А.Ю. Согомонов

УДК 174

Прикладная этика или предиктивная аналитика (размышление по поводу)

Аннотация. В статье рассматриваются проблемы формулирования этики искусственного интеллекта. Сегодня он активно используется во многих сферах жизнедеятельности человека, в том числе в таких традиционно нравственно чувствительных сферах, как образование, медицина, право, управление. И, судя по всему, масштабы его внедрения будут только увеличиваться, что актуализирует задачу выработки адекватных этико-прикладных подходов в его использовании. Сегодняшний тренд показывает устойчивый курс этико-прикладной мысли на социально-технологический прагматизм и конструирование так называемой этики предсказаний. В ее основании рационализация максимально возможного числа ситуаций столкновения машинных алгоритмов с этическими ценностями человечества, а также выработка собственной «машинной» этической стратегии.

Ключевые слова: цифровая этика, искусственный интеллект, этико-прикладная рефлексия, этика предсказаний.

Я не Спиноза какой-нибудь,

чтоб выделывать ногами кренделя.

А.П. Чехов «Свадьба»

Открытие искусственного интеллекта (далее: AI) уже более пятидесяти лет тому назад, было встречено в свое время, с огромным воодушевлением и породило большие надежды на будущее. По мере развития AI как технологии замещения человеческого разума страсти не только не поутихли, но и разгораются все с большей силой. Только по одному этому аффективному показателю вся передовая часть человечества сегодня разделена на две неравные половинки, вступившие в жесткий дискурсивный клинч по поводу активного внедрения AI в разные сферы современной жизни. Причем это не столько идейно-политическая борьба, как можно было ожидать, между сторонниками и противниками AI, сколько необычное противостояние инструменталистов и прикладников. В задачу первых входит программная разработка, в задачу вторых – включение AI в социальные и культурные контексты. И все чаще именно прикладники высказывают опасения и сомнения относительно адекватности масштаб-

ного внедрения AI в нашу жизнь, хотя и среди них немало тех, кто все еще уверен в возможности полного контроля над AI и открыто выступают против разворачивающейся цифровой паники и киберфобии. Похоже, что число скептиков растет активнее их оппонентов, хоть я и не уверен в точности этой количественной оценки. А знакомство с новейшими исследованиями, подобно недавно вышедшей книге Ника Бострома «Искусственный интеллект. Этапы. Угрозы. Стратегии» [1], даже самых заядлых циников способно переориентировать в сторону более вдумчивого отношения к проблеме и сдержанного энтузиазма. Недавно в эту поистине глобальную дискуссию стали включаться именитые «любители» – важные политики и авторитетные общественные деятели.

Если все же оставить в стороне операционные сложности и угрозы, а сконцентрироваться, прежде всего, на гуманитарно-профессиональной составляющей внедрения AI, то на передний план выйдут вопросы именно этико-прикладного характера. Кстати, именно с этих позиций выступают значимые публичные фигуры. Учитывая же тот факт, что AI сегодня довольно активно внедряется в медицине, правоприменении, образовании, управлении, медиа и коммуникациях, то несложно предположить, что эти поля цифрового прогресса станут ареной новых, в первую очередь моральных конфликтов и трений. В этих сферах использование AI не только весьма болезненно воспринимается действующими акторами, но и создает массовое переживание, что рано или поздно произойдет катастрофическое вытеснение человека из самого процесса принятия решений. И тем самым все будет переподчинено «умным» машинам. Впрочем, оставим в стороне эти публичные страхи и вернемся к более узкой проблематике цифровой этики.

Как избежать, а точнее даже – как не допустить новых нравственных конфликтов? Как остаться в привычных рамках мира человека социально действующего и морально ответственного? Как все-таки сохранить право выбора и контроль только за человеком? Как, наконец, запрограммировать AI на предотвращение конфликтов между собой, с миром людей и вещей, их гуманитарными смыслами и духовными исканиями? Одним словом, если на заре разработки AI на него возлагались великие планы по переустройству всего цивилизационного порядка, то сегодня, оказавшись наедине с ним и трезво оценивая его технологические выгоды и социальные последствия, человечество вынуждено параллельно решать исторически беспрецедентные философско-этические проблемы цивилизационного сосуществования разума и машины. И лишь в итоге окончательно определиться с тем, является ли AI угрозой для человечества или нет?

