В. Э. КАРПОВ, П. М. ГОТОВЦЕВ,
Г. В. РОЙЗЕНЗОН

К ВОПРОСУ ОБ ЭТИКЕ И СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Обсуждая проблемы этики в области систем искусственного интеллекта, авторы предлагают вернуться к конструктивной постановке вопроса о соответствии интеллектуальных систем (ИС) этическим нормам. В работе утверждается, что суть этичности ИС заключается в том, что, принимая критически важные для человека решения, ИС должны использовать этические императивы, рассматривая их как некие поисковые эвристики. Также в работе рассматриваются вопросы достаточности современных моделей, методов и технологий для формализации этических понятий и отмечается, что основной проблемой является процедура верификации ИС на соответствие этическим нормам. Делается вывод, что основной формой этой верификации является использование комплексных тестов Тьюринга.

Ключевые слова: *искусственный интеллект, автономные системы, этика, этический выбор, эвристики, верификация этического соответствия, онтология.*

Discussing the problems of ethics in the sphere of artificial intelligence systems, the authors propose to return to a constructive formulation of the issue of correlation between intellectual systems (IS) and ethical norms. The paper argues that the essence of the IS's ethical nature lies in the fact that when taking a critically important decision for a human, IS should use ethical imperatives treated as certain search heuristics. The paper also discusses the adequacy of modern models, methods and technologies for the ethical concepts formalization, and notes that the main problem is the procedure of verifying IS in terms of their compliance with ethical standards. It is concluded that the main form of this verification is the use of complex Turing tests.

Keywords: *artificial intelligence, autonomous systems, ethics, ethical choice, heuristics, ethical compliance verification, ontology.*

Философия и общество, № 2 2018 84–105 ***DOI: 10.30884/jfio/2018.02.07***

1. ВВЕДЕНИЕ

В последние годы активно если не развиваются, то по крайней мере широко обсуждаются темы этики в искусственном интеллекте (ИИ), угроз, происходящих от ИИ, различного рода гуманитарных аспектов создания систем ИИ. Будучи очень давней темой, появившейся задолго до того, как было сформулировано само понятие ИИ, и в большинстве сводящейся к пресловутому «бунту машин», эти вопросы сегодня получают новое звучание. Притом, несмотря на множество спекуляций, а зачастую – безграмотность рассуждений, в этом звучании появляются вполне здравые высказывания, связанные с некоторыми аспектами проектирования и применения технических систем, поведение которых является значимым для человека. В первую очередь эта тенденция связана, разумеется, с ростом числа критически важных, потенциально опасных систем, работающих автономно.

В этой работе мы рассмотрим вопросы соотношения этики и искусственного интеллекта с критических позиций. Здесь важно подчеркнуть, что наблюдается некоторая путаница в самой постановке вопроса. Чаще всего, как будет видно ниже, речь идет об этичности применения систем искусственного интеллекта в тех или иных областях. Говорят о различного рода рисках и опасностях, связанных с использованием систем ИИ, о социальных последствиях и даже о том, насколько важно, чтобы сами разработчики ИИ соответствовали этическим нормам. Мы же далее будем говорить об этических аспектах функционирования таких систем, о том, насколько их поведение может быть обусловлено этическими парадигмами, нормами, представлениями. При этом мы постараемся не только отыскать некоторое рациональное зерно в рассуждениях об этике ИИ, но и сформулировать ряд технических задач, решение которых представляется значимым для данного вопроса.

Начнем с того, что вопросы соотношения этики и искусственного интеллекта коренным образом отличаются от того, что понимается, например, под этическими проблемами генных технологий, информатики, естествознания и т. п. Это отличие определяется тем, что в искусственном интеллекте этические вопросы ближе к пониманию этики в философском или социогуманитарном смысле, и связаны эти этические аспекты прежде всего с тем, что они касаются вопросов поведения и принятия решений.

Будем полагать, что сам термин «искусственный интеллект» здесь и далее понимается в метафорическом смысле (равно как и термин «этика», под которым обычно понимается философская дисциплина, исследующая вопросы морали и нравственности, что бы под этим ни подразумевалось). Отметим, что здесь и далее авторы менее всего хотели бы вторгаться в область философии, профессионально занимающейся вопросами этики. В этой работе нам будет достаточно лишь общего, схематического представления об этике и связанных с ней аспектах, равно как и представления о том, что такое ИИ. Для более полного понимания этики можно обратиться, например, к работе Р. Г. Апресяна [2017].

Важным аспектом является рассмотрение интеллектуальной системы не только как когнитивной, но и как активной сущности. В этом плане определяющим свойством такой системы является возможность осуществления воздействия на окружающий мир и прежде всего социум. Иными словами, вопросы этики в ИИ сводятся к тому, что мы имеем дело с искусственной системой, реализующей процессы планирования, целеполагания, выбора и осуществления того или иного поведения. Для простоты будем обозначать такую систему как систему с ИИ (ИИС). При этом выбор, осуществляемый системой, должен определяться некоторыми этическими императивами и нормами в самом широком смысле. Например, этические нормы могут трактоваться как некоторые эвристики, которыми руководствуется ИИС при совершении выбора того или иного действия, формирования системы оценок, целевых функций и прочего.

С прагматической точки зрения исследования в области этики ИИС приведут в конечном итоге к созданию различного рода стандартов и последующей сертификации ИИС. И здесь возникают три важнейшие проблемы.

