ЭТИКА

А. В. РАЗИН

ЭТИКА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В статье очерчивается круг проблем, с которыми человек столкнулся при первых попытках создания искусственного интеллекта, способного в той или иной мере принимать самостоятельные решения. Поднимается вопрос об этических ограничениях, которые могут быть заложены в искусственные интеллектуальные системы при программировании. Далее отмечается, что это само по себе еще не может считаться этикой искусственного интеллекта, так как для того, чтобы решать этические задачи, надо обладать свободой воли. В данной связи рассматривается вопрос о свободе воли у человека, так как наличие таковой подвергается сомнению в некоторых современных аналитических исследованиях. Мы доказываем, что человек обладает свободой воли, он может создавать произвольные образы, связанные с разными уровнями отражения реальности, и манипулировать ими. Это оказывается необходимым для успешного ориентирования. Однако из этого же следует допущение принципиальной возможности ошибки, как в рассуждениях, так и в действиях. Этика непосредственно начинается тогда, когда появляется способность реагировать на собственные ошибки, осуществлять рефлексию поведения, учитывая при этом мнения других людей. Такая же принципиальная возможность ошибки должна быть заложена и в работу искусственного интеллекта, чтобы можно было говорить о его этике в собственном смысле слова. Должны быть также выполнены условия коммуникации машин, их взаимных оценок и наличия у них феноменального опыта.

Ключевые слова: *сознание, интеллект, воля, этика, ограничения, ошибка, рефлексия, коммуникация, оценка.*

The article outlines the range of problems aroused by the first attempts to create artificial intelligence that is capable to some degree make independent decisions. Here the question of ethical constraints that can be incorporated into artificial intelligent systems in programming is raised. It is further discussed that this cannot yet be considered the ethics of artificial intelligence, since in

order to solve ethical problems it is necessary to possess free will. In this regard, the question of a human free will is considered, since its presence is questioned in some modern analytical studies. We prove that a person possesses a free will and he is able to create conscious images associated with different levels of reflection of reality and manipulate them. This turns out to be necessary for successful orientation. However, this also implies the assumption of a fundamental possibility of error, both in reasoning and in actions. Ethics actually emerges when there originates the ability to react to one's own mistakes, to reflect over a personal behavior taking into account other people's opinion. The same fundamental possibility of error should be incorporated in the work of artificial intelligence, so that we could speak about its ethics in the proper sense of the word. The conditions for the communication among machines, their mutual assessments and their phenomenal experience must also be met.

Keywords: *consciousness, intelligence, will, ethics, restrictions, error, reflection, communication, evaluation.*

Одной из этических проблем цифрового общества может быть вопрос об этике искусственного интеллекта в смысле этических ограничений в деятельности искусственного интеллекта, в тех решениях, которые способна принимать машина, наделенная последним, робот, компьютер, встроенный в другую техническую систему и т. д. Еще более сложный вопрос – о том, способна ли техническая система, решая какие-то позитивные задачи, например, спасения людей, брать на себя риски принятия решений, отвечать за эти решения перед людьми и другими техническими системами, насколько такая машина (с искусственным интеллектом) будет способна к научению, развитию в этическом плане и насколько общество готово взять на себя риски деятельности такой машины, если допустить, что искусственный интеллект в принципе будет способен принимать разные решения.

То, о чем я пытаюсь сейчас говорить, во многом может быть отнесено к достаточно отдаленному будущему, но, строго говоря, рассуждать об этике искусственного интеллекта, не допуская возможностей принятия им различных решений, а следовательно, и не обсуждая вопрос о свободе воли, нельзя. Ибо, если мы сведем этику искусственного интеллекта только к заложенным в программу его деятельности этическим ограничениям, это будет не этика ис-

кусственного интеллекта, а этические правила создания интеллектуальных систем, необходимые при программировании.

Но даже если ограничиться только этой задачей, и здесь возникает масса проблем.

Конечно, все мы хорошо знаем законы создания робототехники, предложенные Айзеком Азимовым, но это только базовые принципы, которые могут оказаться недейственными в конкретных ситуациях.

Скажем, уже сейчас в США ведется разработка таких компьютерных систем, которые в критических ситуациях будут блокировать действия водителя автомобиля и принимать самостоятельные решения.

