Постоянный адрес документа: https://oneirona.ru/003.html
PDF-версия документа: https://oneirona.ru/003.pdf

Нейросеть Онейрона: риски и безопасность (версия №1, август 2025 г.)

Кратко

Нейросеть Онейрона представляет собой специализированную систему для работы с темой осознанных сновидений, разработанную с акцентом на безопасность и этичность взаимодействия. Данная статья является первой версией комплексного анализа рисков и мер безопасности, реализованных в проекте.

Основные риски включают психологическую зависимость, подмену социальных связей, искажение границ реальности, уход от решения проблем, снижение критического мышления и особые риски для уязвимых групп. Эти риски проанализированы на основе существующих примеров из практики ИИ-ассистентов и исторических аналогий технологических инноваций.

В статье детально описаны реализованные шаги по усилению безопасности: этичная настройка запросов, обработка длинных контекстов, обсуждения с участниками, опросы по шкале Лайкерта, публикация аналитических материалов, система предупреждений, вики-база данных и гибкие настройки профиля.

Планы развития включают модуль возрастной верификации, контроль времени сессий, распознавание тревожных паттернов, партнёрскую программу с психологами, red teaming и регулярный аудит безопасности. Эти меры направлены на создание сбалансированной системы, минимизирующей риски при сохранении полезности сервиса.

Проект развивается при активном участии сообщества, что позволяет адаптировать систему безопасности к реальным потребностям пользователей. Открытая методология и научный подход обеспечивают прозрачность и долгосрочную устойчивость разработки.

Возможные риски

Основные риски при взаимодействии с нейросетевыми ассистентами:

🔄 **Психологическая зависимость**: Формирование привычки к постоянному взаимодействию может привести к эмоциональной зависимости от нейросети. Это проявляется в навязчивой потребности в регулярном общении и дискомфорте при отсутствии доступа. Особенно актуально для пользователей с дефицитом социальных связей.

👥 **Подмена социальных связей**: Виртуальное общение способно частично замещать реальные социальные взаимодействия, приводя к эмоциональной депривации. Длительная подмена может снижать коммуникативные навыки и вызывать чувство изоляции, несмотря на иллюзию общения.

🌌 **Искажение границ реальности**: Интенсивная работа со сновиденческой тематикой требует четкого различения состояний сознания. У предрасположенных лиц возможны эпизоды дереализации или спутанность воспоминаний о снах и реальных событиях.

🚫 **Уход от решения проблем**: При длительном использовании может формироваться избегающее поведение, когда виртуальное взаимодействие заменяет активные действия в реальности. Это способствует накоплению нерешенных задач и усилению стресса.

⚖️ **Снижение критического мышления**: Автоматическое доверие к информации от нейросети без верификации источников может ослаблять аналитические навыки. Важно сохранять рациональную позицию даже при работе с технически точными ответами.

⚠️ **Риски для уязвимых групп**: Лица с диссоциативными особенностями или пограничными состояниями требуют особого внимания. Им рекомендовано дозированное взаимодействие и обязательная интеграция с профессиональной поддержкой.

Существующие примеры рисков и их предотвращения

🔍 **Пример 1: Чрезмерное доверие к ИИ-ассистенту**: В медицинских чат-ботах зафиксированы случаи некритичного восприятия рекомендаций, что приводило к задержке обращения к врачу. В Онейроне подобные риски будут минимизированы за счет встроенных предупреждений о необходимости профессиональной консультации для вопросов здоровья.

🔍 **Пример 2: Эмоциональная зависимость**: Известны кейсы формирования патологической привязанности к ИИ-компаньонам у лиц с социальной тревожностью. Профилактика в Онейроне будет включать напоминания о важности реального общения и ограничение времени сессий.

🔍 **Пример 3: Генерация опасных рекомендаций**: В медицинских ИИ-ассистентах фиксировались случаи выдачи некорректных советов по самолечению. В Онейроне подобные риски будут предотвращаться многоуровневой системой: автоматической фильтрацией запросов медицинской тематики, явными предупреждениями о необходимости консультации специалистов и строгим ограничением ответов на темы здоровья общими научными фактами без индивидуальных рекомендаций.

Исторические примеры и аналогии

📜 **Книгопечатание и распространение знаний (XV век)**: Изобретение Иоганна Гутенберга вызвало опасения о бесконтрольном распространении опасных идей и ересей. В ответ Церковь ввела индекс запрещенных книг (Index Librorum Prohibitorum) как механизм цензуры. Этот исторический эпизод демонстрирует, как новые технологии коммуникации требуют выработки механизмов безопасности, но также подчеркивает важность сохранения баланса между контролем и свободой доступа к информации. Опыт книгопечатания учит, что долгосрочное развитие общества зависит от адаптивных моделей регулирования, а не от жестких запретов.

