Два поколения объяснительного искусственного интеллекта XAI 2.0

Аверкин Алексей Николаевич
ФИЦ «Информатика и Управления» РАН

averkin2003@inbox.ru

Проблемы с текущим XAI (XAI 1.0)

Ограничения XAI 1.0:

Апостериорные объяснения могут не в полной мере отражать сложность модели.

Отсутствие стандартизации метрик оценки для объяснимости.

Трудности в балансе точности и интерпретируемости.

Ограниченное внимание к дизайну, ориентированному на пользователя, что приводит к объяснениям, которые не являются действенными или интуитивно понятными для конечных пользователей

Потребность в развитии:

По мере того, как искусственный интеллект становится все более интегрированным в повседневную жизнь, растет спрос на более надежные, надежные и удобные для пользователя решения для объяснимости

XAI 2.0 – новое поколение XAI 1.0

XAI 2.0 представляет собой следующее поколение объяснимого искусственного интеллекта, ориентированное на интеграцию прозрачности и интерпретируемости непосредственно в архитектуру моделей, а не только на post hoc анализ , на удобство использования и надежности. XAI 2.0 включает слтдующее:

1. Упреждающая объяснимость: Встраивание объяснимости непосредственно в архитектуру модели во время обучения.
2. Дизайн, ориентированный на пользователя: Адаптация пояснений к конкретным потребностям и контекстам пользователя.
3. Динамическая адаптация: Предоставление контекстно-зависимых объяснений в режиме реального времени при изменении входных данных.
4. Совместимость: Обеспечение совместимости между различными фреймворками и платформами ИИ.

Преимущества XAI 2.0

Для бизнеса:

- Улучшенное соответствие нормативным требованиям
- Повышение удовлетворенности и лояльности клиентов за счет прозрачного взаимодействия.

Для разработчиков:

- Упрощение отладки и оптимизации сложных моделей.
- Стандартизированные инструменты и фреймворки для создания объяснимых систем.

Для конечных пользователей:

- Большая ясность и контроль над решениями, принятыми на основе искусственного интеллекта, влияющими на их жизнь.

Основы XAI 2.0

1. Прозрачность:

Четкое понимание того, как работает система искусственного интеллекта, включая ее источники данных, алгоритмы и логику принятия решений.

2. Интерпретируемость:

Создание удобочитаемых объяснений, которые соответствуют знаниям предметной области и ожиданиям пользователей.

3. Надежность:

Укрепление доверия к системам ИИ с помощью строгих процессов тестирования, проверки и сертификации

4. Возможность действовать

Предоставление пользователям возможности предпринимать осмысленные действия на основе предоставленных объяснений.

5. Масштабируемость:

Поддержка масштабных развертываний без ущерба для производительности и понятности.

Manifesto of Explainable Artificial Intelligence (XAI) 2.0

Поскольку системы, основанные на непрозрачном искусственном интеллекте (ИИ), продолжают успешно применяться в различных реальных приложениях, понимание моделей черного ящика приобретает первостепенное значение. Объяснимый ИИ (XAI) превратился в область исследований, приносящую практические и этические преимущества в различных областях.

Эксперты из разных областей для выявляют нерешенные проблемы, стремясь синхронизировать исследовательские программы и ускорить внедрение XAI в практику. Поощряя совместные обсуждения и междисциплинарное сотрудничество, они стремятся продвигать XAI вперед, способствуя его дальнейшему успеху. С этой целью выдвинуто комплексное предложение по продвижению XAI. Для достижения этой цели представлен манифест XAI 2.0 из 27 открытых проблем, разделенных на девять категорий.

Manifesto of Explainable Artificial Intelligence (XAI) 2.0

Longo, Luca; et al. (2024). "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions". Information Fusion. 106. doi:10.1016/j.inffus.2024.102301.

Этот манифест должен послужить дорожной картой для будущих исследований, направленных на синхронизацию усилий между различными дисциплинами и ускорение разработки приложений XAI

3.1.Создание объяснений для новых типов искусственного интеллекта

Создание объяснений для генеративных моделей и больших языковых моделей

Создание объяснений для концептуальной осведомленности

3.2.Улучшение существующих методов XAI

Усиление и улучшение методов атрибуции

Устранение артефактов в синтезированных объяснениях

Создание устойчивых объяснений

3.3.Оценка методов XAI и пояснения

Упрощение оценки объяснений человеком

Создание системы оценки методов XAI

Преодоление ограничений исследований с участием человека

Manifesto of Explainable Artificial Intelligence (XAI) 2.0

3.4.Содействие оценке объяснений человеком

Разъяснение основных понятий

Разъяснение взаимосвязи между XAI и доверием

Поиск полезной информации для понимания.

3.5.Поддержка многоразмерности объяснения

Создание многогранных объяснений

Обеспечение междисциплинарной работы в XAI

3.6.Поддержка ориентированности объяснений на человека

Создание объяснений, понятных людям

Облегчение объяснимости через концептуальные объяснения

Решение проблемы объяснений, оторванных от реальности

Выявление причинно-следственной связи для получения полезных практических объяснений

Manifesto of Explainable Artificial Intelligence (XAI) 2.0

3.7.Адаптация методов и объяснений XAI

Адаптация объяснений к различным заинтересованным сторонам

Адаптация объяснений к различным областям применения

Адаптация объяснений к различным целям

3.8. Смягчение негативного воздействия XAI

Смягчение неэффективной поддержки с помощью XAI

Разработка критериев фальсификации объяснений

Защита объяснений от злоупотреблений со стороны злонамеренных человеческих агентов