Но какой подход, какую этическую доктрину надо будет принять при программировании – абсолютистскую или утилитарную? Скажем, если возникнет дилемматическая ситуация: можно спасти жизни нескольких человек ценой жизни одного, но того, кто в принципе не виноват в возникновении самой критической ситуации. Абсолютистский подход, конечно, однозначно запрещает приносить в жертву какого-либо одного случайно попавшего в ситуацию человека ради спасения жизней нескольких людей. А утилитарный подход при каких-то обстоятельствах это в принципе допускает.

Можно ли свернуть с дороги ради спасения жизни пешехода, можно ли пересечь сплошную линию ради предотвращения аварийной ситуации, можно ли при каких-то обстоятельствах заехать на тротуар? Все эти вопросы неизбежно будут стоять перед разработчиками сложных интеллектуальных управляющих систем.

Наши коллеги из Курчатовского института, с которыми мы некоторое время назад пытались разобраться в проблеме этики искусственного интеллекта применительно к задаче создания охранных роботов, отмечают следующие задачи в области разработки этики искусственного интеллекта.

«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 будущем эти стандарты и не будут ли они нести сугубо ограничительную роль, тормозящую развитие ИИС» [Карпов и др. 2018: 86–87].

Последнее предупреждение звучит, на мой взгляд, крайне актуально, так как если мы разработаем жесткие этические стандарты для всех систем искусственного интеллекта, это и будет означать, что никакого дальнейшего развития этики искусственного интеллекта не будет.

Авторы цитированной статьи отмечают сложность формализации этических систем, неопределенность этических критериев решений, принятых искусственным интеллектом в ряде ситуаций. Они предлагают для решения этой проблематики обращение к многозначным логикам и так называемой нечеткой логике, которую можно считать некоторым обобщением многозначной логики.

«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.» [Там же].

Но возникает вопрос: зачем мы будем проводить такую дифференцированную оценку, зачем будем определять степень отклонения от норм, существенность их нарушения? Если просто чтобы решить вопрос о том, выключить ли эту машину или нет, направлять ли ее на перепрограммирование, это выглядит как очень ограниченная задача. Если же мы будем проводить анализ с целью научения машины, разъяснения искусственному интеллекту, какие особенности ситуации не были учтены, задача приобретает гораздо более широкий смысл.

Однако ставя такие задачи, допуская такой тип отношения человека с искусственным интеллектом, мы уже должны допустить в деятельности искусственного интеллекта некоторую степень неопределенности, сходную с понятием свободы воли.

Необычность ситуации, связанной с обсуждаемым вопросом, заключается в том, что есть целая группа аналитических философов, которые отрицают наличие свободы воли и у человека.

В разных формах свободу воли отрицают такие аналитические философы, как Дерк Перебум, Ван Иген, Джон Мартин Фишер. Их аргументация выглядит довольно просто: на каждое решение воздействует предшествующее состояние, а на это состояние – то, которое предшествовало ему, и таким образом выбор оказывается предопределенным. Этой позиции противопоставляется либертарианство, которое отказывается от детерминизма. Роберт Кейн (либертарианец) считает, что свобода воли несовместима с детерминизмом, но совместима с индетерминизмом. Однако ответственности индивида за свои действия не получается ни в том ни в другом случае, ведь как можно отвечать за события, от тебя не зависящие?

Одним из аргументов тех, кто отрицает свободу воли и возможность сознания влиять на принятие решений, послужили известные эксперименты Б. Либета, который показал, что переживание, сопровождающее решение что-то сделать, запаздывает за мозговыми процессами, то есть возбуждение в мозге предшествует осознанию того, что совершается некоторое действие. Эти исследования получили дальнейшие подтверждения и обсуждаются до сих пор. Но Д. И. Дубровский справедливо отмечает, что в эксперименте действие осуществляется по программе, заданной экспериментатором (надо будет поднять руку), а значит, оно уже предварительно осмысливается. Кроме того, Дубровский ставит вопрос: а что, собственно, доказывают подобные эксперименты, даже если бы можно было допустить их полную корректность?

«Но давайте условно примем, что эксперименты Либета и Суна были безупречны. Что же тогда им удалось доказать? Только то, что некоторое простейшее действие (одно из двух), заданное согласно инструкции, начинается и выполняется до того, как испытуемый осознает свое решение произвести действие, то есть как бы осуществляется на бессознательном уровне. Но при этом, повторю,

у испытуемого было сформировано ясное понимание и осознание поставленной перед ним задачи, а также согласие выполнить ее. Трактовать все это как опровержение свободы воли, или даже как существенный довод в пользу такого заключения, по меньшей мере, наивно. Хорошо известно, что в обыденной жизни сплошь и рядом четкая и важная сознательная установка может развязывать цепь действий, часть которых "проскакивает" на неосознаваемом уровне, обеспечивая, однако, достижение сознательно поставленной цели» [Дубровский 2017: 741].