Мы отчетливо понимаем, что проигрываем машинам в когнитивном плане, не говоря уж об их возможностях хранения бесконечной информации. Здесь процесс передачи большей части наших действий и полномочий под контроль AI приобрел необратимый характер. И поэтому речь сегодня все чаще идет о том, как не допустить грубых ошибок в будущем? Как использовать имеющийся опыт для более тщательного и этически взвешенного программирования? Благо пока ситуация позволяет нам смягчить потенциальные конфликты и превентивно обезопаситься от новых коллизий.

Мировая литература на тему цифровой этики множится из года в год чуть ли не кратно. Разумеется, в основном это многочисленные международные англоязычные издания. На русском же языке подобные работы – большая редкость. А исследований, подытоживающих опыт, буквально считанные единицы. Одна из таких, я бы сказал, «продвинутых» книг была опубликована в прошлом году. Она весьма примечательна по нескольким обстоятельствам. Во-первых, книга носит обобщающий характер. Во-вторых, представляет собой творческую коллаборацию технолога и философа. В третьих, она чрезвычайно точно демонстрирует нам тренды в актуальном развитии цифровой этики. Оба автора книги «Прикладные проблемы внедрения этики искусственного интеллекта в России», Диана Гаспарян и Евгений Стырин, аффилированы к Высшей школе экономики, на издательской базе которой собственно книга и вышла [2]. В ней собран большой эмпирический материал, обработанный мастерски, хотя, как мне показалось, с ложной теоретической отсылкой. И поэтому, если рассматривать само это исследование как примечательный знаниевый «кейс», не сложно будет определить тот вектор, в котором развивается сегодня цифровая прикладная этика.

Всякий читатель, так или иначе погруженный в философские контексты фундаментальной и прикладной этики, буквально с первых страниц книги с удивлением обнаружит для себя принципиальную методологическую нестыковку. Вначале авторы предлагают интерпретировать AI как сложную программную систему, базирующуюся на антропоразмерном интеллекте и функционирующем автономно. К этому нет вопросов. Очевидно, именно так и следует понимать технологическую природу AI. Однако именно к ней, то есть к технологии, они предлагают применять процедуру этической экспертизы. Как это понимать? Читаем с недоумением: «Этическая экспертиза – тестирование технологии с точки зрения ее релевантности этическим нормам, причастности ценностям и нормативным предписаниям, а также психологической безопасности» [2, *3*].

Нет сомнений в том, что AI – технология особенная, но разве можно *технологию* тестировать нравственно-психологически? Ведь ни газ и ни закрытая камера по отдельности, и даже ни газовая камера сами по себе были аморальными и бесчеловечными технологиями. А ценностно-рациональный и, главное, вполне осознанный тип их государственно-политической эксплуатации и ее этическая рационализация фашистскими идеологами. Вопрос о возможности приложения этической оптики к технологии, конечно же, риторический. Объектами прикладной этической экспертизы могут выступать только люди, создающие эти программирующие системы, а также вовлеченные в их орбиту пользователи или рядовые бенефициарии от их применения. Казалось бы, совершенно банальное утверждение, ведь вы же не можете применить критерий добра и зла к самой цифровой технологии, в противном случае вы скатываетесь на уровень бытового шельмования всего виртуально – нового. А если система некорректно спрограммирована или дает сбои, да еще с негативными и долгосрочными последствиями, то этическую претензию следует выдвигать не существующей системе, как таковой, а к тем, кто ее задумал, практически реализовал, получил от ее применения бонусы.

И все-таки я готов допустить, что такой «причудливый» этико-прикладной подход приемлем, если понимать «этическую экспертизу» как метафору, но при этом иметь в виду нечто совершенно иное. Авторы книги выделяют две взаимосвязанные проблемы использования AI с точки зрения этики. Во-первых, согласование работы AI с существующими в обществе ценностными установками. Во-вторых, формализацию данных ценностных установок [2, *9*]. На первый взгляд все кажется логичным. Однако опять же получается, что авторы допускают саму возможность вынесения этической претензии непосредственно технологическому продукту, как если бы он сам принимал этически взвешенные решения по им придуманному модулю. Когда-то, может быть, до этого и дойдет дело, но в книге речь идет о сегодняшнем дне, а пока AI полностью контролируется человеком.

