ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ КАК ДРАЙВЕР СОЦИАЛЬНЫХ ТРАНСФОРМАЦИЙ

Юрий Михайлович Батурин

Герой России, член-корреспондент Российской академии наук,
профессор, главный научный сотрудник Института истории
естествознания и техники им. С.И. Вавилова
Российской академии наук,
Москва, Россия;
e-mail: yubat@mail.ru

О построении стандартов искусственной этики роботов

УДК: 004.896
DOI: 10.24412/2079-0910-2024-4-7-22

В статье ставится задача построения правил, согласно которым осуществляются действия робота, наделенного искусственным интеллектом, и задаются ограничения на них. Обсуждаются этические проблемы отношений человека и искусственного интеллекта. Пара «человек–искусственный интеллект» рассматривается с позиции теории очень больших (сложных) систем. Предлагается строить стандарты систем искусственного интеллекта для их взаимодействия с человеком и основывать их на ролевых обязанностях. При этом в центре внимания оказываются не действия искусственного интеллекта, а его роль в этих отношениях. Приводится рекомендуемая структура стандарта взаимодействия человека и искусственного интеллекта. Указывается, что на границе двух сред — искусственного и естественного интеллектов — происходит потеря и искажение информации: различие в понимании и смещение смыслов при переводе с естественного языка на машинный, что чувствительно для стандартизированных действий.

***Ключевые слова*:** искусственный интеллект, робот, этика, стандарт, ролевая обязанность, взаимодействие.

Будем исходить из определения, которое дается в Национальной стратегии развития искусственного интеллекта на период до 2030 г., с изменениями, утвержденными Указом Президента РФ № 124 от 15 февраля 2024 г.: «Искусственный интеллект — комплекс технологических решений, позволяющий имитировать когнитив-

ные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их» (п. 5а) [1].

Концепция развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 г., утвержденная распоряжением Правительства РФ от 19 августа 2020 г. № 2129-р[2], совершенно справедливо не разделяет искусственный интеллект (ИИ) и роботов. Дадим операциональное (рабочее) определение: робот — функциональная конструкция, автономный агент, снабженный инструментальными средствами, управляемыми искусственным интеллектом данного агента [*Батурин, Полубинская*, 2022]. Подчеркнем, что в этом определении искусственный интеллект является управляющей частью робота, а инструментальные средства лишь обеспечивают выполнение заданных функций. Будем также называть его ИИ-робот. Когда говорят об этике искусственного интеллекта (робота), имеют в виду один или несколько типов этих правил:

— правила проектантов, конструкторов, программистов ИИ (прототипы и единичные образцы);
— правила производителей ИИ (выпуск серии);
— правила поставщиков и продавцов ИИ;
— правила собственников и пользователей ИИ;
— правила, заложенные в ИИ.

Первые четыре группы правил — правовые нормы (законодательство) и нормы этики. Пятая группа правил — ни то и ни другое, и именно она станет предметом нашего рассмотрения. Концепция Правительства РФ устанавливает важное требование: «Развитие технологий искусственного интеллекта и робототехники должно основываться на базовых этических нормах» (п. I-3). Возникает вопрос: какие нормы имеются в виду — регулирующие поведение человека (первые четыре группы) или поведение искусственного интеллекта, заложенные непосредственно в него (пятая группа)? Системы искусственного интеллекта, согласно концепции Правительства РФ, характеризуются «неспособностью непосредственно воспринимать этические и правовые нормы, учитывать их при осуществлении каких-либо действий» (п. I-1). Следовательно, в основу действий искусственного интеллекта должны быть заложены нормы, по природе своей напоминающие этические, но это «этика» не людей, а ИИ-робота. Европейская сеть исследований робототехники (European Robotics Research Network — EURON), организация, занимающаяся обменом знаниями об исследованиях в области робототехники и, в частности, поиском стандартной процедуры оценки этических проблем, связанных с разработками

---

[1] Национальная стратегия развития искусственного интеллекта на период до 2030 г., с изменениями. Утверждена Указом Президента РФ № 124 от 15 февраля 2024 г. Режим доступа: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=http://www.kremlin.ru/acts/bank/50326&ved=2ahUKEwi04u_ln9SJAxUiFBAIHcUlKZMQFnoECBMQAQ&usg=AOvVaw2-gR-OC3OrkSreGGTo8DS8 (дата обращения: 11.11.2024).

[2] Концепция развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 г. Утверждена Распоряжением Правительства РФ от 19 августа 2020 г. № 2129-р. Режим доступа: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=http://government.ru/docs/all/129505/&ved=2ahUKEwj6jYuyn9SJAxUOKhAIHb8mMTwQFnoECBAQAQ&usg=AOvVaw3gF16XUYygp6xYDlilL4rn (дата обращения: 11.11.2024).

в области робототехники, в 2006 г. опубликовала «Дорожную карту роботоэтики», в которой ввела понятие «искусственная этика» [*EURON*, 2006, p. 7], которое мы будем использовать далее.

Что такое искусственная этика

Чаще всего, говоря об «этике ИИ» или о «законах робототехники», не проводят различия между нормами закона, нормами этики и техническими нормами (алгоритмами), подразумевая некоторые абстрактные правила.

Что такое «этика ИИ», роботоэтика или, как мы условились, искусственная этика?

Искусственная этика — набор внутренних (заложенных проектантами) правил (ограничений), за нарушение которых ИИ-робот ответствен перед собой. Что означает «перед собой»? Смысл простой: нарушение внутренних (заложенных) правил влечет внутреннюю же реакцию робота на нарушение. Следовательно, этическое поведение (действия) ИИ-робота — режим обратной связи в его управляющем звене — искусственном интеллекте. Таким образом, этика робота (искусственного интеллекта) — это система заложенных в контур управления ИИ-робота обратных связей (положительных и отрицательных). Внутренние, заложенные в искусственный интеллект, правила суть нормативные ограничения действий робота, нарушение которых влечет его самостоятельную реакцию. Но существуют и нормативные ограничения (правила), за нарушение которых искусственный интеллект «ответствен» перед внешним миром (миром людей). И то, и другое — нормативные ограничения поведения ИИ-робота. Выход за эти нормативные ограничения означает, соответственно, «неэтичность» и неправомерность поведения робота — внутренний нормативный сбой ИИ. Понятие «нормативный сбой» введено в работе: [*Держивицкий и др.*, 2016, с. 34, 44] для обозначения противоречия между правом и моралью, несоответствия правового порядка изменившимся реалиям, но, как представляется, его можно обобщить на случай нарушения искусственной этики.

