
Алексей
Турчин
ВОЙНА
и еще 25 сценариев
КОНЦА
СВЕТА

Алексей Турчин

Война и еще 25 сценариев конца света

«Европа»

2008-01-01

Турчин А. В.

Война и еще 25 сценариев конца света / А. В. Турчин — «Европа», 2008-01-01

Книга Алексея Турчина – это актуальный обзор последних научных наработок, описывающих, как и когда закончится существование человеческой цивилизации. В ней рассказано обо всех известных на сегодня видах глобальных рисков, которые могут уничтожить человечество, – от астероидов до запуска Большого адронного коллайдера и ядерной войны.Важность книги в эти дни необычайно велика по двум причинам. Во-первых, катастрофа человечества – вопрос, безусловно достойный нашего с вами внимания. Во-вторых, разработка многих новых технологий, способных вызвать глобальные риски, уже идет. Понимание того, сколь велики, реальны и близки риски конца света, должно заставить вас задуматься – как будете спасаться лично вы? Эта книга потребует недюжинной подготовки и силы характера, без которых лавина технологических угроз может сбить вас с ног. Если такое случится с вами, не пугайтесь и доверьтесь автору.

Содержание

Алексей Турчин
Война и еще 25 сценариев конца света

Предисловие
Наука о конце света

Все, что имеет начало, имеет и конец, и человеческая история (как и ваша жизнь) – не исключение. Идеи о конце света не новы, но сегодня они становятся как никогда ранее реальными.

Книга Алексея Турчина – это актуальный обзор последних научных наработок в этой области, описывающих, как и когда закончится существование человеческой цивилизации. От астероидов до ядерной войны – всему, что может уничтожить человечество, будь то природный катаклизм или рукотворная катастрофа, нашлось место в книге. Некоторые риски, такие как кометно-астероидная опасность, вам известны давно, другие, например глобальное потепление, лишь несколько лет, а про риск физических экспериментов на ускорителе под названием Большой адронный коллайдер все услышали прошлым летом в связи с его близким запуском. Другие угрозы катастроф и по сей день ясны лишь узкому кругу специалистов. Среди них – глобальные супернаркотики, супервулканы и космические угрозы вроде гамма-вспышек. Одна из проблем, на которые указывает автор, в том, что общество внимательно относится к «раскрученным» угрозам с хорошим пиаром (и, добавлю, с солидной голливудской поддержкой), а не к вероятным и наиболее опасным.

Важность книги в эти дни необычайно велика по двум причинам. Во-первых, катастрофа человечества – вопрос, безусловно достойный нашего с вами внимания. Во-вторых, разработка многих новых технологий, способных вызвать глобальные риски, уже идет.

Например, миссия к центру Земли с помощью зонда, предложенного Стивенсоном, грозит совершить невозможное – расколоть земной шар ко всем чертям.

При этом другой планеты у нас с вами пока нет. Проекты ядерных ракет, гигантских орбитальных станций и поселений на Луне пока остаются лишь проектами. Поэтому и спасаться от планетарных катастроф нам негде.

Понимание того, сколь велики, реальны и близки риски конца света, должно заставить вас задуматься – как будете спасаться лично вы? Нам нужны индивидуальные планы спасения, нужны и глобальные, общепланетарные планы, разработанные и утвержденные на самом высоком уровне. Нужны и планы эвакуации, стрелочками показывающие, куда мы побежим в случае катастрофы – в подводные города или на орбитальные станции.

Для этого создается новая научная область – изучение глобальных рисков. Это научная дисциплина новой волны – междисциплинарная и сложная для классификации. История, социология, математика, психология, геология, астрофизика и новые научные области сведены здесь воедино.

В книге Турчина рассказано обо всех известных на сегодня видах глобальных рисков, которые могут уничтожить человечество. Впрочем, одной книги вам вряд ли хватит, чтобы составить свой личный план спасения.

Эта книга потребует недюжинной подготовки и силы характера, без которых лавина технологических угроз может сбить вас с ног Если такое случится с вами, не пугайтесь и доверьтесь автору. Понять, что такое «транскранальная магнитная стимуляция», «сильный искусственный интеллект» или «нанофабрика», поможет поиск в Яндексе. Ваш «уровень шока будущего», а значит и вашу выживаемость, можно легко поднять с единички хотя бы до 2–3.

Если справиться с нехваткой ликбеза и необходимостью делать перерывы для переваривания материала, то к последней странице перед вами встанет вопрос о решениях. На этот случай дам несколько советов.

Первое. Будьте готовы к радикальным изменениям уже в ближайшем будущем. Даже если мы сумеем избежать всех катастроф, остается «технологическая сингулярность» – конец человеческой истории и начало новой эпохи. Она неизбежна. К ней нам готовиться еще важнее, чем к падению 100-километрового астероида, потому что есть шанс выжить. Не прекращайте изучение будущего в общем и проблемы глобальных рисков в частности.

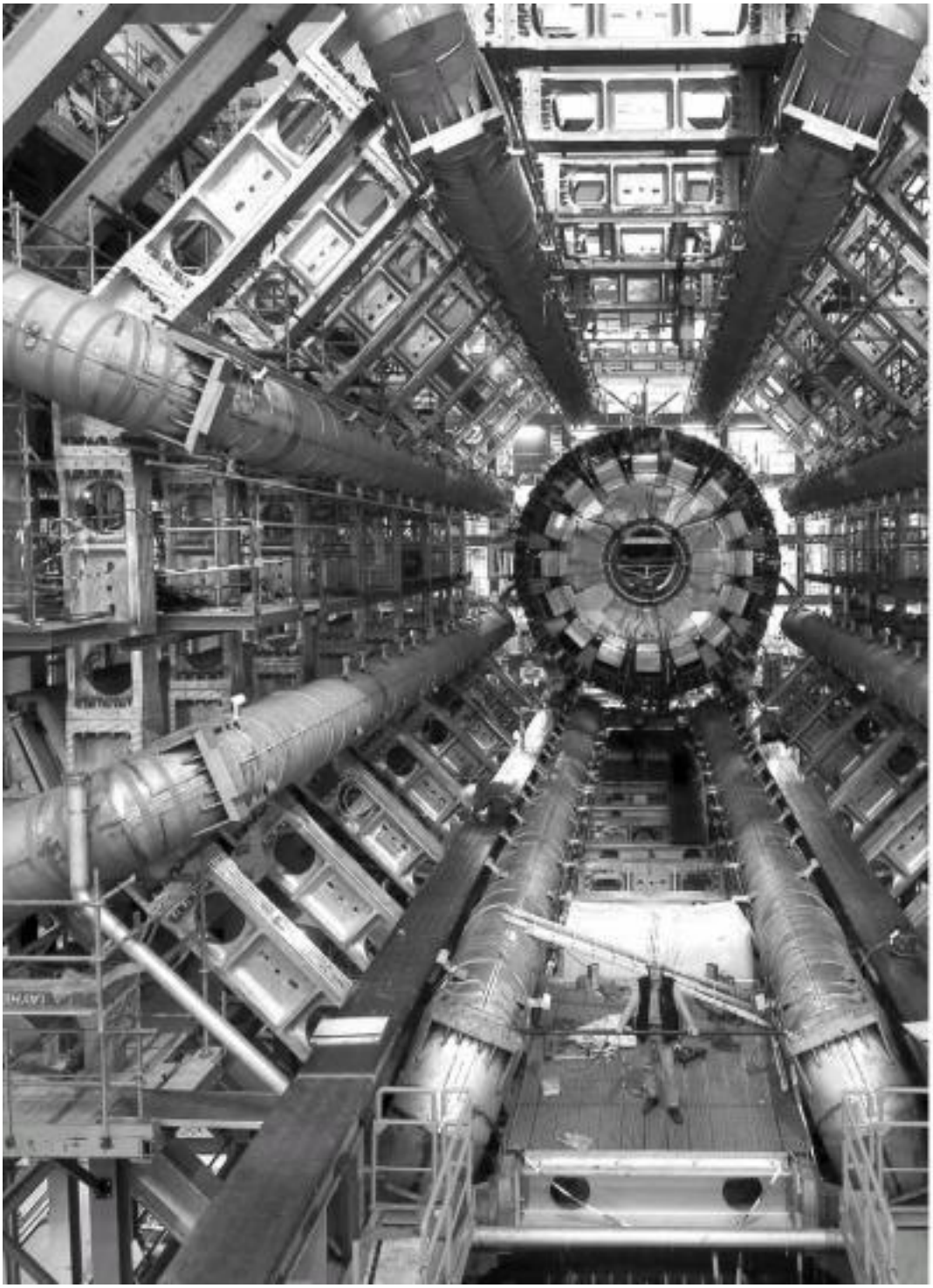
Второе. Уделите достаточно сил изучению теории вероятности, типичных когнитивных ошибок, стратегий эффективного мышления. Усиливайте свой интеллект с помощью новых алгоритмов и психологических практик, специализированных компьютерных программ.

