

ПРЕДИКТИВНАЯ МЕТОДОЛОГИЯ МАШИННОГО ОБУЧЕНИЯ И ПОИСК МЕХАНИЗМОВ ПОЗНАНИЯ И ПОВЕДЕНИЯ

Сущин Михаил Александрович – канд. филос. наук, старший научный сотрудник, Институт научной информации по общественным наукам РАН (ИНИОН РАН), Москва, Россия; sushchin@bk.ru; ORCID 0000-0002-8805-6716

Аннотация. В статье рассматриваются возможности методологического влияния современной области машинного обучения на психологию и нейронауку для улучшения предсказывающего потенциала нейропсихологических моделей. Обсуждается аргумент сторонников предиктивных теорий, что преодоление проблем исследований, вызвавших кризис воспроизводимости в психологии, может быть достигнуто посредством использования больших датасетов, аккуратной оценки предсказательных ошибок модели, а также упрощения модели за счет устранения несущественных коэффициентов. В этой связи указывается, что критикуемые предиктивистами исследования с небольшой выборкой не являются образцом объяснительных психологических моделей. Теории в когнитивной психологии, когнитивной нейронауке, нейропсихологии и психофизиологии в большей степени имеют отношение к идентификации причинных механизмов познания и поведения. Вероятно, в большей степени точности и в больших возможностях манипуляции каузальными факторами познания и поведения кроется важнейшее различие между теориями механистических объяснений и вдохновленными машинным обучением предиктивными моделями. Последние в отрыве от поиска реальных причинных механизмов могут ухватывать только поверхностные корреляции между событиями. При принятии важных решений в контексте образовательной, биомедицинской и правовой деятельности лучше опираться на знание каузальных механизмов познания и поведения. Соответственно, предиктивные модели не могут вытеснить теории, ищущие механистические объяснения познания и поведения. Вместо выбора между предсказательной и объяснительной методологиями более продуктивным было бы нахождение оптимального баланса между ними.

Ключевые слова: психология; машинное обучение; предсказание; объяснение; методология науки.

Для цитирования: Сущин М.А. Предиктивная методология машинного обучения и поиск механизмов познания и поведения // Научоведческие исследования. – 2026. – № 3. – С. 97–112. – DOI: <http://www.doi.org/10.31249/scis/2026.03.08>

Статья поступила в редакцию: 09.06.2026

Принята к публикации: 26.06.2026

PREDICTIVE METHODOLOGY OF MACHINE LEARNING AND THE SEARCH FOR MECHANISMS OF COGNITION AND BEHAVIOR

Mikhail A. Sushchin – Candidate of Sciences (Philosophy), Senior Researcher, Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences (INION RAN), Moscow, Russia; sushchin@bk.ru; ORCID 0000-0002-8805-6716

Abstract. This article considers the potential methodological impact of the modern field of machine learning on psychology and neuroscience to improve the predictive potential of neuropsychological models. It discusses the argument advanced by proponents of predictive theories that overcoming the research problems that have led to the reproducibility crisis in psychology can be achieved through the use of large datasets, careful evaluation of model prediction errors, and model simplification by dropping irrelevant coefficients. In this respect, it is noted that small-sample studies criticized by predictive theorists cannot be considered a paragon of explanatory psychological models. Theories in cognitive psychology, cognitive neuroscience, neuropsychology, and psychophysiology are more deeply involved in the business of identifying causal mechanisms of cognition and behavior. Theories of mechanistic explanations and machine-learning-inspired predictive models may be distinguished most importantly by the former's greater precision and greater potential for manipulating the causal factors of cognition and behavior. The latter, when disconnected from the search for real causal mechanisms, might only capture correlations between events. When making important decisions in the context of educational, biomedical, and legal practices, it is better to rely on knowledge of the causal mechanisms of cognition and behavior. Accordingly, predictive models cannot replace theories that strive for mechanistic explanations of cognition and behavior. Rather than choosing between predictive and explanatory methodologies, it would be more fruitful to search for an optimal balance between them.

Keywords: psychology; machine learning; prediction; explanation; methodology of science.

For citation: Sushchin, M.A. Predictive methodology of machine learning and the search for mechanisms of cognition and behavior // *Naukovedcheskie issledovaniya = Science Studies.* – 2026. – N 3. – P. 97–112. – DOI: <http://www.doi.org/10.31249/scis/2026.03.08>

The paper was submitted: 09.06.2026

Accepted for publication: 26.06.2026

О недавнем влиянии области искусственного интеллекта на науку

Недавние громкие успехи области искусственного интеллекта находят у всех на слуху. Менее чем за десять лет большие языковые модели (БЯМ) превратились из рудиментарных экспериментальных моделей в мощный инструмент, по своему влиянию на общественную жизнь сопоставимый со ставшими привычными поисковыми системами, электронной почтой, видеохостингами и т.д. Данное влияние, разумеется, не могло не распространиться и на науку и систему образования. За короткий срок обсуждение возможностей и рисков применения БЯМ в научном и образовательном процессах породило обширную литературу.

