ПРОБЛЕМА МОРАЛЬНОЙ ОТВЕТСТВЕННОСТИ АВТОНОМНЫХ РОБОТОТЕХНИЧЕСКИХ СИСТЕМ

Доманькова В.В., Лащенко Н.А., студенты

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Ратникова И.М. – канд. фил. наук, доцент

В работе анализируется проблема моральной ответственности автономных робототехнических систем. На основе классических критериев морального агентства (Кант, Стросон, Сёрл) обосновывается невозможность наделения роботов подлинной моральной ответственностью. Рассматриваются аргументы сторонников функционального подхода (Флориди, Матиас), указывающие на феномен «разрыва ответственности» в условиях эмерджентного поведения нейросетевых алгоритмов. В качестве решения предлагается модель распределённой ответственности между разработчиками, производителями, пользователями и регуляторами.

Стремительное развитие технологий искусственного интеллекта и робототехники ставит перед современной философией ряд фундаментальных вопросов, среди которых основное место занимает проблема возможности наделения искусственных агентов моральной ответственностью за их действия. Если ранее технические устройства рассматривались исключительно как инструменты в руках человека, то появление автономных систем, способных принимать решения без непосредственного вмешательства существенно размывает традиционную границу между субъектом и объектом действия, вынуждая нас пересматривать устоявшиеся этические категории. В данной работе предпринимается попытка философского анализа условий моральной ответственности и их применимости к робототехническим системам.

В классической этической традиции, восходящей к Аристотелю и получившей развитие в работах Канта, моральная ответственность неразрывно связана с понятием морального сознания, предполагающим наличие у субъекта ряда специфических характеристик. Чтобы индивид мог нести моральную ответственность, ему необходимы: способность быть причиной события, свобода воли, то есть реальная возможность выбора между альтернативными вариантами действий, и понимание последствий своих поступков и наличие осознанных намерений. Кроме того, он должен быть восприимчив к моральным нормам, то есть уметь их понимать и следовать им. Ответственность подразумевает подверженность санкциям: возможность наказания или поощрения имеет смысл лишь в том случае, если субъект способен переживать последствия своих действий – испытывать раскаяние, страдание или удовлетворение. Как подчёркивал Иммануил Кант в «Основах метафизики нравственности», «воля есть способность действовать согласно представлению о законах» [2, с. 172], и именно эта способность ставит разумное существо в особое положение в мире, делая его ответственным за свои поступки. Применительно к современным роботам и системам искусственного интеллекта подавляющее большинство этих критериев остаётся невыполненным: роботы представляют собой детерминированные вычислительные системы, лишённые феноменального сознания, не обладающие подлинными намерениями и неспособные переживать моральные эмоции, что делает проблематичным рассмотрение их в качестве полноценных моральных агентов.

В современной философии техники преобладает точка зрения, согласно которой роботы не могут нести моральную ответственность в строгом смысле этого понятия, и аргументация этой позиции опирается на не

сколько ключевых соображений. Во-первых, роботам недостаёт подлинной интенциональности: как показал Джон Сёрл в своём знаменитом аргументе «Китайская комната», оперирование символами согласно формальным правилам не эквивалентно пониманию их значения. «Никакая чисто синтаксическая операция сама по себе не гарантирует семантического содержания, — пишет Сёрл в работе «Сознание, мозг и наука», – и именно поэтому компьютер, каким бы сложным он ни был, не может обладать подлинной интенциональностью» [5, с. 192].

Во-вторых, моральная ответственность традиционно связана с возможностью возмездия и наказания виновного, однако наказание робота лишено смысла, поскольку машина не способна страдать, и это делает саму концепцию вины применительно к искусственному агенту ошибочной. Как отмечал П. Ф. Стросон в своей программной статье «Свобода и негодование», «человеческие чувства негодования, обиды и благодарности являются центральными для практики моральной ответственности, и эти чувства уместны лишь по отношению к существам, способным участвовать в межличностных реактивных установках». Поскольку робот не способен ни испытывать негодование, ни становиться объектом подобных чувств в подлинном смысле, его включение в сферу моральной ответственности оказывается неправомерным.