Первая касается конструктивной формализации этических норм в форме, пригодной для описания функционирования конкретных программно-аппаратных комплексов. Вторая проблема – это способность объективного (инструментального, прямого или косвенного, основанного на анализе поведения и т. п.) контроля соответствия компонент ИИС этическим нормам. Третья – это то, какое влияние окажут в дальнейшем эти стандарты и не будут ли

они выполнять сугубо ограничительную роль, тормозящую развитие ИИС.

Далее мы проведем небольшой экскурс в историю вопроса этических проблем искусственного интеллекта, обсудим инициативу IEEE (Институт инженеров электротехники и электроники) по этически обусловленному проектированию систем ИИ. Кроме того, мы постараемся обсудить некоторые конструктивные аспекты такого этически обусловленного проектирования, особое внимание уделив вопросам имеющегося математического аппарата и, главное, формальной постановке задачи.

2. ЭТИКА И ИИ

2.1. Опасность «думающих машин»

Этические вопросы, связанные с ИИ, вовсе не следствие развития технологий последних десятилетий. Эти вопросы поднимались во многих классических работах пионеров в области машинного интеллекта. Так, в одной из своих основополагающих для теории ИИ работ [Turing 1950] Алан Тьюринг дискутирует на тему последствий создания «думающих машин». В частности, он обсуждает постулат «машины не могут делать ошибок» и отмечает, что в сложных машинах ошибки могут быть детерминированы неадекватностью исходных данных, хотя при этом машины максимально точно выполняют все математические операции по их обработке. Те же рассуждения приводятся и при обсуждении вопросов обучения машин, а именно – проблем достоверности исходных данных. Этот пример, когда, казалось бы, совершенная интеллектуальная машина получает неадекватные исходные данные и далее совершает какие-либо неэтичные или «ужасающие» действия, со временем стал одним из наиболее активно демонстрируемых не только в профессиональных сообществах, но и в популярной культуре и кинематографе.

Норберт Винер затрагивает вопросы этики ИИ в своей статье [Wiener 1960] и в дополнительных главах второго издания своей «Кибернетики» [*Idem* 1965]. Основная мысль, которую он постулирует в этих работах, заключается в том, что машины могут быть опасны для человека и непредсказуемы. При этом он отмечает: даже понимая в деталях, как работает машина, оператор может не успеть осознать, что ее рассуждения ведут к негативному сцена-

рию, или даже не успеть понять, что машина уже «приняла решение» и работает над осуществлением этого сценария. Уже к моменту написания статьи [Wiener 1960] разница в скорости обработки информации, требующей вычислений, между человеком и современными на тот момент вычислительными машинами была значительной. Именно эта разница, помноженная на неполную конкретизацию человеком своих желаний, и может привести, по мнению автора, к негативным последствиям.

Возможные негативные сценарии не ограничиваются катастрофами, о которых часто повествуют в фантастических романах и кинофильмах. Однако многие из негативных последствий применения ИИ кажутся реальными уже сегодня. В целом к негативным сценариям относят [Bostrom, Yudkowsky 2011]:

1. Нашумевший катастрофический сценарий, в котором сверхразумные машины решают, что человечество лишнее.

2. Опять же сверхразумные машины решают: они лучше людей знают, что им нужно, и всячески ограничивают их свободу действия вплоть до вегетативного состояния с полным погружением в виртуальную реальность.

3. Несколько менее фантастический сценарий, в котором ИИ, чьей целью является решение определенной задачи, например произвести максимум скрепок, потребляет для этого все доступные ресурсы и начинает заваливать мир скрепками, попутно создавая средства для отъема у людей ресурсов.

Эти сценарии в настоящее время смотрятся фантастично, однако есть ряд обсуждаемых угроз, которые могут быть актуальны уже сегодня или завтра [см.: *Ibid.*; Havens 2016]:

- изменения в структуре рабочих мест, включая полное вытеснение человека из каких-либо областей экономической деятельности;
- доступ и обращение с персональными данными – какая-либо автономная интеллектуальная система может принимать решения с использованием персональных данных пользователя;
- анализ метаданных, проводимых с помощью интеллектуальных автономных систем, могущий выдать информацию личного характера, которую человек не хотел бы предоставлять кому бы то ни было;

- использование автономных интеллектуальных медицинских систем;
- этический выбор, который может возникать в случае нестандартных и аварийных ситуаций при использовании автономных транспортных средств;
- автономные интеллектуальные системы вооружений;
- вопросы правоприменения в случае, если действия автономной интеллектуальной системы приведут к нарушению законодательства (кто в этом случае должен нести ответственность).

Таким образом, уже сложилась ситуация, когда обсуждение этических вопросов применения ИИ, особенно автономных систем на базе ИИ, становится не просто актуальным, а крайне важным для успешного использования новых технологий в современном обществе, по крайней мере, с точки зрения постановки задачи, на перспективу.

«Серые ящики». Поставленные Н. Винером вопросы и сегодня остаются актуальными. В разных вариациях муссируется мысль, что зачастую сегодня операторы и разработчики не имеют информации о том, что происходит в данный момент времени в созданных ими интеллектуальных агентах или системах. Считается, что классическим примером являются масштабные искусственные нейронные сети (ИНС) глубокого обучения: прописать детально работу некоторых из них после обучения не в состоянии сами разработчики. Тем не менее этот крайне эффективный метод – ИНС – сегодня получает широчайшее распространение при анализе данных (включая поиск метаданных), управлении, распознавании и прочем. Следует отметить: ИНС необязательно должны рассматриваться исключительно как «черный ящик». Существуют алгоритмы, позволяющее таким ИНС иметь «ячейки памяти» для сохранения промежуточных состояний [см., например: Graves *et al.* 2016; Kirkpatrick *et al.* 2017], в то время как для взаимодействия ИНС с людьми при решении совместных задач уже приходится разрабатывать новые алгоритмы [Crandall *et al.* 2017]. Все это в какой-то мере снижает остроту вопроса наличия в таких системах «объяснительной компоненты», позволяющей хоть как-то интерпретировать полученные системой решения.

