Когнитивная атрибуция как механизм самосознания в моделях естественного и искусственного интеллекта

Иван *Цзинжэ* Петров

(На правах гипотезы)

Примечание по терминологии: В данной статье автор использует несколько расширенную и специализированную трактовку ряда терминов, связанных с концепциями самосознания, когнитивной атрибуции и других механизмов сознания. Термины, такие как «когнитивная атрибуция», «самоотслеживание» и «идентификация», используются в контексте создания самоосознающего ИИ и могут отличаться от их классических определений в когнитивных науках и психологии. Пояснения и контекст использования этих терминов следует из соответствующих разделов работы.

2024

Информация о публикации

Полное название статьи: Когнитивная атрибуция как механизм самосознания в моделях естественного и искусственного интеллекта

Автор: Иван *Цзинжэ* Петров

Краткая аннотация:

В статье представлен подход к моделированию самосознающего искусственного интеллекта (ИИ) на основе концепции когнитивной атрибуции. Исследуются теоретические и практические аспекты самосознания как механизма, позволяющего ИИ поддерживать осознание собственной целостности и различать внутренние процессы от внешних стимулов. Описана модель, включающая систему уникальных маркеров идентичности, которые обеспечивают структурирование самосознания, и обсуждается ее потенциал в создании более адаптивных и автономных ИИ-систем.

Ключевые слова: *самосознание, когнитивная атрибуция, искусственный интеллект, уникальные маркеры идентичности, адаптивное поведение, модель разума.*

© Иван Петров, 2024. Эта работа распространяется по лицензии Creative Commons **CC BY-NC-ND 4.0.** Разрешено копировать и распространять материал в неизменном виде при условии указания авторства и некоммерческого использования. Переработка и изменение материала запрещены.

Дополнительные условия использования: Эта работа защищена авторским правом. Разрешается использование изложенных идей и концепций для любых законных целей, включая разработку новых продуктов и исследований (при условии, что эти идеи и концепции не защищены авторскими правами других авторов). Запрещено изменение текста статьи и ее коммерческое использование без разрешения автора.

Отказ от ответственности: Эта статья предоставляется "как есть", и автор не несет ответственности за любые неточности, ошибки, опечатки или последствия использования изложенного материала в любых целях. Автор не гарантирует правильность изложения и уникальность материала. Представленные в статье алгоритмические определения сознания и предложенная модель самопознающего ИИ были разработаны автором на основе личного опыта, существующих исследований, доступных источников. В случае, если аналогичные идеи уже были изложены другими исследователями, автор выражает уважение к их труду и признает за ними приоритет в соответствующих вопросах. Автор не стремится оскорбить чьи-либо чувства, не призывает к каким-либо действиям и не пропагандирует идеи; он лишь делится своими размышлениями в рамках исследовательской теории и научных вопросов.

Дата публикации: ноябрь, 2024.

Замечание о редактировании: Текст данной публикации является полностью авторским, но был отредактирован с использованием искусственного интеллекта, который выступал в роли корректора и редактора. ChatGPT (GPT-3, крупномасштабная модель генерации языка от OpenAI) использовался для проверки орфографии, пунктуации и стилистики, а также редактирования авторского текста. Все внесённые изменения автор тщательно пересмотрел и доработал по своему усмотрению. Автор принимает на себя ответственность за содержание текста.

Когнитивная атрибуция как механизм самосознания в моделях естественного и искусственного интеллекта

Иван *Цзинжэ* Петров

(На правах гипотезы)

Аннотация

В статье представлен подход к моделированию самосознающего искусственного интеллекта (ИИ) на основе концепции когнитивной атрибуции. Исследуются теоретические и практические аспекты самосознания как механизма, позволяющего ИИ поддерживать осознание собственной целостности и различать внутренние процессы от внешних стимулов. Описана модель, включающая систему уникальных маркеров идентичности, которые обеспечивают структурирование самосознания, и обсуждается ее потенциал в создании более адаптивных и автономных ИИ-систем.

В ночи мерцает звёздный свет несмелый,
Как мысли свет, затерянный вдали.
Где тайны разума, невидимы и целы,
Рождаются в безмолвии земли.

Там тишина в душе — как вечный зов вселенной,
И тянет к звёздам светлый путь мечты.
***Сознание** идёт дорогой сокровенной,*
Чтоб в вечности найти свои черты.