Так, почти сразу после ошеломляющего успеха ChatGPT и Github Copilot в 2022 г. многие авторы и авторские коллективы стали обсуждать те значительные возможности, которые открылись перед учеными, преподавателями и учащимися с появлением БЯМ. В этой связи чаще всего отмечается способность БЯМ моментально делать краткие обзоры каких-либо отдельных работ или даже целых областей знания, отталкиваясь от уровня пользователя, разъяснять простым языком сложные вещи (скажем, основы программирования, высшей математики или квантовой физики), корректировать тексты и переводить их на другие языки (а также давать консультации по тончайшим аспектам перевода), генерировать (и разъяснять) компьютерный код на всех основных языках программирования и др. Таким образом, у представителей научного и образовательного сообществ появился легкодоступный мощный персональный ассистент, способный облегчить выполнение многих видов работ.

Равным образом значительное внимание было приковано и к рискам, связанным с использованием БЯМ в науке и образовании. Корень проблем может быть во многом связан с природой БЯМ. Как известно, в своей основе БЯМ представляют собой автоматы, осуществляющие генерацию следующего наиболее *вероятного* в данном контексте символа. Ввиду этого с их ответами связана принципиальная ненадежность, невозможность гарантировать предоставление исключительно корректной информации, подчас проявляющаяся в том, что даже на элементарный вопрос машина может дать неверный ответ. (Часто генерация всецело вымышленной информации проявляется при запросе списка литературы по какой-либо теме.) Данная проблема усугубляется возрастающей степенью неотличимости ответов БЯМ от ответов простых людей и даже экспертов. Наконец, само обстоятельство активного распространения БЯМ ставит проблему излишней зависимости пользователей от такого рода интеллектуальных инструментов. Так, способность БЯМ в считанные мгновения делать обзор отдельных работ или целых областей знания создает искушение пренебрежения традиционными аналитическими видами работ с текстами, что может иметь губительные последствия для формирования навыков критического анализа информации и творческих способностей.

Впрочем, эти и иные схожие вопросы применения БЯМ в науке и образовании затрагивают лишь поверхностные аспекты влияния разработок искусственного интеллекта (ИИ) на эти области. В этом отношении наука испытывает такое же влияние, как другие области жизни современного человека. Между тем ажиотаж последних лет оттеснил на второй план гораздо более радикальные и глубокие попытки концептуального влияния современной области машинного обучения на научные исследования. Так, еще за некоторое время до возникновения БЯМ в психологии появился новый тип предсказывающих моделей, вооружившихся ключевыми методологическими принципами области машинного обучения. В этой связи один из главных сторонников моделей нового типа психолог

Т. Яркони выступил с чрезвычайно смелым тезисом, что психология должна отдать приоритет способности предсказания поведения человека, а не главенствующим уже долгое время попыткам его объяснения [Yarkoni, Westfall, 2018]. По мнению Яркони и других адептов этого вида предиктивизма, приоритет предсказаниям может быть отдан даже *в ущерб* объяснительной активности. Вдохновленное машинным обучением предиктивное моделирование уже было охарактеризовано как «поворот» в психологии и нейронауках [Weiskopf, 2022]. Вместе с тем обоснованность некоторых утверждений сторонников этого направления вызывает сомнения. В этой статье будет предпринята попытка оценить статус предиктивных моделей нового типа в связи с более традиционными глубокими причинными объяснительными теориями в психологии и нейронауке.

Современные предиктивные модели: теоретическая мотивация и конкретные примеры

Традиционно значительная часть психологических исследований (особенно в области психологии личности и социальной психологии) фокусировалась на нахождении статических показателей, факторов, которые позволяли бы предсказывать поведение человека (predictors). Например, как добросовестность может влиять на производительность труда или академическую успеваемость, как психологические травмы в детстве могут обуславливать депрессию во взрослом возрасте, как размер подкупа может влиять на самовнушение [Festinger, Carlsmith, 1959] и т.п.

Подобного рода исследования обычно проводились на основе небольшой выборки данных (небольшого количества участников) с использованием экспериментального протокола, именуемого проверкой значимости нулевой гипотезы. В основе этого протокола лежит принятие в качестве базовой так называемой нулевой гипотезы H_0 – предположения об отсутствии причинной связи между изучаемыми феноменами. Одновременно с этим выдвигается противоположная гипотеза H_1 о наличии связи между данными феноменами. После сбора данных проводится статистический тест для расчета вероятности нулевой гипотезы p . Если $p < 0.05$ (менее одного шанса из двадцати), заключают о верности гипотезы H_1 , а нулевую гипотезу отбрасывают. Как отмечают в этой связи Д. Иглесиас, М. Соррель и Р. Олмос: «Если обнаруживаются статистически значимые результаты (т.е. $p < 0.05$), делается вывод, что рассматриваемая теория прошла проверку на соответствие действительности (passed the truth test) и, таким образом, предлагает объяснение на основе эмпирических свидетельств» [Iglesias, Sorrel, Olmos, 2025].