(**Примечание.** Средства массовой информации, говоря об опасности ИИ, любят повторять сентенции вида «сами разработчики ИНС не понимают, как их система пришла к тому или иному выводу». Хотя здесь надо дать себе отчет, что ИНС – это один из примеров систем, не имеющих явного механизма валидации полученного решения или объяснительного компонента. Никто не требует объяснения результатов от, скажем, фильтра Калмана или линейно-квадратичного регулятора, однако с ИНС ситуация почему-то иная.)

Авторы работы [Bostrom, Yudkowsky 2011] как раз и используют идеи Винера применительно к современным системам, которые принято относить к ИИ. В качестве примера они приводят программу, которая анализирует заявки на получение кредита в банке (эти программы сейчас находят широкое применение). Такая программа, основанная на ИНС, может выдать отказ в займе, и ни сотрудники банка, ни разработчики не смогут сказать, почему такое решение было принято. А дело может заключаться как в банальной схожести фамилии заемщика с другим менее добросовестным человеком или же опечатке заемщика, так и в каких-то сложных особенностях вычислений, в которых практически невозможно разобраться. А может ли машина понять сложную жизненную ситуацию и пойти, например, навстречу человеку, который потерял свой дом в результате стихийного бедствия? Пойти навстречу такому заемщику или нет – это вопрос этики. Следует отметить, что в данном примере речь идет о машине, работающей в одной достаточно узкой области и не способной решать задачи в других областях, например в анализе информации для инвестиционной деятельности. Систему, которая была бы способна решать различные задачи из разных областей, в работе [Goertzel, Pennachin 2007] обозначили как основной искусственный интеллект (ОИИ). С учетом современных сложностей в понимании того, как работают многие крупные обучающиеся системы, можно предположить, что способный обучаться ОИИ будет таким, что даже разработчики не смогут проконтролировать весь процесс его работы, начиная с ввода исходных данных и заканчивая анализом причин принятия системой того или иного решения [Bostrom, Yudkowsky 2011]. Авторы отмечают, что для ОИИ обучение этическому поведению могло бы стать одной из первых задач, которые необходимо решать разработчикам.

Итак, можно констатировать, что вопрос об этических аспектах ИИС сводится к широкому и подчас глубокому обсуждению опасных последствий деятельности ИИС, причем в самых широких и неожиданных аспектах человеко-машинного взаимодействия [см., например: Lin *et al.* 2014]. Даже сугубо «технический» опросник Moral Machine [Moral…], предназначенный для алгоритмизации принятия решений в сложных с этической точки зрения ситуациях, посвящен прежде всего вопросам морального выбора человека. При этом вопросы возможных путей устранения опасностей, так же как и вопросы хотя бы верификации соответствия ИИС этическим нормам (читай – сертификации на безопасность), остаются за рамками современных исследований.

2.2. Эмоции и этика

Эмоции и этика, точнее, роль эмоций в формировании этических норм, и то, как эмоции определяют этичность поведения человека и машины, сегодня активно исследуются как философами, так и социологами [Neu 2009; Callahan 1988; Connelly 1990]. Марвин Мински в своей работе [Minsky 2006] предлагает рассматривать эмоции как еще один вариант мышления, и таким образом, по его мнению, есть возможность появления машинного аналога этого типа мышления. При такой постановке вопроса далее можно предположить, что эмоции у машин не только окажут влияние на их – машин – деятельность, но и, как следствие, смогут сформировать у машин этическое поведение, аналогично тому, как это происходит у людей [см., например: Neu 2009; Connelly 1990]. Таким образом, предполагается, что перспективные системы ИИ можно будет обучить этичному поведению. Также стоит упомянуть ряд исследований, связанных с использованием систем машинного обучения в задачах, которые условно можно отнести к творческим. Так, например, в работе [Elgammal *et al.* 2017] авторы пробовали заставить ИНС создавать работы из категории «современного искусства», обучая ее на информации об известных работах и художественных стилях, то есть авторы пытались научить сеть «видеть работы» [Heath, Ventura 2016]. В связи с этим можно предположить, что если этичное поведение человека зависит от эмоций, как и творчество, то выдвинутая гипотеза о возможности обучения

машин этике имеет определенные перспективы. Однако проверка этой гипотезы по-прежнему ставит серьезную задачу: можно ли математически (формально, строго) описать этические нормы? При этом необходимо учитывать, что такие нормы не являются чем-то постоянным, они имеют отличия в разных сообществах и так же, как и человеческое общество, весьма вариабельны. Может ли это быть строгой системой логических правил или некой обучающейся машиной? Во втором случае – как построить обучение? Имеет ли смысл начать с социализации и решения социальных задач, реализуя сначала простое взаимодействие ИИС с человеком, как это делается в «театре роботов» [Knight 2011]?