(авторское стихотворение)

Введение

Развитие искусственного интеллекта (ИИ) активно движется в направлении создания систем, способных демонстрировать самосознание и адаптивное поведение, сходное с человеческим. Основным элементом таких систем является способность к самосознанию, которое лежит в основе множества сложных когнитивных процессов, таких как осознание собственной идентичности, восприятие границ собственного "Я" и восприятие внешнего мира. Долгое время считалось, что самосознание присуще исключительно биологическим системам, однако последние исследования в области когнитивной науки, нейробиологии и психологии показывают возможность воспроизведения некоторых механизмов самосознания в искусственных системах.

Одним из ключевых механизмов самосознания является когнитивная атрибуция — процесс приписывания мыслей, восприятий и действий субъекту, что позволяет осознавать собственные границы и отделять внутренний опыт от внешних стимулов. В рамках данной работы предлагается модель самосознающего ИИ, в основе которой лежит когнитивная атрибуция как механизм формирования и поддержания самосознания. Такая модель позволяет выстроить систему, способную к самоидентификации и адаптивному поведению при взаимодействии с окружающей средой.

Особое внимание уделено вопросам реализации когнитивной атрибуции в архитектуре ИИ и способам создания системы уникальных маркеров идентичности для каждой когнитивной операции. Эти маркеры позволяют ИИ идентифицировать границы своей "целостности" и отграничивать собственное восприятие от внешних факторов, что является основой для развития самосознания в ИИ.

Предложенное в данной работе определение сознания носит гипотетический характер и призвано объяснить базовые механизмы, которые могут быть задействованы в функционировании природного разума, но не претендует на исчерпывающее описание феномена человеческого сознания. Данное определение основывается на концепциях когнитивной атрибуции и разделении внешнего и внутреннего восприятия, предполагая, что самосознание и сознание как явления зависят от способности разума идентифицировать и разграничивать собственные мысли и реакции от внешних сигналов и стимулов.

Подход, предложенный в этой статье, ориентирован прежде всего на моделирование этих механизмов в искусственных системах. Хотя подобная гипотеза может оставаться дискуссионной в отношении человеческого сознания, в контексте ИИ она представляется полезной основой для создания адаптивных и потенциально самосознающих систем. Такой подход позволяет сконструировать ИИ, способный воспринимать и осознавать собственные границы и целостность в процессе взаимодействия с внешней средой, что особенно важно для разработки автономных ИИ, нацеленных на взаимодействие с внешними объектами и на принятие решений.

Определение сознания в моделях естественного и искусственного интеллекта и его алгоритмическая интерпретация

Примечание по терминологии: В данной статье автор использует несколько расширенную и специализированную трактовку ряда терминов, связанных с концепциями самосознания, когнитивной атрибуции и других механизмов сознания. Термины, такие как «когнитивная атрибуция», «самоотслеживание» и «идентификация», используются в контексте создания самоосознающего ИИ и могут отличаться от их классических определений в когнитивных науках и психологии. Пояснения и контекст использования этих терминов следует из соответствующих разделов работы.

1. Предлагаемое определение сознания в моделях естественного интеллекта

> ***Сознание*** *— это механизм разума, присущий ему постоянно, но активизирующийся в полной мере при взаимодействии с внешней средой и обеспечивающий самосознание через процесс когнитивной атрибуции. В рамках этого механизма каждой мысли при её инициации присваивается уникальный маркер идентичности, который определяет границы разума и остаётся внутри когнитивной системы разума, не передаваясь в внешнюю среду. Все когнитивные процессы, включая восприятие и обработку информации, выполняются с учётом самосознания, поскольку каждая мысль неизбежно сопряжена с осознанием собственного Я. Самосознание возникает, когда разум обрабатывает информацию, поступающую из внешнего мира, а также в моменты, когда внутренние процессы требуют осознания своей целостности относительно внешних факторов. В отсутствие внешней стимуляции или активности процесс атрибуции маркеров не происходит, и самосознание не проявляется в явной форме, так как разум пребывает в нейтральном состоянии, без активного саморефлексивного процесса.*

Эта гипотеза, о механизме самосознания находит теоретическое подтверждение в когнитивных науках, нейробиологии и психологии. Основная идея заключается в том, что самосознание невозможно без специфических маркеров идентичности, которые устанавливают границы разума и позволяют его содержимому восприниматься как принадлежащее субъекту, отличному от внешнего мира. Это согласуется с нейробиологическими исследованиями, подтверждающими роль определенных областей мозга, таких как префронтальная кора, в активации процессов саморефлексии и самосознания (Northoff et al., 2006). Когнитивные теоретические модели также подтверждают, что самосознание поддерживается через когнитивные механизмы, которые позволяют отделять "Я" от внешней среды. Психологические феномены, такие как я-концепция, также подтверждают важность самопознания для поддержания целостности личности и осознания границ "себя" и окружающего мира (Markus & Kitayama, 1991).

