Модель динамики смыслового поля как концептуальная основа для Explainable AI

Аннотация
В статье предлагается новый концептуальный подход к проблеме объяснимости искусственного интеллекта (XAI), основанный на модели динамики смыслового поля. Решение нейросети рассматривается как аналог интуитивного акта, а его объяснение post factum — как структурная экспликация логического следа. Модель опирается на гносеологическую концепцию сверхлогики и использует аппарат теории динамических систем (странный аттрактор). Показано, что предложенный подход позволяет по-новому описать задачу XAI: не как «вскрытие чёрного ящика», а как переход от интуитивного схватывания паттерна к рациональному обоснованию решения. Обсуждаются ограничения модели и перспективы её применения.

Ключевые слова: XAI, объяснимый искусственный интеллект, смысловое поле, странный аттрактор, сверхлогика, логический след, структурная экспликация, чёрный ящик.

1. Введение: проблема «чёрного ящика»

Современные системы искусственного интеллекта, основанные на глубоких нейронных сетях, достигают впечатляющих результатов в распознавании образов, обработке естественного языка, медицинской диагностике и многих других областях. Однако их внутренняя работа остаётся непрозрачной. Мы знаем входные данные, знаем выходной результат, но не понимаем, как именно система пришла к данному решению. Это — проблема «чёрного ящика».

Направление Explainable AI (XAI) стремится сделать процесс принятия решений искусственным интеллектом прозрачным и доступным для человеческого понимания. Существующие подходы можно разделить на две группы: ante-hoc (создание изначально интерпретируемых моделей) и post-hoc (объяснение уже обученной модели). Второе направление особенно близко к тому, чем занимается настоящая работа.

С гносеологической точки зрения задача post-hoc объяснения нейросетевого решения изоморфна задаче экспликации интуитивного акта. Нейросеть «схватывает» паттерн мгновенно и неосознанно, не имея доступа к механизму своего решения — подобно тому, как человек переживает интуитивное озарение. Задача XAI — восстановить логический след post factum.

2. Модель динамики смыслового поля: ключевые понятия

Модель динамики смыслового поля опирается на гносеологическую концепцию сверхлогики, определяемой как интуиция, у которой обнаруживается логический след [1]. Логический след — это возможность рационального обоснования интуитивного акта, вытекающая из его структурных оснований [2].

В рамках модели смысловое поле рассматривается как открытая нелинейная динамическая система со странным аттрактором — сложной, фрактальной областью смыслового схождения [3]. Основные свойства модели:

· Нелокальность: связь схватывается целостно, без пошагового вывода.

· Резонансность: связь отзывается на опыт субъекта.
· Множественность: одна конфигурация элементов порождает различные, но равно валидные траектории смыслов.

Модель носит инструментальный характер и не утверждает онтологического тождества между смыслом и физической системой. Это способ описания, а не заявление о природе реальности.

3. Нейросеть как «интуитивный субъект»

С точки зрения предложенной модели, обученная нейросеть может рассматриваться как интуитивный субъект. Она принимает решение быстро, без осознанного логического вывода, на основе «опыта» (обучения на данных). Результат этого решения — аналог «Да-эффекта»: система «уверена» в ответе, но не может объяснить, почему.

Сходство между нейросетевым решением и интуитивным актом человека не поверхностно, а структурно. Оба процесса характеризуются быстротой и неосознаваемостью. Оба опираются на предшествующий опыт — у человека это накопленные знания и практика, у нейросети — обучение на данных. В обоих случаях результат дан субъекту непосредственно, тогда как механизм его получения остаётся скрытым. И в обоих случаях результат может быть эксплицирован post factum — у человека через сверхлогику, у нейросети через post-hoc объяснение.

Существующие методы post-hoc объяснения включают LIME (локальную аппроксимацию модели интерпретируемым суррогатом), SHAP (оценку вклада признаков на основе теории игр), методы на основе внимания (attention-based explanations) и контрфактические объяснения [4].При всём разнообразии эти методы решают одну задачу: восстановить основания, на которых нейросеть приняла решение. Именно эта задача — экспликация логического следа — является центральной для модели динамики смыслового поля.

4. Структурные основания нейросетевого решения

В модели динамики смыслового поля выделяются четыре типа структурных оснований, на которых держится интуитивно схваченная связь: фонетический, ассоциативный, культурный и индивидуально-психологический [3]. Следует подчеркнуть, что на сегодняшний день типология структурных оснований разработана автором в рамках частного случая — применительно к восприятию поэтического текста. Её перенос на нейросетевые решения носит предварительный, поисковый характер и требует дальнейшей теоретической и экспериментальной проработки. Предлагаемые ниже аналогии призваны показать саму возможность переноса логики модели на новый материал, а не дать окончательную классификацию.