В разрешении проблемы о свободе воли я принимаю теорию эмерджентной причинности, то есть представление о том, что одни нейронные сети могут управлять другими и это позволяет интерпретировать идеальное на основе материального. Идеальное в такой интерпретации оказывается некоторым смыслом, порожденным сложноорганизованными взаимодействиями мозговых сетей.

Правда, это возвращает нас к известной теории тождества, а один из наших крупнейших специалистов по сознанию В. В. Васильев считает, что теорию тождества нельзя верифицировать и поэтому она должна быть отвергнута. Выражая свою позицию, он пишет: «Простой довод, сразу подрывающий теорию тождества, состоит в том, что ее главный тезис – если не интерпретировать его в элиминативистском смысле, когда он очевидно ложен, – попросту бессмыслен: это и есть та карусель иллюзий, на которой помещалась вся теория тождества. И бессмыслен этот тезис потому, что его нельзя верифицировать. В самом деле, как верифицировать утверждение, что некое ментальное состояние – скажем, чувство, возникающее при просмотре фильма Lost in Translation, – тождественно нейронному процессу в мозге?» [Васильев 2009: 80].

Элиминативистский смысл – это просто лишение сознания онтологического статуса, что пытается сделать Д. Деннет. Но что может означать этот онтологический статус, насколько идеальное может онтологически быть чем-то противоположным материальному? Без решения этого вопроса невозможно говорить об искусственном интеллекте, ведь тогда действительно будет непонятно, о тождестве чего идет речь.

Теория тождества сталкивается с одной, казалось бы, непреодолимой сложностью. В логике тождества на первый взгляд оказыва-

ется невозможным понять, что такое идеальное, какой вообще смысл имеет данное понятие, если все сводится к материальным процессам. Тем не менее при таком подходе не нарушается принцип казуальной замкнутости физического. Если же рассматривать идеальное и ментальное как нечто радикально отличное от материального, принцип казуальной замкнутости физического нарушается.

В. В. Васильев предлагает решить эту проблему через утверждение, что принцип казуальной замкнутости физического может нарушаться на локальном уровне, но присутствовать на глобальном уровне, предполагающем связь мозга и порожденного им сознания с миром в целом. Для этого вводится понятие глобальной супервентности. Такая супервентность предполагает, что сознание как таковое не заключено в мозге, а обусловлено связью мозговых процессов с миром в некотором глобальном смысле, например – резонансными отношениями с микромиром [Васильев 2009: 224–227].

Так как в микромире имеет место принцип неопределенности, неопределенность или свобода оказывается заложенной и в работу сознания человека.

Тем не менее на данный момент это не более верифицируемо, чем то, что предполагается в теории тождества. Кроме того, можно заметить, что резонансные отношения могут быть между какими-то схожими сущностями, а процессы в мозге человека вряд ли схожи с процессами, которые можно наблюдать в микромире.

И если мы примем данную теорию, то уж точно не сможем создать искусственный интеллект в смысле способности искусственных систем к каким-то аналогам сознательной деятельности. Ведь смоделировать связь между искусственным мозгом, созданным на небиологическом носителе, и иными природными процессами гораздо сложнее, чем создать способную к мыслительным действиям машину.

Но вернемся к теории тождества и зададим вопрос: действительно ли нельзя верифицировать, что чувству соответствует некоторый нейронный процесс? На мой взгляд, вполне можно.

Допустим, мы воспринимаем некоторый пространственный образ. Это может сделать только целая цепь нейронов. Тому, кто знаком с аналитической геометрией, понятно, как пространственный образ может быть задан в двоичном коде. Это будет целая матрица

значений. Но на этом работа мозга не заканчивается. По-видимому, данная цепь имеет и своего представителя в виде одного контрольного специализированного нейрона. Так как нейроны не транзисторы, а клетки, они, как уже показано в исследованиях, могут быть специализированы. Это именно те нейроны, которые, как даже уже снято на видео, приходят в особое возбужденное состояние при работе мозга по различению пространственных образов. Это сигнал: образ опознан – это кошка, ее можно погладить, и далее этот сигнал, получающий эмоциональную окраску, передается либо непосредственно к миндалевидному телу, отвечающему за движения, либо, в случае неопределенной ситуации, в неокортез, и уже затем к миндалевидному телу. В своей книге «Эмоциональный мозг» Д. Гоулман показал, что эти два пути имеют большое эволюционное значение. Один вызывает более быструю реакцию, а другой – более долгую, но зато более продуманную. Одно дело, когда вы видите кошку, другое дело – когда змею. В последнем случае уже нет времени на размышления [Гоулман 2009: 39].

