

### НЕЛИНЕЙНОЕ ПРОСТРАНСТВО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ФИЛОСОФСКО-МИРОВОЗЗРЕНЧЕСКОЕ ОСМЫСЛЕНИЕ

*В статье автор с позиции методологии нелинейного прогнозирования раскрывает перспективы и риски возникновения искусственного интеллекта. Анализируются причины как оптимистичных прогнозов, так и комплекс угроз, исходящих от «горизонтов» сингулярности, обусловленной созданием машинного разума. Особое внимание уделяется концепции дружественной разумной машины в условиях интеллектуального взрыва.*

*Осуществлена попытка философской рефлексии нескольких вариантов пагубного отказа, а именно: порочной реализации и инфраструктурной избыточности.*

**Ключевые слова:** *искусственный интеллект, сингулярность, нелинейность, антропоморфизация, «бог в ящике», тест Тьюринга, Три закона робототехники, «черный ящик», генетические алгоритмы.*

Одним из основных факторов, влияющих на относительно устойчивое развитие человека и социальных систем, выступает преобразовательная деятельность, которая по мере усложнения форм, средств и способов инноваций постепенно превращалась в технологии. Их негэнтропийный потенциал и техническая реализация в разных сферах жизнедеятельности с древних времен оказывали существенное влияние на формирование научной рациональности современного типа, что несомненно коррелирует с возникновением информационного общества с характерным для него усилением роли статистических закономерностей и фактора нелинейности. С середины XX века наука стала играть ведущую роль в системе общественного производства, а наукоемкие технологии стали претендовать на роль нового аттрактора, детерминирующего новые гносеологические векторы и аксиологические горизонты развития социальных систем. Но чем сложнее социальная система, тем более она подвержена влиянию стохастических факторов, создающих нелинейное пространство для дальнейших путей развития, формирующихся в результате актуализации фазовых переходов – вынужденных ответов неравновесной структуры на угрозу снижения устойчивости.

В этом контексте прогресс приобретает характер не самоцели, не самообусловленной ценности, а выступает в роли способа сохранения

относительно сложной целостности. Пути достижения такого состояния нелинейны, так как заранее просчитать их количество, степень детерминации и опасности на какой-либо период времени возможным не представляется. Будущее может стать как и «лучше» настоящего по строго обозначенным параметрам, так и «хуже» по другим параметрам и технологии, особенно наукоемкие, которые могут стать причиной экзистенциального кризиса глобальных масштабов, играют в этом смысле не последнюю роль. Решение одних противоречий «запускает» нелинейную цепь множества других, новых, еще более неоднозначных проблем. В дальнейшем это обуславливает возникновение векторов эволюционных изменений: от более стохастических («естественных») к менее вероятным состояниям. В соответствии с нелинейной моделью прогресс как «удаление от природной ниши» означает восстановление относительной устойчивости системы на все более высоком уровне неравновесия.

В свете вышесказанного необходимо подчеркнуть, что «технологизация» и «прогрессизация» общества представляют собой неразрывное единство, но, увы, не всегда органическое, как бы того хотелось субъекту исторического процесса. На сегодняшний день ученые и философы признают, что именно наукоемкие технологии тот фактор и аттрактор, которые детерминируют развитие социальных систем, а наука давно превратилась в могучую производительную силу, которая время от времени, игнорируя гуманистические идеалы, загоняет себя в циничное поле, описываемое как «цель оправдывает средства».

Иллюстрацией подобного сценария может служить классический роман М. Шелли «Франкенштейн, или Современный Прометей» [14], лейтмотивом которого является идея о том, что наука, как и научный метод, без аксиологических оснований и морально-этических мотиваций рискует превратиться в орудие мясника, пытающимся примерять покровы высоких целей на высохший остов личных амбиций и стремлений. Любая наукоемкая технология в состоянии как запустить нелинейное пространство новых рисков, так и сама по себе является потенциальным источником угрозы. И в этом смысле, на наш взгляд, особый интерес представляет проблема искусственного интеллекта.

Создание думающей машины, как и она сама является процессом с обостренным режимом протекания, одной из отличительных сторон которого является разрастание микрофлуктуаций, вследствие чего «мышь родит гору». В нелинейной и неустойчивой социальной системе в планетарном масштабе это может иметь катастрофические последствия, вопреки многочисленным мнениям апологетов этой технологии. Ведь любую техническую инновацию можно использовать как во зло, так и во благо. И история учит, что, как правило, первое является закономерным следствием второго или же эти ипостаси выступают двумя сторонами одной медали. Риски и степень

ответственности вырастают в сотни раз, когда речь заходит о технологии, детище которой будет иметь возможность совершенствовать само себя. А одним из вариантов реализации сценария творения машинного разума может стать интеллектуальный взрыв, вследствие которого сверхразумный агент, вырвавшийся в мировую сеть начнет стремительно экспоненциально обучаться и в течение самого короткого интервала времени выйдет из под контроля недалёковидного человечества и, возможно, идентифицировав вид *homo sapiens* как конкурирующий, использует его в качестве ресурса, или попросту уничтожит.

Но прежде чем рассмотреть возможные негативные флуктуации, которые могут возникнуть как в процессе, так и в результате возникновения искусственного интеллекта, проанализируем причины, вследствие которых несмотря на очевидные риски, человечество все же жаждет появления в этом мире универсального всемогущего раба, который с легкостью может превратиться в искусственного бога. Почему же многие представители научного сообщества не замечают или предпочитают игнорировать возможность такой трансформации?

Прежде всего, это ошибка завышенного оптимизма, обусловленного многими факторами, спектр которых довольно широк от психологических до экономических. Ведь искусственный разум, по мнению многих, осуществит или значительно ускорит коренной качественный переворот во многих научных и технологических областях. Безусловно, что появление многих научных открытий приведенных ниже может оказаться невозможным благодаря только одному типу наукоемких технологий: будь то нанотехнологии или разумная машина. Как и не представляется возможным развитие и возникновение разумного агента в рамках исследований одной научной области. Такая ситуация обусловлена корреляционным характером научных революций, большая часть которых началась еще в прошлом веке в сфере информационных, коммуникационных и биологических технологий. Скачкообразное развитие когнитивных наук в последние десятилетия позволили говорить экспертам о приближении новой научной революции. Особый интерес и значение представляет корреляционное взаимопроникновение именно информационных, биологических, нано и когнитивных технологий, получивший название NBIC-конвергенция.

Искусственный интеллект играет в такой синергетической системе одну из ведущих ролей. Например: искусственный суперинтеллект в перспективе может создать самовоспроизводящиеся машины молекулярной сборки, которые получили название наноассемблеры, дав обещание человечеству, что их использование принесет исключительно пользу. Объективно же в дальнейшем суперинтеллект вместо того, чтобы преобразовывать песок в золото, начнет превращать материалы в программированную материю, которую затем он сможет превращать во что угодно: от компьютерных

процессоров до космических мегамостов для колонизации Вселенной. Но все это выглядит пока на уровне сюжета для художественного произведения. Вот только некоторые из выгодных перспектив, открывающихся перед нами после возникновения машинного сверхума:

- способность влиять на генетику органических видов;
- целенаправленное воздействие на структуру органов и тканей; создание органических протезов и донорских органов;
- предупреждение и лечение всех заболеваний на начальной стадии;
- существенное замедление процесса старения;
- прогрессивная революция интеллектуальных возможностей человека благодаря киборгизации и слияния с компьютерным интерфейсом;
- углубление и расширение виртуализации социального и личностного пространства.