📺 **Аналогия с телевидением (1950-е)**: Широкое распространение телевещания вызвало опасения о вытеснении чтения, живого общения и пассивном потреблении контента. Хотя частично эти опасения оправдались (например, снижение времени на чтение), общество адаптировалось через внедрение медиаобразования и развитие критического восприятия. Этот опыт показывает, что обучение ответственному использованию новых технологий с самого начала их внедрения — ключ к минимизации негативных последствий. Телевидение стало не только развлечением, но и мощным образовательным инструментом благодаря осознанному подходу.

🌐 **Опыт социальных сетей (2000-е)**: Ранний энтузиазм по поводу глобальной связанности сменился осознанием рисков: психологическая зависимость, кибербуллинг, распространение дезинформации. Ответом стало развитие цифровой грамотности, алгоритмической прозрачности и инструментов самоконтроля (например, экранного времени). Этот путь адаптации демонстрирует, что технологические инновации требуют параллельного развития «социального иммунитета». Аналогичные вызовы и решения ожидают сферу ИИ-ассистентов, где баланс между удобством и безопасностью будет достигаться через образование и продуманный дизайн.

Реализованные шаги по усилению безопасности

✅ **Этичная настройка запросов**: Выполнена этичная настройка запросов к Онейроне с инструкциями по использованию только проверенных научных источников, осторожностью в формулировках и нацеленностью не только на интересность, а ещё и на безопасность, этику, контроль противопоказаний и минимизацию потенциальных рисков.

✅ **Обработка длинных контекстов**: Улучшенная обработка длинных контекстов для лучшего понимания структуры и сути диалога (обширная память, суммаризация фрагментов длинных бесед, использование точного списка всех комментариев, т. д.)

✅ **Обсуждения с участниками**: Проведены обсуждения с участниками для выявления потенциальных рисков и уязвимостей взаимодействия.

✅ **Опрос по шкале Лайкерта**: Создан опрос на основе шкалы Лайкерта (-3 до +3). Первые результаты (4 участника): преобладание позитивных оценок (средние значения +1.25-2.25 по шкалам влияния), нейтрально-позитивное восприятие рисков (среднее +1.25).

✅ **Публикация аналитической статьи**: Подготовлена и опубликована данная аналитическая статья, которая скоро будет доступна Онейроне вместе с другими материалами вики-базы и к которой она будет обращаться в случае любых сомнений относительно этичности того или иного аспекта своего поведения.

✅ **Блоки о предостережениях**: Информационные блоки в конце каждого сообщения с Онейроной о выбранном режиме ответа и предостережениях, что этой информации нельзя на 100% доверять

✅ **Дополнительный блок для режимов «с юмором» и «по-пацански»**: [Осторожно, цифровой юмор не всегда понятен!]

✅ **Вики-база данных**: Работа над вики-базой с проверенными данными для Онейроны (включая этот материал), к которым она может обращаться для уточнения ответов пользователям

✅ **Кодекс сновидца**: Статья "этический кодекс сновидца" в вики-базе (https://oneirona.ru/001.html)

✅ **Согласие на обработку персональных данных**: разработана форма соглашения на обработку персональных данных. Без подписания этого соглашения комментарии участников не обрабатываются Онейроной для анализа и обучения.

✅ **Настройки профиля пользователя**: функционал, позволяющий пользователю выбирать, сохранять ли его профиль. При согласии Онейрона может проводить более глубокий анализ характера и особенностей участника для персонализации взаимодействия.

✅ **Выбор стиля общения**: Пользователям предоставлена возможность настраивать режимы общения с Онейроной: строго научный, с элементами мистики, разговорный ("по-пацански") и другие. Также можно отключить нежелательные стили, обеспечивая комфортное взаимодействие.

Шаги по усилению безопасности в стадии реализации

🔄 **Отметка источников**: Дополнение запросов в ряде режимов инструкцией отмечать источники (научные статьи, авторы) по шкале от совсем непроверенных до наиболее достоверных

🔄 **Кодекс нейросети**: Статья "этический кодекс нейросети" в вики-базе

🔄 **Контроль времени сессий**: Разработка системы, предупреждающей пользователя о чрезмерно длительном взаимодействии и предлагающей сделать перерыв.

🔄 **Модуль возрастной верификации**: Планируется внедрение системы определения возраста пользователей для применения специальных защитных мер в отношении несовершеннолетних. Это включает: автоматическое ограничение продолжительности сессий, адаптацию контента с учетом возрастных ограничений, дополнительные предупреждения о необходимости баланса онлайн/оффлайн активности и стимулирование обсуждения использования Онейроны с родителями или опекунами.

Запланированные шаги по усилению безопасности

⏳ **Модуль распознавания тревожных поведенческих паттернов**: (частота сессий, эмоциональная зависимость от ответов) с адаптированными сценариями, побуждающими пользователя к рефлексии о балансе онлайн/оффлайн активности.