В теории исследования, проведенные в соответствии с указанными стандартами, должны быть достаточны и для *объяснения*, и для *предсказания* поведения человека. «С этой точки зрения, объяснение необходимым

образом способствует предсказанию: модель, которая в наилучшей степени аппроксимирует ментальные процессы, порождающие наблюдаемое поведение, также лучше всего предсказывает будущее поведение» [Yarkoni, Westfall, 2017, p. 1]. Однако, как повсеместно отмечают критики, на практике исследования с небольшой выборкой столкнулись с серьезными проблемами в аспекте предсказания поведения человека.

Прежде всего, 2010-е гг. были отмечены масштабным кризисом воспроизводимости (*replication crisis*) в психологической науке, поставившим под сомнения основания исследований с небольшой выборкой. Так, по результатам работы проекта «Сотрудничество в сфере открытой науки» (*Open Science Collaboration*) были сделаны весьма неутешительные выводы. В ходе попыток воспроизвести результаты 100 известных исследований из областей социальной и когнитивной психологии, две трети попыток «оказались неуспешными, а когда они были успешными, масштаб эффектов часто оказывался меньше, чем в оригинальных исследованиях.... Как утверждает в комментарии в журнале *Nature* (Baker, 2015): “Не верьте всему, что Вы читаете в психологической литературе. В действительности, две трети [этой литературы], вероятно, следует поставить под сомнение”» [Smith, Little, 2018].

Это обстоятельство позволило критикам заключить, что, хотя объяснение и предсказание в психологии могут быть связаны между собой как две стороны одной монеты, на практике исследователям придется делать ставку либо на объяснительную, либо на предсказательную способность своих теорий. Так, в уже цитировавшемся манифесте Яркони и Вестфолла, громко названном «Выбирая предсказание вместо объяснения в психологии: уроки машинного обучения», указывается, что «со статистической точки зрения, просто не верно, что модель, лучше всего аппроксимирующая порождающий данные процесс (*data generating process*), в общем случае будет успешнее всего предсказывать результаты в реальном мире» [Yarkoni, Westfall, 2017, p. 1]. По их мнению, за образец в этом отношении стоит взять добившуюся резонансных успехов область машинного обучения.

Так, в то самое время, как в психологии разворачивался кризис воспроизводимости исследований с небольшой выборкой, вооруженная массивными датасетами и новыми вычислительными мощностями область машинного обучения поставила, по сути, на конвейер развитие успехов коннекционистских сетей 1980–1990-х гг. С учетом успехов близкой по духу области машинного обучения и ввиду появления больших данных в психологии, Яркони и Вестфолл предлагают реорганизовать психологическую науку по методологическим лекалам области машинного обучения: «с учетом относительно недавней революции в теории и методологии *машинного обучения* – в которой предсказание ненаблюдаемых данных рассматривается как золотой стандарт успеха, а объяснение обычно представляется малоинтересным или неинтересным – а также последовательно увеличивающейся доступностью масштабных датасетов, описывающих

поведение человека <....> возможно продвинуть психологию ближе к предсказывающей науке...» [Yarkoni, Westfall, 2017, p. 2].

В чем же именно с этой точки зрения может заключаться проблема исследований с небольшой выборкой, которая привела к ситуации, когда результаты двух третей известных исследований просто не были воспроизведены повторно? Как уже говорилось выше, в теории исследования, предлагающие убедительные объяснения психологических феноменов, также должны хорошо предсказывать их. Однако на практике этого часто не происходит. Как правило, объяснительные исследования оцениваются по тому, насколько хорошо они описывают исходные данные. Этот показатель известен как «качество соответствия» (goodness of fit). Естественно, что модель, хорошо описывающая свои данные, будет иметь низкую степень ошибок (т.н. ошибок обучения) при проверке этими же исходными данными. Как утверждают Яркони и Вестфолл, проблема заключается в том, что хорошая степень соответствия исходным данным не только не гарантирует того, что модель будет хорошо предсказывать при проверке новыми данными (out-of-sample data), но и зачастую этому откровенно препятствует.

Почему же модель, хорошо описывающая исходные данные, часто имеет слабые предсказательные характеристики? Так, данная проблема особенно остро стоит для моделей с *большим* количеством предикторов (тех самых факторов, позволяющих предсказывать поведение), подгоняемых к *небольшим* датасетам. Существо проблемы многие авторы связывают с тем, что в процессе обучения на небольших данных модель ухватывает не только паттерны для предсказания, но и характерный для этого конкретного небольшого датасета шум (скажем, имеющиеся у данной небольшой группы работников, а не населения в целом, особенности, позволяющие им показывать высокую производительность труда). Поэтому при тестировании такой модели на новом датасете ее предсказательный потенциал оставляет желать лучшего. Этот феномен плохой обобщаемости модели на новые данные получил название переподгонки (или переобучения, overfitting).