В более далеком будущем возможен переход к принципиально новым ценностно-смысловым паттернам бытия:

- целенаправленное изменение структуры материи;
- стирание грани между искусственным и естественным, природным и социальным;
- нивелирование дихотомии между материальным и идеальным в онтологическом и гносеологическом аспектах;
- трансформация представлений о живом и мертвом, благодаря достижению ревитализации.

В свете вышесказанного становится понятным тот факт, что человечество в тщеславной погоне за лавровой веткой первенства мало перед чем остановится, тем более что далеко не последнюю роль в такой гонке играет финансовая сторона дела. Ведь ни для кого не секрет, что одним из основополагающих составляющих производительных сил является прибыль. А искусственный интеллект, возможно, не только ускорит этот процесс, но и сам станет новым видом способа производства. Вопрос только в том будет ли в этот процесс включен сам человек?

Бескрайний оптимизм подпитывает также и комплекс сложных психологических причин, совокупность которых получило название антропоморфизм – экстраполяция человеческих качеств на возможности сверхразумного агента. «Чтобы разговор о сверхразуме получился осмысленным, необходимо заранее осознать, что сверхразум – это не просто еще одно техническое достижение, еще одно орудие, которое увеличит человеческие возможности. Сверхразум – нечто принципиально иное. Этот момент нужно всячески подчеркивать, поскольку антропоморфизация сверхума – плодороднейшая почва для заблуждений» [4, 224].

Искусственный интеллект – качественно новое в технологическом понимании, по словам Н. Бострома, в силу того, что его возникновение изменит

суть прогресса, исключив из этого процесса человека. Машинный разум будет ни на что не похож. Будучи результатом человеческой деятельности, он будет стремиться к саморазвитию по одному только ему известному сценарию, который не зависит от человека. У него не будет личностных мотивов, потому что не будет человеческой сущности.

Итак, антропоморфизация машины является источником ошибочных представлений о том, что можно создавать безопасные машины, наделенные разумом и что это рано или поздно не приведет к катастрофе. Отчасти такая вера в «дружелюбных роботов» основана на трех законах робототехники, которые почему-то широкая публика воспринимает как данность и слепо полагается на них. Тогда как на самом деле эти три закона в рассказе «Хоровод» [2, 122–164] были предложены фантастом Айзеком Азимовым. По сюжету они намертво упрочены в нейронные сети «позитронного» мозга роботов:

1. Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред.
2. Робот должен повиноваться командам человека, если эти команды не противоречат Первому закону.
3. Робот должен заботиться о своей безопасности до тех пор, пока это не противоречит Первому и Второму законам.

По сути данные законы являются своеобразной адаптированной интерпретацией заповеди «Не убий», христианского понимания того, что к греху можно прийти, как совершая поступки, так и «сидя на берегу реки», клятвы Гиппократов. Важно другое: в литературном наследии А. Азимова эти законы так или иначе дают сбой. Да и в большинстве своем творчество известного фантаста направлено на то, чтобы продемонстрировать насколько самонадеян человек в своих попытках контролировать сущности, превосходящие его практически по все параметрам. В вышеупомянутом рассказе «Хоровод» геологическая экспедиция на Марсе доверила роботу перевозку токсичного для него вещества. Но миссия оказывается невыполненной в силу того, что робот попадает в ловушку обратных связей второго и третьего законов. Он бы так и остался в этом лабиринте сомнений, не приди ему на помощь геологи. Рассказы А. Азимова изобилуют сюжетами, в которых достичь цели можно только в обход трех законов (то есть изначально выдвигается постулат о сомнительном характере этих «трех заповедей»).

Оригинальные и захватывающие сюжеты в художественной литературе и безопасность в реальном мире – вещи разные. Многообразие и сложная нелинейность реального мира более непредсказуема и стохастична нежели любой вымысел гения. Случайность в литературе фиксирована и является «узником» пределов данного семантического поля (рассказа, повести, романа). Случайность же в реальном мире, с одной стороны, «вынашивает» следующий миг, а с другой, предыдущим мигом может быть «выношена» и растворена в

потоке событий, не поддается фиксации и контролю до реализации из потенции и бывает трудно выявляема и описываема после. Особенно когда мы имеем дело с объектами возникновения и поведения которых не поддается стандартным методам. К их числу несомненно относится и появление искусственного интеллекта. Несомненно, что вышеупомянутых трех законов недостаточно. Прежде всего, дело в их недостаточно четкой формулировке и расплывчатой семантике понятий, которыми данные законы оперируют. Так, привычная для нас сегодня дихотомия между живым (человеком) и неживым (роботом), по всей вероятности будет размыта, когда наука научится усиливать тело и мозг человека при помощи, например, компьютерного интерфейса. И что тогда будет называться человеком? Аморфны и другие понятия законов: «вред», «безопасность», «команды».

Справедливости ради стоит сказать, что А. Азимов позже дополнил свои три закона своего рода «приквелом», нулевым законом, который запрещает роботам вредить человечеству в глобальном масштабе. Но качественно это проблему не разрешает. Эти законы являются монументальным обелиском великому гению автора художественной прозы и самому жанру, но они являются наиболее цитируемыми, когда осуществляются попытки выстроить стратегию будущего сосуществования вида *homo sapiens* и сверхразумного агента. Возникает очевидный, но пугающий вопрос: неужели законы Три закона это все, что мы имеем на сегодня?

В то время как «темные стороны» стороны роботизации уже дают о себе знать и реальное положение дел заставляет задуматься. В мире более чем в пятидесяти странах на сегодня ведутся разработки милитаризованных думающих машин. И уже есть первые проявления некачественного программирования, повлекшие за собой страшные последствия: во время военных действий на Ближнем Востоке боевые дроны, оснащенные автоматическим оружием, после того как использовали его против своих, были выведены из строя. Сбой в работе программы роботизированной зенитной пушки повлек за собой гибель девяти и тяжелые ранения пятнадцати солдат в 2007 г. на африканском континенте [13, 128]. Очень симптоматичен тот факт, что продолжительность этого инцидента была менее секунды. Таким образом, возникает мнение, что дискуссии об этике и о технических инновациях идут на разных планетах.

Концепция техно-гуманитарного баланса постулирует необходимость наличия соответствующего духовного уровня цивилизации для адекватной «притирки» к технологическим инновациям разной степени сложности [8]. Технология искусственного разума, как и деление ядер, как и многие другие, – технология двойного назначения. Расщепление ядра является источником освещения городов, тем самым питая цивилизацию, но также может ее и испепелить. Трагедия Хиросимы и Нагасаки является чудовищным примером того, как «технологическое семя» упало в «непаханое гуманитарное

поле» цивилизации. В истории таких примеров более чем достаточно. Если скачок от искусственного интеллекта человеческого уровня к искусственному суперинтеллекту произойдет по сценарию стремительного интеллектуального взрыва, то человечеству еще раз придется столкнуться с угрозой планетарного масштаба. Вопрос в том – хватит ли нам мудрости и пережить очередной «вызов».

Второе искажение в осмыслении рисков возникло вследствие популярности темы искусственного разума в мире развлечений. Как правило, обсуждение опасностей, коими нам сулит искусственный интеллект происходят в отрыве от контекста, с подменой понятий и не отличаются профессиональной оценкой и аналитической глубиной. Разумеется, что в академических научных и философских кругах данную проблему не обходят стороной, но зачастую ей не уделяют должного внимания. Большинство специалистов считают создание искусственного суперразума только делом времени и находятся в приятной эйфории от тех перспектив, которые вырисовываются на горизонте.