Но все-таки, с каким этическим кодексом должен соотноситься AI, если такого не существует в современном мире, а ценностный плюрализм воспринимается человечеством как вполне нормальное цивилизационное состояние? И, кроме всего прочего, что в цифровом формате может означать сама технология по «выработке этического решения»? На подобные вопросы невозможно ответить, если в основу этической экспертизы будет заложена технология.

Причина подобного когнитивного диссонанса кроется в том, что программист AI чаще всего тоже воспринимается как *чистая функция*, оторванная от реального и взаимообразно заменяемого челове-

ка, у которого не может быть изначально своего этического резона. Он действует строго прагматически, как если бы сам был запрограммирован только на цифровое программирование. Поэтому многие прикладники полагают, чтобы избежать каких-то последующих культурных или нравственных сбоев, в самого программиста необходимо «завести» этическую программу.

Да, именно так: настроить его на введение цифровых табу любых гомофобных или неадекватных сегодняшней культурной политике рацио-подобных действий AI. Иными словами: настроить саму технологию через настройку программиста на системное предупреждение потенциальных этических конфликтов. И поскольку у «живого» программиста может быть снижена нравственная чувствительность по ряду волнующих общество вопросов, следовательно, набор таких табу должен быть спущен «сверху-вниз» – совершенно *директивно*. Чем, собственно, и занимаются разные правительственные или нанятые властью подведомственные комиссии. А именно – *предиктивным анализом* (диагнозом и прогнозом). Они озадачены решением непростого «квеста» – как просчитать все возможные варианты и выставить адекватные цифровые барьеры на пути их возникновения.

Похоже, что именно так и рассуждают сегодня многие международные эксперты в области AI и цифровой культуры в целом. Для них важно и актуально решать множество частных вопросов формализации ценностей и норм, культурно-политических представлений, чтобы минимизировать потенциальную конфликтность, которая неизбежно складывается при соприкосновении машинного и человеческого разума, а точнее: между возможностями первого и ожиданиями второго. Авторы упомянутой книги прекрасно владеют этой литературой, умело пользуются имеющимися мировыми наработками. Однако, как мне кажется, с прикладной этикой такой подход имеет мало общего. И весьма показательно, что ни в книге, ни в ее библиографии нет отсылок ни на одну из существующих традиций прикладной этики. Так же как не упомянут ни один из известных исследователей того, что мы собирательно именуем “practical ethics”, как если бы прикладная этика вообще не имела бы своей предыстории и долгой философской проработанности, а каждый раз конструировалась бы «с нуля» и под ad hoc задачи.

Разумеется, цифровое пространство – совершенно новый, но вполне привычный, социокультурный контекст для этика, профессионально занимающегося прикладными темами. Многое в этом пространстве зависит, конечно же, от миллиардов рядовых пользователей, но когда речь заходит о сложных программных системах, моральная ответственность переносится на плечи разработчиков. Как я

уже предположил, требования к ним могут предъявляться лишь извне, ибо только так внешняя среда, регулируемая прикладниками, пытается себя морально обезопасить от угроз гиперцифровизации. Если есть заданное человеком целеполагание, то и цифровые средства должны быть морально ратифицированы. И поскольку глупо предполагать, что программист может предвосхитить все возможные будущие конфликтные ситуации, то сообщество пользователей и настраивает его на хорошо известный нам этический алгоритм: *не навреди*!

Этот алгоритм, очевидно, самый простой, легко реализуемый и отработанный веками, поэтому он столь часто востребован в разных моделях профессиональной этики. И, действительно, мы видим, что курс в большей степени к разработчикам AI применяется прямолинейный этико-*рестриктивный* подход. Им вменяется в обязанность, например, не допускать дискриминационных практик или не провоцировать цифровую агрессию, не использовать данные во вред человеку и т.д. Хотя чаще всего сообщества или власти пекутся в первую очередь о защите своей индивидуальной приватности и свободы. А в этом деле разработчик окружен множеством «запретов», которые, по массовому заблуждению почему-то традиционно именуются «этико-прикладными». Но замечу, все они не требуют ни этической рефлексии, ни решения сложных моральных дилемм, – а лишь четкого следования инструкциям. Сам же программист зачастую и не участвует в их разработке.