Выделенные автором четыре типа оснований описывают различные аспекты восприятия поэтического текста. Применительно к нейросетевым решениям можно предложить их рабочие аналоги. Так, фонетическому типу, связанному со звуковой организацией, может соответствовать признаковый уровень — активация определённых фильтров, чувствительность к текстурам и частотам. Ассоциативному типу, связанному с коннотативными полями, может соответствовать эмбеддинговый уровень — близость векторов в скрытом пространстве представлений. Культурному типу может

соответствовать контекстуальный уровень — зависимость решения от более широкого контекста входных данных. Индивидуально-психологическому типу — стохастический уровень, связанный с влиянием случайных факторов обучения.

5. Экспликация логического следа как задача XAI

В рамках предложенной модели задача XAI может быть переформулирована следующим образом: не «вскрыть чёрный ящик», а восстановить логический след нейросетевого решения — то есть эксплицировать его структурные основания post factum.

Это означает не попытку «заглянуть внутрь» каждого нейрона, а построение рациональной реконструкции: какие именно признаки, паттерны, аналогии послужили основанием для данного решения.

Существующие методы XAI — LIME, SHAP, saliency maps, attention visualization — можно интерпретировать как различные стратегии структурной экспликации. Модель динамики смыслового поля даёт для них общую концептуальную рамку.

6. Ограничения

Модель не утверждает, что всякое нейросетевое решение может быть эксплицировано. В некоторых случаях логический след не обнаруживается — и тогда мы имеем дело с аналогом «чистой интуиции». Это ограничение не является дефектом модели; оно отражает реальное положение дел в XAI, где полная объяснимость остаётся недостижимой целью.

Кроме того, предложенная аналогия между человеческой интуицией и нейросетевым решением не является строгой. Нейросеть не переживает «Да-эффект» в человеческом смысле. Модель предлагает функциональную, а не феноменологическую аналогию.

Кроме того, следует разграничить область применимости модели. Предложенная аналогия между нейросетевым решением и интуитивным актом релевантна прежде всего для сложных, нестандартных решений — тех, которые требуют целостного схватывания паттерна и не сводятся к рутинной классификации. Значительная часть решений современных нейросетей относится именно к рутинным (например, классификация изображений на заранее заданные категории). Для таких решений модель избыточна: они не требуют «интуиции» и могут быть объяснены более простыми средствами. Модель динамики смыслового поля ориентирована на те случаи, где решение возникает как целостное схватывание сложной конфигурации признаков — подобно тому, как эксперт принимает решение в неопределённой ситуации.

7. Заключение и перспективы

Модель динамики смыслового поля предлагает новый язык для описания проблемы объяснимости искусственного интеллекта. Она позволяет переформулировать задачу XAI в гносеологических терминах: от «вскрытия чёрного ящика» к «структурной экспликации логического следа».

Дальнейшие исследования могут быть направлены на формализацию понятия «структурных оснований» применительно к нейросетевым архитектурам, а также на разработку методов экспликации, вдохновлённых моделью четырёх типов.

Литература

1. Смокотина О.Ф. Сверхлогика как гносеологическая категория. — Препринт. — PREPRINTS.RU, 2026.
2. Смокотина О.Ф. Логический след: определение и характеристики. — Препринт. — PREPRINTS.RU, 2026.
3. Смокотина О.Ф. О введении гипотезы динамического поведения системы смыслов. — Препринт. — PREPRINTS.RU, 2026.
4. Samek W., Montavon G., Vedaldi A., Hansen L.K., Müller K.-R. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. — Springer, 2019. — 439 p.

References

1. Smokotina O.F. Sverkhlogika kak gnoseologicheskaya kategoriya [Superlogic as a Gnoseological Category]. Preprint. PREPRINTS.RU, 2026. (In Russian)
2. Smokotina O.F. Logicheskii sled: opredelenie i kharakteristiki [Logical Trace: Definition and Characteristics]. Preprint. PREPRINTS.RU, 2026. (In Russian)
3. Smokotina O.F. O vvedenii gipotezy dinamicheskogo povedeniya sistemy smyslov [On the Introduction of the Hypothesis of Dynamic Behavior of the System of Meanings]. Preprint. PREPRINTS.RU, 2026. (In Russian)
4. Samek W., Montavon G., Vedaldi A., Hansen L.K., Müller K.-R. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019. 439 p.

Сведения об авторе:
Смокотина О.Ф. — независимый исследователь. Сфера научных интересов: гносеология, эпистемология, искусственный интеллект, XAI. Автор концепции сверхлогики.