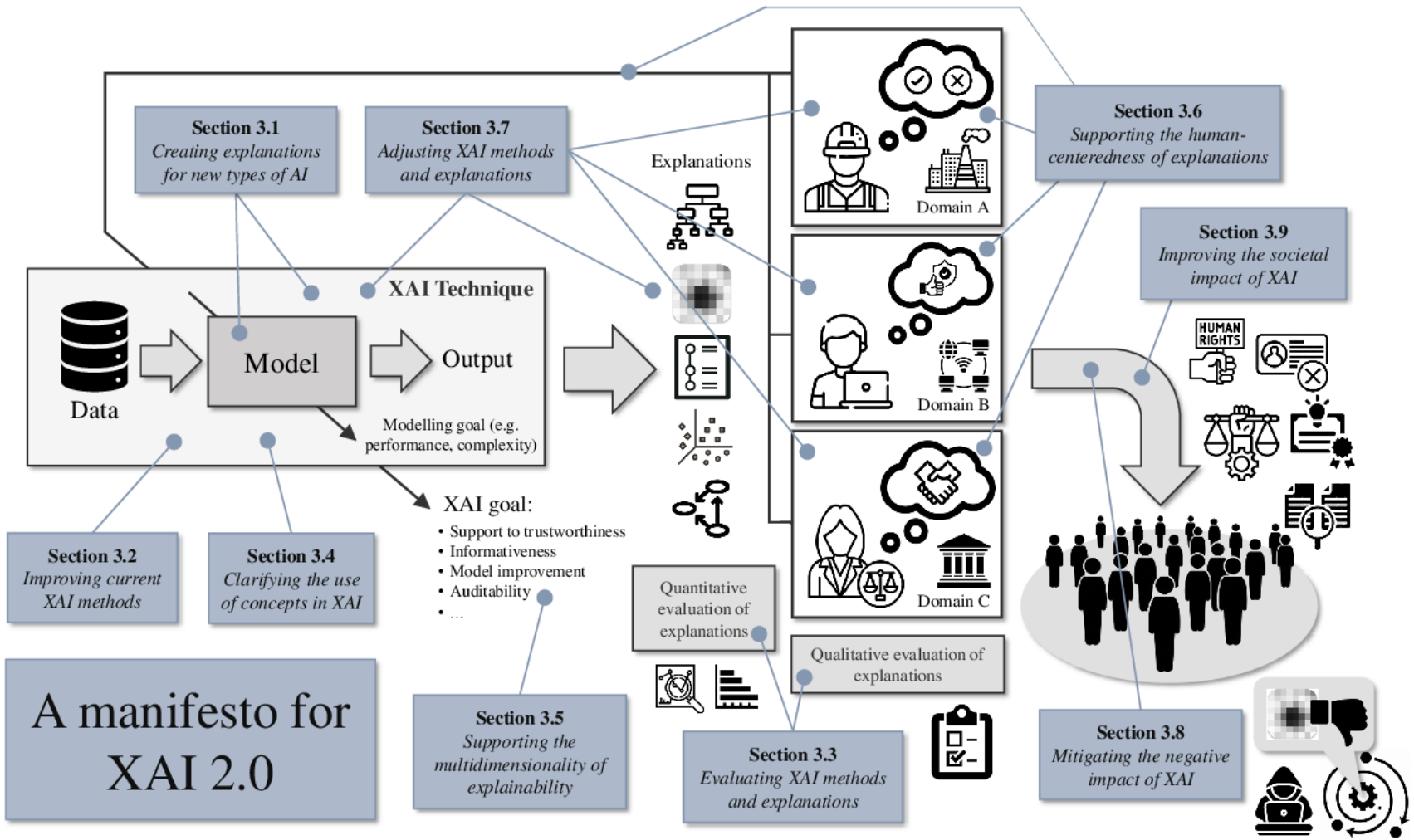
Защита объяснений от злоупотреблений со стороны злонамеренных сверхинтеллектуальных агентов

3.9 Улучшение общественного воздействия XAI:

Определение авторства искусственных данных и выявление плагиата

Поддержка права на забвение

Решение проблемы дисбаланса власти между индивидами и компаниями


A manifesto for XAI 2.0
Section 3.1
Creating explanations for new types of AI
Section 3.7
Adjusting XAI methods and explanations
XAI Technique
Data
Model
Output
Modelling goal (e.g. performance, complexity)
Explanations
XAI goal:
• Support to trustworthiness
• Informativeness
• Model improvement
• Auditability
• …
Domain A
Domain B
Domain C
Section 3.6
Supporting the human-centeredness of explanations
Section 3.9
Improving the societal impact of XAI
HUMAN RIGHTS
Section 3.2
Improving current XAI methods
Section 3.4
Clarifying the use of concepts in XAI
Section 3.5
Supporting the multidimensionality of explainability
Quantitative evaluation of explanations
Qualitative evaluation of explanations
Section 3.3
Evaluating XAI methods and explanations
Section 3.8
Mitigating the negative impact of XAI

Gadekallu, Thippa et al.(2024). XAI for Industry 5.0 -Concepts, Opportunities, Challenges and Future Directions. IEEE Open Journal of the Communications Society. PP. 1-1. 10.1109/OJCOMS.2024.3473891.

Достижения в области искусственного интеллекта вызывают преобразования, которые заставляют все больше компаний вступать в Индустрию 4.0 и 5.0. Во многих случаях эти преобразования происходят постепенно и по принципу «снизу вверх».

Это означает, что на первом этапе промышленное оборудование модернизируется для сбора как можно большего количества данных без фактического планирования использования информации. Кроме того, инфраструктура хранения и обработки данных подготавливается для хранения больших объемов исторических данных, доступных для дальнейшего анализа.