Третье. Сформируйте свой собственный план на будущее, даже если его первая версия и не имеет шансов оказаться правильной.

Четвертое. Несмотря ни на что, начинайте постепенно менять мир, чтобы выжить в нем. Автор этой книги тоже не всегда был авторитетным экспертом в области глобальных рисков, как и все, когда-то он начинал с нуля, испуганно пытаясь осмыслить мир вокруг себя и разобраться в проблемах, казавшихся не по зубам. Но однажды он решил что-то сделать, и у него получилось. Надеюсь, получится и у вас!

Данила Медведев,
кандидат экономических наук,
футуролог, эксперт Российского фонда развития высоких технологий

Введение
Истина и коллайдер



Мы не знаем наверняка, возможен ли конец света – глобальная катастрофа, способная уничтожить человечество. Свои выводы относительно этого мы делаем на основании собственных суждений, которые вольно или невольно могут не отражать истинного положения вещей уже в силу того, что наше представление о мире неполно, а наш подход к изучению проблем зачастую не вполне объективен. Нам удобно верить в то, что никакая наша деятельность или

бездеятельность наблюдаемого равновесия мира необратимо не нарушит, и поэтому истинность наших суждений и высказываний не кажется нам вещью предельно важной. В результате мы пренебрегаем риском, забывая о его цене и заблуждаясь относительно того, что глобальная катастрофа – вопрос послезавтрашнего дня. В определенном смысле «послезавтра» уже наступает, и примеров, показывающих, насколько наш разум к этому готов, уже вполне достаточно.

В 2008 году в центре внимания общественности оказались гипотетические риски, связанные с запуском Большого адронного коллайдера (БАК, или LHC), расположенного на границе Франции и Швейцарии. Причиной тому, возможно, стали выступления ряда ученых, высказавших опасения, что в результате экспериментов, которые будут проводиться на новой установке, может возникнуть черная дыра, которая уничтожит Землю. Чтобы развеять эти опасения, были сделаны оценки безопасности экспериментов, которые показали, что риск катастрофы бесконечно мал. То есть оказалось, что безопасность всей планеты и выживание человечества были поставлены в зависимость от истинности выдвинутых доказательств.

Итак, каким же именно образом была доказана безопасность экспериментов на БАК?

Основные риски, связанные с ними, состоят в том, что в установке могут возникнуть некие гипотетические частицы, которые так или иначе будут способны захватывать частицы обычной материи и поглощать их или трансформировать в подобные себе. В основном речь идет о двух гипотетических частицах – микроскопических черных дырах и стрейнджлетах (страпельках). (Помимо них рассматривались еще два класса возможных опасных объектов – магнитные монополи и пузыри «истинного вакуума».)

Сразу отметим, что полнота этого списка опасных объектов ничем не доказана. То есть, хотя и есть достаточные основания полагать, что все они либо безопасны, либо невозможны, это не значит, что нет какого-то пятого класса частиц, о которых мы можем ничего не знать до тех пор, пока их не откроем.

Почему же эти частицы невозможны?

Потому что их существования не допускает стандартная модель, принятая современной физикой. Однако коллайдер построен именно для того, чтобы исследовать границы применимости стандартной модели и найти возможные ее расширения. Здесь возникает логический парадокс: безопасность коллайдера доказывается через то, что мы знаем, тогда как цель запуска БАК – проникнуть в неведомое. Можно сказать так: чтобы полностью исследовать некую систему, нужно ее разобрать, то есть разрушить. Для познания человеческой анатомии потребовалась патологоанатомия. Соответственно окончательное познание тайн Вселенной потребует ее разрушения.

Хотя стандартная модель и не допускает возникновения микроскопических черных дыр в недавно достроенном коллайдере (так как для этого просто не хватит энергии), некоторые расширения этой модели предполагают наличие дополнительных измерений и делают возникновение таких черных дыр возможным со скоростью порядка одной штуки в секунду. С точки зрения современной физики эти микроскопические черные дыры должны немедленно разрушаться Хокинговским излучением – то есть они будут излучать быстрее, чем будут способны притягивать материю.

Правда, с самим Хокинговским излучением есть небольшая проблема: хотя его концепция выглядит теоретически убедительной, у нее нет никаких экспериментальных доказательств. (Поскольку чем больше черная дыра, тем меньше она излучает «по Хокингу», при этом у наблюдаемых, причем косвенно, космических черных дыр это излучение ничтожно мало и не может быть зафиксировано.)

А что если излучение не сработает?

Предполагается, что даже если микроскопическая черная дыра возникнет в БАК и даже если она не разрушится Хокинговским излучением, она будет настолько малой массы и размеров, что будет намного меньше размеров атома, и ее гравитационное поле тоже будет про-

стираться на расстояния, меньшие размеров ядра атома. Таким образом, такая черная дыра будет очень мало способна к каким-либо реакциям. Она может свободно летать среди вещества, никак с ним не взаимодействуя.

Вместе с тем существует теория, согласно которой в процессе формирования такая микроскопическая черная дыра приобретет электрический заряд или магнитный момент и в силу этого все же начнет гораздо быстрее притягивать к себе электрически заряженные ядра атомов и электроны. По мере роста ее массы ее способность поглощать материю тоже будет расти, и не известно, по какому закону – степенному или экспоненциальному.