В-третьих, робот, будучи творением человека, всегда будет лишь средством, а не самостоятельным субъектом. Его действия – это продолжение воли и решений его создателей и пользователей. Перекладывание вины на машину – это опасная тенденция, которая позволяет людям

дистанцироваться от своих этических обязательств. Мы можем назвать это «моральной буферизацией», когда технология становится удобным оправданием для избегания личной ответственности.

Тем не менее, ряд современных исследователей, среди которых Л. Флориди, А. Матиас и Дж. Данэхер, предлагают пересмотреть традиционные критерии морального сознания в контексте развития автономных технологий, выдвигая аргументы в пользу функционального подхода к моральной ответственности роботов. Сторонники этой позиции указывают на то, что современные нейросетевые алгоритмы часто функционируют как «чёрные ящики», поведение которых возникает в процессе машинного обучения и не может быть полностью предсказано или детерминировано разработчиками, что порождает феномен «разрыва ответственности» – ситуации, когда вред причинён автономной системой, но ни один из человеческих участников процесса не может быть признан виновным в силу отсутствия прямого контроля или предвидения. Как констатирует Андреас Матиас в статье «Задача ответственности для автономных систем», «чем более автономной становится система, тем сложнее приписать её действия какому-либо конкретному человеку, и в предельном случае мы сталкиваемся с ситуацией, когда моральная ответственность буквально растворяется в технологической сложности» [4, с. 179]. В таких условиях автономность робота становится функциональной реальностью, требующей наделения его моральным статусом не в онтологическом, а в прагматическом и юридическом смысле. Лучано Флориди, предлагая концепцию «информационной этики», вводит понятие «распределённой моральной ответственности», утверждая, что «в эпоху гиперсвязанных систем моральные агенты должны пониматься не как изолированные субъекты, а как узлы в сетях взаимодействий, где ответственность распределяется между людьми и технологическими артефактами» [1, с. 214]. Такой подход не утверждает, что роботы обладают сознанием или свободой воли, но предполагает, что для целей регулирования, определение ответственности и защиты прав пострадавших будет целесообразно рассматривать определённые классы автономных систем как ограниченных моральных агентов, способных нести функциональную ответственность в рамках специально разработанной нормативной базы.

Однако даже при условии принятия функционального подхода, остаётся открытым вопрос о том, как именно должна быть организована система ответственности в отношении робототехнических систем. Наиболее перспективной является модель распределённой ответственности, согласно которой моральная и юридическая ответственность за действия автономного робота распределяется между различными участниками технологического цикла: разработчиками, производителями, пользователями и регуляторами. Это позволяет избежать несправедливого возложения вины на неодушевленный объект, неспособный к моральному осмыслению, и одновременно не освобождает от ответственности людей, чьи решения и действия определяют поведение робота.

На текущем этапе технологического развития роботы и системы искусственного интеллекта не обладают статусом полноценных моральных субъектов. Это обусловлено их неспособностью к осознанию, свободе воли и переживанию моральных последствий. Быстрое развитие автономных технологий порождает серьезные этические и правовые проблемы, требующие разработки новых подходов и правил. Функциональный подход к моральной ответственности, сочетающийся с моделью распределённой ответственности между человеческими участниками технологического процесса, представляется наиболее сбалансированным решением, позволяющим обеспечить защиту прав пострадавших, стимулировать ответственное развитие технологий и сохранить моральную значимость человеческого выбора в эпоху искусственного интеллекта. Как справедливо замечает Ханна Арендт в работе «О насилии», «ответственность – это плата, которую человеческая свобода платит за свою способность начинать нечто новое в мире» [3, с. 87].

Список использованных источников:

1. Floridi L. The Ethics of Information / Luciano Floridi. — Oxford : Oxford University Press, 2013. — 357 p.

2. Кант И. Основы метафизики нравственности / Иммануил Кант ; пер. с нем. Л. Д. Брагиной // Кант И. Сочинения : в 8 т. — Москва : Чоро, 1994. — Т. 4. — С. 153–246.

3. Арендт Х. О насилии / Ханна Арендт ; пер. с англ. Г. М. Дашевского. — Москва : Новое издательство, 2014. — 148 с.

4. Matthias A. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata / Andreas Matthias // Ethics and Information Technology. — 2004. — Vol. 6, no. 3. — P. 175–183.

5. Сёрл Дж. Сознание, мозг и наука / Джон Сёрл ; пер. с англ. А. Ф. Грязнова // Путь. — 1993. — № 4. — С. 185–206.