Кроме статистической переподгонки, положение дел усложняется еще и склонностью некоторых исследователей слишком вольно обращаться с данными, чтобы получить желаемое значение нулевой гипотезы (*p*-значение, *p*-value, ниже конвенционального порога 0.05). Например, распространена практика произвольной остановки исследования (optional stopping) во время сбора данных – для этого периодически подсчитывают *p*-значение и, если оно укладывается в 0.05, дальнейшие подсчеты просто останавливают. Такого рода практики вольного обращения с данными для получения нужного *p*-значения обрели широкую известность как *p*-хакинг (подгонка данных, *p*-hacking). По мнению Яркони и Вестфолла, *p*-хакинг можно рассматривать как особого рода форму переподгонки: «ее можно понимать как форму *процедурной переподгонки* (курсив авторов – М.С.), которая происходит до или параллельно с оценкой модели – например, во время очистки данных, отбора модели или выбора анализа для отчета. По-

добно тому, как статистическая модель может учитывать специфические закономерности в учебных данных, которых на самом деле нет в генеральной совокупности, исследователи-люди часто замечают и действуют на основе специфических различий в результатах, полученных при использовании различных аналитических методов» [Yarkoni, Westfall, 2017, p. 5].

Итак, для обуздания этой излишней гибкости психологических теорий и превращения психологии в успешную предсказывающую науку Яркони и Вестфолл предлагают обратиться к ключевым методологическим правилам современной области машинного обучения, позволившим ей добиться тех впечатляющих успехов, которые, собственно, находятся сейчас у всех на слуху. В общем смысле важнейшим фактором успешной минимизации ошибок предсказаний должно стать нахождение оптимального баланса между тем, что в статистике и машинном обучении именуется смещением (bias) и дисперсией (variance). Например, при игре в дартс может случиться так, что все дротики попадут только в какую-то одну половину мишени, будучи там относительно свободно распределены (ситуация высокого смещения и высокой дисперсии) или, скажем, тесно сгруппированы (высокое смещение и низкая дисперсия). Либо же они могут свободно распределиться относительно центра мишени (низкое смещение и высокая дисперсия) или вообще все попасть близко к «яблочку» (низкое смещение и низкая дисперсия). Чаще всего на практике исследователи сталкиваются с двумя средними случаями: высоким смещением / низкой дисперсией и низкой дисперсией / высоким смещением. В последнем случае модель может быть гибкой и хорошо описывать имеющиеся данные, но плохо предсказывать при проверке новым датасетом (характерная особенность многих психологических моделей, полученных на небольшой выборке). Либо же модель может пожертвовать свойством аккуратного соответствия тренировочным данным в пользу большей простоты (высокого смещения), но и одновременно лучшей обобщаемости и предсказательной способности. Именно данному подходу, отмечают Яркони и Вестфолл, чаще всего и отдается предпочтение в области машинного обучения.

В общем и целом, авторы манифеста предлагают три непосредственные меры для улучшения предсказательной способности психологических моделей, широко применяющиеся в области машинного обучения. Таково, во-первых, использование больших данных (больших датасетов) для улучшения обобщаемости модели. Во-вторых, требуется применение техник аккуратной оценки ошибок предсказаний моделей для их диагностики и улучшения. В-третьих, это способность желаемым образом настроить соотношение между смещением и дисперсией (преимущественно в пользу смещения), чтобы избежать переподгонки и улучшить предсказывающие характеристики модели.

Итак, первой непосредственной мерой в борьбе с переподгонкой и излишне гибкими исследовательскими практиками должно стать использование больших данных. Главное достоинство массивных датасетов – их

большая репрезентативность: чем больше выборка, тем меньше сюрпризов будет при проверке предсказаний модели новыми данными. При наличии больших датасетов даже гибкие полиномиальные модели могут выдавать меньшую ошибку, чем более простые линейные модели, что крайне маловероятно, когда исследователи оперируют небольшими данными. Эта особенность больших датасетов, по замечанию Яркони и Вестфолла, стала основанием для известного афоризма, а именно что «больше данных лучше, чем лучшие алгоритмы»: «Очень часто, – пишут они, – самая лучшая вещь, которую исследователь может сделать для улучшения обобщающей способности модели – и едва ли вообще не самая простая и дешевая – это собрать больше данных» [Yarkoni, Westfall, 2017, p. 9].

Одним из интересных и неожиданных следствий постепенного внедрения больших данных в психологическую науку стал пересмотр результатов, полученных в исследованиях с небольшой выборкой. Так, с увеличением датасетов оценки эффектов, полученных в прежних исследованиях, снизились. Предполагая, что сами психологические феномены в целом не изменились, Яркони и Вестфолл утверждают, что наиболее правдоподобным объяснением снижения оценок эффектов в эпоху больших данных является, то обстоятельство, что прежние исследования, вероятно, страдали от переподгонки. В силу своей природы исследования с небольшой выборкой могут давать более вариативные оценки тех или иных феноменов, которые имеют тенденцию сглаживаться при появлении более широкой перспективы. По Яркони и Вестфоллу, для преодоления кризиса воспроизводимости психология должна отдавать предпочтение небольшим эффектам из исследований с большой выборкой, а не наоборот [Yarkoni, Westfall, 2017, p. 11].