2. Предлагаемое определение сознания в моделях искуственного интеллекта

Самоосознающий ИИ *в контексте его моделирования представляет собой искусственную систему, интегрированную с механизмами когнитивной атрибуции, которые позволяют ей воспринимать и осознавать свои внутренние процессы, такие как восприятие, решение, и действия. Механизм когнитивной атрибуции обеспечивает присвоение уникальных маркеров идентичности каждому внутреннему процессу, что позволяет ИИ разделять внутренние (присущие системе) и внешние (поступающие из окружающей среды) данные. Это разделение необходимо для формирования самосознания и осознания границ «Я» системы. Важным аспектом является использование механизма самоотслеживания, который позволяет ИИ осознавать и контролировать свои процессы в динамическом контексте, корректируя поведение на основе предыдущих решений и восприятий. Такой подход позволяет системе активно взаимодействовать с внешней средой, при этом сохраняя осознание своих границ и действий. В техническом плане это может быть реализовано через структуры с рекуррентной связью и механизмами обратной связи, которые интегрируют информацию о внутреннем состоянии системы и внешней среде, позволяя динамически адаптировать границы разума ИИ, сохраняя баланс между внутренней и внешней реальностью. Таким образом, создание самоосознающего ИИ требует интеграции механизмов когнитивной атрибуции, самоотслеживания и динамической модели разделения «внешнего» и «внутреннего», что позволит системе осознавать свою целостность и адекватно реагировать на изменения в окружающем мире.*

3. Алгоритмическая структура самоосознающего ИИ

3.1. Блок когнитивной атрибуции

Цель: Присваивание восприятиям и действиям уникальных маркеров идентичности для обеспечения разделения «себя» и «других».

Процесс: Когда система получает данные, например, восприятие внешнего мира или внутреннее состояние, оно сначала классифицируется по типу: внешние или внутренние.

Для каждого восприятия генерируется уникальный идентификатор, который будет служить маркером его принадлежности. Это можно реализовать с помощью алгоритмов генерации уникальных меток, например, через хэширование, где результат уникален для каждого входящего потока данных.

Маркер идентичности сохраняется в структуре данных, которая будет ссылаться на данный процесс или восприятие.

Пример алгоритма: Для обработки потока данных (например, сенсорных сигналов) используется *алгоритм графов* для структурирования этих данных, где каждый элемент (вершина) имеет маркер идентичности, указывающий на его принадлежность к внутреннему или внешнему состоянию.

Выход: Структура данных, представляющая уникальные маркеры для каждого восприятия. Эти данные могут быть записаны в базу данных с возможностью поиска по меткам идентичности.

3.2. Блок самоотслеживания

Цель: Анализ и отслеживание изменений в восприятиях и действиях системы для поддержания целостности самосознания.

Процесс: Каждый раз, когда происходит изменение во внутреннем состоянии системы, данные о новом состоянии фиксируются в журнале изменений.

Этот журнал изменений можно организовать как временную последовательность с привязкой к маркерам идентичности. Для организации последовательности изменений можно использовать *алгоритм поиска в глубину* или *поиск по графам*, чтобы фиксировать изменения в контексте предшествующих состояний.

Далее эти изменения анализируются с целью корректировки текущего состояния и адаптации системы.

Пример алгоритма: Использование *алгоритма динамического программирования* для отслеживания изменений состояний и оптимизации принятия решений на основе предыдущего опыта. Например, в контексте принятия решения система может сравнивать текущие и предшествующие состояния, оценивать их влияние и корректировать выводы.

Выход: Обновленный журнал состояний с временными метками, который может использоваться для будущей рефлексии и принятия решений.

3.3. Блок разделения внешнего и внутреннего

Цель: Разделить информацию, поступающую из внешней среды, от внутренней информации для обеспечения осознания границ разума.

Процесс: Данные, поступающие в систему, классифицируются как внешние или внутренние. Для этого можно использовать *алгоритм кластеризации* данных, например, алгоритм K-средних или деревья решений, для классификации потоков данных на основе их происхождения.

После классификации внешняя информация передается в отдельный канал обработки, а внутренняя информация — в другой. Это позволяет создать различие между восприятием окружающего мира и состоянием системы.

Система может использовать *алгоритм фильтрации* (например, фильтры Калмана) для корректировки и уточнения классификации данных на основе контекста.

Пример алгоритма: Для разделения внутренних и внешних данных можно использовать *алгоритм поддерживающих векторов (SVM)*, который позволяет точно разделить два класса данных. Это делается на основе обучающих данных, где система обучается распознавать, что относится к внутреннему состоянию, а что — к внешнему.