Между тем подобные внутренние рестрикты проистекают, прежде всего, из самой природы профессиональной деятельности программистов. Внутренние табу на пикантные темы, скабрезные акценты или намеки, ненормативную лексику, на использование чужой информации, диффамацию, на тактику недоговаривания или сокрытия информации, и т.п. – все это элементы их профессиональной культуры, поддержанные цеховыми сообществами и логически вытекающие из смыслов профессионального разделения труда в современном «обществе знаний». Да, они имеют определенную ценностную подоплеку. Безусловно, коррелируют с репрезентативным типом культуры (что принято, а что отвергается в большом обществе). Но чаще всего они вводятся «директивно» во избежание социальных конфликтов, а поэтому регулируются правом, политикой и лишь отчасти цеховыми нормами. Могут ли они быть успешно кодифицированы? Не думаю, слишком уж быстротечны социокультурные перемены в мире. Да и нет в этом особой необходимости, проще кадастр рестриктов постоянно дополнять и редактировать.

Все это собственно и составляет особый *профессиональный этикет* программиста и первичных пользователей AI. Это уже ближе к существу прикладной этики, но по-прежнему логически не проистекает ни из одной известной этико-философской парадигмы. Можно, конечно же, утверждать, что предиктивный подход базируется на этическом *консеквенциализме*. Но следует помнить, что в его орбите большая группа моральных теорий, соотносящих нравственное действие с их последствиями. Консеквенциализм своей теоретической логикой роднит такие этически полярные теории, как утилитаризм и эвдемонизм, гедонизм и разумный эгоизм. Впрочем, это уже довольно высокий уровень философского теоретизирования. Для цифровой этики, как показывает мировой опыт, пока достаточно *простого социально-нравственного прогноза*. И поскольку в основу, как полагают многие исследователи, в том числе и авторы анализируемой монографии, помещен «рациональный агент», следовательно, он оперирует критерием полезности, на чем, собственно. и следует базировать всю цифровую *этику предсказуемости*. Показательно, что эта философская сложность осознается почти всеми акторами этико-цифрового взаимодействия, но ее разрешение упорно адресуется будущим поколениям.

А пока предиктивная аналитика в AI исходит из нескольких простых этических посылов: (1) не навредить действием или бездействием; (2) безоговорочно подчиняться приказам человека; (3) стремиться сохранить себя, но обязательно с учетом первых двух условий [2, *11*]. Из этого вытекает, что программисты и прикладники должны следовать двум фундаментальным правилам: ориентироваться на этический разум человека, но при этом формализовать свои собственные этические стратегии. Круг замкнулся. Логическая нить разорвана. Самому AI переданы этические «верительные грамоты». Впрочем, в этой ситуации философская рефлексия сложившегося морального тупика становится еще интереснее, интригующей и очень азартной «игрой» двух разумов – человеческого и машинного. Но, ведь, это – излюбленная тема писателей старшего поколения мировой научной фантастики.

Post Scriptum

В замечательном чеховском водевиле «Свадьба» жених, Апломбов, в ответ на предложение станцевать, произносит ставшую впоследствии крылатой фразу «Я не Спиноза какой-нибудь, чтоб выделывать ногами кренделя». Имел ли он в виду этика и философа Бенедикта Спинозу? Конечно же, нет. Подразумевался известный в последней трети XIX века испанский танцовщик Леоне Эспинозе, ко-

торый, в частности, с триумфом гастролировал в Большом театре. Для гротеска Чехов упрощает его фамилию до узнаваемого в просвещенных кругах варианта созвучного имени, подчеркивая неграмотность и необразованность московских обывателей.

Не хочу остаться неправильно понятым, но своим эпиграфом я не хотел никого задеть, тем более, обидеть, а намеревался лишь подчеркнуть, насколько далеко предиктивная аналитика отстоит от подлинной цифровой прикладной этики. Но, увы, вынужден признать, что таков моральный дух нашего времени.

Список литературы

1. *Бостром Н.* Искусственный интеллект. Этапы. Угрозы. Стратегии. Москва: Манн, Иванов и Фарбер, / Ник Бостром ; пер. с англ. С. Филина. М. : Манн, Иванов и Фербер, 2016. 490 с.

2. *Гаспарян Д.Э., Стырин Е.М.* Прикладные проблемы внедрения этики искусственного интеллекта в России: отраслевой анализ и судебная система. М.: Высшая школа экономики, 2020.