Таким образом, в работе мозга представлены не только чувственные реакции, но и пути этих чувственных реакций, связанные с особыми эмоциями типа страха, чувства опасности или, наоборот, предвкушения чего-то приятного.

Но должны ли мы имитировать подобные реакции при создании искусственного интеллекта? Что, собственно, это может дать для расширения возможностей мышления искусственно созданной машины, которая, скорее всего, никогда не встретится со змеей и реакции которой будут важны для человека именно в плане решения вопросов какой-то ограниченной сферы ее применения? Как это ни парадоксально, эмоциональный настрой человека важен даже при совершении сугубо логических операций или решении математических задач. Неслучайно Л. Фейербах отмечал: «Мышление без желания мыслить, будь то даже самое трезвое, самое строгое, будь то даже математическое мышление, – без ощущения удовольствия или счастья в этом мышлении – это пустое, бесплодное, мертвое мышление» [Фейербах 1995: 437]. Почему? У Фейербаха точного ответа на этот вопрос нет, но с точки зрения современной психологии и физиологии высшей нервной деятельности он очевиден. Интерес и удовольствие от достижения результата обеспечивают сосредото-

чение ресурсов, открывают дополнительные возможности из-за того, что в процесс включаются дополнительные нейронные сети.

Вроде бы к сосредоточению ресурсов способна и машина в том виде, в каком в настоящее время существует искусственный интеллект. Но в действительности не все так просто. Машина или компьютер просто перебирает заложенную в нее информацию, человек же способен классифицировать информацию как важную и неважную, соответственно, способен открывать более быстрый путь решения проблемы, освобождая себя от анализа незначимой или не очень значимой для данного случая информации. Более того, человек обладает, по крайней мере на данный момент, гораздо большей способностью ассоциативного мышления. Это позволяет найти нестандартные решения, к чему машина пока способна в гораздо меньшей степени. С точки зрения возможности усвоения внешней информации из сетей и других источников искусственный интеллект, компьютер обладает огромными возможностями, например, по частоте употребления пользователями некоторых понятий, и с точки зрения не только их количественной, но и качественной оценки. Однако это зависит уже от тех параметров, которые заложены в машину программистом. Для того же, чтобы самостоятельно создавать качественные характеристики, искусственный интеллект должен в значительной степени освоить логику развития культуры и самое главное – иметь собственный феноменальный опыт.

Здесь мы подходим к важнейшей проблеме, возникающей в философии сознания, – проблеме «квалио». Для того чтобы понять, что это такое в окончательном смысле, надо обратиться к способам кодирования образов и операций, которые способно проводить с ними сознание.

Для этого надо прежде всего осмыслить, что феноменальный опыт – это опыт, порожденный поэтапным осмыслением развития нашего тела и нас самих в конкретных жизненных ситуациях, конечно, и при учете влияния общего культурного фона. Такой опыт наделяет все образы нашего сознания особым субъективным смыслом, эмоциональным отношением, которое присутствует тогда, когда образы прошлых восприятий извлекаются из нашей памяти. Когда мы смотрим фильм или театральную пьесу, мы сопережива-

ем героям, отражаем в образах, представленных в нейронных взаимодействиях, образы того, что происходит на сцене. Но не только, в игре образов мы опережаем события, хотим, чтобы они разворачивались в желаемом нами направлении. Это и придает образам сознания состояние «квалио», в котором образы всегда оказываются связаны с нашими эмоциональными реакциями.

Феноменальный опыт оказывается не менее важен, чем способность чисто рационального рассуждения, потому что с ним связана классификация событий. В уже упоминавшейся книге «Эмоциональный мозг» Д. Гоулман приводит пример с человеком, у которого в результате хирургической операции на мозге оказалась повреждена эмоциональная сфера. Он мог прекрасно решать математические задачи, но был беспомощен даже в том, чтобы назначить обычную встречу с врачом, так как всегда находились мешающие обстоятельства и все события казались ему равнозначными [Гоулман 2009: 90–91].