Как только речь заходит о неприятных прогнозах, многие представители СМИ (технические журналисты, блогеры, редакторы) рефлексивно не принимают их всерьез. Подобная реакция детерминирована обычным нежеланием добраться до сути проблемы, что отражено в несостоятельности выдвигаемых ими контраргументов. Большинству обывателей, для которых рассчитана продукция СМИ, проблема искусственного разумного агента кажется чем-то далеким и абстрактным, а следовательно она не является для журналистов делом первостепенной важности. Более привлекательной для технической журналистики является сфера развлечений: плазмы на квантовых точках, более емкостные носители информации и другие новинки программного рынка. Кинематограф подарил нам десятки апокалиптических финалов, которые затеял взбунтовавшийся машинный интеллект («Из машины», «Космическая одиссея 2000 года», «Терминатор», «Матрица», «Я-Робот» и другие), что породило эффект «комфортного абстрагирования» от всерьез нависшей опасности. Эти «техномонстры» доставили зрителю столько приятных часов «безопасного щекотания нервов», что в конце концов в подсознании большинства закрепился эффект надуманной опасности. Ведь, по сути, увиденное на экране – результат чьего-то воображения и не имеет к реальному миру никакого отношения. Иначе говоря, растиражированность художественных интерпретаций опасностей возникновения машинного разума (кино, литература) сыграла роль своеобразной прививки, после которой возможность серьезного осмысления и анализа катастрофических рисков стала практически невозможной. Не вызывает опасений и созданная в начале XXI в. роботизированная имитация известного фантаста Филиппа К. Дика. С роботизированной машиной можно обсуждать творчество автора. «Радушная имитация имитирует радушие». Как в этом контексте не вспомнить: «Самая большая хитрость дьявола – заставить всех поверить, что его не существует».

Когнитивное искажение является следущей причиной, порождающей заблуждение в результате которого возникновение разумной машины редко несет в себе деструктивное начало для человека. И на этой проблеме мы остановимся более подробно.

Дело в том, что человек в процессе своей деятельности принимает решения, руководствуясь, в большей степени, собственным опытом, который зачастую не подвергается рефлексии, а имеет, скорее, стихийно-обыденный уровень. Объективно же всегда остается вероятность того, что исследователь учел не все входные данные эксперимента и флуктуация одного из непросчитанных элементов окажет качественное влияние на конечный результат. А в результате того, что всякий опыт имеет ограниченный характер, в гносеологическом аспекте мы неизбежно сталкиваемся с проблемой неопределенности заблуждения.

Есть вероятность того, что отсутствие контакта с объектом в ретроспективном аспекте делает невозможным построения каузальных связей между возникновением суперразумного агента и исчезновением цивилизации в глобальном масштабе. На сегодня ни один человек не имел дело ни с одним серьезным событием, имеющим трагические последствия, к началу и к основе которого так или иначе был бы причастен искусственный разумный агент. Искусственный интеллект пока не расценивается как источник экзистенциальной угрозы. Создается мнение, для того, чтобы человечеству оценить всевозможные риски, причиной которых может быть искусственный интеллект, ему нужно оказаться на грани жизни и смерти. Столкновение с разумом, превосходящим наш, не будет иметь ничего общего с ограниченными во времени и пространстве террористическими атаками, ядерными зимами и другими катастрофами, имеющими эндо-техногенный характер. В результате актуализации сценария, вырвавшегося неконтролируемого искусственного суперразума, глобальная человеческая раса, по всей видимости, оставит после себя лишь след в виде сказочных историй, которые будут рассказывать роботы своим детям перед сном.

Освободившийся из «ящика» искусственный интеллект имеет еще одно принципиально качественное отличие от техногенных катастроф. На сегодняшний день человечество сталкивалось только с теми событиями, негативные последствия которых были устранимы. В случае же «взбунтовавшегося демона в машине» имеет место самосовершенствующаяся и самовоспроизводящаяся умная программа, которая потенциально может существовать вечно. И само собой разумеется у этой сознающей себя системы будут базовые потребности. Согласно С. Омохундро их четыре: эффективность, самозащита, ресурсы, творчество [9], что дублирует потребности человека и история учит к чему, как правило приводят попытки эгоистичного их удовлетворения. Насколько потенциально опасна каждая из них на сегодня сказать сложно, но если угрозу от первых трех мы можем представить в общих



чертах, то потребность в творчестве при первом осмыслении как бы и не несет никаких рисков, а скорее наоборот. Но именно способность к творчеству является одним из основных условий существования и функционирования думающей машины. Именно оно будет обуславливать возможность самоидентификации машины и осознания ею себя как Я-бытие, а следовательно и личностной автономии. Здесь и возникает существенная проблема: создать искусственный интеллект человеческого уровня – значит наделить машину сознанием, одним из основных характеристик которого является способность к целеполаганию, созданию чего-то нового, что невозможно без относительной самостоятельности функционирования программного обеспечения. Это неизбежно приведет к относительной потере контроля человека над своим детищем, а в условиях непрерывного самосовершенствования машины к полной автономии функционирования разумного агента. Иными словами наличие творчества у сознающей себя машины является его детерминирующей характеристикой, которая в то же время и является ключом к его свободе и этот ключ, так или иначе, мы вручим ему сами. В наших ли силах будет противостоять угрозе, причина которой соотносится с нами в умственном развитии примерно так же как мы и дождевой червь? И сможем ли мы совладать с устранением последствий катастрофы, однажды начавшейся, но длящейся бесконечно?

Также одной из причин некорректной оценки искусственного разума как начала обратного отсчета существования цивилизации является тот факт, что феномен искусственного разумного агента получил свое отражение в другом неоднозначном явлении, получившим название сингулярность.

Понятие «сингулярность» широко используется в научных и философских кругах и имеет различную контекстуально-семантическую окраску. В нашей статье употребление данного понятия имеет значение, вложенный в него Р. Курцвейлом, который понимает под этим некую меру перехода количества в качество (около 2045 г.), при котором скачок технического прогресса принципиально трансформирует бытие человека. Разум перестанет быть исключительно человеческой прерогативой и, по мнению Р. Курцвейла, станет, в большей степени, компьютеризированным, что сделает его в разы более мощным, чем на сегодняшний день. Автор этой концепции настроен оптимистично считая, что такой переход искоренит из человеческого существования такие негативные явления как голод, болезни, а, возможно, в перспективе человек обретет и вечную жизнь [5, 136–147].

По мнению Р. Курцвейла, искусственный разумный агент окажет определяющее влияние на принципиально новый качественный переход мировой цивилизации, но, как уже было обозначено, это повлечет за собой развитие и смежных областей наукоемких технологий – нано [5, 47]. Прогнозы специалистов указывают на то, что интеллектуальный взрыв и последующее появление искусственного сверхразума неизбежно повлечет за собой резкий

скачок в розвитку нанотехнологій. Многие эксперты считают, что приоритет в возникновении должен принадлежать именно искусственный суперразум в силу того, что нанотехнологии слишком стохастический инструмент, контроль над которым может оказаться для человека невозможным. Становится понятным почему оптимизм в контексте сингулярности исходит именно из области разработок нанотехнологий, а не искусственного разума. Инженерия на атомарном уровне, возможно в будущем, даст шанс человеку обмануть смерть.

Но наряду с позитивными прогнозами существует и «ложка дегтя». Например, наниты, способные к самовоспроизводству, превратят окружающий нас мир в так называемую «серую слизь». Сценарий «серой слизи» – один из вероятных апокалиптических, и эта проблема – «темная территория» пространства нанотехнологий. Размышляя о зловещей стороне нанотехнологий, многие упускают из виду принципиальную непредсказуемость, а следовательно, угрозу, исходящую из создания искусственного всемогущего помощника, если тот будет прогрессировать по сценарию стремительного самосовершенствования, в ходе которого машины, превосходящие по всем параметрам человека, выйдут из под контроля и уничтожат глобальную цивилизацию.