Только на последнем этапе разрабатываются методы обработки данных для улучшения или получения более глубокого понимания промышленных и бизнес-процессов. В результате такой схемы многие компании сталкиваются с проблемой огромного количества данных, неполным пониманием того, как существующие знания представлены в данных, при каких условиях эти знания перестают быть актуальными или какие новые явления скрыты в этих данных.

Этот пробел должен быть устранен следующим поколением методов XAI, которые должны быть ориентированными на экспертов и сфокусированными на задачах генерации знаний, а не на отладке моделей. Данная статьяоснована на результатах проекта EC CHIST-ERA по объяснимому предиктивному менеджменту (XPM).

.

Промышленные данные являются черным ящиком

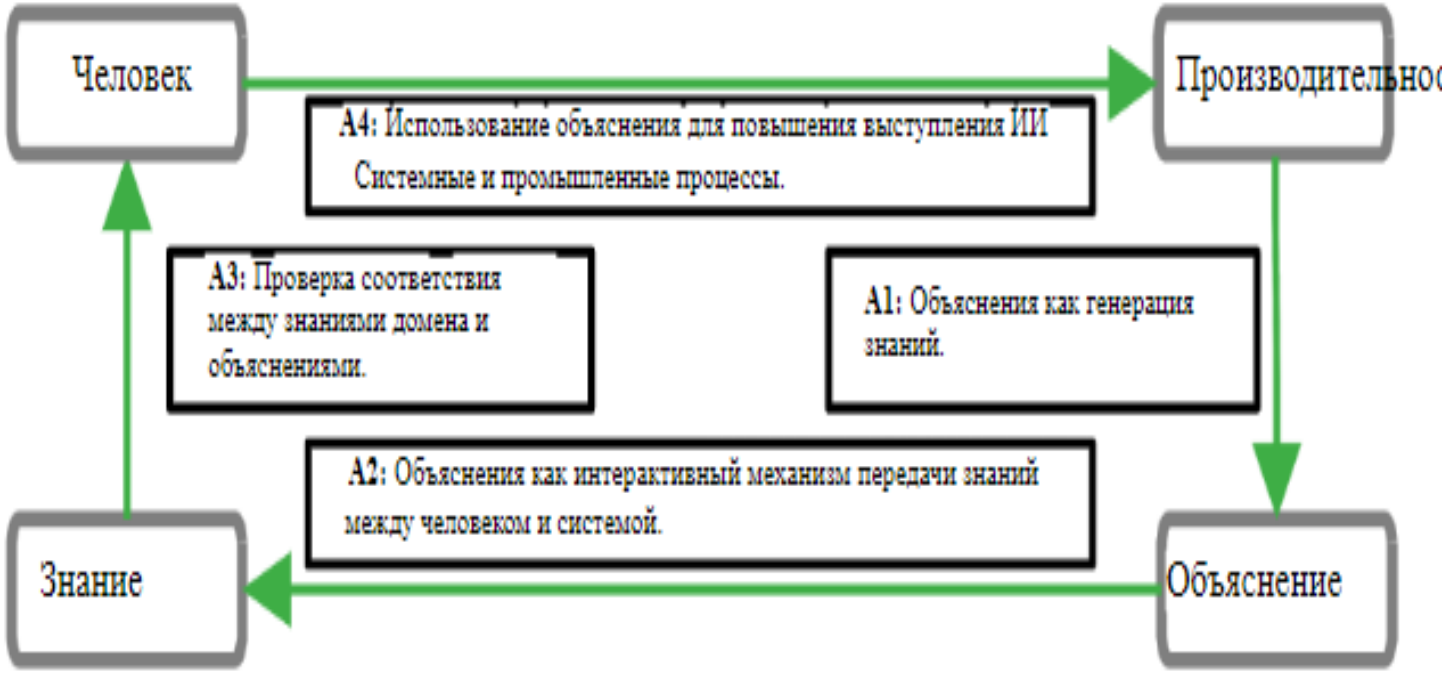
Высокоэффективные модели содержат важные знания, которые можно использовать для лучшего понимания не только самой модели, но и данных и процесса, который их генерирует. Это особенно важно в таких областях, как промышленность, где данные сами по себе являются «черным ящиком» - в большинстве случаев они поступают немаркированными, зашумленными, неполными и, что самое главное, без каких-либо явно сформулированных фоновых знаний о процессах, которые их генерируют.

Это еще более важно для Индустрии 5.0, где сотрудничество между ИИ и человеком является одним из основных предположений. Поэтому, проливая больше света на данные и фокусируясь на генерации знаний о механизме, лежащем в основе источника данных, можно улучшить не только модель, но и весь процесс (т. е. бизнес-процесс, производственный процесс, процесс обслуживания), на котором можно обучить следующее поколение моделей. Эта мотивация никогда не лежала в основе современных алгоритмов XAI. Новые методы XAI для Индустрии 5.0 требуют нескольких существенных изменений в проектных предпосылках.

Новые методы XAI для Индустрии 5.0

- A1 - Представление знаний, уже имеющихся в домене, ближе к подтверждению, чем к объяснению. Новые методы XAI должны быть больше сосредоточены на извлечении знаний из прогностических моделей, а не ограничиваться только объяснениями, которые в основном служат для отладки модели или набора данных.
- A2 - Объяснение - это акт передачи знаний. Это обусловливает необходимость исследования методов, которые позволят осуществлять двунаправленную передачу знаний между людьми и системами ИИ. Новое поколение алгоритмов XAI должно быть изначально спроектировано так, чтобы человек мог формулировать ожидания и потребности, которые должны быть удовлетворены алгоритмом XAI, и получать объяснения на желаемом уровне абстракции.
- A3 - Новые знания, обнаруженные в процессе объяснения, могут не соответствовать знаниям домена. Новое поколение алгоритмов XAI должно генерировать объяснения, которые могут быть проверены на соответствие знаниям домена и оспорены человеком . Поэтому они должны быть тесно связаны с механизмами аргументации .
- A4 - Объяснения полезны, если они пригодны к действию. Новые алгоритмы XAI должны быть разработаны таким образом, чтобы извлекаемые ими знания могли быть непосредственно использованы моделями машинного обучения, бизнес-процессами или мгновенным принятием решений, замыкая цикл объяснения.