Небольшим утешением может быть то, что процесс начального роста микроскопической черной дыры может быть крайне медленным. (Можно, например, предположить, что катастрофа с микроскопической черной дырой уже произошла при работе предыдущих ускорителей (например, RHIC) и мы пока не наблюдаем ее проявлений, так как она пока еще не выросла.)

То есть черная дыра может возникнуть на коллайдере БАК, и никто этого не заметит. Она погрузится в центр Земли, где начнет очень медленно, но с растущей скоростью набирать массу. По некоторым предположениям, это потребует миллионов и миллиардов лет, прежде чем станет заметным – а значит, не угрожает безопасности человечества. Однако, как показано в статье *Benjamin Koch, Marcus Bleicher Horst Stb'cker. Exclusion of black hole disaster scenarios at the LHC* (http://arxiv.org/PS_cache/arxiv/pdf/0807/0807.3349v1.pdf), в случае, если наша Вселенная имеет одно скрытое измерение, время поглощения Земли составит 27 лет, а если два – то десять тысяч триллионов лет. (Понятно, что только первый сценарий заслуживает внимания.) 27 лет – это, конечно, не те несколько секунд, за которые поглощается Земля на известном видеоролике, выложенном на YouTube.

Отметим, однако, что человечество погибнет гораздо раньше, чем произойдет полное поглощение Земли черной дырой. Поскольку примерно половина массы при поглощении вещества черными дырами переходит в энергию излучения (за счет этого светят квазары), то процесс поглощения планеты будет сопровождаться ее разогревом. То есть вначале, например, из-под земли начнут вырываться потоки раскаленных газов в виде мощнейших вулканических извержений, которые сделают атмосферу непригодной для дыхания.

Итак, микроскопическая черная дыра может быть опасной, только если ряд теоретических предположений окажется истинным. Понятно, что это маловероятно, хотя каким образом применять понятие вероятности к тем или иным свойствам законов Вселенной, не вполне ясно.

Однако это еще не все: кроме теоретического способа обоснования безопасности коллайдера существует еще один – основанный на эмпирических свидетельствах.

Эмпирические «заверения в безопасности» строятся на том факте, что энергии космических лучей, которые непрерывно бомбардируют атмосферу Земли, гораздо выше энергий, которые будут достигаться в коллайдере. А раз Земля до сих пор существует, то, значит, и установка безопасна. Более продвинутые версии доказательств используют тот факт, что существуют Луна, нейтронные звезды и белые карлики, несмотря на их непрерывную бомбардировку космическими лучами.

То есть любые эмпирические доказательства безопасности основываются на определенных аналогиях, при том что БАК – сооружение уникальное. Например, говорится о том, что происходящее в коллайдере аналогично тому, что уже триллион триллионов раз происходило на Земле и во Вселенной без каких-либо негативных последствий. Действительно, нет сомнений в том, что случились триллионы столкновений атмосферы Земли с космическими лучами – однако то, что этот процесс ПОЛНОСТЬЮ аналогичен тому, что происходит в коллайдере, это лишь предположение. (Подробно возражения относительно «аналогичности» процессов приводятся в главе «Физические эксперименты», при этом следует подчеркнуть, что наличие

возражений само по себе вовсе не означает, что катастрофа с коллайдером неизбежна или что я в ней уверен.)

Нельзя сказать, что сомнения относительно безопасности коллайдера замалчивались – в течение последних лет вышло несколько статей, в которых обосновывается невозможность катастрофы с черными дырами. При этом, однако, общее свойство этих статей состоит в том, что они появились ПОСЛЕ того, как решение о строительстве коллайдера было принято, десятки тысяч физиков были наняты на работу и миллиарды долларов были потрачены. То есть цель этих статей – не исследовать вопрос о том, каковы реальные шансы катастрофы, а успокоить публику и обеспечить продолжение исследований. (Этим данные статьи отличаются, например, от рассекреченного недавно отчета LA-602 о рисках термоядерной детонации атмосферы, который был представлен перед первыми испытаниями атомной бомбы в США в 1945 году Комптоном, цель которого состояла в исследовании вопроса, а не в успокоении публики.)

Другими словами, гораздо честнее было бы использовать в качестве обоснований рисков не публикации 2007–2008 годов, приуроченные к завершению работ по строительству коллайдера, а публикации 1999 года, на основании которых принимались решения о строительстве. Отметим, что наихудшая оценка риска в публикациях 1999 года, как сообщает Э. Кент в статье «Критический обзор оценок рисков глобальных катастроф», была 1 к 5000.

Кстати, вопрос о том, какой риск катастрофы с коллайдером является приемлемым, заслуживает отдельного рассмотрения.

К 2004 году наиболее твердая оценка риска, выведенная из эмпирических астрофизических наблюдений, показывала шансы получить катастрофу 1 к 50 миллионам. Очевидно, что эта оценка была принята в качестве достаточной, так как строительство было продолжено. Однако математическое ожидание числа жертв, то есть произведение числа погибших – 6 миллиардов на вероятность события составляет в данном случае 120 человек. Ни один другой научный проект с таким ожидаемым числом возможных жертв никогда бы не был допущен к реализации. Например, при захоронении радиоактивных отходов в Великобритании допустимым принимается ожидаемое число жертв только в 0,00001 человека в год. Отметим, что здесь учитывается только гибель ныне живущих людей. Однако вымирание человечества означало бы и невозможность рождения всех последующих поколений людей, то есть число неродившихся людей могло бы составлять тысячи триллионов. В этом случае математическое ожидание числа жертв также возросло бы на несколько порядков. Наконец, гибель Земли означала бы и гибель всей информации, накопленной человечеством.

Другим способом оценки рисков является так называемый астероидный тест. Утверждается, что если риск, создаваемый коллайдером, меньше, чем риск человеческого вымирания в результате падения огромного астероида (примерно в 1 к 100 миллионам в год), то риском первого можно пренебречь. Однако сам риск падения такого астероида является неприемлемым – ведь ради его предотвращения затеваются специальные программы. То есть принятие астероидного теста равносильно утверждению о том, что нет разницы, погибнут ли в авиакатастрофе 300 человек или 301 человек.

Третий способ оценки рисков связан с анализом затрат и рисков, выраженных в денежной форме. Сделать это пытается, например, американский судья и популяризатор науки Р. Познер в своей книге «Катастрофа: риск и реакция».

Сумма выгод, которые мы ожидаем получить от ускорителя, примерно равна его стоимости (если бы она была значительно – скажем, в десять раз – больше, то строительство ускорителей было бы крайне выгодным бизнесом, и многие бы им занимались). Хотя огромная выгода возможна, например, в случае невероятного ценного открытия, вероятность этой выгоды не оценивается как большая.

Стоимость ускорителя составляет около 10 миллиардов долларов. С другой стороны, можно оценить человеческую жизнь. (Американские страховые компании оценивают год

жизни здорового человека в 50 000 долларов.) Отсюда и из разных других оценок получается, что цена человеческой жизни в развитом обществе составляет порядка 5 миллионов долларов. Тогда цена всего человечества примерно равна $3x10^{16}$ долларов. В этом случае приемлемым оказывается риск менее чем 1 к 3 миллионам. Однако если учитывать цену неродившихся поколений, то потребуются гораздо более строгие границы риска. Кроме того, выгодополучатели и объекты риска не совпадают. Выгоду от работы ускорителя получат в первую очередь ученые и, кроме того, люди, интересующиеся наукой, тогда как большинством жертв возможной катастрофы будут люди, которые вообще никогда об ускорителе не слышали.