В контексте вышесказанного, в результате анализа исследовательского опыта по проблеме оптимистического отношения к разумным машинам, условно можно обозначить два подхода. Теоретико-мировоззренческие горизонты первого варианта задают исследования в духе работ Р. Курцвейла [5], в которых будущее антропной компоненты нашей планеты мыслится как крайне позитивное. Негативные флуктуации в магистральном течении событий в таком осмыслении подавлялись бы оптимизмом.

Второй подход основан на работах в стиле Д. Стибела [10], который подходит к осмыслению этой проблемы сквозь призму прагматического практицизма. Авторы концепции рассматривают мировую сеть как все более усложняющийся мозг с миллионами связей и хорош тот делец, который с оптимальным эффектом для себя сумеет лавировать в пространстве интернет-тенденций и извлекать оттуда максимум прибыли.

Большинство экспертов, задействованных в сфере наукоемких технологий не рассматривают более скептический третий вариант, суть которого заключается в том, что финальным этапом разработок разумных агентов, а затем и машин превосходящих по интеллекту человеческий, станет не гармоничное единение искусственного интеллекта с человеком, а превращением последнего, по всей видимости, в сырье для «триумфального шествия» по мирозданию нового субъекта.

Взаимодействие «традиционного» разума человека с суперразумом машины тождественно расширению сфер влияния технологической западной цивилизации на общества аграрного типа, которая или ассимилировала

последние, или превратила в свой ресурс. Для примера достаточно вспомнить следующие антагонизмы и чем они завершились: Колумб – Тиано, Писарро – Инки, европейцы – американские индейцы.

Что далее – homo sapiens против машинного суперинтеллекта?

Вполне вероятно, теоретики в области наукоемких технологий уже осмыслили все «темные стороны» искусственного интеллекта, но проанализировав все «за» и «против», пришли к выводу – цель оправдывает средства. Или же понимают, что точка невозврата пройдена и принимают неизбежность любого исхода, постулируя невозможность что-либо предотвратить. Выдвинута идея, согласно которой человек сможет осмыслить и открыть способы защиты от прогрессирующего машинного разума только в процессе интеграции данного феномена в наше бытие. Это взаимодействие будет происходить постепенно и у человечества будет шанс внедрить разумному агенту «алгоритмы послушания» и создать дружественный суперинтеллект [3, 75]. Созвучно этой идее и понимание невозможности осознания всех истинных рисков искусственного разума из настоящего времени с привычными для нас паттернами бытия [5, 34]. Иными словами, обуздание дикого мустанга на диком Западе не приводит к пониманию специфики управления гоночным автомобилем на горной дороге.

Таким образом, проблема обозначенного подхода заключается в том, что если угроза и признается всерьез и рассматривается объективно, то исходит она от непредсказуемости «темного пространства» интеллектуального взрыва и суперинтеллекта, тогда как риски и угрозы от промежуточных этапов создания искусственного интеллекта анализируются недостаточно серьезно или же не берутся в расчёт вовсе. Иначе говоря, грозная львица, конечно, является источником опасности для туриста в саване, но нельзя забывать о потенциальной угрозе, исходящей от ее милого львенка. Концептуальные построения градуалистов, в той или иной степени, конституируют постепенный характер скачка от искусственного интеллекта уровня человека к искусственному суперразуму, временной интервал которого может растянуться от нескольких лет до десятилетий. Этот прогноз позволяет рассчитывать человеку на период дружественного симбиоза с умными машинами уровня человеческого уровня и делает возможным сценарий, в рамках которого человек сможет создать действенные рычаги контроля над формирующимся суперинтеллектом.

Но существует когорта исследователей, по мнению которых человечество лишено какого-либо временного запаса. Дело в том, что скачок от искусственного разума человеческого уровня к искусственному супермозгу через самосовершенствование может произойти резко. В соответствии с этим сценарием стремительная трансформация онтологических оснований человечества станет отражением быстрого превращения искусственного разума уровня человека в искусственную сверхразумную систему. Этот период

может занять недели, дни, а возможно и часы. Этот сценарий получил название *Busty Child*.

Система, состоящая из нескольких суперинтеллектуальных систем, каждый из которых в несколько тысяч раз будет превосходить самого умного человека, без особых трудностей сможет преодолеть все преграды, созданные нами. Это сравнимо с бескрайним пространством чуждого интеллекта и одной его песчинкой. Для того, чтобы в общих чертах описать те ощущения, которые испытывает человек в процессе взаимодействия всего лишь с отдельной программой (*Deer Blue* – компьютерный шахматист фирмы IBM), а не с самосовершенствующейся группой искусственного суперразума, приведем здесь высказывания двух гроссмейстеров, весьма сходных по смыслу, суть которых сводится к следующему: «Будто стена на тебя надвигается» [6, 48].

Весьма показателен в этом смысле, проведенный в начале XXI века в Силиконовой долине эксперимент, суть которого заключалась в следующем: специалист по изучению искусственного интеллекта заключил весьма специфический спор – кто победит в игре, которую он назвал «искусственный интеллект в ящике». Ставки на игру были достаточно высокими. В ходе этого эксперимента роль машины играл инициатор эксперимента Е. Юдковски [16], в роли цепных псов выступали миллионеры, заработавшие свои состояния на различных интернет-проектах. Каждый из них, в порядке очереди, исполнял роль творца разумного агента, перед которым стояла цель не выпустить из «ящика» искусственный разум. «Узник» и «Страж» поддерживали контакт через онлайн-чат. Игра длилась не более двух часов и всего их состоялось пять. Возможность утомить «Стража» молчанием была предусмотрена, но не использовалась. В результате машина одержала победу в трех сеансах. Каким образом «Пандоре» удалось сбежать неизвестно, так как одним из условий проведения эксперимента была полная конфиденциальность содержания переписки между «Узником» и «Стражем» [16].

Этот эксперимент доказывает наши опасения – если обычный смертный смог посредством слов «выпустить из ящика Пандору», то заключенный суперинтеллект, уровень которого будет в неизвестное количество раз превосходить человеческий, сделает это быстро и гарантированно. К тому же для машины достаточно всего лишь раз совершить побег. И будет лучше для всех нас, если он окажется дружелюбным.

Интересен этот эксперимент еще и тем, что, по сути, является вариантом теста А. Тьюринга, который разработал его в 50 году прошлого века для определения уровня интеллекта в машине. В процессе этого теста программе и оппоненту, в роли которого выступает человек, задаются письменные вопросы. Арбитр должен суметь определить по ответам, кто – человек, а кто – машина. Если идентификация невозможна, то компьютер одерживает победу. Возникает вполне закономерный вопрос: как отличить оригинальный человеческий разум и мышление от достоверных попыток имитировать его?

Где эта грань между объектом и образом объекта в мышлении субъекта? Другими словами, машине совсем не обязательно думать, как человек, чтобы давать ответы, идентифицирующего ее как такового. Машине достаточно сымитировать мыслительный акт, давая «антропоморфные» ответы. Это стало «игрой в имитацию» А. Тьюринга, который считал, что машины способны к деятельности, которую наблюдающий субъект с легкостью примет за разумную [12, 63]. По сути же, этот алгоритм в исполнении машины не имеет ничего общего с мыслительным актом, совершаемым человеком. Тем самым, А. Тьюринг вступает в полемику с Д. Сёрлем, который считал, что если машина не думает подобно человеку значит она не разумна. Но большинство исследователей в области создания искусственного разума выражают солидарность с первым: если разумный агент совершает разумные действия, имитируя сознание человека, то какая разница какие процессы запускают его программы?