Смысл сознания как субъективной реальности заключен не только в образах. Поддерживая идеи К. В. Анохина, А. М. Иваницкий совершенно верно отмечает, что «сознание человека состоит не только из последовательности образов. Оно также способно манипулировать этими образами и символами, формируя мыслительный процесс» [Иваницкий 2010: 449].

Но мыслительному процессу, понятому как рациональная процедура, предшествуют операции с образами, необходимыми при осуществлении сложной ориентировочной деятельности – определении своего субъективного положения в пространстве и времени, а также планировании действий, отнесенных к некоторому моменту в будущем.

С точки зрения П. Я. Гальперина, манипулирование образами, разыгрывание в образах вариантов будущего действия является основой психической организации человека и высших животных. Суть дела заключается в том, что инстинктивная связь с миром не позволяет учитывать действие в ситуации с меняющимися параметрами, например, когда приходится прыгать на жертву с разного расстояния. Здесь невозможна однозначная реакция (то есть по существу одинаковое усилие мышц, необходимое для совершения действия), возникающая в связи с некоторым внешним раздражи-

телем. Поэтому мышцам должен быть дан разный сигнал на основе предварительно переработанной мозгом информации [Гальперин 1999: 164–165].

Д. И. Дубровский также отмечает, что человек оказывается способным управлять своими мозговыми процессами, и это формирует то, что можно назвать психической причинностью. «У человека высшие формы нисходящих детерминаций выступают… в форме психической причинности, которая знаменует активность нашего Я, является функцией Эго-системы головного мозга» [Дубровский 2017: 751].

По существу, к близким выводам приходят те, кто занимается созданием искусственного интеллекта на базе искусственных нейронных сетей.

Игорь Феликсович Михайлов в статье «Искусственный интеллект и когнитивные науки: перспектива антирепрезентационализма», ссылаясь на западные источники, приводит классификацию подходов к пониманию искусственного интеллекта. Она включает восемь позиций: 1) символический подход, в котором символы соответствуют состояниям и намерениям; 2) силовой, в котором увеличение мощности искусственной системы позволяет перебирать больше ситуаций; 3) основанный на знаниях, предполагающий, что в искусственный интеллект надо заложить как можно больше информации; 4) прецедентный, предполагающий опору на прошлые решения; 5) креативный – перекомбинирование известных паттернов; 6) биокомпьютационный, подразумевающий коннекционализм, то есть обучение машины связывать разные явления; 7) динамический, описание состояний системы в терминах пространства, зависимых от переменных времени; 8) воплощенный. «Воплощенный (situated) подход отказывается от классицистского понимания интеллекта как абстрактного, индивидуального, рационального и оторванного от восприятия действия, противопоставляя этому понимание его как отелесненного (embodied), встроенного (embedded) и распределенного (distributed). Иными словами, когнитивные процессы протекают не в мозге, а между мозгом, остальным телом и средой» [Михайлов 2017: 292].

В ряде своих публикаций я показал, что сознание принципиально не может существовать без тела, причем тела, постадийно развивающегося, и без постоянной коммуникации субъекта, обладающего телом с другими субъектами [Разин 2011: 23–32].

Из этого следует, что, строго говоря, чтобы создать искусственный интеллект, обладающий какими-то аналогами человеческого сознания, мы должны создать самопорождающее себя сообщество машин или роботов, способных вступать в коммуникации с другими роботами и проходить разные стадии развития. Это, конечно, не означает, что роботы должны рожать друг друга наподобие человека. Просто в их конструкцию должны быть заложены принципы постадийного развития интеллектуальных схем работы искусственного интеллекта.

Уже очень давно психолог К. Леви предложил мысленный эксперимент, задав вопрос: можно ли будет сформировать у дельфина сознание, если найти код перевода звуковых сигналов дельфинов в человеческую речь? Его ответ был отрицательным. Дельфин не имеет схожего с человеком тела, и его детеныши не захотят подражать человеку в смысле манипуляции предметами, освоения мира по такому же пути, по какому его осваивает человеческий ребенок. Отсюда следует вывод о том, что интеллектуальное сообщество машин для их успешной коммуникации должно будет иметь еще и схожие тела.

Но вернемся к эмерджентной причинности, отвлекаясь пока от глобальной задачи создания сообщества машин.