Таким образом, искусственная этика — установленные (имеющие управленческий генезис) обыкновения (*usage*) искусственного интеллекта, регуляторный образ (модель, гомоморфное отображение) закладываемого в искусственный интеллект поведения робота, или просто — обыкновения искусственного интеллекта [*Батурин, Полубинская*, 2022, с. 148].

Проблемы искусственной этики возникают из-за способности ИИ генерировать реалистичный текстовый и визуальный контент, такой как медицинские отчеты и изображения. В 2024 г. *Google* представила свои смартфоны *Pixel 9*; *Apple* выпустила свой *iPhone 16*, а *Samsung* — телефон *Galaxy S24*. Одной из их ключевых особенностей является способность преобразовывать реальность, точнее, наше восприятие ее. Новые функции камер позволяют пользователям редактировать фото- и видеоизображения, что прежде было доступно только профессионалам. Технические возможности впечатляют, но и вызывают серьезные этические проблемы[3]. Отсутствуют решения этических проблем, возникающих при использовании генеративного ИИ (GenAI), и, тем более, проблем, вытекающих из методов *GenAI* за пределами

[3] AI is working its way into smartphones, but some tools could be subverted for misinformation. Available at: https://theconversation.com/ai-is-working-its-way-into-smartphones-but-some-tools-could-be-subverted-for-misinformation-239063 (date accessed: 14.10.2024).

больших языковых моделей (Large language model — LLM), таких как *ChatGPT*. Примером могут служить генеративно-состязательные сети (Generative Adversarial Network — GAN), которые используются для генерации данных медицинских исследований, таких как медицинские изображения. Из-за своей сложности и расширенных возможностей мультимодальные методы *GenAI* могут потенциально вызвать еще больше этических проблем [*Yilin Ning et al.*, 2024]. Нормативных актов, ГОСТов и руководств недостаточно, поскольку интерпретация этических норм сложна из-за многообразного понимания ключевой терминологии норм и неоднозначности понятия «искусственная этика», используемого в дорожной карте EURON. Вместе с тем связь между искусственной и человеческой этикой довольно тесна, поскольку обе этики пока создаются людьми. И именно этот факт облегчит нам путь к созданию стандартизированной этической структуры ИИ, острую необходимость которой сегодня ощутили все пользователи систем с искусственным интеллектом.

Нереализуемость «законов» робототехники

В 1942 г. американский писатель-фантаст Айзек Азимов в рассказе «Хоровод» ввел три закона робототехники, которые потом широко использовал в других своих рассказах, объединив их в книге «Я — робот» (1950) (см. перевод: [*Азимов*, 1979]); затем эти «законы» пошли гулять по фантастической литературе, а ныне мы наблюдаем небывалый всплеск интереса к ним у юристов, которые весьма некритично подошли к литературной модели, взяв ее за образец для построения правил для искусственного интеллекта. На них нередко ссылаются и ученые, разрабатывающие иные проблемы ИИ. Более того, даже составители важного государственного документа — Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 г., утвержденной распоряжением Правительства РФ, — не смогли выбраться из той же колеи: в ней перечисляются нормы, очень похожие на знаменитые законы робототехники Азимова, хотя в целом это несколько другой комплекс базовых принципов разработки искусственного интеллекта. Сравним их бегло.

Первый закон. Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред. Концепция Правительства РФ налагает запрет на причинение вреда человеку по инициативе систем ИИ и робототехники (п. I-3). Видно, что эта норма существенно отличается от первого закона. В самом деле, робот не может причинить вред человеку по собственной инициативе, но может по приказу другого человека, а также может спокойно наблюдать, как человек подвергается опасности.

Второй закон. Робот должен подчиняться всем приказам, которые дает человек, кроме тех случаев, когда эти приказы противоречат первому закону. В концепции Правительства РФ аналог второго закона сформулирован следующим образом: «Подконтрольность человеку (в той мере, в которой это возможно с учетом требуемой степени автономности систем искусственного интеллекта и робототехники и иных обстоятельств)» (п. I-3).

Третий закон. Робот должен защищать свою безопасность в той мере, в которой такая защита не противоречит первому или второму законам (в концепции Правительства РФ аналога нет).

Можно показать, что на метауровне законы робототехники Азимова противоречивы, поскольку основываются на разных ценностных системах. Точно так же входят в противоречие с собой ввиду несопрягаемости ценностных правил и базовые принципы, изложенные в концепции Правительства РФ. Невозможно получить работающий искусственный интеллект, какую бы искусственную этику мы ни закладывали в него, если предварительно не произведен ценностный, связанный с отношением к добру и злу (т. е. к основам этики), метавыбор между доминирующей этикой предполагаемой рабочей среды разрабатываемого ИИ-робота и альтернативными системами [*Батурин*, 2022, с. 267–268, 273–274].

Надо сказать, что и сам Азимов прекрасно понимал противоречивость введенной им системы норм. В авторском предисловии к «Остальным роботам» (1964) Азимов писал: «В “Трех законах” было достаточно двусмысленности, чтобы обеспечить конфликты и неопределенности, необходимые для новых историй, и, к моему великому облегчению, всегда можно было придумать новый ракурс из 61 слова “Трех законов”» [*Azimov*, 1964]. Хотя эти двусмысленности создают замечательные художественные сюжеты с примерами головоломок для ИИ и людей, они оказываются существенными препятствиями на пути к установлению практически полезных стандартов искусственной этики для ИИ-роботов.