Второй важнейшей мерой для улучшения предсказывающего потенциала психологических теорий должно стать применение техник количественной оценки ошибок предсказаний, получаемых на новых датасетах. В области машинного обучения для этих целей чаще всего используется техника перекрестной проверки (кросс-валидации, cross-validation). Существо метода (в упрощенном виде) заключается в разделении датасета на две части (блока, folds), одна из которых используется для подгонки модели, а другая – для тестирования (имитируя таким образом проверку модели новыми данными). Затем, чтобы продуктивно использовать весь датасет и избежать недоподгонки (underfitting, поскольку изначально тренировочные данные были уменьшены в два раза), проводят еще один цикл подгонки и тестирования, меняя ролями те самые части датасета (блоки). В итоге состоятельность модели оценивается путем усреднения значений, полученных в каждом цикле (или итерации) подгонки и проверки. (В общем виде данный подход именуется k -кратной перекрестной проверкой, k -fold cross-validation, так как датасет разбивается на k равных частей – чаще всего от 3 до 10, – где одна часть используется для обучения, а все остальные – для тестирования.)

Перекрестная проверка представляет собой важнейший инструмент борьбы с переподгонкой моделей. При аккуратном использовании с ее помощью можно получить значения, близкие к тем, которые обычно наблюдаются при проверке модели новыми данными. Также она может помочь и в борьбе с *p*-хакингом. Например, исследователь может попробовать несколько методов обработки данных и выбрать в итоге тот, который лучше всего себя показывает с точки зрения перекрестной проверки [Yarkoni, Westfall, 2017, p. 12].

Наконец, в дополнение к указанным мерам исследователям также нужны непосредственные средства манипуляции моделями, если известно, что их предсказывающий потенциал может быть улучшен. На эту роль авторы манифеста предлагают технику, известную как регуляризация «лассо» (lasso regularization, least absolute shrinkage and selection operator). Суть регуляризации «лассо» заключается в уменьшении до нуля (и, по сути, элиминации) наименее существенных коэффициентов модели, что, собственно, и ведет к ее упрощению. Это достигается за счет добавления к функции ошибок (разнице между предсказаниями модели и наблюдаемыми результатами) так называемого штрафа (penalty term) – числа, пропорционального сумме абсолютных значений коэффициентов модели. Скажем, с увеличением суммы коэффициентов модели возрастает и ее «объяснительная» способность (т.е. ее соответствие данным), однако вместе с этим возрастает и штраф. В духе обсуждавшегося выше подхода к дилемме смещения и дисперсии метод лассо намеренно отдает предпочтение простоте моделей (высокому смещению) за счет обнуления несущественных коэффициентов. Такой подход может быть наиболее полезен, по замечанию Яркони и Вестфолла, как раз для моделей с большим числом предикторов по отношению к размеру тренировочных данных (которые, как мы видели выше, распространены в психологии образования, психологии личности и др.)

Таким образом, Яркони, Вестфолл, Вейскопф и другие адепты этого направления призывают к массовой переориентации психологических исследований с поиска каузальных механизмов поведения на создание предиктивно аккуратных моделей. Непосредственным следствием данной методологической установки оказывается рассмотрение причинных механизмов поведения в качестве своего рода черного ящика. Психолога, принявшего на вооружение методологические принципы области машинного обучения, в первую очередь интересует предсказательная точность его моделей, а не то, насколько хорошо они ухватывают причинные механизмы поведения. В этом отношении Яркони и Вестфолл ссылаются на идею двух статистических культур Бреймана: «Подавляющее большинство статистиков относится к “культуре моделирования данных”, в рамках которой предполагается, что данные являются следствием некоторого процесса порождения данных (data generating process), и главная цель заключается в оценке истинных параметров этого процесса. В противоположность, не-

значительное меньшинство статистиков (и большинство исследователей в области машинного обучения) принадлежит к “культуре алгоритмического моделирования”, в которой предполагается, что *данные представляют собой результат некоторого неизвестного и, возможно, непознаваемого процесса* [курсив мой – М.С.], а основная цель заключается в нахождении алгоритма, который при тех же входных данных дает те же результаты, что и этот процесс» [Yarkoni, Westfall, 2017, p. 19].

Помимо, собственно, теоретических манифестов, с этим направлением ассоциируется ряд конкретных экспериментальных исследований, выдержанных в духе обсуждавшейся выше методологии. Одним таким исследованием является экспериментальная работа Риссмана, Грили и Вагнера [Rissman, Greely, Wagner, 2010]. В этом исследовании была предпринята попытка оценить, является ли зафиксированная с помощью фМРТ активность мозга надежным маркером (предиктором) воспоминаний. Так, двадцати участникам предъявлялся ряд изображений лиц, а час спустя их способность распознать предъявлявшиеся изображения была оценена во время сканирования мозга с помощью фМРТ. Зафиксированные паттерны активности мозга субъектов с помощью алгоритма мультивоксельного анализа паттернов (МВАП) использовались для предсказания (1) субъективной уверенности участников (их убежденности, в том, видели ли они то или иное изображение ранее), а также (2) того, насколько они реально помнят то или иное изображение. Оказалось, что предиктивный алгоритм намного увереннее справляется с первой задачей, корректно предсказывая субъективную уверенность участников с вероятностью примерно в 90 процентов, тогда как действительные воспоминания успешно предугадывались с вероятностью чуть более шанса (около 56 процентов). Эти результаты дали основание авторам исследования заявить, что метод фМРТ имеет существенные ограничения для определения действительных воспоминаний субъектов (что крайне важно, например, в контексте рассмотрения использования фМРТ в судебных разбирательствах, так как с его помощью, вероятно, нельзя достоверно определить, видел ли свидетель подозреваемого или нет).