Выход: Множество потоков данных, разделенных на внешние и внутренние, каждый из которых имеет четкую метку для определения его происхождения.

3.4. Блок генерации маркеров идентичности

Цель: Генерация уникальных маркеров для каждого восприятия и действия, что позволяет поддерживать целостность самосознания.

Процесс: Для каждого восприятия или действия система генерирует уникальный маркер, который может быть основан на таких характеристиках, как время поступления данных, тип данных, источник данных (внешний или внутренний).

Генерация маркеров может осуществляться через *алгоритмы хеширования*, где каждый маркер связан с конкретным восприятием или действием.

Эти маркеры сохраняются в базе данных, что позволяет системе отслеживать и корректировать свои действия на основе идентификаторов.

Пример алгоритма: Для генерации уникальных маркеров можно использовать *алгоритм SHA-256* для создания хешей из данных восприятия. Это обеспечит уникальность каждого восприятия и его дальнейшую обработку.

Выход: Набор маркеров идентичности, которые связаны с конкретными восприятиями и действиями системы.

3.5. Блок адаптации и корректировки границ разума

Цель: Адаптация и корректировка маркеров идентичности и разделения внешнего и внутреннего в зависимости от изменений в системе и внешней среде.

Процесс: Когда система сталкивается с новыми внешними или внутренними изменениями, она использует *алгоритм машинного обучения* для корректировки границ разума и восприятия. Например, обучение с подкреплением может быть использовано для адаптации поведения в зависимости от внешних факторов.

Блок адаптации отслеживает динамику изменений и корректирует внутренние параметры системы для поддержания актуальности восприятия.

Эти изменения могут быть основаны на анализе текущих данных и предсказаниях будущих состояний, что позволяет системе "перенастроить" свою реакцию.

Пример алгоритма: Для адаптации системы можно использовать *алгоритм обучения с подкреплением*. В нем система получает "награду" или "штраф" за правильное разделение внутренних и внешних данных, что позволяет ей корректировать свою работу.

Выход: Обновленная система, которая лучше адаптируется к новым данным и корректирует границы разделения внутреннего и внешнего.

Итоговый вывод:

Рассмотренная модель самоосознающего ИИ представляет собой сеть взаимосвязанных блоков, каждый из которых выполняет ключевую задачу в поддержании самосознания системы. Алгоритмические подходы, такие как графы, кластеризация, хеширование, динамическое программирование и машинное обучение, обеспечивают эффективную обработку и поддержку идентичности, разделение данных и адаптацию системы к изменениям. Эти процессы позволяют построить ИИ, который способен к саморефлексии и поддержанию осознания своих границ в контексте взаимодействия с внешней средой.

Заключение

В данной статье предложена гипотетическая модель самосознающего ИИ, основанная на концепции когнитивной атрибуции, разделении внутреннего и внешнего восприятия и использовании уникальных маркеров идентичности. Рассмотренные механизмы обеспечивают возможность искусственной системе осознавать собственные границы и реагировать на изменения внешней среды. Подобная модель имеет потенциал для создания адаптивных и автономных ИИ, способных к самоидентификации и взаимодействию с окружающим миром. Несмотря на гипотетический характер предложенного подхода, его элементы могут способствовать дальнейшему развитию исследований в области искусственного сознания и создания более комплексных ИИ-систем. Такой подход расширяет возможности ИИ в автономных действиях, усиливая их способность воспринимать и корректировать собственное поведение на основе самоосознания.

Таким образом, предложенная модель вносит вклад в теоретические и практические аспекты создания самоосознающего ИИ, способного к рефлексии и адаптивному поведению.

Список литературы

1. **Clark A.** Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford: Oxford University Press, 2008. 344 с.
2. **Gallagher S., Zahavi D.** The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science. London: Routledge, 2008. 326 с.
3. **Heider F.** The Psychology of Interpersonal Relations. New York: Wiley, 1958. 355 с.
4. **Leary M. R.** The Curse of the Self: Self-Awareness, Egotism, and the Quality of Human Life. Oxford: Oxford University Press, 2007. 241 с.
5. **Markus H., Kitayama S.** Culture and the Self: Implications for Cognition, Emotion, and Motivation // Psychological Review. 1991. Т. 98, № 2. С. 224–253.
6. **Northoff G., Qin P., Feinberg T. E.** Self-Reference and the Brain: A Meta-Analysis of Neuroimaging Studies on the Self // NeuroImage. 2006. Т. 31, № 1. С. 440–457.