Итак, мозг работает по принципу взаимосвязи сетей разного уровня. Более того, эти сети, наиболее вероятно, не являются чем-то постоянным. Они подвижны, подчинены определенным функциям, в том числе – отвечающим данному моменту активности субъекта. Особенностью сетевых взаимодействий в принципе является то, что единый управляющий центр отсутствует. Для решения какой-то задачи подключаются новые ресурсы, создаются новые нейронные связи. В принципе на это способен и процессор компьютера, так что не составляет загадки то, что в человеческом мозге мы не находим какого-то специального центра сознания. Но ком-

пьютер не обладает телом, без которого оказывается невозможен феноменальный опыт.

В то же время из сказанного выше ясно, что именно феноменальный опыт придает образам нашего сознания статус «квалио», без которого оказывается невозможным обратное влияние того, что производится сознанием, разыгрывается в сигналах головного мозга в связи с предположениями о будущих состояниях реальности и наших возможных действиях в этих будущих состояниях, невозможным оказывается также и преобразование этих состояний в смысле соотнесения их с нашими собственными состояниями, то есть эмерджентная причинность.

Таким образом, можно сказать, что идеальное не абсолютно противоположно материальному, но связано с несколькими уровнями рефлексивного отношения:

– отражение того, что происходит в мире, и оценка этого с точки зрения возможности удовлетворения некоторых первичных потребностей субъекта;

– разыгрывание в аналоговых образах того, что может произойти в мире, и перезаписывание этого в цифровых образах, что позволяет обращаться к эмоциональной памяти;

– рефлексия своего возможного положения и действий в соответствии с тем, что могло бы произойти при моих успешных действиях в мире;

– наконец, исследование объективных свойств реальности при отсутствии непосредственного заинтересованного действия в некоторых обстоятельствах. Это будет связано уже с потребностями нового уровня. У животных это просто ориентировочный инстинкт. У человека – те эмоциональные состояния, которые сопровождают процесс познания, открытие истины (иногда это также называют любопытством, по крайней мере, любопытство составляет часть этого процесса).

Итак: наша свобода воли заключена, во-первых, в произвольной модельнотворческой (или гипотезотворческой) активности мозга. Это условие ориентации на основе идеальных образов, так как мозг прорабатывает разные варианты и выбирает из них наилучшие.

Во-вторых – в том, что отбираемые управляющей инстанцией решения в принципе никогда не являются окончательными. Каждый из нас сталкивался с феноменом позднего решения, когда, например, наиболее удачные ответы оппонентам приходят уже после завершения дискуссии. Это связано с тем, что мозг продолжает работать и после свершившегося события. Он продуцирует ответы, которые могут быть учтены в будущем.

Мозг может ошибаться, в непосредственной ситуации действия какие-то решения могут казаться равнозначными, но действовать все равно приходится. Дж. Сёрль пытался решить эту проблему, введя понятие разрыва. Разрыв, с его точки зрения, ничем не заполненный, как раз связывается им с понятием свободы воли [Сёрль 2004: 28]. Свободные решения поэтому всегда оказываются как бы недодетерминированными. Они принимаются в условиях неполной информации и являются вероятностными. Но наши неправильные или не до конца правильные решения могут создавать новые ситуации. Таким образом, фаллибилизм (принципиальная возможность ошибки) также лежит в основании свободы воли.

Наконец, в-третьих, свобода воли связана с саморефлексией. Что бы ни говорили какие-то философы о предопределенности нашего характера прошлыми обстоятельствами жизни, человек ответственен за свой характер. У него есть возможность размышлять над своим поведением, выяснять, к чему он склонен, и, если есть необходимость, бороться с этими склонностями. Это прекрасно понимал уже Аристотель.

Итак, человек обладает свободой воли. С наличием его свободной воли связано понятие моральной и правовой ответственности. Насколько все это можно отнести к искусственному интеллекту? Для этого, даже если речь идет о каких-то самых элементарных формах человекоподобного искусственного интеллекта, должны быть выполнены определенные условия. В искусственный интеллект должны быть заложены:

– произвольная игра образами;

– критерии выбора оптимальных решений вместе со встроенными этическими ограничениями (например, такими, как законы Азимова + ответственность перед человечеством в целом);

– возможность оценки степени рисков при опоре на разные теоретические концепции;

– феноменальный опыт как основа классификации событий, причем – индивидуальный. Вряд ли есть смысл создавать универсальные искусственные интеллекты;