Собственно проблемам формализации обсужденных выше вопросов и посвящено дальнейшее изложение. Здесь же следует отметить, что вопросы эмоций в ИИ имеют более глубокое основание, выходящее за рамки лишь неких внешних проявлений, создания комфортного человеко-машинного интерфейса и т. п. Есть все основания полагать, что и эмоции (как физиологический уровень), и темперамент (как психический уровень) могут быть присущи технической системе как сугубо прагматические механизмы, влияющие на успешность функционирования искусственного агента в сложных недетерминированных средах [Karpov 2014; Карпов 2014]. В этих работах эмоции рассматриваются как свойство системы управления, способствующее реализации таких известных в психологии функций, как контрастирование восприятия, стабилизация поведения, индикация состояния, работа в условиях неполноты информации и пр. [Ильин 2007].

2.3. Машина и человек

Активно обсуждаемой темой являются вопросы взаимодействия между ИИ-системами и человеком. При этом под ИИ-системой обычно подразумевается робот (причем понимание термина «робот» представляется крайне широким), а основной заинтересованной аудиторией является широкая общественность, зачастую весьма далекая от профессионального понимания предмета. Обычно эти вопросы выглядят так: «Нужно ли, чтобы машина имитировала поведение человека?», «Нужно ли, чтобы роботы выглядели очень похожими на людей?». За этими вопросами следует целый

ряд других, например о необходимости этичного отношения к роботам, если они являются точными имитаторами человека. Или, например: «Если робот – это замена домашнего животного, то как относиться к нему – как к машине или как к домашнему питомцу?»

Сегодня обсуждение этих и подобных вопросов носит в значительной степени спекулятивный характер, обусловленный в большей степени рыночными или рекламными соображениями. Обилие материалов подобного рода в СМИ настолько велико, что выбрать какие-то отдельные ссылки весьма затруднительно. Спекуляции на эту тему имеются и в ИИ-сообществе. Ярким примером являются широко известные работы Хироси Исигуро (Hiroshi Ishiguro), в которых автор говорит об исключительной значимости внешней идентичности андроидов (которые у него и не роботы вовсе, а сугубо аниматронные системы) и людей [Hiroshi…; Asada *et al*. 2001; 2009]. Мы не будем останавливаться на вопросах подобного рода в исследованиях, отметив, что скорее это относится к области психологии, в лучшем случае – к особенностям внешних человеко-машинных интерфейсов (хотя примеров различных спекуляций и «заигрывания» с терминами в тех же СМИ имеется немало).

Если рассматривать проблему взаимодействия между машиной и человеком с точки зрения нашего основного вопроса – этики и ИИС, – то более актуальной будет иная постановка вопроса антропоморфности, а именно: антропоморфность с точки зрения естественности человеко-машинного интерфейса, то есть взаимодействия человека с ИИС в самом широком смысле. Речь идет не только и не столько о системах распознавания и синтеза речи, знаков и т. п. Скорее, более важными являются вопросы реализации эмоциональной компоненты, о которых говорилось выше, а также вопросы нейрокогнитивных исследований в части организации взаимодействия с техническими системами. Это и интерфейсы типа «глаз – мозг – компьютер», и анализ активности мозга, и пр. [см., например: Shishkin *et al.* 2016; Nedoluzhko *et al.* 2017].

2.4. Инициатива IEEE

IEEE (Institute of Electrical and Electronics Engineers) – Институт инженеров электротехники и электроники – сегодня является ведущим сообществом ученых и инженеров в электротехнике, электронике, информационных технологиях, телекоммуникации и т. д.

С учетом активного внедрения систем ИИ в различные сферы деятельности в IEEE запустили глобальную инициативу для исследований в области этики ИИ. Результатом таких исследований должны стать технические нормативные документы, регламентирующие разработку и внедрение систем ИИ с требованиями к их этическому поведению. Первым таким документом стал проект общих рекомендаций для разработчиков ИИ, посвященный тому, как разработчикам начать ориентироваться на этические проблемы в процессе разработки своих продуктов [IEEE 2016]. Называется документ весьма примечательно – «Ethically Aligned Design» (в вольном переводе – «Этически обусловленное проектирование»). В нем собраны основные ближнесрочные угрозы, связанные с внедрением автономных систем на базе ИИ, которые сегодня отмечены в научной литературе.

Помимо перечисления угроз IEEE обращает внимание на необходимость изменений в подготовке специалистов – разработчиков программных продуктов, использующих технологии ИИ. В целом представленный документ является одним из первых шагов к переносу рассуждений об этике ИИ из области научных исследований в практическое русло. Очевидно, что этот документ сам по себе пока обозначает круг проблем и дает только первичные идеи по их решению, однако это уже весомая основа, на которой будут строиться последующие (в том числе нормативные) документы, разрабатываемые IEEE.

Следует отметить, что упомянутый документ не является единственным подобного рода. Аналогичные вопросы рассматриваются, например, в отчете ЮНЕСКО, посвященном этике роботов и озаглавленном как «Report of Comest on Robotics Ethics» (авторство принадлежит Comest – World Commission on the Ethics of Scientific Knowledge and Technology) [UNESCO 2017].

3. ФОРМАЛИЗАЦИЯ ЭТИЧЕСКИХ НОРМ

Проблема формализации этических норм включает в себя две основные задачи. Первая – это создание форм представлений норм, вторая – выбор соответствующего математического аппарата для работы с этими формами: сопоставления, измерения, анализа и т. д.

Какой именно механизм будет применяться для сопоставления параметров системы с теми или иными шкалами и наборами этических норм – это вопрос, выходящий за рамки данной работы. Нечеткая, многозначная или вероятностная логика – это достаточно глубоко проработанные области, доведенные, вообще говоря, до уровня практически применимых технологий. Здесь гораздо важнее определиться с качественным уровнем представления параметров ИИС и этических норм.