Наверное, будет лишним здесь объяснять, что «законы Азимова» — никакие не законы в юридическом смысле слова, и даже не нормы этики, а всего лишь технические правила — нормативные ограничения, искусственная этика. Тем не менее, полезно будет обсудить некоторые их особенности, учитывая многочисленные попытки использовать «Три закона» робототехники Азимова в качестве руководства для создания кодекса искусственной этики ИИ-роботов.

Первые два «закона» ориентированы на человека, это механизм, обеспечивающий человеческое превосходство, в то время, как третий закон смещен в сторону ИИ-робота, которому предписана обязательная самозащита. Но можно представить ситуацию, когда робот становится опасным для человека, и нужна ли для этого программа самоуничтожения, или, скажем мягче, самоотключения? Самозащита робота или защита им человека могут включать ситуации, в которых робот не согласен с человеком. Например, описанная Владимиром Высоцким в «Песне самолета-истребителя»:

Что делает он, ведь сейчас будет взрыв!..
Но мне не гореть на песке, —
Запреты и скорости все перекрыв,
Я выхожу из пике.

Если самолет врезается в землю, автопилот должен отказаться от приказа летчика и попытать уклониться от столкновения с землей. Благополучие человека не всегда имеет приоритет над выполнением задачи. Поэтому третий закон, гласящий, что робот должен защищать себя, не имеет смысла.

Азимов, формулируя свои «Три закона», подразумевал возможность перехода робота в полностью автономный режим, в котором, действуя независимо от любых команд человека, он способен принимать собственные решения, основанные на его «знаниях» и «рациональности». Вместе с тем, в сущности, «Три закона» ограничивают не столько роботов, сколько их разработчиков.

Остается множество вопросов по практическому обеспечению вводимых нормативных ограничений. Если использовать законы Азимова как руководство к действию, то как мы обеспечим соблюдение нормы о том, что роботы не должны допускать причинения вреда человеку своим бездействием (Первый закон)? Как гарантировать выполнение Второго закона, учитывая далекие от совершенства результаты по обучению компьютеров глубоко и надежно понимать сложный естественный язык и значения жестов, сопровождающие содержательные разговоры? А как быть в ситуации, когда ИИ-робот выполняет неверную команду человека?

Если двигаться по пути коррекции и развития «трех законов», то разработчики довольно скоро окажутся в бифуркации, где линия проектирования разделяется на две: создание рассуждающего ИИ или оснащение ИИ-роботов все возрастающим числом нормативных ограничений для обеспечения необходимого уровня надежности усложняющегося доверенного интеллекта. Жизнь с ее свойством создавать непредвиденные ситуации в сочетании с постоянным изменением обстоятельств и среды делает любую кодификацию норм искусственной этики принципиально неполной, хотя бы даже по причине нехватки доступных вычислительных мощностей. Наделение же ИИ-роботов способностью рассуждать (что, кстати, точно соответствует изначальной идее: *artificial intelligence* — машина, способная на логические рассуждения, а вовсе не интеллект), во-первых, требует обучения искусственного интеллекта понимать абстрактные выражения, метафоры, неопределенные слова, обычные для человеческого общения, различать многие смыслы одного слова (например, «кровопускание» — акт насилия, старинная лечебная процедура, восточная медицинская практика...), и во-вторых, означает предоставление им возможности выбирать, как реагировать на стимулы (входные сигналы). В какой-то момент такие решения потребуют искусственного этического «мышления» — способности проводить различие между правильным и неправильным, между добром и злом. Сигео Хиросэ приводит экстремальный пример: ИИ-роботы могут быть запрограммированы на совершение убийств при определенных обстоятельствах, основываясь на желаниях человеческого большинства [*Hirose*, 1989]. Такие примеры приводят к философским вопросам вроде того, как заставить роботов следовать трудновыполнимым или некорректно сформулированным правилам, которые может установить человек. На подобные вопросы может оказаться несколько возможных решений или не найтись ни одного не вызывающего убедительных возражений ответа. Поэтому построение искусственной этики посредством поиска проверенного на полноту и непротиворечивость конечного множества правил, скорее всего, потерпит неудачу, как и попытка имитировать принятие решений человеком. Люди часто принимают решения нелогично, нерационально и не приравнивая рациональность к логике. Случается, первыми срабатывают инстинкты, чувства и эмоции. Выбор человеком решения зависит от сложного сочетания результатов воспитания, уровня человеческой зрелости, внутренней системы ценностей и выработанных принципов, понимания ситуации и многих других факторов, которые трудно сформулировать и тем более свести к алгоритмам и программным кодам. Компьютеризованный робот оперирует символами (знаками), но не понимает их смысла, потому что смысл не вложен в строки символов, а считывается через человеческое восприятие. Таким образом, разрыв между архитектурой ИИ-роботов, их алгоритмическими ограничениями и потребностями человека в нужных действиях робота за пределами четко определенных и хорошо понятных сред, описанных на языке вычислительных ал-

горитмов, непреодолим с помощью правил типа «законов Азимова» и даже неограниченного их расширения.

Поиск новых подходов

Сознавая бесперспективность улучшения «законов Азимова», реализуемых алгоритмическими ограничениями поведения роботов, Робин Мерфи и Дэвид Вудс попробовали разработать правила взаимозависимой деятельности искусственного интеллекта и человека [*Murphy, Woods*, 2009, p. 14–18]. Они исходили из модели работы искусственного интеллекта и человека в одном контуре, то есть как подсистем одной системы. Мерфи и Вудс предложили вместо триады Азимова три альтернативных правила для ИИ-роботов, которые основывались на их взаимозависимости и оперировали понятиями «роль» и «обязанность».

1. Робот не может использоваться вне системы «человек–робот», отвечающей установленным правовым и профессиональным стандартам безопасности и этики.
2. Робот должен реагировать на людей так, как уместно для их ролей.
3. Робот должен обладать степенью автономности, достаточной для защиты своего функционирования, по крайней мере, пока такая защита обеспечивает не нарушающее первый и второй закон восстановление связей управления в системе «человек–робот».