Почему Индустрии 5.0 нужен XAI 2.0?



XAI 2.0 должен быть явно ориентирован на извлечение знаний из модели и генерацию объяснений в более целостной манере.

Контекстуальная осведомленность в программе DARPA AI Next

Программа AI Next, запущенная Агентством перспективных оборонных исследовательских проектов (DARPA), направлена на изучение и разработку технологий искусственного интеллекта следующего поколения.

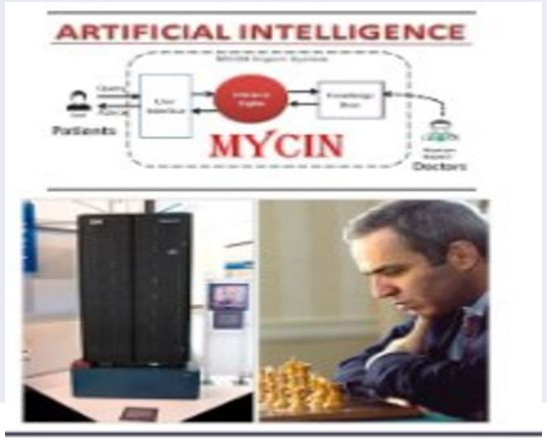
Она фокусируется на продвижении ИИ за пределы узких приложений для создания систем, которые являются более надежными, адаптируемыми и объяснимыми

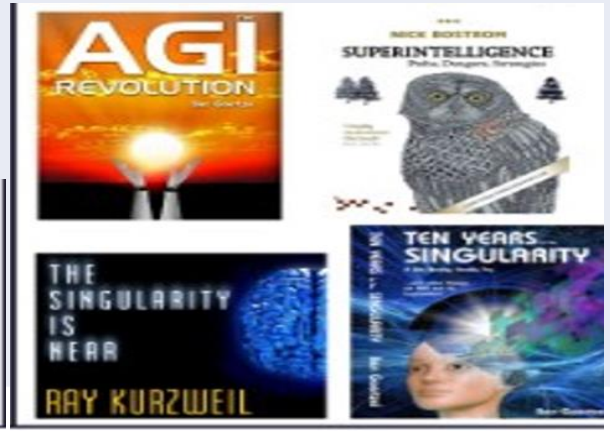
Контекстуальная осведомленность является ключевым направлением в рамках прогаммы AI Next, в которой особое внимание уделяется разработке систем искусственного интеллекта, которые адаптировать свои объяснения на основе ситуационного контекста, потребностей пользователя и факторов окружающей среды.

Это тесно согласуется с принципами XAI 2.0, особенно в том, что касается повышения прозрачности и адаптивности.