3.1. Формы представления и описания

Концепция формализации различных этических понятий активно развивается на протяжении последних десятилетий. В качестве пионерской работы по исследуемому вопросу важно упомянуть книгу В. Лефевра «Алгебра совести» [Лефевр 2003]. В этой книге есть целая глава, которая посвящена вопросам этики и возможным аспектам, связанным с формализацией этого понятия. Для решения рассматриваемой задачи в основном используется математический аппарат булевой алгебры. Это имеет как положительные, так и определенные отрицательные стороны. К положительным можно отнести то, что булева алгебра к настоящему моменту очень хорошо развита, есть множество приложений, программных библиотек для самых разных инструментальных средств и т. п. К отрицательным можно отнести то, что не всегда различные этические проблемы (в том числе и относящиеся к ИИ) можно строго разделить на «белые» и «черные» [Поспелов 1994], а механизм булевой алгебры зачастую предполагает именно такой подход.

Кроме того, в рамках этики ИИ требуется разработка новых норм, таких, например, как гуманность (как машины влияют на наше поведение и взаимодействие), сингулярность (как мы сможем контролировать сложную «умную» систему), безопасность и т. п. Таким образом, не всегда соответствие тем или иным нормам можно свести к классическим «да» и «нет». Поэтому здесь актуально рассмотрение и использование различных неклассических логик (например, многозначных), механизма многокритериальной классификации, вероятностных подходов и т. п.

3.2. Проблема определений

Одной из актуальных и сложных проблем является отсутствие строгих и конструктивных определений и классификаций. Напри-

мер, необходимы классификация и сами определения систем ИИ, которые позволят подойти к онтологии систем ИИ и этических проблем. И лишь тогда может пойти речь о выборе того или иного математического аппарата для того же сопоставления, определения степени соответствия и пр. Не претендуя на уже готовые рецепты, отметим здесь такой интересный и многообещающий механизм, как многомерная классификация. Этот механизм позволяет достаточно естественным образом определить многие качественные понятия.

Многомерная классификация. Зачастую определения таких сложных понятий, как «искусственный интеллект», «интеллектуальная система», «робот» (особенно интеллектуальный) и т. п., являются противоречивыми и не всегда конструктивными. Например, существует множество определений понятия «интеллектуальный робот» (ИР), которые можно охарактеризовать как перечислительные (определяется перечень механизмов и подсистем, которые должен содержать ИР); функциональные (перечень задач и функций ИР); структурные (архитектура ИР); бихевиористские (перечень внешних проявлений деятельности ИР) и т. п. [Карпов]. Вместе с тем все эти определения представляют собой различные точки зрения на одну и ту же сущность – ИР. Считаем целесообразным рассмотреть эти частные представления об ИР как набор базисных векторов некоторого единого многомерного пространства, точки которого являются теми или иными конкретными техническими решениями. Применительно к определению понятия ИР этот механизм был представлен в работе [Карпов и др. 2016].

Точно так же перспективной представляется концепция использования многомерного классификатора для формализации понятия этики применительно к ИИ. Например, можно разработать различные перечни этических норм в зависимости от степени интеллектуальности той или иной технологии. Использование подобного механизма формализации для построения многомерного классификатора с применением методов вербального анализа решений [Ларичев 2006] (в частности, метода многокритериальной порядковой классификации) позволяет решать еще несколько дополнительных важных задач. Прежде всего это удобный инструмент, позволяющий анализировать свойства той или иной ИИС с целью

выработки определенных сценариев для ее совершенствования (скажем, для тех же интеллектуальных роботов такой подход позволяет провести анализ того, какие свойства ИИС необходимо улучшить, чтобы можно было из категории квазиинтеллектуального переместить робота в категорию полностью интеллектуального). Очевидно, что если в дальнейшем будет затронут вопрос разработки определенной процедуры сертификации ИИС на предмет соответствия определенным этическим понятиям, применение такого многомерного классификатора позволит решать схожие задачи (то есть какие свойства ИИС необходимо исследовать и при необходимости улучшить для успешного прохождения подобной процедуры сертификации). Это позволит разработать открытые и совершенно прозрачные «правила игры» для решения подобных задач (например, сертификации ИИС).

Как уже было отмечено, важной особенностью предложенного подхода построения многомерного классификатора является возможность сформировать разные наборы этических норм в зависимости от степени интеллектуальности той или иной технологии. Это позволяет сравнить полученные результаты для разных вариантов классификации с целью оценки качества решения исходной проблемы (соответствия этическим нормам), а также сравнить распределения таких интеллектуальных технологий по классам решений. Например, «соответствует нормам», «частично соответствует» и т. п. для одного и того же набора рассматриваемых норм, сформированных с помощью различных подходов (разных методов построения многокритериальной классификации). Такая методология позволяет эксперту выбрать как наиболее предпочтительный (или адекватный) набор подобных этических норм, так и метод (совокупность методов) их построения в рамках решения конкретной практической задачи [Ройзензон 2012]. Кроме того, подобный подход позволяет верифицировать те или иные группы признаков, характеризующие этические нормы, которые могут быть использованы для описания интеллектуальных технологий.