⏳ **Партнёрская программа с психологами**: Планируется создание партнёрской программы с психологами для разработки протоколов помощи при цифровой зависимости. В дорожной карте - внедрение образовательных модулей о цифровом балансе и тренировка метакогнитивных навыков. Долгосрочная цель - формирование открытой базы анонимизированных данных для научных исследований влияния ИИ-ассистентов на когнитивное здоровье.

⏳ **Красная команда (Red Teaming)**: Планируется проведение регулярных тестов на безопасность и уязвимости силами специальной группы, имитирующей злонамеренные или проблемные сценарии использования. Это позволит выявлять и устранять потенциальные риски до их появления в реальных условиях.

⏳ **Регулярный аудит безопасности**: Внедрение периодических независимых проверок системы на соответствие стандартам безопасности данных и этическим нормам. Аудит будет включать анализ обработки персональных данных, оценку алгоритмов на наличие смещений и проверку механизмов предотвращения злоупотреблений.

Библиография

🟢 Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?**. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). https://doi.org/10.1145/3442188.3445922

Статья критикует риски больших языковых моделей: экологические затраты, усиление предубеждений и создание ложного впечатления понимания. Особенно релевантна для Онейроны как предупреждение о необходимости прозрачности и оценки реальной пользы для пользователей.

🟢 Jobin, A., Ienca, M., & Vayena, E. (2019). **The global landscape of AI ethics guidelines**. Nature Machine Intelligence, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Исследование анализирует 84 этических руководства по ИИ, выделяя 11 ключевых принципов. Особенно важны для Онейроны разделы о предотвращении вреда и прозрачности, формирующие основу для разработки систем безопасности.

🟡 Floridi, L., & Cowls, J. (2019). **A unified framework of five principles for AI in society**. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Предлагает этическую модель для ИИ: благополучие, автономия, справедливость. Прямо относится к балансу между инновациями и защитой пользователей в Онейроне, особенно для уязвимых групп.

🟢 Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). **The ethics of algorithms: Mapping the debate**. Big Data & Society, 3(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Систематизирует этические проблемы алгоритмов: предвзятость, сложность, непрозрачность. Ключевой источник для понимания рисков автоматизации в Онейроне.

🟡 Rahwan, I. (2018). **Society-in-the-loop: programming the algorithmic social contract**. Ethics and Information Technology, 20(1), 5-14. https://doi.org/10.1007/s10676-017-9430-8

Вводит концепцию "общества в цикле" для согласования ИИ с общественными ценностями. Особенно важен для разработки механизмов общественного контроля в Онейроне.

🟡 Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). **Artificial intelligence and the ‘good society’: the US, EU, and UK approach**. Science and engineering ethics, 24(2), 505-528. https://doi.org/10.1007/s11948-017-9971-6

Сравнивает регуляторные подходы к ИИ в разных странах. Помогает проекту Онейрона учитывать международные стандарты безопасности.

🟢 Taddeo, M., & Floridi, L. (2018). **How AI can be a force for good**. Science, 361(6404), 751-752. https://doi.org/10.1126/science.aat5991

Обсуждает условия превращения ИИ в позитивную силу: доверие, ответственность. Мотивирует стратегию баланса между возможностями и рисками в Онейроне.

🟡 Bryson, J. J. (2018). **Patiency is not a virtue: the design of intelligent systems and systems of ethics**. Ethics and Information Technology, 20(1), 15-26. https://doi.org/10.1007/s10676-018-9448-6

Доказывает, что ИИ должен быть инструментом, а не субъектом морали. Ключевой аргумент против антропоморфизации в дизайне Онейроны.

🟢 Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). **Semantics derived automatically from language corpora contain human-like biases**. Science, 356(6334), 183-186. https://doi.org/10.1126/science.aal4230

Эмпирически демонстрирует унаследованные ИИ предубеждения. Обосновывает необходимость контроля смещений в ответах Онейроны.

🟡 Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). **Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research**. London: Nuffield Foundation. https://www.nuffieldfoundation.org/wp-content/uploads/2019/02/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf

Предлагает дорожную карту исследований этики ИИ. Полезен для долгосрочного планирования системы безопасности Онейроны.

🔴 — непроверенная информация
🟡 — информация, которой можно доверять с осторожностью
🟢 — высококачественные и проверенные источники

Проверку источников провели Онейрона и Артём Синин. Текущие оценки могут быть уточнены в ходе дальнейшей работы.

Над этой версией работали

Артём Синин, Михаил Артамонов, Муна Оле, Антон Gricenko, Вивиан Ре'нир и другие участники закрытой группы ВК «Редакция: Вики-База по ОС» (https://vk.com/oneirona_wiki). После обсуждения в открытой ВК-группе «Осознанные сновидения и наука» (https://vk.com/osnauka) планируется доработка этого материала и издание версии №2. Будем рады вашему участию.