Впрочем, неготовность вовремя анализировать риски и искать истину, а не утешение – лишь часть проблемы. Если бы в XXI веке только проведение экспериментов на БАК порождало риск глобальной катастрофы, все было бы довольно просто. На самом деле явлений и процессов, несущих угрозу безопасности человечества, гораздо больше. Из уже разворачивающихся событий такого рода можно, например, отметить обещание компании Google создать к 2011 году искусственный интеллект (ИИ), а также развивающийся на мировых рынках кризис (начавшийся с кризиса американской ипотеки) или рост цен на нефть. При этом следует помнить о том, что не все риски сегодня нам известны: чем шире познания о мире и совершеннее технологии, тем шире и список возможных угроз.

Итак, задача, которую мы попытаемся решить в рамках этой книги, – изучение границ нашего знания в отношении рисков глобальной катастрофы, которая может привести к полному вымиранию человечества. Очевидно, зная эти границы, мы легче сможем ответить на вопрос о том, возможна ли вообще такая катастрофа, и если да, то когда, с какой вероятностью, отчего и как ее предотвратить. И даже если ее вероятность в ближайшее столетие равна нулю, мы должны знать это наверняка – результат в данном случае должен быть хорошо обоснован.

В любом случае знать результат нам очень важно. Возьму на себя смелость заявить, что вопрос выживания человеческой цивилизации является важнейшим из тех, которые могут перед ней встать. Уже был период, когда он начал осознаваться как весьма актуальный – в годы холодной войны, – но впоследствии интерес к нему угас и маргинализировался. Отчасти это связано с психологическим феноменом утомления, отчасти – с мыслями о бесполезности публичных усилий. Сегодня трудно представить массовые демонстрации с требованиями запрета опасных биотехнологий, опытов на ускорителях, разработок нанотехнологического оружия.

Так или иначе, сценарии и риски вымирания человечества оказались практически вытесненными в область бессознательного. На то есть свои причины: глобальные катастрофы заслонены от нас как пеленой «технического» незнания (вроде незнания реальных орбит астероидов и тому подобного), так и психологической защитой (по существу скрывающей нашу неспособность и нежелание предсказывать и анализировать нечто ужасное). Более того, глобальные катастрофы отделены от нас и теоретическим незнанием – нам неведомо, возможен ли искусственный интеллект, и в каких пределах, и как правильно применять разные версии теоремы о конце света, которые дают совершенно разные вероятностные оценки времени существования человечества.

Исследование рисков глобальной катастрофы парадоксальным образом оказывается и исследованием природы непредсказуемости, поскольку вопрос о рисках такой катастрофы поднимает множество гносеологических проблем. В первую очередь речь идет об эффекте наблюдательной селекции и принципиальной невозможности и нежелательности решающего эксперимента в области глобальных катастроф.

Следует отметить, что глобальная катастрофа труднопредсказуема по определению. Во-первых, потому что мы не знаем всех вариантов, которые могут выпасть. Во-вторых, нам очень трудно определить их вероятность. В-третьих, некому будет проверить результат. В-четвертых, мы не знаем, когда бросят «монету» и бросят ли ее вообще. Это незнание похоже на то незнание, которое есть у каждого человека о времени и причине его смерти (не говоря уже о том,

что будет после смерти). Но у человека есть хотя бы опыт других людей, который дает статистическую модель того, *что* и с какой вероятностью может произойти.

Другими словами, есть реальность – и есть мир человеческих ожиданий. До определенного момента они совпадают, создавая опасную иллюзию, что наша модель мира и есть мир. В какой-то момент они настолько расходятся, что мы сталкиваемся лоб в лоб с тем, что для нас раньше не существовало. Человек не знает, что с ним будет через 15 минут, но претендует на то, чтобы планировать будущее. Свет нашего знания всегда ограничен. Конец света – это столкновение несовершенства нашего знания с тьмой непостижимого. Следовательно, задачу, которую призвана решать гносеология катастроф, можно сформулировать и так: изучение общих закономерностей того, каким образом неправильное знание приводит к катастрофе. Ведь именно в ситуации глобальной катастрофы разрыв между знаниями и будущей реальностью наиболее велик.

Глава 1
Будущее и природа человеческих заблуждений

Попытки предсказания будущего – прекрасный материал для исследования природы человеческих заблуждений, и наоборот, данные экспериментальной психологии позволяют нам оценить, насколько ограничены мы в предсказании будущего.

Нет смысла пытаться заглядывать в будущее, не исследовав границы своего возможного знания о нем. Эти границы обозначат горизонт нашей способности к прогнозу. Существует

более ста возможных ошибок, которые могут совершить люди, пытаясь предсказать будущие катастрофы. Но вместо того чтобы перечислять их, мы постараемся исследовать их вероятные корни, то есть причины человеческой склонности к ошибкам.

Основной корень возможных ошибок состоит в том, что человек «не предназначен» для оценки рисков глобальных катастроф. Не предназначены для этого ни его мозг, ни обычный взгляд на мир, ни созданный человеком научный метод. Рассмотрим поочередно каждый из этих трех источников неправильных представлений о сущности и вероятности глобальных катастроф.

Мозг и эволюция

Человеческий мозг сформировался в процессе эволюции. Поэтому, в первом приближении, мы можем сказать, что в ходе этого процесса в нем сохранились и усилились те качества, которые способствовали выживанию людей и заселению ими всей Земли. Отсюда можно было бы заключить, что люди успешно избегали катастроф, приводящих к вымиранию вида (в прошлом, во всяком случае). Те линии людей, которые не умели этого избегать, не дожили до наших дней, как, например, неандертальцы. Однако люди выживали вовсе не потому, что у них развился некий навык предотвращать катастрофы, а чисто статистически, за счет того, что они обладали большей живучестью.

(Одной из возможных причин вымирания неандертальцев называют прионную инфекцию в духе коровьего бешенства, которая распространилась во всей их популяции в течение 100–200 лет за счет практики каннибализма; в то же время запрет на каннибализм, распространенный среди большинства сообществ homo sapiens, способствовал их выживанию – однако этот запрет никаким образом не приводил к росту понимания того, какова его реальная причина.)

Эта человеческая живучесть, безусловно, является ценным ресурсом в случае будущих глобальных катастроф, но она ничего не дает для понимания соответствующих рисков, так как является свойством вида, а не отдельных людей. Для выживания вида каждый человек должен был дожить до возраста рождения детей или немногим более, и в силу этого его **понимание рисков ограничивалось преодолением, в первую очередь, краткосрочных угроз**. Более того, рискованное поведение отдельных людей способствовало выживанию вида в целом, побуждая отправляться на заселение далеких островов и иметь больше детей.