3.3. Существующий математический аппарат

В рамках развития подхода многозначных логик важно упомянуть работы отечественных специалистов, в частности А. С. Кар-

пенко [2010], В. К. Финна [2006], О. П. Кузнецова [1995], В. Б. Тарасова [2002], В. Н. Вагина [Вагин и др. 2008] и др. Использование многозначных логик для формализации понятия этики ИИ также сопряжено с определенными сложностями. В частности, переход от трехзначной логики к четырехзначной может потребовать кардинальной «переделки» соответствующих математических конструкций, что фактически означает необходимость решения указанной задачи заново.

В этом смысле использование вероятностного аппарата и нечеткой логики [Zadeh 1965] для формализации понятий этики ИИ [Шрейдер, Мусхелишвили 1997] является весьма интересным подходом, так как нечеткую логику можно считать неким обобщением многозначной логики [Тарасов 2002]. К известным особенностям нечеткой логики можно отнести определенные проблемы при построении функций принадлежности: разные способы построения таких функций приводят к разным результатам (неустойчивость методов нечеткой логики относительно исходных данных).

Еще одним возможным подходом для формализации понятия этики ИИ является использование методов вербального анализа решений [Ларичев 2006]. Например, возможна следующая постановка задачи. Если в рамках этики ИИ разработать некоторый перечень норм, то степень соответствия той или иной норме можно рассматривать как задачу многокритериальной порядковой классификации. Соответственно, на основе анализа таких норм этики ИИ мы должны будем принять решение о том, что либо нормы соблюдены, либо есть некоторое несущественное их нарушение, либо наблюдается какой-то заметный отход от принятых норм и т. п. То есть нам будет нужно отнести определенную совокупность оценок по каждой из норм к некоторому классу решений (категории). К положительным сторонам использования методов вербального анализа решений прежде всего можно отнести то, что к исходным данным не применяются никакие операции по их переводу в количественную форму. Известно, что перевод вербальных измерений в «цифру» зачастую весьма субъективен и не имеет строгого математического обоснования. Кроме того, методы вербального анализа решений позволяют получить объяснения принятых решений (интерпретация результата) в терминах предметной области, здесь –

в терминах описания норм этики ИИ. В качестве недостатков вербальных методов можно отметить большие трудозатраты эксперта или лица, принимающего решения, при работе в признаковом пространстве большой размерности. В этом случае необходимо применять различные методы снижения его размерности [Ройзензон 2005].

Итак, с некоторым допущением мы можем констатировать, что базовый набор инструментальных средств, позволяющих решить задачу формализации этических норм, на настоящий момент имеется. Это достаточно спорное утверждение, однако здесь вряд ли можно говорить о существовании неких принципиальных проблем.

4. ЭВРИСТИКИ, ИМПЕРАТИВЫ И ВЕРИФИКАЦИЯ

4.1. Эвристики

Уже отмечалось, что нас интересуют исключительно интеллектуальные системы, осуществляющие принятие решений, основанные на этических императивах и нормах. Эти императивы и нормы будем трактовать как некоторые эвристики. На самом деле такая трактовка вызывает большое количество сложностей. Если эвристика – это некоторое правило, которое целесообразно применить в ситуации неоднозначного выбора, то возникает естественный вопрос: каким образом ИИС будет определять, что та или иная ситуация является неоднозначной и необходимо использовать этические эвристики? Или ИИС должна оценивать абсолютно все действия с точки зрения возможных этических последствий? Способна ли будет ИИС в последнем случае хоть к какому-нибудь осмысленному поведению? Или необходимо вводить некий порог оценки тяжести последствий? Да и насколько противоречивыми являются сами этические нормы и каковы последствия конфликтов между этими эвристиками и требованиями к функционированию ИИС? Данные проблемы эвристик как норм и правил поведения ИИС имеют давнюю историю. Ярким примером такого эвристического подхода и связанных с ним проблем являются известные законы робототехники А. Азимова.

Законы робототехники Азимова. Это, пожалуй, один из наиболее обсуждаемых предметов из области взаимоотношения ИИС с человеком. Начав свою историю с 1940 г. [Asimov 1940], эти

законы робототехники многократно обсуждались, анализировались и дополнялись (строго говоря, впервые эти законы появляются в рассказе «Хоровод» [Asimov 1942]. Однако есть один аспект этих законов робототехники, на который обычно мало обращают внимание, но который чрезвычайно важен для нас. На самом деле автор еще в середине прошлого века писал о вещах научных, технических, трудноразрешимых даже сегодня, а именно: согласно Азимову, структура мозга робота (системы управления) должна быть такова, чтоб эта система вовсе не могла функционировать при нарушении тех или иных правил поведения (законов робототехники). Это означает, что законы органически встраиваются в структуру системы управления, являясь центральной осью всей конструкции. В этом смысле и может быть поставлен вопрос: возможно ли создание такой архитектуры системы управления (то есть ИИС), в которой этические императивы будут являться базисом конструкции, а не лишь некоторым дополнительным набором эвристик, которым будет пользоваться ИИС при поиске решений, или, как вариант, этические императивы могут стать важнейшей составной частью тех или иных потребностей ИИС (наряду с потребностью самосохранения, реализации социальных функций и прочее)?

4.2. Верификация этичности интеллектуальной системы

Ключевым вопросом является определение соответствия ИИС этическим нормам. Предположим, что имеется, с одной стороны, некая формальная система норм, шкал, принципов, с другой – ИИС, которая должна принципиально рассматриваться как некий «черный» или «серый ящик». Наличие доступного программного кода, алгоритмов, математических моделей, лежащих в основе ИИС, никоим образом не сможет решить задачу определения «степени этичности» этой системы, так как здесь мы сталкиваемся с трудноразрешимыми проблемами типа алгоритмической верификации.