В другом известном исследовании паттерны нейронной активности отдельных индивидов использовались для предсказания поведения целых групп населения [Falk, Berkman, Lieberman, 2012]. Так, группе курильщиков, желавших покончить с этой вредной привычкой, были показаны видеоролики из трех рекламных кампаний против курения (по несколько роликов из кампаний, условно названных А, В и С). В это время активность их мозга (медиальной префронтальной коры) фиксировалась с помощью фМРТ. После этого им давали возможность субъективно оценить убедительность этих кампаний. В результате участники наиболее убедительной назвали кампанию В, затем кампанию А, а затем кампанию С. Между тем анализ данных нейронной активности показывал иной рейтинг убедительности: $C > B > A$. В итоге именно нейронная активность субъектов оказа-

лась более состоятельным предиктором эффективности рекламных кампаний в масштабе всего населения: по результатам звонков на горячую линию был получен аналогичный рейтинг эффективности – $C > B > A$.

Несмотря на использование отдельных предиктивных техник и алгоритмов, указанные исследования все-таки основывались на стандартных небольших выборках. Так, недавно созданная генеративная модель Delphi-2M является примером полноценной предсказывающей модели, созданной на основе больших данных [Shmatko et al., 2025]. Эта модель позволяет предсказывать индивидуальные риски развития более тысячи заболеваний на срок до двадцати лет вперед. Delphi-2M представляет собой нейросеть-трансформер на основе модели GPT-2, натренированную на анонимизированных данных 402799 пациентов из британской биомедицинской базы данных UK Biobank. Подобно тому, как большие языковые модели предсказывают последовательности символов (токенов) с учетом поступившего запроса, так и Delphi-2M предсказывает развитие того или иного заболевания на основе имеющейся клинической картины, где отдельными токенами выступают события из жизни пациента (пол, состояние здоровья в определенном возрасте, вредные привычки в определенном возрасте и т.п. – всего 1258 токенов; например, если на входе имеется «пациент 41 год, курит и употребляет алкоголь», то на выходе может быть получен ответ «44 года, варикозное расширение вен»). Модель получила независимую проверку на данных почти двух миллионов пациентов из датского национального реестра пациентов. В итоге Delphi-2M достигла предсказательной точности, сопоставимой с показателями предиктивных моделей для одного заболевания, воплощая в себе обсуждавшиеся методологические принципы и в то же время делая шаг на пути к прозрачности «черного ящика» внутренних вычислений [Learning the natural history, 2025].

Предиктивизм черного ящика и поиск причинных механизмов в науке

Итак, мы рассмотрели теоретическую мотивацию, методологические основания и конкретные примеры недавних предсказывающих моделей в психологии и нейронауке. Как можно оценить их статус в общем контексте развития и структурной организации психологии и нейронауки? Действительно ли появление предиктивной методологии является «поворотным моментом» в истории этих дисциплин? Действительно ли предиктивная методология обладает автономией по отношению к стандартным объяснительным практикам и должна выйти из их тени [Weiskopf, 2022]? Насколько оправданы призывы выбрать предсказание вместо объяснения в качестве ведущего методологического ориентира в психологии и нейронауке?

Прежде всего попробуем разобраться с уместностью альтернативы «предсказание или объяснение» для психологии и нейронауки. Здесь следует начать с того, что критикуемые адептами методологического предиктивизма распространенные исследования с небольшой выборкой в социальной психологии или психологии личности не исчерпывают собой все стили психологических исследований. Более того, было бы преувеличением говорить, что именно теории в этих областях служат парадигмой научных работ, направленных на поиск причинных механизмов поведения, психических функций и познания. Исследования в нейропсихологии, психофизиологии, когнитивной психологии и когнитивной нейронауке куда в большей степени связаны с поиском реальных причинных механизмов познания и поведения.

Говоря о причинных механизмах познания и поведения, я буду во многом опираться на концепцию механистических объяснений в нейронауке философа науки К. Крейвера [Craver, 2007]. Согласно Крейверу, объяснение представляет собой прежде всего описание причинных механизмов. Механизмы состоят из сущностей (entities), обладающих теми или иными свойствами, которые позволяют им принимать участие в определенных процессах, или активностях (activities). Например, механизм везикулярного высвобождения нейромедиаторов включает в себя множество компонентов (т.е. сущностей: ионы кальция, везикулы с нейромедиаторами, поры слияния, нейронные мембраны и др.) и их разнообразных активностей (открытие, фиксация, диффузия, слияние, прайминг и т.д.) [Craver, 2007, p. 5]. Коротко говоря, по Крейверу, механизм есть «ряд сущностей и активностей, организованных таким образом, что они демонстрируют (exhibit) подлежащий объяснению феномен» [Там же]. Используя терминологию философии науки, в концепции Крейвера описание механизма представляет собой эксплананс (объяснение), а порождаемый механизмом феномен – экспланандум (нечто, подлежащее объяснению).