Первое правило Мерфи–Вудса прямо относится к человеку и фактически содержит норму, которая требует правового закрепления. Она категорична в отношении ИИ-робота — объекта, изготовленного человеком, предназначенного для достижения установленных человеком целей и тем самым ведомого разумом человека, каким бы автономным робот ни казался внешнему наблюдателю.

Второе правило Мерфи–Вудса также обращено к человеку и требует от него вывести принципы и нормы искусственной этики для искусственного интеллекта.

Третье правило Мерфи–Вудса требует от проектировщиков ИИ-робота на случай нештатных ситуаций создать резервный режим автономности и предусмотреть в системе норм его искусственной этики порядок восстановления связей управления в системе «человек–робот».

Базовым их положением является необходимость установления стандартов, в нашей терминологии, искусственной этики ИИ-робота, работающего в системной взаимозависимости с человеком, которая отвечает их ролевым функциям. Роль всегда исполняется в отношениях с другим (другими). Следовательно, необходимо определить вид и характер отношений. При этом в центре внимания оказываются не действия ИИ-робота, а его роль в этих отношениях, причем она не должна выходить за рамки его вычислительных возможностей, о которых человек должен иметь ясное представление, чтобы избежать завышенных ожиданий. ИИ-робот выполняет (решает) задачу, поставленную человеком. Поэтому должна быть оценена постановка задачи, ее корректность, строгость, однозначность, то есть выполнение своей роли человеком, и эта его роль не может быть передана ИИ-роботу, равно как и наоборот. Хотя перераспределение ролей, точнее соответствующих задач, может происходить. Например, уточнение заданной цели по мере получения промежуточных результатов. Ролевые взаимодействия порождают доверие, но не слепое, а основанное на знаниях и мониторинге действий робота, гарантирующих его предска-

зуемость и надежность. Такова новая постановка вопроса, меняющая классическую правовую и этическую антропогенную установку на установление норм для субъектов отношений на новую — установление правил (стандартов) для отношений.

Ролевое взаимодействие человека и искусственного интеллекта

С точки зрения теории систем [*Дружинин, Конторов*, 1985] пара «человек — ИИ» рассматривается однозначно: человек отдает команду — ИИ ее выполняет (пока еще не наоборот), то есть исполняет роль инструмента, расширяющего возможности человека. И это беспокойства не вызывает. Когда системотехника вышла на уровень больших систем [*Касти*, 1982], стало понятно, что в одних отношениях сложность человека превышает сложность ИИ, но в некоторых отношениях уступает ему, то есть и в решении ряда задач. С этого времени ИИ-роботов психологически стали рассматривать как соперников и совершенно необоснованно разжигать страхи будущего порабощения искусственным интеллектом человека. В последние годы был выявлен новый тип систем, для которых свойственно коллективное (когерентное) поведение элементов, начинающих как бы «чувствовать» друг друга даже при взаимодействии на больших расстояниях. За счет этого феномена такие системы приобретают новое качество, отличающее их от как от простых, так и от больших систем: они образуют в процессе эволюции (развития) временно устойчивые структуры. Такие системы получили название «очень большие системы» (ОБС) [*Доброчеев*, 2019]. Одними из наиболее известных примеров ОБС являются сплошные среды, например, вода и воздух (так, море превышает размеры пробирки с водой, атмосфера — объем воздушного шара). В ОБС физической природы временно устойчивые структуры — это ячейки Бенара, вихри или когерентное лазерное излучение. Для биологических ОБС — организмы и сообщества организмов. Или социально-экономические ОБС масштабами от семьи до государства и объединений государств. Наконец, ментальные и знаниевые образования, интеллектуальные проекты и построения, к которым логично отнести и системы, которые мы называем искусственным интеллектом [*Батурин*, 2024].

Коль скоро это так, целесообразно выйти из колеи бесперспективной дискуссии о наделении искусственного интеллекта рядом субъективных прав и рассматривать пару «человек–искусственный интеллект» с позиции теории очень больших (сложных) систем, в которых возникает эффект когерентных действий внутри указанной пары. Профессор И.М. Рассолов предложил для искусственного интеллекта конструкцию «отношения соучастия», или «соучастные отношения» [*Рассолов*, 2021, с. 179]. Идею И.М. Рассолова понятие «отношения соучастия» («соучастные отношения» с филологической точки зрения менее удачный термин), по существу отражает верно, но, к сожалению, на него отбрасывает тень уголовно-правовое понятие соучастия, которое предполагает единство умысла соучастников. Но очевидно, что конструкция вины, разработанная наукой для человека, непригодна для искусственного устройства. В предисловии к монографии автор этих строк предложил альтернативный термин: «когерентные отношения» [*Батурин*, 2021, с. 7]; будем пользоваться им в дальнейшем, предполагая, что они осуществляются через ролевые обязанности каждого. То есть ИИ рассматривается не как носитель прав,

а как носитель ролевых обязанностей, взаимодействующий с человеком в рамках когерентных отношений, в коллективном режиме как партнер.

Идея предоставления прав ИИ-роботам появилась именно тогда, когда ИИ стал рассматриваться как партнер. Но наделение ИИ-роботов правами — не лучшая идея. Концепция прав — конкурентна, состязательна, чревата коллизиями, наводит на мысль о будущих конфликтах между людьми и искусственным интеллектом и даже внушает страх перед будущей «всемогущей» искусственной цивилизацией ИИ-роботов. Напротив, партнерство человека и ИИ-робота подразумевает командную работу, в которой и тот и другой имеют ролевые обязанности — безусловные для выполнения действия, направленные на достижение целей команды. Ролевая обязанность — обязанность выполнять свои функции. Ролевые обязанности ИИ поощряют командную работу (взаимодействие с человеком).