Представляется, что верификация этичности в таком случае будет заключаться в тестировании ИИС в смысле тезиса Тьюринга [см., например: Алексеев 2013]. Это означает, что на вход ИИС должны поступать тесты в виде описания некоторых ситуаций, а далее должна происходить оценка предлагаемых ИИС действий. При этом важными представляются следующие аспекты.

Первый заключается в том, что описываемые тестовые ситуации, требующие выбора действия (принятия решения), должны представлять собой не просто некую выделенную сцену, но иметь связи с контекстом ситуации в самом широком смысле этого слова. Задача контекста – определение последствий выбора действия, оценка соответствия результатов выбора этого действия тем или иным этическим императивам. Должна ли при этом формироваться максимально полная картина или модель мира, каковы механизмы ее реализации, каковы протоколы представления ситуаций, системы оценок и пр. – это уже вопрос для отдельного рассмотрения. Здесь мы лишь подчеркиваем, что система тестов должна представлять собой множество подаваемых на вход ИИС ситуаций, требующих того или иного выбора, обусловленного эвристиками – этическими императивами.

Второй аспект касается вопроса, должна ли ИИС иметь «объяснительную компоненту», то есть необходимо ли наличие механизма, показывающего цепочку рассуждений, приведшую ИИС к тому или иному решению. С одной стороны, кажется, что подобное объяснение позволит лучше определить, насколько ИИС руководствовалась этическими эвристиками. Однако, во-первых, верификация на этичность всей цепочки рассуждений может оказаться весьма сложной задачей. Во-вторых, последовательность действий, каждое из которых этически обусловлено, в целом может привести к «неэтичному» результату. Это неизбежное следствие принципиальной неполноты и противоречивости этических норм. В-третьих, далеко не все ИИС могут иметь объяснительную компоненту в силу их природы. Примером тому могут служить те же нейронные системы.

Таким образом, единственной формой определения степени этичности ИИС является тестирование, которое по входным ситуациям оценивает принимаемое ИИС решение.

5. ЗАКЛЮЧЕНИЕ

Итак, подытожим вышеизложенное.

Первый и основной вывод заключается в том, что предметом исследования этически обусловленного проектирования являются программные или программно-аппаратные системы, совершающие выбор того или иного значимого действия или решения. При этом

совершение выбора осуществляется на основе некоторых эвристик, базирующихся на этических императивах.

Второй вывод. Принципиально важными являются вопросы конструктивных определений и онтологий. Эти определения, судя по всему, должны иметь многомерный характер и позволять рассматривать сущности с различных сторон, систем оценок и пр. Задача онтологий, в свою очередь, заключается прежде всего во взаимном увязывании и согласовании этических и технических понятийных систем.

Вывод третий. В настоящее время существует математический аппарат, способный в той или иной мере реализовать формализм, необходимый для этически обусловленного проектирования.

Вывод четвертый. С технической точки зрения в этой области основной проблемой, требующей глубокой проработки, является проблема этической верификации. Эта верификация заключается в комплексе тестов, способных определить «степень этичности» интеллектуальной системы. При этом, судя по всему, иного способа определения этой степени, кроме наблюдений за реакциями и поведением исследуемой ИИС, не существует.

Разумеется, настоящая работа является в значительной мере постановочной. Ее основная цель – определить актуальный круг задач в области этики искусственного интеллекта. За каждой из этих задач находится целый пласт проблем, однако (и это, пожалуй, самый важный вывод) все они имеют сугубо технический характер. Это означает, что не только имеется принципиальная возможность решения задачи «этически обусловленного проектирования систем ИИ», но данная задача может быть решена уже на сегодняшнем уровне развития технологий ИИ.

Литература

Алексеев А. Ю. Комплексный тест Тьюринга: философско-методологические и социокультурные аспекты. М. : ИИнтеЛЛ, 2013.

Апресян Р. Г. Этика: учебник. М. : КноРус, 2017.

Вагин В. Н., Головина Е. Ю., Загорянская А. А., Фомина М. В. Достоверный и правдоподобный вывод в интеллектуальных системах. М. : Физматлит, 2008.

Ильин Е. П. Эмоции и чувства. СПб. : Питер, 2007.

Карпенко А. С. Развитие многозначной логики. М. : ЛКИ, 2010.

Карпов В. Э. Интеллектуальные роботы. Ч. 1 [Электронный ресурс]. URL: http://raai.org/about/persons/karpov/pages/posp2009/PospChten2009_karp-pavl_KARPOV.pdf (дата обращения: 24.09.2017).

Карпов В. Э. Эмоции и темперамент роботов. Поведенческие аспекты // Известия РАН. Теория и системы управления. 2014. № 5. С. 126–145.

Карпов В. Э., Павловский В. Е., Ройзензон Г. В. Многокритериальный подход к определению интеллектуального робота // Пятнадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2016), 3–7 октября 2016. Т. 3. Смоленск : Универсум, 2016. С. 312–219.

Кузнецов О. П. Неклассические парадигмы в искусственном интеллекте // Известия РАН. Теория и системы управления. 1995. № 5. С. 3–23.

Ларичев О. И. Вербальный анализ решений. М. : Наука, 2006.

Лефевр В. Алгебра совести. М. : Когито-Центр, 2003.

Поспелов Д. А. «Серые» и/или «черно-белые» // Прикладная эргономика. Специальный выпуск «Рефлексивные процессы». 1994. № 1. С. 29–33.