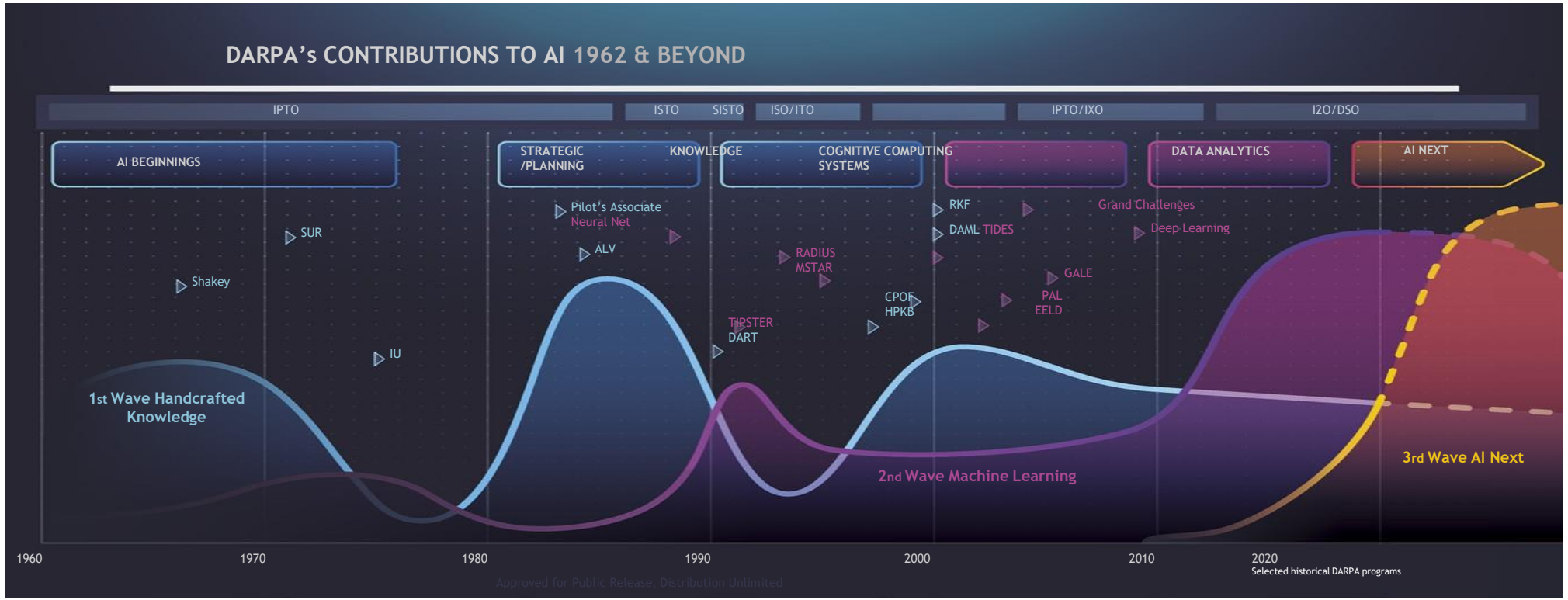
Пять волн искусственного интеллекта

Первая волна	Вторая Волна	Третья Волна	Четвертая и пятая волна
1970 – 1990 гг.	2000-2020 гг.	2020-2030 гг.	2030 ------→
Хорошо рассуждает, но есть проблемы с обучением и обобщением. Символический, эвристический, основанный на правилах	Хорошо обучается и воспринимает, но слаб в рассуждении и обобще-нии. Статистическое обучение, глубокое обучение, обработка текста и изображений	Прекрасно обучается, обобщает и рассуждает. Контекстуальная адаптация, способен объяснять решения. Способен общаться на ЕЯ. Нужны меньшие объемы данных для обучения и мини-мальный внешний контроль	Способен решать те же интеллектуальные задачи, что и человек. Сильный искусственный интеллект, ведущий к Суперинтеллекту и «технологической сингулярности»





3 волны инвестиций в ИИ агества DARPA с 1962 года



На графике показано три волны инвестиций DARPA в ИИ, высота каждой волны соответствует величине инвестиций DARPA.

Почему важна контекстуальная осведомленность?

Повышенная актуальность:

- Подбирает пояснения к конкретной задаче, предметной области или роли пользователя (например, врач или пациент).

Улучшенное удобство использования:

- Предоставляет практическую аналитику с учетом базовых знаний пользователя и целей принятия решений.

Динамическая адаптация:

- Динамически корректирует пояснения по мере появления новых данных или изменяющихся условий.

Построение доверия:

- Демонстрирует понимание реального контекста, повышая доверие пользователей к системе ИИ

Примеры контекстуальной осведомленности в действии

Здравоохранение:

Модель искусственного интеллекта, объясняющая диагноз по-разному медицинскому работнику и неспециалисту.

Финансы:

Предоставление регулирующим органам подробных причин отказа в кредите, упрощенное резюме для заявителей.

Автономные транспортные средства:

- Предоставление кратких объяснений во время рутинного вождения и более подробное обоснование во время критических маневров.

Как XAI 2.0 обеспечивает контекстуальную осведомленность?

Профилирование пользователей:

- Анализирует характеристики пользователя (например, уровень знаний, предпочтения) для создания персонализированных объяснений.

Включение окружающей среды:

- Включает в процесс объяснения внешние переменные, такие как время, местоположение или операционные ограничения.

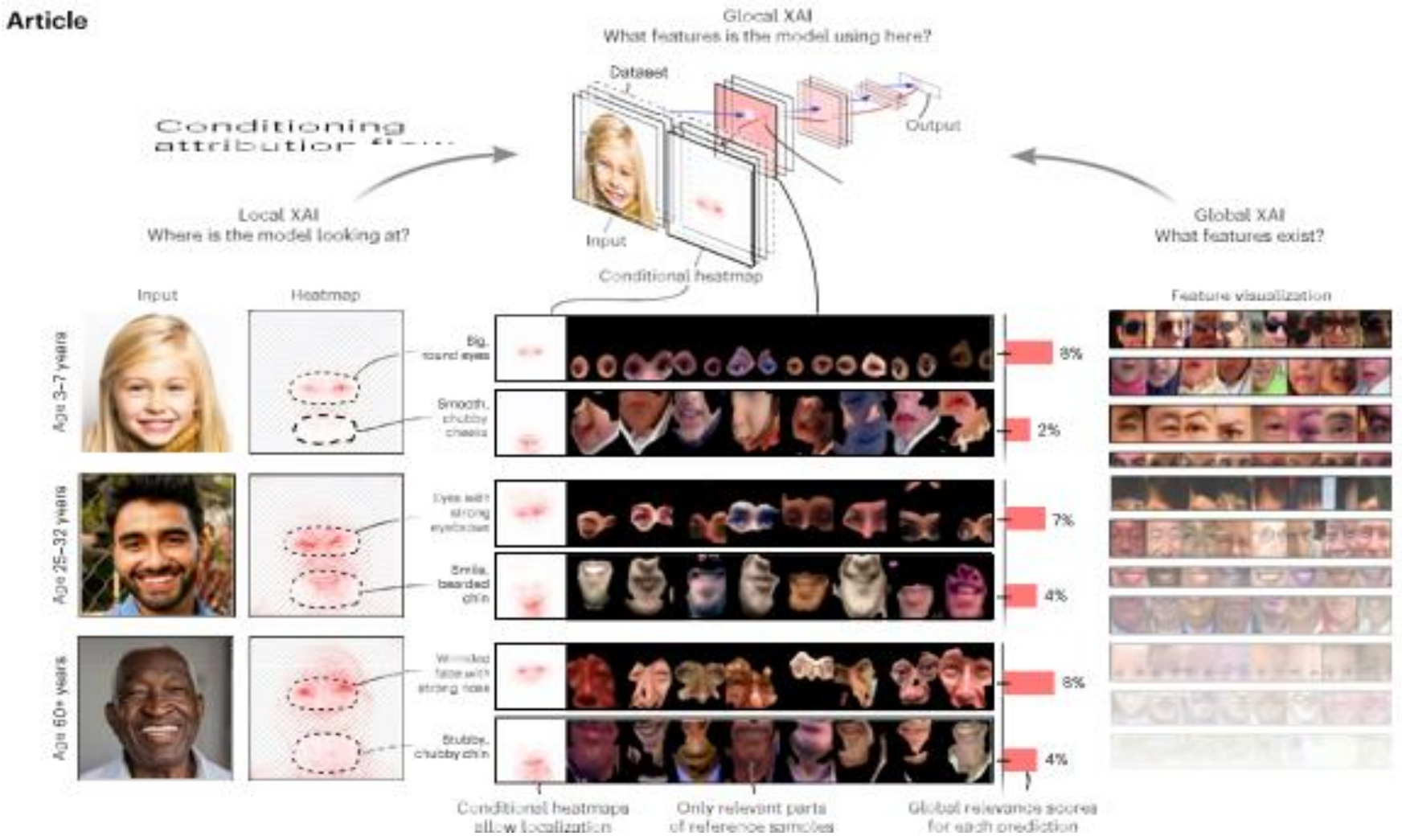
Интерактивные интерфейсы:

- Позволяет пользователям запрашивать дополнительные сведения или уточнения в зависимости от их текущих потребностей.

Петли обратной связи:

- Учится на взаимодействии с пользователем для уточнения контекстуального понимания с течением времени

Контекстуальная осведомленность. Glocal XAI объединяет локальный и глобальный XAI.



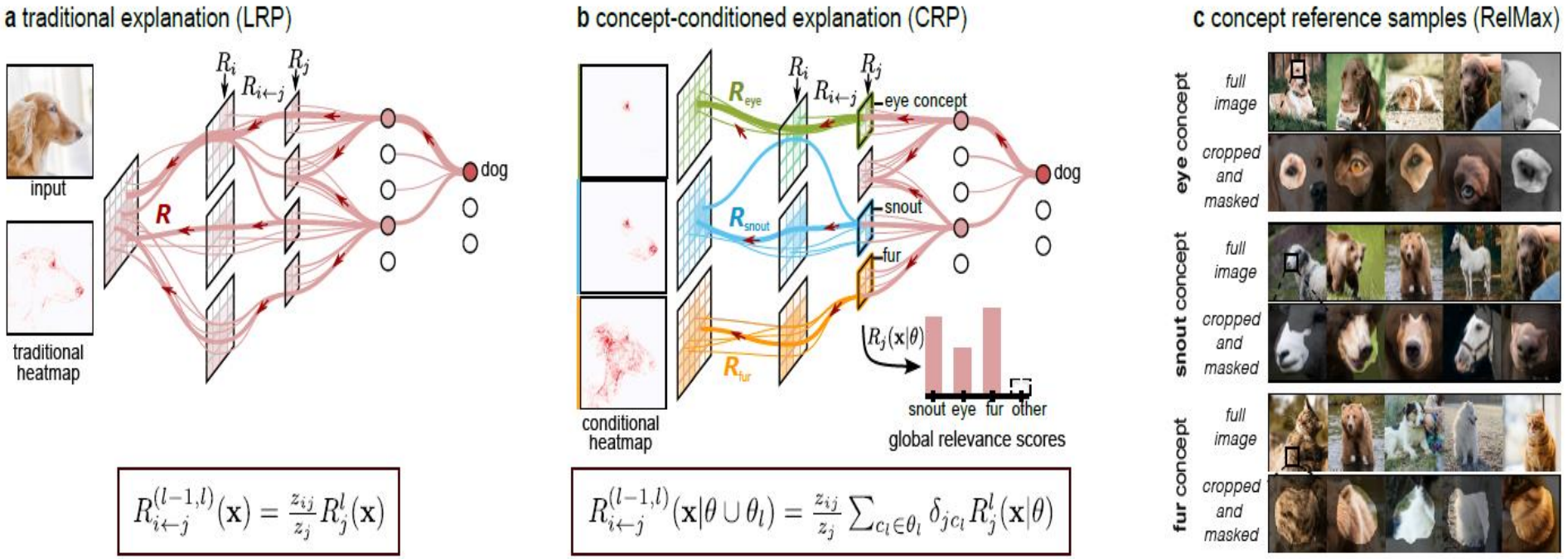
Локальные и глобальные объяснения

- Glocal XAI может определить, какие функции существуют и как они используются для прогнозирования, объединив локальный и глобальный XAI.
- Слева: локальные объяснения визуализируют, какие входные пиксели релевантны для прогноза. Здесь модель фокусируется на области глаза для всех трех прогнозов. Однако какие именно особенности модель распознала в этих регионах, остается открытым для интерпретации пользователем.
- Справа: глобальные объяснения путем нахождения референсных изображений, которые максимально представляют конкретные (групп) нейронов, глобальные методы XAI дают представление о концепциях, обычно закодированных моделью. Однако глобальные методы сами по себе не информируют о том, какие концепции распознаются, используются и комбинируются моделью при выводе для каждой выборки.

Глокальные объяснения

Центр: Центр: glocal XAI может идентифицировать соответствующие нейроны для конкретного предсказания (свойство локального XAI), а затем визуализировать, какие концепции эти нейроны кодиру (свойство глобального XAI).). Кроме того, используя концептуальные условные объяснения в качестве фильтрующей маски, определяющие части концепций могут быть выделены на референсных изображениях, что значительно повышает интерпретируемость и ясность. Здесь самая верхняя выборка была отнесена к возрастной группе 3–7 лет из-за больших радужных оболочек и круглых глаз выборки, в то время как средняя выборка прогнозируется как 25–32 года, так как видна большая часть склер и брови более заметны. Для нижней выборки модель предсказала класс 60+ на основе распознавания тяжелых морщин вокруг глаз и на веках, а также ярко выраженных слезных мешков рядом с большим узловатым носом.может идентифицировать соответствующие нейроны для конкретного предсказания (свойство локального XAI), а затем визуализировать концепции этих нейронов encode (свойство глобального XAI).