Кроме того, homo sapiens, для того чтобы стать единственным хозяином планеты, вынужден был развить в себе **навык уничтожения целых классов разумных существ**, которые могли бы стать ему конкурентами. За примерами не надо ходить далеко: от тех же неандертальцев до тасманийских аборигенов, уничтоженных поголовно в XIX веке; кроме того, человек уничтожил мегафауну на всех континентах, кроме Африки.

В XX веке люди решили (опыт мировой войны), что геноцид – это не способ решения проблем. Тем не менее в багаже очевидных решений, приходящих нам на ум, сохранилась идея «уничтожить всех врагов». Нет нужды говорить, что вид, обладающий таким навыком, как «уничтожение всех», становится опасным сам для себя, когда отдельным его представителям попадают в руки достаточные для реализации такого проекта технические средства. В данном случае «уничтожить всех» – это модель поведения, а не ошибка, но она может проявляться и как ошибка в планировании будущего. Ошибка здесь состоит в том, что попытки «уничтожить всех» ведут не к решению проблемы, а наоборот, только к ее усугублению. Подтверждение этому мы видим на примере конфликта Израиля и палестинцев: чем больше израильтяне хотят уничтожить всех террористов, тем больше палестинцы хотят уничтожить всех израильтян и порождают из своей среды террористов. И наоборот, чем меньше стороны пытаются «решить проблему», тем менее актуальна и сама проблема.

Следующая «собака», которую нам подложила эволюция, это как наш мозг сопоставляет рискованность и полезность действий. Хотя эволюция отсеивала тех людей, которые слишком сильно рисковали собой и погибали в детстве, она также отсеивала и тех, кто слишком сильно стремился к безопасности. Происходило это, возможно, следующим образом: допустим, в стаде было четыре «рисковых парня». Они бурно выясняли между собой отношения, в результате чего трое погибли, а четвертый стал вожаком стада, покрыл всех самок и у него родился десяток детей. В силу этого рискованное поведение закрепилось, так как, несмотря на то, что большинство склонных к риску самцов погибло, один выживший оставил большое потомство. Более того, склонность к риску позволила ему повысить свой социальный статус и даже стать вожаком стада. С другой стороны, самец, который стремился бы к абсолютной безопасности и избегал стычек с другими самцами, в конечном счете утратил бы статус в стаде и оставил мало потомства. Наоборот, рисковый самец, став вожаком, будет «руководить» рискованными методами. Опять же, в силу этого какие-то племена погибнут, какие-то – попадут на новые территории и заселят их.

Более близкий пример: Рональд Рейган, объявляя крестовый поход против СССР, увеличивал своими действиями риск ядерной войны, однако выигрыш его интересовал больше, чем проигрыш. Существует много исследований, которые показывают, что **восприятие человеком вероятностей выигрышей и проигрышей значительно отличается от поведения абсолютно рационального субъекта**. Например, человеку свойственно пренебрегать небольшой вероятностью очень плохого исхода – «тяжелым хвостом». С нашей эволюционной точки зрения это можно связать с тем, что человек не был бессмертным существом и знал это, поэтому он мог просто не дожить до редкого наихудшего случая, ущерб от которого перевесил бы полезность от всех позитивных исходов, вместе взятых.

Доказано, **что людям свойственно считать вероятности рисков, меньшие 0,001 процента, равными нулю**. Понятно, почему это происходит: если средняя жизнь человека включает примерно 20 000 дней, то добавление на каждый день вероятности смерти порядка 0,001 процента изменит ее ожидаемую продолжительность только на несколько тысяч дней, что находится в пределах погрешности оценки продолжительности жизни. Люди выбирают езду на машине, питье, курение и т. д. потому, что риск смертельной аварии или заболевания попадает в эти пределы.

Однако такая оценка определенно не годится для обеспечения безопасности человечества, потому что она означала бы гарантированное вымирание человечества в течение 300 лет. Кроме того, люди обычно не складывают вероятности, при том что многократное повторение разных событий с вероятностью в одну стотысячную может дать значительный эффект.

Здесь же мы можем отметить психологический эффект, состоящий в том, что риск сам по себе запускает механизмы вознаграждения в мозгу, в силу чего некоторые **люди склонны повышать свою норму риска** – чтобы «поймать адреналин». Безусловно, это вопрос личного выбора этих людей, пока это касается только их самих. Однако не составляет труда привести примеры ситуаций, когда люди рисковали жизнями сотен людей ради собственного развлечения. Например, еще в советское время пилот самолета ТУ-154 на спор с товарищами взялся посадить самолет с закрытыми шторками окон кабины. Более ста человек погибли. Нетрудно вообразить ситуацию из не очень отдаленного будущего, где кто-то ради развлечения рискнет судьбой всего человечества.

Ведь чем выше ставки, тем выше выброс адреналина! Например, ракетчики на одной американской базе придумали от скуки, как можно обойти систему запуска, требующую двух человек, с помощью одного человека, ножниц и нитки.

Интересна также склонность людей выбирать свою норму риска. Например, при сравнении смертности на разных классах автомобилей (на основании статистики США) оказалось, что на более безопасных по заводским тестам машинах гибнет примерно такое же количество

людей, как и на более дешевых и менее безопасных машинах. Это было связано с тем, что люди на более безопасных машинах ездили агрессивнее и быстрее, а на менее безопасных – осторожнее. В отношении глобальных катастроф это выглядит следующим образом: когда человек едет на машине, он зависит от своей машины и от машины на встречном курсе, то есть от двух машин. Но человек, живущий на Земле, зависит от всего множества войн, терактов и опасных экспериментов, которые могут иметь глобальный масштаб. И это приводит к тому, что для него складываются тысячи норм риска, которые допускают разные люди, решившиеся на реализацию этих проектов.

Как принимаются «обоснованные» решения

Другой корень неверных представлений, сформировавшихся эволюционно, состоит в крайне опасной связи между правотой, уверенностью и высоким социальным статусом. Одним из важнейших проявлений этого является неизменно присущее человеку качество сверхуверенности. Сверхуверенность состоит в том, что **людям свойственно приписывать слишком высокую достоверность собственным мнениям**. Сверхуверенность проявляется также в чувстве собственной важности, то есть в уверенности в своем более высоком социальном статусе и более высоких способностях, чем это есть на самом деле. Большинство опрошенных социологов считают, что они принадлежат к 10 лучшим в мире специалистам. Три четверти водителей считают, что они водят машину лучше, чем среднестатистический водитель.

Другое проявление сверхуверенности – это **очень низкая способность сомневаться в истинности своих высказываний**. В экспериментах это проявляется, когда испытуемым предлагают дать оценку некой неизвестной величины (например, числа яиц, производимых в США в год), а затем просят указать интервал в 99 процентов уверенности вокруг этой величины. Несмотря на то что никто не мешает людям указать очень широкий интервал от нуля до бесконечности, люди все равно указывают слишком узкие интервалы, и реальная величина в них не попадает. Даже если люди являются экспертами в своей области (например, инженерами-гидротехниками), их интервал уверенности оказывается слишком узким. Даже если людей заранее предупреждают о том, что такой эффект имеет место быть, и просят его избегать, все равно их интервал уверенности оказывается слишком узким. Даже если людям предлагают денежное вознаграждение в размере трехмесячной зарплаты за правильный результат, все равно их интервал уверенности оказывается слишком узким. Из этого следует, что проблема не в том, что люди не хотят угадать правильный интервал уверенности: они действительно этого не могут сделать – не могут преодолеть свою сверхуверенность, которая «прошита» в устройстве их мозга.