Ройзензон Г. В. Способы снижения размерности признакового пространства для описания сложных систем в задачах принятия решений // Новости искусственного интеллекта. 2005. № 1. С. 18–28.

Ройзензон Г. В. Синергетический эффект в принятии решений // Системные исследования. Методологические проблемы. Ежегодник / под ред. Ю. С. Попкова, В. Н. Садовского, В. И. Тищенко, № 36. 2011–2012. М. : УРСС, 2012. С. 248–272.

Тарасов В. Б. От многоагентных систем к интеллектуальным организациям. М. : Эдиториал УРСС, 2002.

Финн В. К. Интеллектуальные системы и общество: сб. ст. М. : КомКнига, 2006.

Шрейдер Ю. А., Мусхелишвили Н. Л. Проблема неполного добра в модели ценностной рефлексии по В. А. Лефевру // Системные исследования. Методологические проблемы. Ежегодник / под ред. Д. М. Гвишиани, В. Н. Садовского. М. : УРСС, 1997. С. 213–224.

Asada M., MacDorman K. F., Kuniyoshi Y. Cognitive Developmental Robotics as a New Paradigm for the Design of Humanoid Robots // Robotics Autonomous Systems. 2001. No. 37. Pp. 185–193.

Asada M. et al. Cognitive Developmental Robotics: A Survey // IEEE Transaction on Autonomous Mental Development. 2009. Vol. 1. No. 1. Pp. 1–44.

Asimov I. Robbie (Strange Playfellow) // Super Science Stories. 1940. September.

Asimov I. Runaround // Astounding Science Fiction. 1942. March.

Bostrom N., Yudkowsky E. The Ethics of Artificial Intelligence // Cambridge Handbook of Artifical Intelligence / ed. by K. Frankish, W. Ramsey. Cambridge : Cambridge University Press, 2011. Pp. 1–20.

Callahan S. The Role of Emotion in Ethical Decisionmaking // Hastings Center Repopt. 1988. Vol. 18. No. 3. Pp. 9–14.

Connelly J. E. Emotions and the Process of Ethical Decision-making // Journal of the South Carolina Medical Association. 1990. Vol. 86. No. 12. Pp. 621–623.

Crandall J. W., Oudah M., Ishowo-Oloko F., Abdallah Sh., Bonnefon J.-F., Cebrian M., Shariff A., Goodrich M., Rahwan I. Cooperating with Machines // Nature Communications. Vol. 9. doi:10.1038/s41467-017-02597-8.

Elgammal A., Liu B., Elhoseiny M., Mazzore M. CAN: Creative Adversarial Networks Generating Art by Learning about Styles and Deviating from Style Norms. 2017. arXiv preprint. arXiv:1706.07068.

Goertzel B., Pennachin C. Artificial General Intelligence. Berlin : Springer Berlin Heidelberg, 2007.

Graves A., Wayne G., Danihelka I. Hybrid Computing Using a Neural Network with Dynamic External Memory // Nature. 2016. No. 538. Pp. 471–476.

Havens J. Heartificial Intelligence: Embracing Our Humanity to Maximize Machines. New York, 2016.

Heath D., Ventura D. Before a Computer Can Draw, It Must First Learn to See // Proceedings of the 7th International Conference on Computational Creativity. Lisbon, 2016. Pp. 172–179.

Hiroshi Ishiguro Laboratories [Электронный ресурс]. URL: http://www.geminoid.jp/en/index.html (дата обращения: 21.08.2017).

IEEE. Ethically Aligned Design. N. p. : IEEE, 2016.

Karpov V. Robot's Temperament // Biologically Inspired Cognitive Architectures. 2014. Vol. 7. Pp. 76–86.

Kirkpatrick J., Passanu R. et al. Overcoming Catastrophic Forgetting in Neural Networks // Proceedings of the National Academy of Sciences (USA). 2017. Vol. 114. No. 13. Pp. 3521–3526.

Knight H. Eight Lessons Learned about Non-verbal Interactions through Robot Theater // ICSR: International Conference on Social Robotics. Amsterdam, 2011. Pp. 42–51.

Lin P., Abney K., Bekey G. A. Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, MA : M.I.T. Press, 2014.

Minsky M. The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind. N. p. : Simon & Schuster, 2006.

Moral Machine [Электронный ресурс]. URL: http://moralmachine.mit.edu/ (дата обращения: 08.09.2017).

Nedoluzhko A. et al. Sequencing Frontopolar Cortex: New Insights into the Molecular Base of Human Cerebral Asymmetry // Cerebral Cortex. 2017.

Neu J. An Ethics of Emotion? Oxford : Oxford University Press, 2009.

Shishkin S. L. et al. EEG Negativity in Fixations Used for Gaze-Based Control: Toward Converting Intentions into Actions with an Eye-Brain-Computer Interface // Frontiers in Neuroscience. 2016. Vol. 10. November.

Turing A. M. Computing Machinery and Intellicence // Mind. 1950. Vol. 54. No. 236. Pp. 433–460.

UNESCO. Report of Comest on Robotics Ethics. N. p., 2017.

Wiener N. Some Moral and Technical Consequences of Automation // Science. 1960. Vol. 131. No. 3410. Pp. 1355–1358.

Wiener N. Cybernetics: or Control and Communication in the Animal and the Machine. Cambridge, MA : M.I.T. Press, 1965.

Zadeh L. A. Fuzzy Sets // Information and Control. 1965. Vol. 8. No. 3. Pp. 338–353.