Концептуальное объяснение закодированное каналом скрытого слоя сети



$$R_{i\leftarrow j}^{(l-1,l)}(\mathbf{x}) = \frac{z_{ij}}{z_j} R_j^l(\mathbf{x})$$

$$R_{i\leftarrow j}^{(l-1,l)}(\mathbf{x}|\theta \cup \theta_l) = \frac{z_{ij}}{z_j} \sum_{c_l \in \theta_l} \delta_{jc_l} R_j^l(\mathbf{x}|\theta)$$

а) Традиционные методы, основанные на обратном распространении

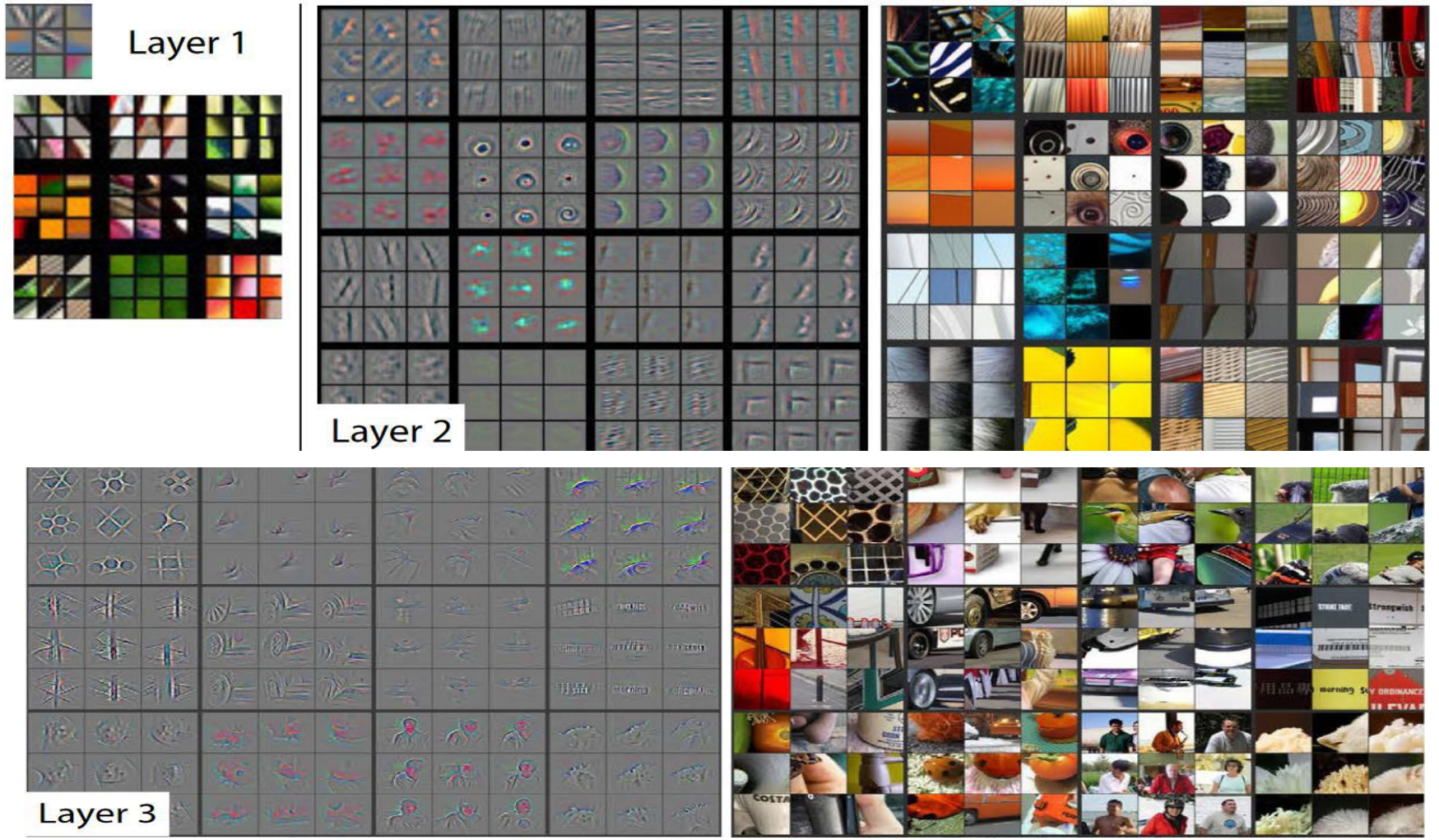
б) Концептуальное объяснение закодированное каналом скрытого слоя сети

в) Визуализировация входных выборк, в которых латентная структура релевантна для прогноза. Мы можем дополнительно выделить семантику, отображая только релевантные входные части в соответствии с концептуальными пояснениями, представленными в b.