Ситуацию можно сравнить с игрой хорошей футбольной команды. Казалось бы, любой из игроков может перемещаться по полю, находиться в любом месте, отдавать пас и бить по мячу. Но у каждого футболиста есть своя ролевая обязанность в командной игре. У нападающих — забивать голы. У форварда — второго игрока в линии нападения — создание голевых ситуаций, помощь нападающему, взаимодействие с полузащитниками, которые тоже различаются по ролевым обязанностям: центральный полузащитник должен создавать голевые ситуации для нападающих, крайние полузащитники противостоят крайним защитникам соперника, отбирают мяч у команды соперника, выполняют передачи и навесы в штрафную площадку. Защитники отвечают за то, чтобы блокировать попытки соперника забить гол. Но и у них разделение ролевых обязанностей: крайние защитники опекают на флангах нападающих команды соперника и блокируют их один на один, поддерживают вратаря, останавливая удары до того, как они попадут в зону, близкую к воротам. Ролевая обязанность центрального защитника схожа с ролью крайних защитников в том, что они отвечают за блокирование ударов до того, как они дойдут до ворот. Он также должен перехватывать передачи или завладевать мячом. Последняя линия обороны и единственный игрок на поле, которому разрешено использовать руки, — вратарь. Его ролевая обязанность — останавливать удары соперника у ворот. Невозможно себе представить, что на разборе игры вдруг заговорили о нарушении прав, например: «Ты не имел права отбивать мяч на угловой» или «Я имел право получить пас». Но совершенно обычны такие оценки как: «Твоя обязанность была — забить штрафной» или «Если ты хотел, чтобы команда выиграла, то обязан был не сам пытаться забить гол, а отдать пас нападающему, который находился в выгодной позиции». Помимо нелепости координации игры на языке прав, это очевидно подрывает спортивный дух.

Правда, механизм взаимодействия человека и ИИ-робота значительно сложнее, чем футболистов в сыгранной команде, но особенность в том, что оба они при совместной работе нуждаются друг в друге. Во-первых, любая роль выполняется в определенном контексте, а искусственный интеллект «понимает» контекст хуже человека. Поэтому человек должен корректировать модель мира ИИ. Во-вторых, способность искусственного интеллекта распознавать аномалии невысока, поскольку он пытается делать это, опираясь на опосредованные стимулы, в то время как человек быстрее обнаруживает аномалии. В-третьих, чувствительность к изменениям у ИИ ниже, чем у человека, который, распознавая аномалии, видит и изменения. С другой стороны, ИИ-робот может работать в агрессивной среде, недоступной че-

ловеку, анализировать изображения с большей детальностью. Существуют и пары условий, в которых ИИ-робот и человек оказываются в ситуации дополнительности. Например, ИИ-робот может обнаруживать физические факторы, скрытые от восприятия человека, который лучше, чем ИИ, фиксирует социальные сигналы. ИИ быстрее вычисляет, а человек лучше обрабатывает неявные и интуитивные знания и принимает решения с помощью эвристик, освоение которых выходит за пределы вычислительных возможностей машины. Перечисленные обстоятельства способствуют установлению когерентного режима взаимодействия. Чем выше сложность и неопределенность среды, в которой приходится работать паре «человек–робот», тем выше степень когерентности взаимодействия, что позитивно сказывается на достижении общей цели.

Учитывая когерентное взаимодействие (партнерство) в ОБС «человек–ИИ», целесообразно не правами наделять роботов, а устанавливать стандарты (регламенты) взаимодействия человека с ними в соответствии с их функциональными возможностями. Как в нашем примере, помимо официальных стандартов спортивной подготовки для футбольных игроков существуют своего рода «стандарты», по которым тренеры подбирают команду (табл. 1), и «стандарты» взаимодействия (табл. 2):

Табл. 1. «Стандарты» футбольных игроков

Table 1. Football players' standards

Роль	**Ролевые обязанности**
Нападающий	Метко бить по воротам, хорошо видеть поле, быстро концентрироваться, использовать физическую силу
Форвард (второй нападающий)	Быстро менять тактику, пасовать и метко бить по воротам, использовать физическую силу
Центральный полузащитник	Быстро и много бегать, контролировать мяч, владеть дриблингом, точно пасовать, хорошо читать игру и язык тела соперников
Крайний полузащитник	Владеть дриблингом, давать сильные пасы, хорошо владеть мячом, бить по воротам
Крайний защитник	Понимать позиции, быстро и много бегать, бороться за мяч
Центральный защитник	Читать игру, предвидеть складывающиеся ситуации, читать язык тела соперников, использовать физическую силу
Вратарь	Ловить мяч, молниеносно реагировать, хорошо прыгать, выбивать мяч далеко в поле

Табл. 2. «Стандарты» взаимодействия игроков

Table 2. Standards of players' interaction

Команда владеет мячом	**Команда потеряла мяч**
Построение атаки от своих ворот	Контроль за построением игры соперника
Контроль мяча и продвижение	Пресечение возможностей соперника диктовать ритм игры
Проникающие взаимодействия и передачи	Блокирование голевых ситуаций и ударов по своим воротам
Быстрое завершение и прострелы	Воспрепятствование проникновению соперника и его передач в штрафную площадку своей команды
Создание голевых ситуаций	Защита своих ворот от голевых ситуаций

Возвратимся к идее «когерентных прав». Используя ее, мы уходим от привычной схемы «праву субъекта А корреспондирует обязанность субъекта В», и наоборот, и рассматриваем «права» искусственного интеллекта, которые когерентны правам взаимодействующего с ним человека и осуществляются через него. В определенном отношении регуляция когерентных взаимодействий напоминают конфуцианскую традицию в восточном праве, где ритуал *ли* (в случае искусственного интеллекта — ролевая обязанность) работает вместе с законом *фа*, причем *ли* — средство управления, *фа* — помогает управлению, *ли* и *фа* дополняют друг друга, позволяя делать упор то на *ли*, то на *фа*. *Ли* устанавливает гармонию, *фа* восстанавливает нарушенную гармонию. Восточная (конфуцианская) традиция толкует поведение так: «Я уважаю ваши действия в силу вашей роли в нашем взаимодействии» (позиция «мы», то есть коллективное «я»). Концепция взаимных обязанностей — основа конфуцианства. Безусловно, такой подход существенно отличается от западного (и российского) юридического принципа, согласно которому «я уважаю ваше право и не посягаю на него, но оно не должно вступать в противоречие с моим правом» (позиция «я»).