Долгое время пройти тест А. Тьюринга оказывалось неразрешимой задачей для машинного интеллекта и исследователей, его создающих. Но летом 2016 года во многих интернет изданиях появилась информация о том, что впервые компьютерной программе, разработанной российскими авторами, удалось это сделать. Эта программа выдавала себя за 13 летнего мальчика Евгения Густмана и ввела в заблуждение 33% оппонентов, специально приглашенных из британского Университета Рединга. Один из членов команды, который разрабатывал программу рассказал, что главной идеей было убедить комиссию в якобы широком кругозоре, но который ограничен возрастом имитации [11].

Эта новость еще раз подтверждает тот факт, что создание искусственного разумного агента – дело недалекой перспективы и хитроумных уловок в его запасе будет более, чем достаточно, а способность имитировать человеческие качества, вдобавок ко всему вышесказанному, отражает уже вполне реальную возможность достигать своих целей самыми разными путями.

Возможно ли при наличии накопленного за время развития науки прогностического потенциала предсказать эти цели, способы и средства, которые машина для себя изберет; а также уровень угрозы, исходящий от перспектив появления искусственного мыслящего разума? Оценивая возможность появления искусственного разума человеческого уровня, а затем и сверхразумного агента с позиции однозначной положительной рефлексии, исследователи, тем самым, сужают свой методологический инструментарий линейному, жестко детерминированному прогнозированию. Наличие в постнеклассической науке паттернов линейности обусловлено ориентацией классики и неклассики опираться на динамические законы, описывающие мир с позиций жесткой детерминации закрытых систем. В развитии таких систем не учитывались флуктуационные воздействия среды, привносящие в их упорядоченную структуру фактор случайности, относительной неустойчивости.

Прогноз на будущее поведение подобного типа систем был относительно прост – изучив и зная причину, мы однозначно можем предсказать и следствие.

Существенный прорыв в методологии системного прогнозирования был осуществлен в рамках нелинейной методологии, инициированной теорией самоорганизации сложных систем. В зависимости от того, насколько открытые системы различной природы способны преобразовывать внутренние неоднородности своей структуры в полезный для себя потенциал, сохраняя и увеличивая уровень организованной сложности и относительной устойчивости, исследователям удалось систематизировать специфику протекания режимов с обострением, а также способы ликвидации флуктуационных отклонений, ведущих к кризисным состояниям. Как уже было сказано выше, процесс создания машинного разума, как и факт его наличия в будущем относятся к типу таких режимов, в которых случайное микроотклонение повлечет за собой череду непредсказуемых макропоследствий. Анализируя перспективы развития и возникновения искусственного разума с позиций нелинейного подхода, необходимо внести в прогностическую модель данные о специфике искусственной системы, тенденциях развития, уровне угрозы, а также векторах нелинейности. Прогностическое моделирование с неучтенным вектором нелинейности, по своей сути, не отражает всей сложности процесса и является линейным. С подобными недостатками мы имеем дело, сталкиваясь с прогностическими моделями, связанных с технологией искусственного интеллекта. В связи с этим не стоит недооценивать творческий потенциал эволюции, стохастическое влияние которой не является чем-то постоянным и устойчивым. И с этой позиции, процесс самосовершенствования искусственного разума, по всей видимости, будет так или иначе отвечать критериям эволюционного развития. А. Азимов в своих работах выдвигал тезис о том, что когнитивные модели могут быть тождественны компонентам психики иными словами он предположил, что роботы смогут эволюционировать. Его идея о «призраках в машине» в настоящее время обретает новое смысловое наполнение. Он говорит о возможности возникновения случайно сформированных протоколов, которые в перспективе могут развиваться в то, что мы называем поведением, а непредвиденные свободные радикалы могут положить начало креативности, свободе выбора, а возможно и душе [1, 253].

Также стоит учитывать, что возникновение искусственного суперинтеллекта относится к разряду тех событий, последствия от которых могут быть глобальными, а вероятность возникновения низка в силу того, что ничего подобного просто не происходило и эмпирическая компонента человечества здесь равна нулю. Прийти к однозначному решению, имея на вооружении традиционные статистические методы задача не из простых.

В процессе разработки искусственно разума следует иметь в виду, что специалисты не могут себе позволить права на ошибку. В условиях процесса с

обостренным режимом протекания это может иметь последствия «эффекта бабочки» и в результате мы получим нечто совершенно чуждое. Это сравнимо с действиями медвежатника по взлому суперсложного сейфа под сигнализацией в банке. Если он из двадцати цифр кода нажмет правильно девятнадцать, то дверь не приоткроется на пять процентов, она так и останется запертой. Взвоят серена и все закончится плачевно. При создании искусственного интеллекта ошибка даже в одну сотую процента повлечет за собой на сто процентов непредсказуемый результат. Он не будет почти хорош, он будет полностью плох.

Большинство привычных для человека технологий потенциально для него опасны, как и любая инновация «палка о двух концах», о чем уже шла речь выше. По всей видимости, искусственный интеллект не исключение. Он не будет испытывать к вам негатива, но вашим атомам он может найти другое применение. Создавая искусственный разум с благими намерениями и экстраполируя эти положительные эмоции на машину, специалисты ошибочно считают, что это является гарантом появления дружественного характеристик. Такая недальновидность связана с уже упомянутым выше демоном антропоморфизма и линейным детерминизмом, не учитывающим флуктуаций различной степени важности.

Вид *homo sapiens* возник в процессе самоорганизации материи, которая предполагает борьбу за свободную энергию, информацию и вещество. И, если на заре своего становления в бытие человека доминировали процессы стохастического характера, то с развитием общества стихийные процессы были оттеснены на периферию благодаря сознательному фактору и социальным институтам. Бифуркационные фазы преодолевались благодаря возобладанию техно-гуманитарного баланса и осознанию сути и причин проблемы. Оглядываясь назад и анализируя сущность большинства антропологических кризисов, можно прийти к выводу, что их природа заключается в недооцененной степени угрозы последствий или же в непонимании сути причин технологического взлета. Развитие искусственного интеллекта, по всей видимости, тоже будет предполагать его активное включение в процесс естественного отбора, в борьбу за энергию, вещество, и информацию. Да и вообще, за все то, что посчитает ресурсом

Вызывает, мягко говоря, опасения тот факт, что многие исследователи не совсем понимают, как работает вся система в целом. И в этом контексте явно недостаточно создавать разумную машину с чистым сердцем и благими намерениями, надеясь на благополучный результат и чудесное появление дружественного искусственного интеллекта. И в этом нет вины специалистов. Корень проблемы не заключается в незнании специалистов как создать дружественную думающую машину. Причиной, повлекшей за собой прекращение бытия человека, во всяком случае, в привычных для нас формах, может стать убеждение, что искусственный интеллект будет обязательно

дружественным. Прежде всего потому, что нельзя навязать пути развития сложной системе, нелинейной среде, которые ей имманентно не присущи. Имея дело с созданием сознания в машине, человек, как уже было сказано выше, не понимает как работает вся система в целом и поэтому ему не известен тот вектор нелинейных векторов, аттракторов, к которым будет стремиться искусственный разум. Человеку в принципе неизвестно, что будет присуще этой системе!