При этом объяснение не сводится только к описанию активностей наиболее элементарных компонентов на клеточно-молекулярном уровне: психологи и нейрочлены в своей работе (а Крейвер в первую очередь настаивает на дескриптивной аккуратности своей концепции) принимают во внимание реальную каузальную значимость более высоких уровней, включая активность непосредственно на уровне психологических событий и процессов. В качестве примера многоуровневых объяснений Крейвер ссылается на исследования пространственной навигации у грызунов в водном лабиринте Морриса [Craver, 2007, pp. 165–170]. В этих исследованиях чаще всего различают четыре уровня. (1) На верхнем уровне пространственной памяти осуществляются задачи навигации и выучивания пути к скрытой под водой платформе. (2) На более низком уровне формирования пространственных карт ищут и исследуют непосредственный нейронный субстрат навигационной активности (связывая его с деятельностью гиппокампа и других височных и лобных областей коры мозга).

(3) «Этажом» ниже, на клеточно-электрофизиологическом уровне лежат механизмы долговременной потенциации (long-term potentiation), выступающие основой для навигационной активности гиппокампа. (4) Наконец, на нижнем молекулярном уровне находятся «механизмы, делающие возможными химические и электрические активности нервных клеток» [Craver, 2007, p. 169]. Именно все в совокупности эти уровни обеспечивают объяснение феномена пространственной навигации у грызунов.

В конечном счете нейронаука, по Крейверу, имеет в качестве фундаментальных целей как механистическое объяснение тех или иных психофизиологических феноменов (восприятие цвета, пространственная навигация, развитие нервной системы и т.д.), так и контроль над мозгом и нервной системой для диагностирования и лечения болезней, улучшения мозговых функций и др. [Craver, 2007, p. 1].

По-видимому, именно в большей степени точности и в больших возможностях контроля реальных каузальных факторов познания и поведения кроется важнейшее различие между теориями и моделями механистических объяснений, с одной стороны, и вдохновленных машинным обучением современными предиктивными моделями, с другой стороны. Последние при всех их достоинствах в борьбе с перепогонкой и *p*-хакингом могут ухватывать поверхностные корреляции между событиями, в действительности не имеющими причинных взаимосвязей. (Подобно тому, как, скажем, активное использование кондиционеров коррелирует, но не является причиной всплеска продаж мороженого.)

Например, авторы рассмотренной выше модели Delphi-2M открыто заявляют: «Несмотря на то, что Delphi-2M представляется способной моделировать направленные во времени зависимости, мы предостерегаем от интерпретации этих зависимостей в качестве причинно-следственных связей, которые можно было бы использовать для изменения будущего состояния здоровья» [Learning the natural history, 2025]. По их мнению, результаты работы с Delphi-2M говорят в пользу осторожного подхода «при использовании моделей ИИ для осуществления вывода и прогнозирования в гетерогенных наборах данных здравоохранения, потенциально указывая на то, что они могут быть полезным дополнением к существующим диагностическим конвейерам, а не заменой им» [Там же].

Представляется, что только что приведенные суждения содержат в себе, в целом, взвешенную оценку характера и особенностей практического использования результатов, полученных с помощью предиктивно-статистической методологии. Очевидно, что предельная осторожность требуется при отсылке к результатам таких исследований в областях образования и медицины, при рассмотрении правовых вопросов. Во всех этих контекстах гораздо более надежной опорой для принятия решений является знание конкретных каузальных механизмов поведения. Поэтому, вероятно, вместо альтернативы «предсказание или объяснение» более продуктивным выбором было бы нахождение оптимального баланса между

этими исследовательскими стилями с пониманием того, что исследования каузальных механизмов не могут быть вытеснены методологией, опирающейся на идею черного ящика.

Впрочем, это признают и сами авторы манифеста предсказательных моделей нового типа. Так, они пишут, что они вовсе не призывают к «полному отказу от попыток механистически объяснить поведение человека. <...> Мы, однако, охотно признаем, что предиктивно ориентированные (prediction-focused) подходы не подходят для всех видов исследовательских вопросов в психологии. В частности, хорошо организованные, масштабные, рандомизированные, контролируемые эксперименты есть и должны оставаться золотым стандартом для получения каузальных заключений о работе человеческого разума. <...> Таким образом, мы отстаиваем точку зрения, что даже в тех случаях, когда каузальное объяснение есть (надлежащим образом) главная цель, понятия и методология машинного обучения все же могут оказать неоценимую помощь при их инструментальном использовании – посредством минимизации *p*-хакинга, увеличения исследовательской эффективности, способствования оценке работы моделей и увеличения интерпретируемости» [Yarkoni, Westfall, 2017, p. 19].