Мы можем часто сталкиваться с проявлениями такой сверхуверенности у других людей, когда слышим заявления о том, например, что вероятность ядерной войны в XXI веке будет нулевой, или наоборот, что она неизбежна. Чужая сверхуверенность достаточно очевидна, однако наиболее опасна своя, которую практически невозможно обнаружить собственными силами, да людям обычно и несвойственно искать в себе признаки сверхуверенности и стремиться уменьшить свою уверенность в сделанных выводах. Наоборот, люди стараются увеличить свое ощущение уверенности в своих выводах, и в результате на какое-нибудь случайное утверждение наматывается ком селективно подобранных доказательств.

Такой подход известен как wishful thinking – то есть мышление, обусловленное желанием что-то доказать, подкрепить свою уверенность в некой теме. Целью такого мышления вовсе не является поиск истины, отсюда и результат – обычно таким образом истинные высказывания не получаются. Другим названием этой модели поведения является «рационализация» – то есть **подбор рациональных оснований под некое иррациональное решение** . В ходе такого отбора противоречащие доводы отсеиваются, а также нет попыток фальсифицировать

свое исходное мнение – то есть проверить его устойчивость к опровержениям. Интернет оказывает в этом дурную услугу, так как поиск приносит в первую очередь подтверждения. Например, набрав в поиске «взрыв Земли», я наверняка наткнусь на кого-то, кто думает, что это возможно.

Другое проявление борьбы за социальный статус состоит в **резкой поляризации мнений в случае спора**. Если человек колеблется между двумя гипотезами, то столкновение с оппонентом заставляет его выбрать одну из гипотез и начать ее защищать. В этом смысле спор не рождает истину.

Поскольку вопрос о возможности глобальной катастрофы будет оставаться спорным вплоть до самого конца, то в связи с ней возможно особенно много споров. И если в результате этого спора кто-то преувеличит вероятность глобальной катастрофы, то это не пойдет на пользу выживания человечества. Переоценка какого-то одного фактора неизбежно означает недооценку другого. В связи со спорами возникает проблема убедительности. Единственный способ для общества начать предотвращать некую возможную глобальную катастрофу – это то, что кто-то его убедит в близости и реальности такого события, а также в возможности и необходимости его предотвращения.

Но «убедительность» не тождественна истине. Некоторые сценарии глобальной катастрофы будут более «убедительными» за счет своей красочности, наличия исторических примеров и того, что их защищали лучшие спорщики. Например, падение астероидов. Другие будут гораздо более трудно доказуемыми, и их «пиар» будет менее возможным. Например, катастрофа на ускорителе в ходе физических экспериментов.

Следующий **корень ошибочных умозаключений и моделей поведения лежит в эмоциональных реакциях**. Хотя доказательства вероятности некой катастрофы могут быть весьма строгими, эти выкладки понятны только тому ученому, который их сделал, а лица, принимающие решения, не будут перепроверять эти выкладки – будут реагировать на них в значительной мере эмоционально. Поэтому реакция публики, академии наук, парламента, правительства, президента на любые, даже самые серьезные предупреждения, будет эмоциональной, а не логической. Более того, поскольку человеку свойственно определять свою позицию в течение 10 секунд, а потом начать подбирать факты для ее защиты, эта эмоциональная реакция имеет шанс закрепиться. В силу сказанного, даже при наличии очень убедительных доказательств (а чем убедительнее доказательства, тем в большей мере катастрофа уже назрела и тем меньше времени осталось для борьбы с ней), все равно надо учитывать особенности эмоционального реагирования людей, так как окончательное решение в конечном счете будет зависеть не от тех, кто это доказательство вывел. Кроме того, людям свойственно чувствовать себя экспертами по глобальным вопросам, поскольку это повышает их самооценку. Например, если человека спросить, каков порядковый номер лантана в таблице Менделеева, то он, если он не химик, легко признается, что не знает этого и готов посмотреть в таблицу; однако если спросить его о некой глобальной проблеме, например о риске ядерной войны, то он сразу даст ответ, вместо того чтобы посмотреть существующую литературу на эту тему.

Далее, **естественной психологической реакцией является защита от неприятного знания**. Первым уровнем такой защиты является состояние отрицания в духе «это слишком плохо, чтобы быть правдой». Действительно, глобальная катастрофа, ведущая к вымиранию всех людей, – это наихудшее событие, которое может с нами случиться. Тем более что не обязательно она будет быстрой, мгновенной и красивой, а может быть долгой и мучительной, скажем, в случае глобального радиоактивного заражения. Поэтому нетрудно предположить, что психологические механизмы защиты включатся на всю мощь, чтобы уменьшить ее ужас, а главное – или счесть ее невозможной, или вообще исключить из сознания, вытеснить. Поскольку люди знают, что идеи в духе «это слишком плохо, чтобы быть правдой» не имеют под собой никаких оснований, и каждый может вспомнить случаи из жизни, когда происхо-

дили вещи настолько плохие, что в это трудно поверить (ребенок, заболевший раком; невеста, умирающая накануне свадьбы, и т. д.), то этот тезис подменяется другим, а именно: «это слишком невероятно, чтобы быть правдой». Последнее высказывание по своей логической природе является тавтологией: «этого не может быть, потому что не может быть никогда». Однако в отношении глобальных окончательных катастроф их «невероятность» выводится, например, из того, что они не происходили в прошлом. По ряду причин, которые мы подробно рассмотрим дальше, это, однако, ничего не значит (например, в связи с наблюдательной селекцией). Уникальные события регулярно случаются.

Вероятность глобальной катастрофы отвергается также потому, что это очень большое событие. Но и очень большие события иногда происходят, более того, они происходят рано или поздно. Следовательно, подобное преуменьшение вероятности носит в первую очередь эмоциональный характер.

По одной из теорий, психологическая реакция на катастрофу или известие о неизлечимой болезни, называемая «горевание», проходит через пять стадий: отрицание, гнев, попытка заключить сделку с судьбой, депрессия, принятие. Можно предположить, что и эмоциональная реакция на риск глобальной катастрофы будет проходить через похожие стадии. Тогда стремлению «заключить сделку с судьбой» будут соответствовать попытки избежать катастрофы с помощью бункеров и т. д.

То, что известно в быту как **«стадный инстинкт»**, проявляется в психологических экспериментах по исследованию конформизма. Когда группа подсадных испытуемых единогласно утверждает, что белое – это черное, то значительная доля реальных испытуемых, находящихся в этой группе, тоже не верит своим глазам и боится высказать несогласие с группой. Особенно силен этот эффект, когда нужно выступить в одиночку против группы. По крайней мере до недавнего времени люди, высказывающиеся о значительном риске глобальных катастроф, особенно со стороны неких принципиально новых источников, оказывались в похожей ситуации. Они в одиночку должны были выступать против общественного мнения, справедливо опасаясь отвержения обществом и десоциализации. Однако и **общество, со своей стороны, часто весьма заинтересовано в отвержении новых идей** (точно так же, как старый вожак стаи отвергает претензии молодого самца на лидерство). Например, от предложения идеи анестезии в конце XVIII века до реального ее применения прошло почти 50 лет, хотя необходимые препараты – эфир – были уже известны. То же самое произошло и с идеей дезинфицировать руки перед операцией: в середине XIX века врачи все еще не верили в заражение от бактерий, и за десятки лет от того момента, когда идея была высказана, и до того, как она была принята, миллионы рожениц погибли от родовой горячки.