Вера в благонамеренность машинного интеллекта становится еще более опасной после того, как машинный интеллект человеческого уровня, нарушив меру, через скачок интеллектуального взрыва перейдет на принципиально новое качество – суперинтеллект. Но тем не менее в своем исследовании Е. Юдковски выдвигает тезис о том, что дружественность искусственного интеллекта может базироваться на так называемой функции полезности – синтез ценностей, предпочтений, облаченных в удовлетворение от достижения цели, который внедрен в определение пользы в алгоритмических паттернах [16].

Какую же семантическую специфику вкладывают исследователи в понятие «дружественный», употребляя его в контексте искусственного разума? Прежде всего, искусственный интеллект не должен быть никогда враждебно или амбивалентно настроен в отношении человеческого вида, к какой цели бы машина ни стремилась и сколько бы ступеней самосовершенствования ни прошла. Все это невозможно без глубокого понимания машиной человеческой природы (понимает ли ее сам человек?), чтобы в дальнейшем не причинить человеку вред даже в результате случайных, опосредованных последствий своих действий (о чем мы уже говорили в контексте осмысления Трех законов А. Азимова). При этом мы не хотим получить искусственного разумного агента, выполняющий краткосрочные задачи при помощи мер, которые бы оказались для человечества вредоносными впоследствии.

В качестве примера реализации непредвиденных последствий Н. Бостром пишет о так называемых пагубных отказах [4, 153–176]. Здесь мы приведем некоторые варианты пагубного отказа – порочную реализацию и инфраструктурную избыточность. Первый из сценариев, по которому искусственный сверхразум достигает поставленной цели способом оптимальным с его позиций, но противоречащим общечеловеческой шкале ценностей. Например: желаемый результат, к которому хочу прийти я – постоянная улыбка на моем лице; способ достижения этой цели в понимании машины – напрямую задействовать нерв лица, что неизбежно повлечет за собой паралич мимики в результате чего улыбка не будет сходиться с лица.

Самое пугающее, что выбор данного способа реализации ни коим образом не обусловлен стремлением машины навредить человеку. В данном случае порочной реализации – манипуляции на лицевом нерве – для машинного



интеллекта гораздо оптимальнее, нежели привычные методы человека, в силу того, что это наиболее полный способ достичь конечной цели. Существуют ли какие-либо окольные пути, позволяющие разрешить эту проблему? Возможно это удастся благодаря конкретизации формулировки цели? Цель: без воздействия напрямую на лицевой нерв сделай обеспечить мне постоянную улыбку. Способ побочной реализации: активизация зон коры мозга, отвечающих за функции лицевого нерва. Вечно сияющая улыбка на лице вам обеспечена.

Может быть формулировка конечной цели затруднительна вследствие того, что мы использовали привычный для человека понятийно-категориальный аппарат? Попробуем задать конечную цель, суть которой связана с позитивным феноменологическим состоянием, счастьем, субъективным ощущением комфорта, без описания поведенческих моделей. Гипотетически мы допускаем возможность реализации специалистами «вычислительного» представления идеи счастья и дальнейшего его внедрения в эмбрион искусственного интеллекта. Цель весьма сложная и спорная, если не невозможная. И в рамках данной статьи пути ее решения рассматриваться не будут. Но предположим, что программистам удалось поставить перед машинным разумом задачу осчастливить нас. Тогда порочная реализация может приобрести такие очертания: внедрение электродов в центры удовольствия мозга.

Следующий пример является формой реализации другого вида пагубных отказов – инфраструктурная избыточность – такой процесс, при котором разумный агент для достижения конкретно поставленной цели, превращает все известные для него виды энергии, вещества и информации в ресурс, производственно-техническую базу для воплощения этой цели, вследствие чего реализации сущностного потенциала человечества становится невозможной. В рамках этого сюжета запрограммированный искусственный разум, которому в качестве конечной цели было задано штамповать канцелярские скрепки, делает ровно то, что от него требовалось, вне системы человеческих ценностей. В результате суперинтеллект преобразует доступное пространство и вещество в фабрики по производству скрепок.

Отсутствие заостренных, догматических ценностей – еще одно существенное, принципиально важное качество дружественного искусственного интеллекта. Аксиологические ориентиры машинного сознания должны претерпевать соответствующие трансформации в тесной корреляции с изменениями таковых в обществе. К примеру, если бы функция полезности гипотетического разумного агента была ориентирована на ценностные паттерны подавляющего большинства населения Европы 18 века и не поддавалась изменениям в соответствии с развитием общества, то и сегодня этот искусственный интеллект за одну из основ своего рабочего алгоритма имел систему архаических пережитков, среди которых рабовладение, расовая и

половая дискриминация, публичные казни и прочее. Система ценностей, внедренных в дружественную машину не должна быть заданной раз и навсегда.

Такие теоретические построения выглядят, по меньшей мере, утопично. Из всего вышеизложенного становится ясно, что тема дружественного искусственного интеллекта требует дальнейшей конкретизации и развития, но ее сторонники мыслят крайне оптимистично. На сегодня наука не гарантирует однозначную перспективу изложения концепции дружественного сверхразума на языке математики, как и нет гарантии, что создание такого разума вообще возможно или реально его интегрирование в перспективные архитектуры компьютерного сознания. Но теперь, когда Трем законам робототехники заслуженно присвоен статус принципа построения сюжета, а не средства выживания, – концепция дружественного машинного сверхразума, по всей видимости, лучшее, что может предложить человечество перед лицом потенциальной экзистенциальной угрозы. Однако, дружественная машина не создана, а проблем с ее конструированием более чем достаточно.

Одна из них связана с тем, что большое количество организаций по всему миру работают над созданием искусственного разума уровня человека и в области смежных технологий. В погоне за веткой первенства ни одна из этих компаний не приостановит свою деятельность на этом поприще в ожидании дня создания дружественной разумной машины. Слишком много стоит на кону. Более того, мало кто из участников этой гонки вовлечен в научно-философский дискурс, касающийся проблем дружественного искусственного интеллекта.

В число организаций, в круг приоритетов которых входит создание думающей машины человеческого уровня входят: AGIRI, CYC, Google, IBM (несколько проектов), LIDA, Nell, Numenta, SNERG и Vicarious. Также существуют минимум с десятков проектов, источники финансирования которых не отличаются надежностью: NARS, Novamente, Sentience, SOAR, а также DAPRA, финансирующая напрямую или через посредников проекты, связанные с разработкой искусственного разумного агента, а также смежные технологии [13, 36]. Это далеко не полный перечень. В контексте этого возникает совершенно логичный вывод: вероятность того, что первый искусственный разум в рамках легальных проектов увидит мир именно в лабораториях MIRI ничтожно мала, а следовательно, достаточно мизерна возможность внедрения в эту когнитивную архитектуру модуля дружественности. По всей видимости, разработчиков первой разумной машины будет мало волновать проблема дружественности программного обеспечения. Но существуют и стратегические векторы, которые возможно в дальнейшем поспособствуют блокированию враждебного сверхразума. На сегодня существует образовательная программа для передовых университетов и математических конкурсов, в рамках которой MIRI и CEAR организовали так называемые «тренировочные лагеря разума». В этих ячейках обучают потенциальных творцов разумных программ и

руководителей, формирующих дальнейшую техническую политику, инновационному мышлению. В перспективе это должно помочь избежать тупиков в развитии и ловушек, созданного разума. Разумеется этих мер недостаточно, но MIRI и CEAR удалось обратить внимание общественности на важный фактор рисков, благодаря чему все больше работ выходит по проблеме сингулярности, авторы которых таким образом расширяют и углубляют научные и мировоззренческие аспекты феномена рисков искусственного разума.