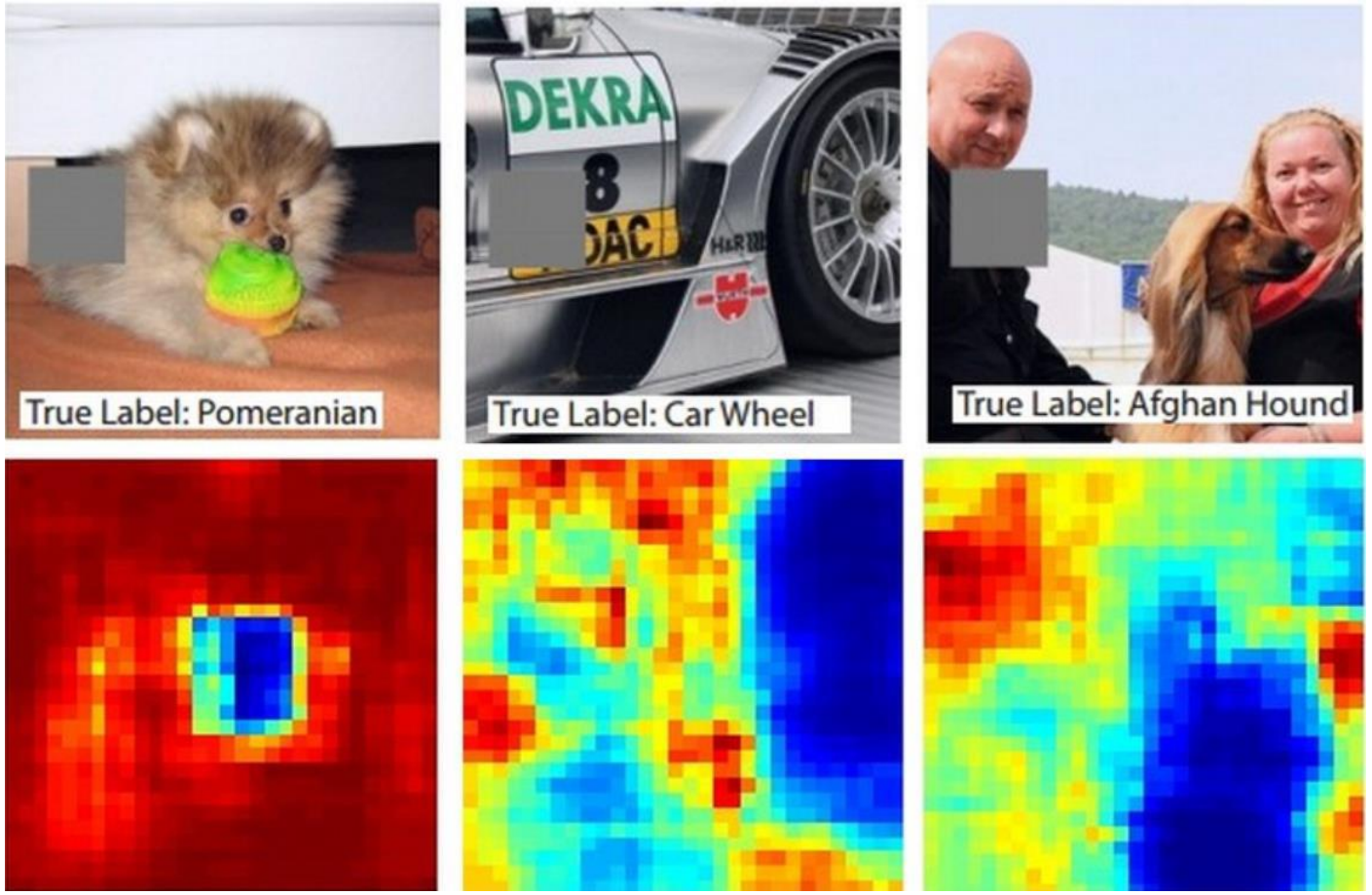
Методы визуализации признаков. Деконволюционные методы.

Деконволюционные методы. В отличие от первой группы методов деконволюционные методы рассматривают сверточную сеть как «белый ящик» и напрямую используют структуру сети для визуализации. Основу этих методов составляет идея о том, что необходимо определить вклад каждого пикселя входного изображения, начиная с активации на интересующем слое и итеративно вычисляя вклад каждой единицы в предшествующем слое на текущую активацию. Таким образом, перемещаясь назад по сети до тех пор, пока не будет достигнут входной слой, могут быть получены значения вклада каждого пикселя, которые вместе образуют визуализацию признаков, наиболее значащих для активации на интересующем слое. Для реализации описанной идеи применяются глубокие развертночные сети. Изначально указанные глубокие модели применяются для обучения сверточных нейросетей без учителя, но оказывается, что они могут быть использованы и для восстановления пути в сети с целью определения наиболее важных паттернов и патчей.

Визуализация сверточных сетей. Деконволюционные методы.



Визуализация сверточных сетей. Методы изменения входа



Визуализация сверточных сетей.

В основе методов лежат техники модификации входа и измерения результирующих изменений на выходе сети или на промежуточных слоях. При этом набор слоев сети, расположенных до интересующего слоя, рассматривается как «черный ящик». Цель методов – визуализация свойств функции, которая представляется этим «черным ящиком». Идея метода состоит в том, что на вход сети подается изображение, у которого перекрываются разные области серым квадратом, и измеряется изменение активации на интересующем слое. Очевидно, что сокрытие важных частей изображения приводит к значительному изменению значений функции активации, и наоборот. Перебор различных положений серого квадрата (подобно «бегущему окну») обеспечивает построение тепловой карты, отражающей важность той или области входного изображения .

Заключение

Современные модели глубокого обучения и большие языковые модели демонстрируют впечатляющие результаты в обработке естественного языка, компьютерном зрении и других задачах. Однако их работа остаётся во многом непрозрачной для пользователя, что затрудняет доверие к их выводам и их применение в критически важных областях. Системы объяснительного искусственного интеллекта (ОИИ) позволяют интерпретировать и объяснить алгоритм принятия решений, даже если он имеет природу черного ящика.

В докладе проводится краткий обзор и анализ эволюции существующих методов ОИИ. Особое внимание будет уделено концепции ОИИ второго поколения (или ОИИ 2.0), которая фокусируется на улучшении качества, адаптивности и персонализации объяснений. В отличие от ОИИ 1.0, который в основном предоставлял постфактум интерпретации («почему модель приняла такое решение?»), ОИИ 2.0 интегрируется в сам процесс обучения и принятия решений модели, делая её более интерактивной, контекстно-осведомлённой (contex-aware) и ориентированной на пользователя.