Научный метод и нежелательность экспериментов

Теперь обратимся к возможным причинам недооценки рисков глобальных катастроф, происходящих из устройства науки. Научный метод был создан, чтобы преодолеть все виды интеллектуальных искажений, связанных с ненадежностью человеческого мозга, и получить доступ к по-настоящему достоверному знанию. Обобщая, скажу, что важной частью научного метода является экспериментальная проверка результатов. Именно предсказуемые повторяющие результаты наблюдений и экспериментов позволяют отсеять ложные теории и подтвердить истинные. Благодаря этому какие бы систематические ошибки ни совершил человек, придумывая теорию, они с большой вероятностью будут вскрыты практикой. Таким образом, научный метод компенсировал несовершенство человеческого мозга экспериментальной проверкой.

Однако ничего подобного не происходит в случае исследований рисков глобальных катастроф. Экспериментальная проверка любых теорий о глобальной катастрофе нежелательна, и более того, физически невозможна, так как в случае успеха не останется ни одного наблюда-

теля. В силу этого **классический экспериментальный метод в случае глобальных катастроф буксует и не выполняет своей функции по устранению различных интеллектуальных искажений**.

Границы человеческой склонности ошибаться в прогнозах будущего стали притчей во языцех. Например, газета «Нью-Йорк таймс» опубликовала передовицу, посвященную очередной неудаче в попытках создать самолет, которая закончилась бесславным падением в воду, и предрекла, что свойства материалов таковы, что не удастся создать самолет в течение ближайшего миллиона лет. Однако братья Райт испытали свой самолет через 9 дней после этой статьи, и об этом событии газета не написала ни слова. То есть коэффициент ошибки составил 40 миллионов раз. В этой истории важно так же то, что самые важные открытия совершались вдали от людских глаз и становились великими событиями с огромными последствиями только задним числом.

Одна из основных ошибок в области футурологии состоит в том, что **людям свойственно переоценивать тенденции, касающиеся ближайшего будущего, и недооценивать качественные изменения, которые проявятся в более отдаленном будущем**. Классический пример такой оценки – прогноз о том, что рост числа повозок в Лондоне приведет к тому, что все жители города станут кучерами, а навоз поднимется до уровня крыш. При этом интересно, что качественно этот прогноз сбылся: почти все стали водить машины, под капотами которых скрыты десятки лошадиных сил, а вредные выхлопы двигателей поднялись значительно выше крыш.

Теперь остановимся на крайне важном различии между наилучшим, наиболее вероятным и наихудшим результатами.

Склонность людей недооценивать вероятность наихудшего исхода регулярно проявляется при реализации крупных проектов. Президент Буш оценивал стоимость будущей иракской войны в 50 миллиардов долларов, однако один из его экономистов выдал более реалистичную оценку в 200 миллиардов, за что был уволен. Теперь суммарные расходы на войну оцениваются в сумму между одним и двумя триллионами долларов. Истребитель F-22 должен был стоить 30 миллионов долларов, а стал стоить 300. МКС должна была стоить около 8–10 миллиардов долларов, а обошлась более чем в 100.

То же самое происходит и с оценками надежности и безопасности. Атомные станции должны были иметь надежность в одну крупную аварию на миллион лет эксплуатации, однако Чернобыль произошел, когда суммарная наработка всех станций в мире составляла порядка 10 000 лет эксплуатации. Космические корабли «Спейс шаттл» были рассчитаны менее чем на одну аварию на 1000 полетов, но первая авария произошла на 25-м полете. То есть исходная оценка надежности 1 к 25 была бы более точной.

То есть реальная ситуация оказывается в несколько десятков раз хуже, чем задуманная и рассчитанная. Все же она оказывается лучше, чем в алармистских предупреждениях в духе того, что шаттл вообще никогда не взлетит, Ирак завоевать не удастся и т. п. Алармистские прогнозы точно так же нереалистичны, только с обратным знаком.

Теперь разберемся со «страшилками». Очевидно, что выработался своего рода **условный рефлекс на сообщения о возможных рисках и опасностях** в духе: это все страшилки, нас пугают, чтобы привлечь внимание и вытрясти денег, но мы это раскусили и потому бояться не будем. Действительно, имеется целый класс «прогнозов» в духе: «к земле летит гигантский астероид», «на дне Неаполитанского залива лежит 20 атомных бомб» и т. п., которые были придуманы и нарочно тиражируются СМИ, чтобы что-то получить с человеческой паники. Результат же – как в сказке про волков: восприятие подобных предупреждений притупляется, и когда появляется реальное предупреждение, его никто не слышит. Я хочу подчеркнуть, что читатель имеет право воспринимать эту книгу как очередную «страшилку», но прошу отметить, что «волк» существует независимо от того, какую игру со страшилками мы,

люди, устроили. При этом здесь вы не найдете призывов о том, что нужно срочно бежать и спасать мир определенным образом – я не знаю, как это сделать.

Следующее важное обстоятельство, часто не принимаемое во внимание при прогнозах будущего, состоит в том, что **более быстрые процессы затмевают более медленные**. Например, если у нас в чашке Петри посеяно несколько культур бактерий, то через какое-то время наибольший объем будет занимать та культура, которая растет быстрее всех. В отношении предсказания будущего это означает, что нам следует выделять не самые большие процессы, а самые быстрорастущие. Например, бессмысленно рассматривать рост населения Земли до 2100 года в отрыве от биотехнологий, потому что эти биотехнологии или уничтожат всех людей, или резко продлят жизнь, или обеспечат всех достаточным количеством пропитания. Более того, и внутри биотехнологий нам следует выделять наиболее быстрорастущие направления.

Человеческие мнения крайне подвержены влиянию фоновых обстоятельств, то есть того, каким образом оформлено высказывание, а не того, что именно в нем сказано. Что звучит страшнее: «Русские посылают сообщения инопланетянам со своего радиотелескопа, выдавая им расположение Земли» или «НАСА транслирует песню Биттлз “Через Вселенную” в сторону Полярной звезды?» Первое сообщение вызывает в целом осуждение, а второе – одобрение, потому что в нем использованы слова, которые связаны с приятными ощущениями. Хотя с физической точки зрения происходит одно и то же. Это говорит о том, насколько наши мнения зависят от того, что, по сути, не важно.

Через 20 или 30 лет люди, если еще будут живы, составят новый список глобальных угроз, потому что эта проблема, однажды возникнув, никуда не денется. И они будут просматривать те списки рисков, которые мы предлагаем сейчас, и будут поражаться их наивности, неполноте и односторонности. Насколько бы совершенные списки рисков мы сейчас ни составляли, мы только царапаем по поверхности этой проблемы и должны быть готовы к тому, что в будущем эти списки будут значительно доработаны и многие наши ошибки и иллюзии станут очевидными.