Таким образом, регулирование отношений по поводу пользования таким сложным объектом, как ИИ, и взаимодействия с ним не стоит искать на привычных юридических путях и как вариант — перейти к принципу установления ролевых обязанностей искусственного интеллекта вместо предоставления прав в его взаимодействии с человеком [*Tae Wan Kim, Strudler*, 2023, p. 79–80], а права рассматривать как когерентные. Регламентацию когерентных прав и ролевых обязанностей целесообразно осуществлять через разработку стандартов взаимодействия (норм «этикета») искусственного интеллекта с человеком.

Регламент взаимодействия человека и искусственного интеллекта

Нормативный регламент взаимодействия ИИ-робота с человеком — набор связанных действий, имеющих символическое значение и подтверждающих ценность (качество) происходящего взаимодействия. В качестве строительного «кирпичика» регламента логично выбрать элементарное взаимодействие (определяемое через выбранные элементарные ролевые обязанности), а не отдельную сущность — человека или ИИ либо их права. Регламент взаимодействия должен гарантировать, что любое решение или действие — результат именно взаимодействия, а не автономное решение робота.

Какой должна быть примерная структура регламента взаимодействия человека и ИИ-робота? Начнем с того, что ИИ-роботу и человеку нужно войти в режим выполнения нормативного регламента взаимодействия.

Первым шагом с обеих сторон подтверждается понимание **извещения** и согласия с содержащимися в нем значениями параметров рабочей среды, их пороговыми значениями, ограничениями компетенций партнеров, возможных пределов вмешательства, а также каталог стоп-сигналов, по которым взаимодействие немедленно приостанавливается («понимание» и «согласие» для ИИ-робота и человека означает несколько разные операции, как и другие элементы регламента ниже; термин «сигнал» включает не только вербальную, но и невербальную информацию — действия, жесты, неязыковые звуки, такие как крик боли, и т. п.).

Шаг второй — **приветствие** — состоит в обмене токенами, означающими установление взаимодействия и подтверждающими компетенции человека и ИИ-робота для участия в решении поставленной задачи.

Шаг третий — **одобрение** планируемого взаимодействия, как минимум, двустороннее, но в случае, например, медицинского ИИ-робота одобрение трех и более сторон.

Шаг четвертый — **проверка полномочий** (сертификация робота для выполнения предопределенной роли, допуск человека к работе с искусственным интеллектом).

Шаг пятый — **начало** ролевого взаимодействия.

Шаг шестой — **контроль** ролевого взаимодействия. ИИ-робот не должен предпринимать никаких действий без ссылки на ролевые обязанности.

Дальнейшие шаги будут делаться в зависимости от решаемой задачи и развития процесса ролевого взаимодействия. **Запрос** на действие, полученный от человека, будет отклонен, если не соответствует роли, и приведет к **аудиту** взаимодействий. **Запрос** должен иметь приоритет над выполнением задачи. Если **аудит** установит нарушение ролевых обязанностей, **приветствие** отменяется, взаимодействие приостанавливается до нового **приветствия** или блокируется. **Аудит** проводится не только при отклонении **запроса**, но и в регулярном режиме. **Информирование** ИИ-роботом человека должно происходить каждый раз, когда запрошенное человеком действие нарушает установленные ограничения. И наоборот: любое указание человека прекратить какое-либо действие должно выполняться роботом безоговорочно, чтобы не быть расцененным как попытка установления доминирования в паре «человек–робот», если только окружающая среда, условия и физические параметры не указывают на серьезную физическую или даже смертельную угрозу для человека в ролевых взаимодействиях. В этом случае ИИ-робот может осуществить **коррекцию** рамок, в которых человек осуществляет контроль взаимодействия с ним, но только после прохождения цепочки **извещение** — **информирование** — **аудит**. В каталог стоп-сигналов также входит разрыв ролевого взаимодействия: ИИ-робот не может работать полностью в автономном режиме. Особо следует оговорить в **извещении** ситуацию, сложную даже для искусственной этики, — при каких угрозах человеку ИИ-робот может или должен пожертвовать собой. В силу сложности возможные условия формулируются с некоторой степенью неопределенности и допускают **коррекцию** со стороны как ИИ-робота, так и человека. Таков вкратце каркас регламента, в котором акцент перенесен с недостижимых правил искусственной этики типа «законов Азимова» на качество и безопасность ролевого взаимодействия [*McBride, Hoffman*, 2016].

Мы пока оставили без рассмотрения важную проблему регламента взаимодействия — потерю информации на границе двух систем — робота и человека. Более точно, различие в понимании, потерю деталей, смещение смыслов при переводе с естественного языка на машинный. Искажение информации создает неоднородность двух сред на границе их раздела, что весьма чувствительно для таких упомянутых в регламенте действий, как информирование, коррекция, аудит и т. д. Важный вопрос: будет ли ИИ-робот понимать суть: зачем вообще нужны ролевые взаимодействия, если можно все сделать проще: поставить задачу и ожидать решения. Сегодня под искусственным интеллектом понимают нейронные сети, позволившие сделать прорыв на этом направлении. Они действительно способны распознавать глубокие и скрытые ассоциации в обучающих наборах данных, особенно в блоч-

ной модели *Deep Learning*. Но человеческий разум работает также в соответствии с абстрактными механизмами символов и правил, техническим аналогом которых является «старый добрый искусственный интеллект» (“Good Old Fashioned Artificial Intelligence” — GOFAI) — термин, введенный философом Джоном Хоугеландом [*Haugeland*, 1985]. Такой символический ИИ в 1960-х гг. смог успешно имитировать процесс рассуждений высокого уровня, включая логическую дедукцию, алгебру, геометрию, пространственные рассуждения и анализ «средства–цель», причем все это на хорошем естественном языке, который используют люди при рассуждениях. Символический ИИ необходим для того, чтобы ИИ-робот правильно понимал ролевые обязательства, представленные в виде набора правил. Таким образом, по-видимому, чтобы перейти к когерентному взаимодействию ИИ-робота и человека, потребуется создание нейросимволического искусственного интеллекта [*Tae Wan Kim, Strudler*, 2023, p. 84–85].