Заключение

Итак, в данной статье был рассмотрен конкретный проект методологического влияния современных исследований искусственного интеллекта (машинного обучения) на такие научные дисциплины, как психология и нейронаука. Мы рассмотрели основные аргументы сторонников использования предиктивной методологии и предиктивных моделей, сводящиеся к тому, что преодоление основных проблем исследований с небольшой выборкой (переподгонка, *p*-хакинг и, как следствие, плохая воспроизводимость) может быть достигнуто посредством (1) использования больших датасетов, (2) аккуратной оценки предсказательных ошибок модели (перекрестная проверка, или кросс-валидация), а также (3) упрощения модели за счет обнуления несущественных коэффициентов (регуляризация лассо). Были также рассмотрены конкретные примеры таких моделей из областей психологии, нейронауки и биомедицины.

Оценивая этот проект, я обосновывал следующие положения. (1) Исследования с небольшой выборкой, критикуемые в связи с кризисом невоспроизводимости, не исчерпывают собой все методологические стили в психологии и нейронауке. Неверно рассматривать их в качестве образца объяснительных теорий и моделей в этих дисциплинах. Такие дисциплины, как когнитивная психология, когнитивная нейронаука, нейропсихология и психофизиология, в большей степени имеют отношение к установлению и использованию каузальных механизмов познания и поведения. (2) Вероятно, в большей степени точности и в больших возможностях ма-

нипуляции каузальными факторами познания и поведения кроется важнейшее различие между теориями механистических объяснений и вдохновленными машинным обучением предиктивными моделями. Последние в отрыве от поиска реальных каузальных механизмов могут ухватывать только поверхностные корреляции между событиями. (3) При принятии важных решений в контексте образовательной, биомедицинской и правовой практик лучше опираться на знание каузальных механизмов познания и поведения. Предельная осторожность требуется при отсылке к результатам предиктивных исследований во всех этих областях. (4) Соответственно, предиктивные модели не могут вытеснить теории, ищущие механистические объяснения познания и поведения. Вместо выбора между предсказательной и объяснительной методологиями более продуктивным было бы нахождение оптимального баланса между ними.

Список литературы

- Craver C. F. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. – Oxford : Clarendon Press ; New York : Oxford University Press. – 2007. – 308 p.
- Iglesias D., Sorrel M. A., Olmos R. Cross-validation and predictive metrics in psychological research: Do not leave out the leave-one-out // *Behavior Research Methods*. – 2025. – Vol. 57. – No. 3. – P. 85.
- Falk E. B., Berkman E. T., Lieberman M. D. From neural responses to population behavior: neural focus group predicts population-level media effects // *Psychological science*. – 2012. – Vol. 23. – No. 5. – P. 439–445.
- Festinger L., Carlsmith J. M. Cognitive consequences of forced compliance // *The journal of abnormal and social psychology*. – 1959. – Vol. 58. – No. 2. – P. 203–210.
- Learning the natural history of human disease with generative transformers / Shmatko A. et al. // *Nature*. – 2025. – Vol. 647. – No. 8088. – P. 248–256.
- Rissman J., Greely H. T., Wagner A. D. Detecting individual memories through the neural decoding of memory states and past experience // *Proceedings of the National Academy of Sciences*. – 2010. – Vol. 107. – No. 21. – P. 9849–9854.
- Smith P. L., Little D. R. Small is beautiful: In defense of the small-N design // *Psychonomic bulletin & review*. – 2018. – Vol. 25. – No. 6. – P. 2083–2101.
- Weiskopf D. A. The Predictive Turn in Neuroscience // *Philosophy of Science*. – 2022. – Vol. 89. – No. 5. – P. 1213–1222.
- Yarkoni T., Westfall J. Choosing prediction over explanation in psychology: Lessons from machine learning // *Perspectives on Psychological Science*. – 2017. – Vol. 12. – No. 6. – P. 1100–1122.

References

- Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford : Clarendon Press ; New York : Oxford University Press. 308 p.
- Iglesias, D., Sorrel, M. A., & Olmos, R. (2025). Cross-validation and predictive metrics in psychological research: Do not leave out the leave-one-out. *Behavior Research Methods*, 57(3), 85.

- Falk, E. B., Berkman, E. T., & Lieberman, M. D. (2012). From neural responses to population behavior: neural focus group predicts population-level media effects. *Psychological science*, 23(5), 439–445.
- Festinger, L., Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The journal of abnormal and social psychology*, 58(2), 203–210.
- Rissman, J., Greely, H. T., & Wagner, A. D. (2010). Detecting individual memories through the neural decoding of memory states and past experience. *Proceedings of the National Academy of Sciences*, 107(21), 9849–9854.
- Shmatko, A., Jung, A. W., Gaurav, K., Brunak, S., Mortensen, L. H., Birney, E., Fitzgerald, T. & Gerstung, M. (2025). Learning the natural history of human disease with generative transformers. *Nature*, 647(8088), 248–256.
- Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic bulletin & review*, 25(6), 2083–2101.
- Weiskopf, D. A. (2022). The Predictive Turn in Neuroscience. *Philosophy of Science*, 89(5), 1213–1222.
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122.