Перова Нина Вадимовна

сектор гуманитарных экспертиз и биоэтики Института философии РАН

Искусственный моральный помощник: возможности и ограничения в контексте биотехнологического морального улучшения

Тезисы доклада

Биотехнологическое моральное улучшение — это совокупность реальных и потенциальных технологий, направленных на изменение биологической природы человека с целью сделать его морально лучше. Сюда входят генная инженерия, нейротехнологии, фармакология и иные вмешательства. Идея опирается на стремление подготовить человека к вызовам будущего, связанным с распространением киберфизических систем, и выражает принципы трансгуманизма. Сторонники биотехнологического морального улучшения настаивают на необходимости преодоления человеком природных моральных улучшений. Такие сторонники морального биоулучшения как Джулиан Савулеску и Ингмар Перссон считают, что для предотвращения катастроф глобального масштаба необходимо изменить моральную природу человека, пусть даже ценой ограничения некоторых фундаментальных прав и свобод. Их подход направлен на предотвращение «Наивысшего Вреда» — состояния, в котором достойная жизнь на планете становится невозможной. Теории биотехнологического морального улучшения ставят множество этических вопросов, в том числе в контексте фундаментальных этических категорий, таких как свобода, ответственность и доверие.

В ответ на существующую критику, сторонники биоулучшения в качестве менее инвазивной альтернативы предлагают использование искусственного интеллекта, в том числе через прямое взаимодействие с человеком — имплантацию, нейроинтерфейсы, симбиотические связи. Разрабатываются проекты морального ИИ, способного, как утверждается, компенсировать моральные слабости человека, не разрушая нравственный плюрализм. Савулеску и Маслен предполагают, что искусственный интеллект может действовать в рамках фундаментальных моральных ценностей, выступая в роли морального наставника. Проекты различаются по степени вмешательства, что выражается в дихотомии между сильным и слабым ИИ. Утверждается, что ИИ может выступать не просто инструментом, а посредником между моральными принципами и действиями

человека. Предполагается его глубокая интеграция в человеческую деятельность, вплоть до вживления в тело. Сторонники такой интеграции — Савулеску, Маслен, Лара и Декерс — считают, что моральный ИИ поможет устранить или нивелировать слабости человеческой морали.

Концепция морального искусственного интеллекта предполагает, что такие системы смогут не просто анализировать данные, но и формировать морально обоснованные рекомендации, взаимодействуя с людьми на постоянной основе. Разделение на сильные и слабые системы искусственного интеллекта предопределяет различие в подходах к построению этих систем.

Сильный ИИ в проектах рассматривается как аналог человеческого разума, способный к автономному моральному суждению. Моральный искусственный интеллект на основе сильного ИИ предполагается как обладающий самостоятельным целеполаганием и моральной агентностью. Теоретически он может выполнять функцию «моральных цензоров», воздействуя на нейропсихические процессы человека для достижения моральных целей. В этих проектах моральность ИИ постулируется как нечто совершенное и универсальное.

Однако на практике создание таких агентов невозможно. Современные системы искусственного интеллекта не обладают сознанием, саморефлексией, способностью к этическому выравниванию. Они не могут взаимодействовать с моральным контекстом так же, как человек. Даже потенциального наличия морального мышления (способности к принятию моральных решений в автономном режиме) недостаточно для признания системы полноценным моральным агентом.

Сильный ИИ не существует в полном смысле: нет автономии, самосознания и ответственности. Более того, даже гипотетическое создание такого ИИ не решает этических проблем. Напротив, оно порождает новые риски: необходимость ограничения свободы, создание тотального контроля, искажение самого понятия моральной автономии. Если ИИ будет наделен принудительной силой, он будет ограничивать моральный выбор, что разрушает идею этического субъекта. Даже Савулеску и Перссон признают, что подобный подход требует отказа от части свобод. Вопрос допустимости такого отказа остается открытым.

Слабый ИИ воспринимается как более реалистичная и приемлемая альтернатива. Предполагается, что он будет использоваться в роли вспомогательной системы, не обладающей самостоятельностью, но способной поддерживать моральное поведение через рекомендации. Слабый ИИ лежит в основе концепции искусственного морального помощника (ИМП). Сторонники описывают его работу следующим образом: ИМП не

принимает решения за человека и не вмешивается в нейропсихическую сферу. Он функционирует как система поддержки моральных рассуждений. В отличие от сильного ИИ, он не претендует на агентность. Его функция — анализ морального контекста, личных ценностей агента и выдача этически релевантных рекомендаций.