Но даже если потенция станет реальностью и дружественный разумный агент буде создан, нет никакой уверенности, что он останется таковым после интеллектуального взрыва. Иначе говоря, сохранит ли искусственный разум благонамеренные качества, если его интеллект вырастет в тысячи раз, то есть возникновение новых качественных характеристик сотни раз будут обуславливать новый характер количественных изменений. Результатом сотен, тысяч таких скачков может стать трансформация его осмысления феномена дружелюбия и однозначное искажение запрограммированной семантики. В любом случае произойдет изменение представлений о морали, а следовательно исказится и функция полезности.

Е. Юдковски с такой точкой зрения не согласен, считая, что прогресс машинного интеллекта отразится на улучшении эффективности функции полезности.

Реализация такой возможности может иметь место в том случае, если процесс интеллектуального взрыва пройдет без флуктуаций, системного сбоя, природу которых мы в силу своей антропоморфности даже представить себе не можем. У человека и плоского червя большой процент общего ДНК, но мысль о том, что у нас могут быть общие ценности и мораль, как минимум, смешна. Не изменило бы это положение вещей и сенсационная новость, в которой бы плоскому червю отводилась роль нашего творца, наделившего свое детище ценностями и идеалами своего вида. Поначалу это бы спровоцировало у нас отторжение и недоумение, удивление, от которого бы мы быстро оправились и вернулись к своей привычной жизни.

В этом же контексте очень показателен пример персонажа Доктора Манхеттена из графического романа А. Мура, включенного в сотню лучших романов 20 века [7]. Известный физик в результате неудачного эксперимента был разобран на атомы и возродившись, стал другой сущностью. Теперь ему под силу менять структуру материи, природу пространства-времени, он видит прошлое и будущее, его разум уже не заботит проблемы человечества. Радости любви, страх смерти, противоречивость жизни и прочие «мелочи», все то, без чего ни один человек не способен себя помыслить, вещи, являющиеся альфа и омега любого личностного существования, не заботят его больше. После долгих размышлений о сущности бытия во вселенском масштабе, он приходит к выводу, что феномен жизни слишком переоценен. В дальнейшем он решает

проблему угрозы неотвратимой ядерной войны убив два миллиарда человек, но тем самым, спасая семь миллиардов. Человеческий дружественный интеллект развился в Великий Разум, сохранивший некие паттерны человеческой морали, но превратился в нечто чуждое, лишённое личностных характеристик, по-своему интерпретируя понятия оптимальности, помощи, полезности. Иными словами, в данном случае мы имеем дело с дружественной машиной, пережившим интеллектуальный взрыв, в результате которого понимание «дружественности» претерпело качественные трансформации.

Д. Хьюз также выдвигает тезис о несостоятельности идеи устойчивых и неизменных первоначальных вводных искусственного интеллекта в процессе скачкообразного развития машины, основываясь на осмыслении преобразования витальных потребностей человека (пища, убежище, самосохранение) в алгоритмы, отличающиеся от первоначальных наборов целей. Например, человек может избрать стратегией своей жизни аскетизм или целибат, что противоречит генетической программе нашего организма. Или среднестатистический гражданин может стать террористом-смертником для того, чтобы гонорар за теракт после смерти был выплачен семье. Таким образом, человек способен подвергать рефлексии свои цели и выстраивать различные алгоритмы их достижения, которые идут вразрез с привычными моделями рациональности и базовыми инстинктами. В связи с этим, мысль о том, что созданный искусственный разум с открытым и гибким разумом (что, по сути, и является сущностными характеристиками думающей машины) не будет подвержен изменениям в процессе развития, по меньшей мере, наивна [3, 115].

Д. Хиллис также считает, что человечество постепенно передает «бразды правления» компьютерам, тем самым отдаляясь от производства, управления, не вникая, по сути, в процесс создания машинами еще более сложных машин и других вещей. Человек уже не понимает, что и как происходит. Технологии создают технологии все более самостоятельно без участия человека. По его мнению, происходящее напоминает эволюцию простейших организмов в многоклеточные. И в этой эволюции Д. Хиллис отводит человеку роль амебы, не понимающего что и как мы создаем [3, 117].

Существенной преградой на пути к созданию контролируемого искусственного интеллекта стало несовершенное программирование. Уверенность в непогрешимости функционирования компьютерных программ разбивается об айсберг статистических данных, свидетельствующих о том, что 60 млрд. долларов в год американская экономика недополучает в результате дефектного программирования [13, 47]. Удивление вызывает тот факт, что компьютеры, как математические машины, должны быть абсолютно линейно детерминированы и предсказуемы; тогда как в реальности все совершенно иначе и создание компьютерных программ, по сути, одна из самых

вероятностных инженерных задач, сопряженных со многими ошибками и проблемами с безопасностью.

С. Омохундро считает, что способом борьбы против некачественного программирования является создание систем, которые осознают себя и способны к рефлексии над своим поведением в процессе работы достижения поставленных целей. Иначе говоря, они должны уметь самосовершенствоваться. Саморазвивающиеся программы – необходимое качество и неизбежный этап на пути к созданию думающей машины. Однако программное обеспечение, сознающее себя, на сегодняшний день еще не разработано, а программы, модифицирующие себя довольно распространены [9].

Одним из алгоритмов машинного обучения, который использует возможности естественного отбора для поисков ответов является генетическое программирование. Этот алгоритм является важнейшим инструментарием для написания мощных программ. В этом виде программирования в отличие от обычного программирования, где используется человеческая логика, применяется логика компьютерная. В обычном программировании при написании строки кода используется работа программиста, благодаря чему процесс обработки данных в пределах схемы «вход-выход» прозрачен и верифицируем. В случае применения генетического программирования программист лишь описывает задачу, а ее решение передается на волю естественного отбора. Генетическая программа генерирует фрагменты кода, являющихся элементами системы очередного поколения. Наиболее прогрессивные из них синтезируются случайным образом, создавая новую генерацию. Программа тем более оптимальна и перспективна, чем более она сумела приблизиться к решению поставленной перед ней задачей. Такой процесс в компьютерной эволюции, в сущности, является подобием естественного отбора в природе, в течение которого слабые отбрасываются, а лучшие вступают во взаимодействие вновь, приводя к случайным трансформациям отдельных команд и переменных. Такие качественные скачки в развитии природы называются мутациями. Программист запустив такую генетическую программу, в дальнейшем от коррекции в ее работе может устраниваться. По сути, некий компьютерный деизм, в котором программисту отводится роль инициатора, запустившего процесс, но в дальнейшем не влияющего на причины, следствия которых в перспективе, быть может, запустят цепь событий, породивших новую реальность.

Генетические алгоритмы применялись при проектировании антенны для NASA, разработки программ распознавания белков. Более двух десятков раз генетические алгоритмы изобрели электронные компоненты, которые до этого уже успел запатентовать человек. Цели задавались инженерными спецификациями готовых устройств. Но и здесь, казалось бы, линейный, проторенный человеком, путь к цели в исполнении машины становится

нелинейным и принципиально непредсказуемым на некоторых этапах создания нового алгоритма. Дело в том, что необъяснимо почему новая схема работает лучше и каким образом, имея с нашей точки зрения элементы, не обладающие функциональной ценностью для системы. Загадка заключается в том, что код, созданный генетическим алгоритмом нечитаем. Программа генерирует решения, которые специалисты не в состоянии воспроизвести, создавая нелинейные пути достижения конечного результата, понять которые эксперты тоже не могут. Если подобные сложности возникают в начале пути с программами, у которых отсутствует самоидентификация, то о каком контроле самосовершенствующихся программ может идти речь в перспективе? Компьютерная система с известным сигналом на входе и выходе, но неизвестной процедурой его обработки получила название «черный ящик». Нелинейность и непознаваемость этой процедуры еще более сгущает темные краски вокруг проблемы контроля и природы машинного разума, акцентируя внимание на его совершенно чуждой нам сущности. Эта «темная сторона» генетического программирования все более отдаляет нас от мечты создания дружелюбной машины, превращая ее, по сути, в иллюзию.

