Pravo,* 2022, no. 45, pp. 147–158. (In Russian)