Источники

Концепция развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 г. Утверждена Распоряжением Правительства РФ от 19 августа 2020 г. № 2129-р. Режим доступа: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=http://government.ru/docs/all/129505/&ved=2ahUKEwj6jYuyn9SJAxUOKhAIHb8mMTwQFnoECBAQAQ&usg=AOvVaw3gF16XUYygp6xYDlilL4rn (дата обращения: 11.11.2024).

Национальная стратегия развития искусственного интеллекта на период до 2030 г., с изменениями. Утверждена Указом Президента РФ № 124 от 15 февраля 2024 г. Режим доступа: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=http://www.kremlin.ru/acts/bank/50326&ved=2ahUKEwi04u_ln9SJAxUiFBAIHcUlKZMQFnoECBMQAQ&usg=AOvVaw2-gR-OC3OrkSreGGTo8DS8 (дата обращения: 11.11.2024).

AI Is Working Its Way into Smartphones, but Some Tools Could Be Subverted for Misinformation. Available at: https://theconversation.com/ai-is-working-its-way-into-smartphones-but-some-tools-could-be-subverted-for-misinformation-239063 (date accessed: 14.10.2024).

Литература

Азимов А. Три закона робототехники. Сборник научно-фантастических рассказов / Пер. с англ.; предисл. А. Шилейко. М.: Мир, 1979. 400 с.

Батурин Ю.М. Таинственное пространство Интернет (Предисловие к третьему изданию) // Рассолов И.М. Право и Интернет. Теория кибернетического права. М.: Норма ИНФРА-М, 2021. С. 5–11.

Батурин Ю.М. Гамлет как робот // Четвертые Бачиловские чтения: материалы Международной научно-практической конференции / Отв. ред. Т.А. Полякова, А.В. Минбалеев, В.Б. Наумов. М.; Саратов: Амирит, 2022. С. 265–274.

Батурин Ю.М., Полубинская С.В. Искусственный интеллект: правовой статус или правовой режим? // Государство и право. 2022. № 10. С. 141–154.

Батурин Ю.М. Об управлении наукой // Наука и общество в условиях новых вызовов: материалы Международной научно-практической конференции (г. Минск, 26–27 сентября 2024 г.) / Гл. ред. Н.Л. Мысливец. Минск: Альфа-книга, 2024. С. 13–21.

Держивицкий Е.В., Ларионов И.Ю., Перов В.Ю. К вопросу об этике права // Вестник С.-Петерб. ун-та. Сер. 17: Философия. Конфликтология. Культурология. Религиоведение. 2016. Вып. 4. С. 33–45. DOI: 10.21638/11701/spbu17.2016.404.

Доброчеев О.В. Механика очень больших систем природы, жизни и разума. М.: ТЕИС, 2019. 144 с.

Дружинин В.В., Конторов Д.С. Системотехника. М.: Радио и связь, 1985. 200 с.

Касти Дж. Большие системы. Связность, сложность и катастрофы. М.: Мир, 1982. 216 с.

Рассолов И.М. Право и Интернет. Теория кибернетического права. М.: Норма ИНФРА-М, 2021. 304 с.

Azimov I. The Rest of Robots. Doubleday, USA, 1964. 556 p.

EURON Roboethics Roadmap / Project coordinator G. Veruggio. Genoa, 2006. 42 p.

Haugeland J. Artificial Intelligence: The Very Idea. Cambridge, Mass.: MIT Press, 1985. 288 p.

Hirose S. A Robot Dialog // Journal of Robotics Society of Japan. 1989. Vol. 7. Iss. 4. P. 121–126.

McBride N., Hoffman R. Bridging the Ethical Gap: From Human Principles to Robot Instructions // IEEE Intelligent Systems, 2016. Vol. 31. No. 5. P. 76–82. URL: https://www.academia.edu/105272367/Bridging_the_Ethical_Gap_From_Human_Principles_to_Robot_Instructions (date accessed: 20.10.2024).

Murphy R.R., Woods D.D. Beyond Asimov: The Three Laws of Responsible Robotics // IEEE Intelligent Systems. 2009. Vol. 24. No. 4. P. 14–18.

Tae Wan Kim, Strudler A. Should Robots Have Rights or Rites? // Communications of the ACM. 2023. Vol. 66. No. 6. P. 78–85.

Yilin Ning et al. Generative Artificial Intelligence and Ethical Considerations in Health Care: a Scoping Review and Ethics Checklist // The Lancet Digital Health. 2024. Vol. 6. No. 11. P. E848–e856. DOI: 10.1016/S2589-7500(24)00143-2. URL: https://pubmed.ncbi.nlm.nih.gov/39294061/ (date accessed: 14.10.2024).

Concerning the Definition of the Standards of Artificial Ethics of Robot

Yuri M. Baturin

S.I. Vavilov Institute for the History of Science and Technology
of the Russian Academy of Sciences,
Moscow, Russia;
e-mail: yubat@mail.ru

The article aims at formulating the rules, according to which the machine with the artificial intelligence functions and which impose restrictions on such a machine. The article discusses the ethical problems of the relations between a human and an artificial intelligence. The relation "human–artificial intelligence" is considered from the perspective of the theory of huge (complex) systems. It is proposed to build standards of artificial intelligence systems for their interaction with humans and to base them on role responsibilities. At the same time, the focus is not on the acts of the artificial intelligence, but on its role in these relations. The author recommends the structure of the interrelations between man and artificial intelligence. The author points out that on the border of the two spheres (natural and artificial intelligence), loss and distortion of information occur: difference in understanding and sense shift, when translating from natural language to the artificial one, something that is sensitive for the standardized systems.