«Черные ящики» не всегда остаются вне контроля. Но принципиально это суть дела не меняет, ведь если когнитивные архитектуры применяют такие структуры для создания системы суперинтеллекта, то сама его суть будет состоять из пластов принципиально непознаваемого для нас кода, а следовательно, непознаваемость и чуждость станут одними из определяющих качеств идентифицирующей себя саморазвивающейся машины.

С. Омохундро иллюстрируя опасность и непредсказуемость работы систем, способных изменять себя, приводит пример робота-шахматиста. Конечно же, речь не идет о компьютерной программе, установленной в любом компьютере средней производительности. С. Омохундро имеет в виду потенциального робота-шахматиста, самосознающего и самосовершенствующегося. Как поведет себя этот разумный агент в ситуации, если вы сыграете с ним партию в шахматы, а затем дадите ему команду отключиться? И вся проблема заключается в том, что отключение для сознающей себя машины событие весьма значимое в силу того, что включиться самостоятельно она не может. Поэтому для него крайне важна уверенность в совершенно адекватной оценке реального положения вещей. И в стремлении оценить ситуацию как можно глубже, робот может прийти к мысли выделить какие-то ресурсы на познание природы реальности перед тем как выйти из нее, отключившись.

Возникает очень важный вопрос: какое количество ресурсов разумный агент может считать достаточным? Ответ, который дал С. Омохундро заставит задуматься многих оптимистов создания думающих роботов: «Роботизированная программа принять решение потратить на это все доступные человеку ресурсы» [9].

Таким образом, нелинейное пространство развития искусственного интеллекта – отражение нелинейной эволюции человека, который в борьбе за ресурсы не раз ставил себя на грань выживания, но впервые за десятки тысяч лет своего существования человеку, по всей видимости, придётся иметь дело с системами превосходящими его в играх, равных в которых ему доселе не было, а именно в играх разума. Роскоши в виде десятков – сотне лет для техно-гуманитарной притирки может не оказаться, в силу того, что самосовершенствующийся искусственный разум выйдет в своем развитии за пределы способности человека каким-либо образом влиять на ход событий. А искусственный интеллект с легкостью преодолев пропасть между искусственным и естественным, займет почетное место в природной нише, оказывая влияние на процессы Вселенной или попросту заменит ее. Судьба человека в этом контексте весьма туманна.

В этом контексте показательны слова главного С. Шостака, который говорил о том, что цивилизации сами создают себе преемников [15].

По всей видимости, у Человека есть два варианта: думать, что еще есть время интеллектуально и стратегически «сгруппироваться», сделав все возможное, чтобы обуздать «черно-белого мустанга» Искусственный Интеллект, стремительно мчащегося за горизонт событий сингулярности.

И второй – надеяться на то, что С. Шостак недостаточно прозорлив.

#### ЛИТЕРАТУРА

1. Азимов А. Загадки мироздания. Известные и неизвестные факты / Айзек Азимов. – М.: Центрполиграф, 2016. – 432 с.
2. Азимов А. Я - Робот / Айзек Азимов. – М.: Эксмо, 2007. – 1296 с.
3. Баррат Д. Последнее изобретение человечества: Искусственный интеллект и конец эры Homo sapiens / Джеймс Баррат. – М.: Альпина нон-фикшн, 2015. – 304 с.
4. Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии / Ник Бостром. – М.: Манн, Иванов и Фербер, 2015. – 496 с.
5. Курцвейл Р. Эволюция разума / Рэй Курцвейл. – М.: Эксмо, 2016. – 448 с.
6. Ллойд С. Программируя Вселенную: Квантовый компьютер и будущее науки / Сет Ллойд. – М.: Альпина нон-фикшн, 2014. – 256 с.
7. Мур А. Хранители / Аллан Мур. – М.: Азбука, Азбука-Аттикус, 2014. – 528 с.
8. Назаретян А. П. Нелинейной будущее / А.П. Назаретян. – М.: МБА, 2013. – 440 с.
9. Омохундро С. Чем нам угрожают роботы? [Электронный ресурс] // Technowars. – 2015. – Режим доступа до ресурсу: <https://technowars.defence.ru/article/1468/>.

10. Стибел Д. Почему я нанимаю неудачников [Електронний ресурс]. – 2015. – Режим доступу до ресурсу: [http://www.michelino.ru/2016/06/blog-post\\_11.html](http://www.michelino.ru/2016/06/blog-post_11.html).

11. Тест Тьюринга пройден (на детском уровне) [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: <https://habrahabr.ru/post/225599/>.

12. Тьюринг А. Может ли машина мыслить? / Алан Тьюринг. – СПб.: Эдиториал УРСС, Ленанд, 2016. – 128 с.

13. Форд М. Технологии, которые изменят мир / Мартин Форд. – М.: «Манн, Иванов и Фербер», 2014. – 268 с.

14. Шелли М. Франкенштейн, или Современный Прометей / Мэри Шелли. – СПб.: Мартин, 2015. – 256 с.

15. Шостак С. Вероятно вземной разум существует. Вы готовы? [Електронний ресурс]. – 2013. – Режим доступу до ресурсу: <http://www.fassen.net/video/gLZJuPM24M/>.

16. Юдковски Е. Обратное глупости не есть ум [Електронний ресурс]. – 2015. – Режим доступу до ресурсу: [http://lesswrong.ru/w/Обратное\\_глупости\\_не\\_есть\\_ум](http://lesswrong.ru/w/Обратное_глупости_не_есть_ум).

## РЕЗЮМЕ

**І.О. Снегірьов. Нелінійний простір штучного інтелекту: філософсько-світоглядне осмислення.**

*У статті автор з позиції методології нелінійного прогнозування розкриває перспективи та ризики виникнення штучного інтелекту. Аналізуються причини як оптимістичних прогнозів, так і комплекс загроз, що виходять від «горизонтів» сингулярності, обумовленої створенням машинного розуму. Особлива увага приділяється концепції дружній розумної машини в умовах інтелектуального вибуху.*

*Здійснено спробу філософської рефлексії декількох варіантів згубного відмови, а саме: порочної реалізації та інфраструктурної надмірності.*

**Ключові слова:** *штучний інтелект, сингулярність, нелінійність, антропоморфізація, «бог в ящику», тест Тьюринга, Три закони робототехніки, «чорний ящик», генетичні алгоритми.*

## SUMMARY

**I.Snegirjov. The nonlinear space of artificial intelligence: the philosophical and ideological understanding.**

*The author from the perspective of a nonlinear prediction methodology reveals the prospects and risks of artificial intelligence. The reasons as the optimistic forecasts and complex threats posed by "horizons" singularity, due to the creation of machine intelligence. Particular attention is paid to the concept of a friendly intelligent machine in terms of intellectual explosion.*



*An attempt to philosophical reflection several options disastrous failure, namely the realization of evil and infrastructure redundancy.*

**Keywords:** *artificial intelligence, the singularity, nonlinearity, anthropomorphism, "God in a box," Turing test, the Three Laws of Robotics, "black box", genetic algorithms.*