– принципиальная возможность ошибки и связанная с этим рефлексия – обобщение прошлого опыта. При этом, конечно, необходимо запрограммировать степень возможной ошибки. Некоторые действия, связанные с предельным риском для существования человека и человечества, в целом должны быть принципиально исключены. Но и законы Азимова не всегда могут быть приняты как абсолютные ограничения, если мы, например, будем создавать роботов-полицейских, боевых или охранных роботов;

– возможность реакции на случайные события, анализ индивидуально неповторимых ситуаций;

– коммуникации – сеть машин, обладающих взаимными ожиданиями;

– аналог нравственных переживаний, связанный с оценками других. Для этого надо будет создать нечто вроде эмоционального центра, связанного с наслаждениями, возбуждением и торможением нервных реакций, беспокойством и успокоением, а в развитом варианте также способность к эмпатии.

Если мы будем создавать возможность поэтапного, постадийного развития искусственного интеллекта, надо будет создавать стимулы. А здесь без аналога приятного возбуждения (наслаждения) не обойтись.

На данный момент самым сложным вопросом при создании практически применимого искусственного интеллекта является точное различение ситуаций и возможность реакции на непредвиденные события. Возможно, при создании автоматически управляемых автомобилей выходом из этих затруднений будет запрещение ручного пилотирования на тех трассах, где машины будут водить роботы. Роботы способны реагировать и действовать однозначно. В этом смысле сложных непредвиденных ситуаций возникать не будет. Не говоря уже о том, что на таких трассах не может быть пешеходов, выскакивающих на дорогу, и других непредвиденных

препятствий (что на скоростных трассах во многих странах уже практически реализовано).

Завершить эту статью хотелось бы ответом на вопрос об извечных опасениях относительно войны машин против человечества, полного вытеснения киборгами и роботами живых людей, имеющих биологическую природу и т. д. Я полагаю, что все эти опасения совершенно беспочвенны. У машин не больше оснований начать войну против людей, чем у самих людей начать войну друг против друга.

Что же касается вытеснения искусственным интеллектом естественного, то это тоже маловероятно. Даже если когда-то будет создано нечто подобное цивилизации роботов, они, скорее, станут использовать естественно развивающуюся цивилизацию как ресурс для совершенствования собственных способностей, культурных навыков, имитации тех способностей людей, которыми технические системы еще долго будут обделены, а возможно, и никогда не смогут их полностью воспроизвести, ведь с имитацией некоторых способностей людей они приобретут и те недостатки биологического носителя жизни, которые в технических системах как раз пытаются преодолеть.

Так что более вероятный сценарий – это параллельное развитие двух цивилизаций: построенной на естественном биологическом носителе и цивилизации технической, в чем-то – более совершенной, но неспособной полностью заменить естественно развивающуюся цивилизацию.

Литература

Васильев В. В. Трудная проблема сознания. М. : Прогресс-Традиция, 2009.

Гальперин П. Я. Введение в психологию. М. : Университет, 1999.

Гоулман Д. Эмоциональный интеллект. М. : АСТ: АСТ МОСКВА; Владимир : ВКТ, 2009.

Дубровский Д. И. Проблема свободы воли и современная нейронаука // Журнал высшей нервной деятельности. 2017. Т. 67. № 6. С. 739–754.

Иваницкий А. М. Наука о мозге на пути к решению проблемы сознания // Вестник Российской академии наук. 2010. Т. 80. № 5–6. С. 447–455.

Карпов В. Э., Готовцев П. М., Ройзензон Г. В. К вопросу об этике и системах искусственного интеллекта // Философия и общество. 2018. № 2. С. 84–105.

Михайлов И. Ф. Искусственный интеллект и когнитивные науки: перспектива антирепрезентационализма // Философия искусственного интеллекта. Научные труды Всероссийской междисциплинарной конференции. МГУ, 17–18 марта 2016 г. / под ред. В. А. Лекторского, Д. И. Дубровского, Ю. А. Алексеева. М. : ИИНТЕЛЛ, 2017. С. 284–294.

Разин А. В. Тело человека как антропологический константа его общественного бытия // Философия и культура. 2011. № 10. С. 23–32.

Сёрль Дж. Рациональность в действии. М. : Прогресс-Традиция, 2004.

Фейербах Л. Эвдемонизм / Л. Фейербах // Соч.: в 2 т. Т. 1. М. : Наука, 1995.