***Keywords*:** artificial intelligence, machine, ethics, standard, role obligation, interaction.

References

AI Is Working Its Way into Smartphones, but Some Tools Could Be Subverted for Misinformation. Available at: https://theconversation.com/ai-is-working-its-way-into-smartphones-but-some-tools-could-be-subverted-for-misinformation-239063 (date accessed: 14.10.2024).

Azimov, I. (1964). *The Rest of Robots*, Doubleday, USA.

Azimov, I. (1979). *Tri zakona robototekhniki. Sbornik nauchno-fantasticheskikh rasskazov* [Three laws of robotics. A collection of science fiction], transl. from English, Moskva: Mir (in Russian).

Baturin, Yu.M. (2021). Tainstvennoye prostranstvo Internet (Predisloviye k tret'yemy izdaniyu) [Secret space of Internet (Preface to the third edition)], in Rassolov, I.M., *Pravo i Internet. Teoriya kiberneticheskogo prava* [Law and Internet. A theory of cybernetics' law] (pp. 5–11), Moskva: Norma INFRA-M (in Russian)

Baturin, Yu.M. (2022). Gamlet kak robot [Hamlet as a robot], in T.A. Polyakova, A.V. Minbaleev (Eds.), *Chetvertyye Bachilovskiye chteniya: materialy Mezhdunarodnoy nauchno-prakticheskoy konferentsii* [The fourth Bachilov readings: materials of International conference] (pp. 265–274), Moskva, Saratov: Amirit (in Russian).

Baturin, Yu.M., Polubinskaya, S.V. (2022). Iskusstvennyy intellect: pravovoy status ili pravovoy rezhim? [Artificial intelligence: the legal status or the legal regime?], *Gosudarstvo i pravo*, no. 10, 141–154 (in Russian).

Baturin, Yu.M. (2024). Ob upravlenii naukoy [About science management], in N.L. Myslivets (Ed.), *Nauka i obshchestvo v usloviyakh novykh vyzovov: materialy Mezhdunarodnoy nauchno-prakticheskoy konferentsii (g. Minsk, 16–17 sentyabrya 2024 g.* [Science and society in the conditions of new challenges: materials of International conference (Minsk, September 26–27, 2024)] (pp. 13–21), Minsk: Al'fa-kniga (in Russian).

Derzhitetsky, E.V., Larionov, I,Yu., Perov, V.Yu. (2016). K voprosu ob etike prava [About the law ethics], *Vestnik S.-Peterb. un-ta. Ser. 17: Filosofiya. Konfliktologiya. Kul'turologiya. Religiovedeniye*, vyp. 4, 33–45 (in Russian). DOI: 10.21638/11701/spbu17.2016.404.

Dobrocheev, O.V. (2019). *Mekhanika ochen' bol'shikh system prirody zhizni i razuma* [Mechanics of very large systems of nature, life and intellect], Moskva: TEIS (in Russian).

Druzhinin, V.V., Kontorov, D.S. (1985). *Sistemotekhnika* [System engineering], Moskva: Radio i svyaz' (in Russian).

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*, Cambridge, Mass.: MIT Press.

Hirose, S. (1989). A Robot Dialog, *Journal of Robotics Society of Japan,* 7 (4), 121–126.

Kasti, J. (1982). *Bol'shiye sistemy. Svyaznost', slozhnost' i katastrofy* [Large systems. Connectivity, difficulty and catastrophes], Moskva: Mir (in Russian).

Kontseptsiya razvitiya regulirovaniya otnosheniy v sfere tekhnologiy iskusstvennogo intellekta i robototekhniki do 2024 g. Utverzhdena Rasporyazheniyem Pravitel'stva RF ot 19 avgusta 2020 g. No. 2129-r [Conception of the development of relations regulation in sphere of artificial intelligence and robotics up to 2024. Established by the Decree of the Government of Russian Federation in August 19, 2020, No 2129-r]. Available at: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=http://government.ru/docs/all/129505/&ved=2ahUKEwj6jYuyn9SJAxUOKhAIHb8mMTwQFnoECBAQAQ&usg=AOvVaw3gF16XUYygp6xYDlilL4rn (date accessed: 11.11.2024) (in Russian).

McBride, N., Hoffman, R. (2016). Bridging the Ethical Gap: From Human Principles to Robot Instructions, *IEEE Intelligent Systems, 31* (5), 76–82. Available at: https://www.academia.edu/105272367/Bridging_the_Ethical_Gap_From_Human_Principles_to_Robot_Instructions (date accessed: 20.10.2024).

Murphy, R.R., Woods, D.D. (2009). Beyond Asimov: The Three Laws of Responsible Robotics, *IEEE Intelligent Systems, 24* (4), 14–18.

Natsional'naya strategiya razvitiya iskusstvennogo intellekta na period do 2030 g., s izmeneniyami. Unverzhdena Ukazom Prezidenta RF No. 124 ot 15 fevralya 2024 g. [National strategy of the development of artificial intelligence up to 2030, with changings. Established by the Decree of the President of Rus-

sian Federation No. 24 on February 15, 2024]. Available at: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=http://www.kremlin.ru/acts/bank/50326&ved=2ahUKEwi04u_ln9SJAxUiFBAIHcUlKZMQFnoECBMQAQ&usg=AOvVaw2-gR-OC3OrkSreGGTo8DS8 (date accessed: 11.11.2024) (in Russian).

Rassolov, I.M. (2021). *Pravo i Internet. Teoriya kiberneticheskogo prava* [Law and Internet. A theory of cybernetics' law], Moskva: Norma INFRA-M (in Russian).

Tae Wan Kim, Strudler, A. (2023). Should Robots Have Rights or Rites?, *Communications of the ACM, 66* (6), 78–85.

Verrugio, G. (Coordinator) (2006). *EURON Roboethics Roadmap*, Genoa.

Yilin Ning et al. (2024). Generative Artificial Intelligence and Ethical Considerations in Health Care: a Scoping Review and Ethics Checklist, *The Lancet Digital Health, 6* (11), E848–e856. DOI: 10.1016/S2589-7500(24)00143-2. Available at: https://pubmed.ncbi.nlm.nih.gov/39294061/ (date accessed: 14.10.2024).