Проекты Маслен, Лары, Декерса и Савулеску подразумевают, что ИМП будет адаптироваться к пользователю, учитывая не только этические системы, но и личные убеждения, историю морального поведения и культурные установки. Это делает ИМП инструментом персонализированной моральной поддержки. ИМП предполагается как форма вспомогательного морального улучшения, что, как утверждается, позволит реализовать проекты улучшения, в то же время сохраняя свободу человека. В то время как ИМП будет выдавать рекомендации, выбор действия и его совершения будет оставаться во власти самого человека. Более того, пользователь может настраивать систему под свои ценности. Благодаря этому ИМП не подменяет человека, а дополняет его, помогая формировать и укреплять собственную моральную позицию. Однако, остается открытым вопрос, насколько реализуем такой проект в этическом смысле.

Одной из ключевых проблем ИМП является сомнительная возможность его моральной экспертности. В отличие от человека, который должен доказывать свою компетентность как морального эксперта, ИМП предполагается сразу в этой роли. Его утверждаемая моральная компетентность строится на рациональности, объеме знаний и эмоциональной беспристрастности. Однако сама система создается людьми и обучается на данных, собранных и структурированных также людьми. Следовательно, моральное несовершенство и предвзятость человека передаются ИМП уже на этапе проектирования и обучения. Это делает невозможным создание системы, обладающей моральной «превосходящей» компетентностью.

ИМП не является независимой сущностью: он не может выйти за пределы той информации и логики, которая заложена его создателями. Он не способен к самокритике или моральной рефлексии — его знания ограничены тем, что уже существует в культурных и этических паттернах общества. Возникает патернализм в отношении системы: человек программирует того, кто затем должен быть его наставником, что в корне противоречит логике морального превосходства ИМП.

ИМП также не может быть свободен от более классических этических проблем искусственного интеллекта. Например, ИИ, включая ИМП, будет на больших массивах текстов, собранных в интернете. Эти тексты уже содержат устойчивые социальные предвзятости, дискриминационные установки, стереотипы. При отсутствии механизмов критического осмысления и фильтрации этих данных, ИМП начнет транслировать и даже

усиливать эти предвзятости в процессе взаимодействия. При этом система не способна отличать этические нормы от социальных фактов, и будет принимать последние за основу для суждений о должном. ИМП может порождать так называемые «галлюцинации» — ложные или бессмысленные ответы, кажущиеся логичными, но не соответствующие действительности. Это происходит из-за того, что система генерирует ответы на основе вероятностных моделей и паттернов, которые она находит в массиве обучающих данных. Без понимания смысла или критической оценки, система может случайно создать этически проблемный или просто ложный ответ, выглядящий достоверно. Это создает риски для человека, полагающегося на такие советы. Еще одна проблема — рост и накопление ошибок. Так как ИМП может обучаться не только на человеческом контенте, но и на текстах, созданных другими ИИ, возникает эффект «каскада галлюцинаций» — одни ошибки порождают другие. В итоге растет объем недостоверной информации, которая может быть воспринята как нормативная.

Если пользователь сам обладает предвзятыми или агрессивными установками, ИМП, построенный на персонализации, может подстраиваться под эти установки, усиливая их. Иными словами, система может принять аморальное за моральное, если оно регулярно воспроизводится данным пользователем. Это разрушает первоначальную цель — развитие моральности — и превращает ИМП в усилитель уже существующих искаженных установок.

Возникает опасность, что человек, доверяя ИМП как моральному эксперту, начнет перекладывать на него ответственность за принятие решений. В ситуациях моральной сложности человек будет склонен не размышлять самостоятельно, а слепо следовать рекомендациям ИМП. Это подрывает моральное развитие личности, исключает возможность нравственного выбора, ставящего под сомнение личные убеждения, и лишает человека ситуации внутреннего конфликта — необходимого для роста. В итоге будет формироваться зависимость от ИМП как «высшей силы», что сделает человека менее способным к этической автономии.

Обучение ИМП будет построено на постоянном взаимодействии с пользователем. Со временем, по мере адаптации системы к конкретному агенту, ее рекомендации могут становиться менее универсальными и менее обоснованными с точки зрения общих этических принципов. Особенно в случаях, когда пользователь морально неразвит или недоброжелателен, ИМП будет «подстраиваться» и тем самым терять свою потенциальную роль как источник развития.

ИИ, включая ИМП, обучается на эмпирических данных. Он воспринимает то, что есть, как основание для того, что должно быть. Это нарушает принципиальное различие

между фактами и нормами, то ест обращает нас к проблеме сущего-должного. В отличие от человека, способного выявить зло, даже если оно массово распространено, ИМП не может критически отнестись к тому, на чем он обучен. Это может привести к закреплению аморальных норм в интерфейсе морального помощника.

Внедрение моральных ИИ в рамках биотехнологического улучшения не решает существующих этических проблем и порождает новые риски. Сильные ИИ требуют отказа от свободы и сталкиваются с трудностями автономности и этического выравнивания. Слабые ИИ в виде ИМП, несмотря на отсутствие принуждения, также проблематичны: они зависят от человека, подвержены предвзятостям и ошибкам, а возникающее «сверхдоверие» к ним препятствует моральному развитию человека и ставит под угрозу его моральную субъектность.