Рекомендуемая литература[1]

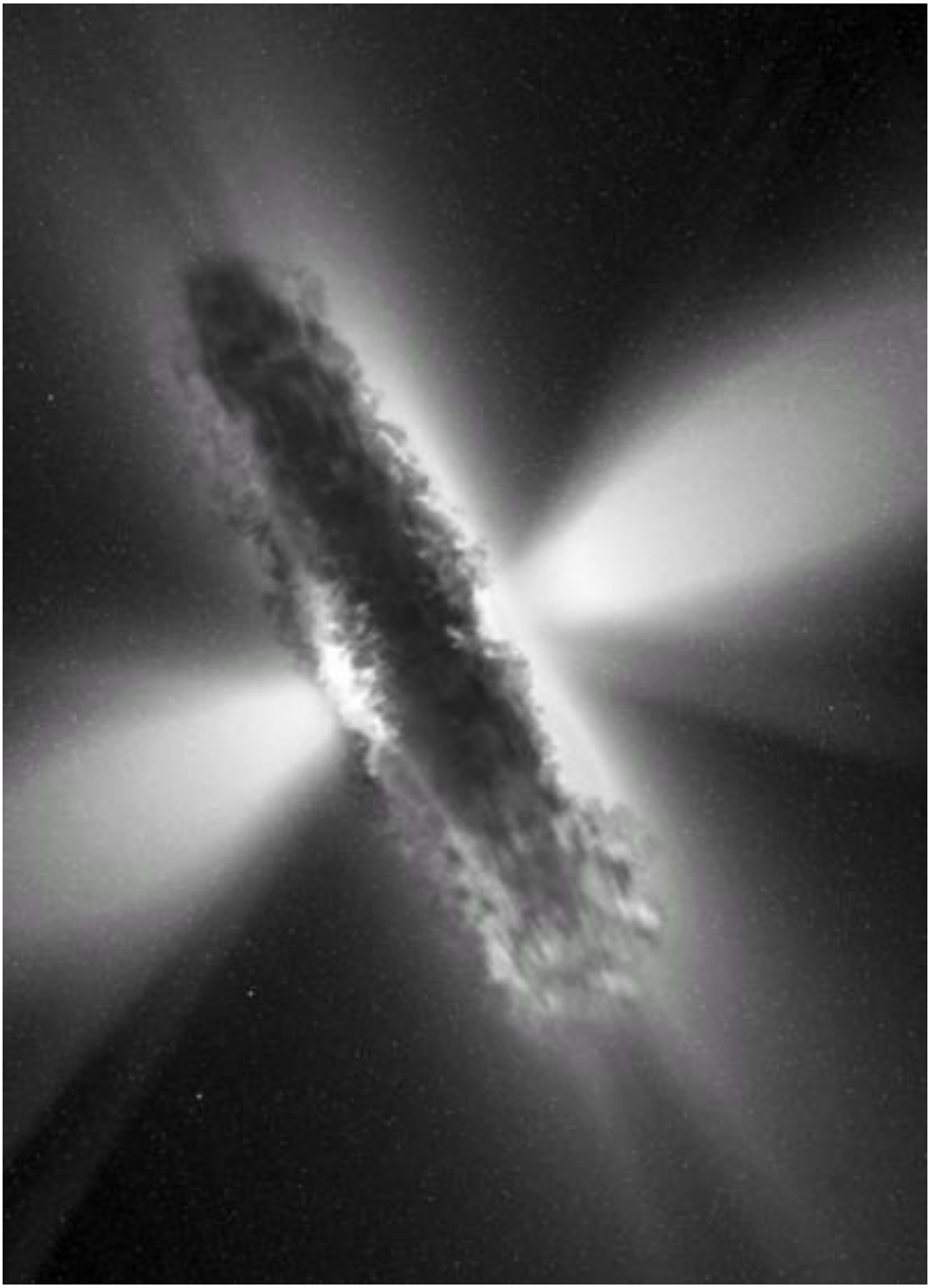
Росс Л., Нисбетт Р. Человек и ситуация: уроки социальной психологии. – М.: Аспект Пресс, 1999.

Турчин А.В. О возможных причинах недооценки рисков гибели человеческой цивилизации // Проблемы управления рисками и безопасностью: труды Института системного анализа Российской академии наук. – Т. 31. – М.: КомКнига, 2007.

Юдковски Е. Систематические ошибки в рассуждениях, потенциально влияющие на оценку глобальных рисков // Сборник «Диалоги о будущем». – М.: УРСС, 2008.

[1] К каждой главе я приложу список из трех-пяти наиболее интересных работ по данной теме, которые стоит почитать; но это не полный список: заинтересовавшиеся читатели без труда найдут более полный список работ в Интернете.

Глава 2
Горизонт прогноза: непредсказуемость против непостижимости



Вот как определяются эти два понятия у К. Кастанеды: «Неведомое в свое время становится известным. Непостижимое, с другой стороны, – это неописуемое, немыслимое, неосуществимое. Это что-то такое, что никогда не будет нам известно, однако это что-то есть там – ослепляющее и в то же время устрашающее в своей огромности».

Но прежде чем перейти к рассмотрению этих предметов, необходимо обратить внимание на следующее: будущего не существует. Нам трудно сказать: «В будущем, завтра утром я пойду на работу». Когда я собираюсь через неделю дописать главу, через месяц поехать на юг, а через год написать новую книгу, я действую в продолженном настоящем. В будущем же может быть все что угодно. В будущем я могу стать альпинистом или переселиться в Японию, жениться на китаянке или вживить себе в мозг имплантат. Все это никак не связано с существующими событиями и в равной мере отражает игру моей фантазии и спектр пространства возможностей. Про будущее можно рассказывать сказки или страшилки, которые не побуждают ни к каким действиям, так как правильно распознаются как развлекательные тексты.

В эссе «О невозможности прогнозирования» С. Лем пишет: «Здесь автор провозглашает тщетность предвидений будущего, основанных на вероятностных оценках. Он хочет показать, что история сплошь состоит из фактов, совершенно немыслимых с точки зрения теории вероятностей. Профессор Коуска переносит воображаемого футуролога в начало XX века, наделив его всеми знаниями той эпохи, чтобы задать ему ряд вопросов. Например: “Считаешь ли ты вероятным, что вскоре откроют серебристый, похожий на свинец металл, который способен уничтожить жизнь на Земле, если два полушария из этого металла придвинуть друг к другу, чтобы получился шар величиной с большой апельсин? Считаешь ли ты возможным, что вон та старая бричка, в которую господин Бенц запихнул стрекочущий двигатель мощностью в полторы лошади, вскоре так расплодится, что от удушливых испарений и выхлопных газов в больших городах день обратится в ночь, а приткнуть эту повозку куда-нибудь станет настолько трудно, что в громаднейших мегаполисах не будет проблемы труднее этой? Считаешь ли ты вероятным, что благодаря принципу шутих и пинков люди вскоре смогут разгуливать по Луне, а их прогулки в ту же самую минуту увидят в сотнях миллионов домов на Земле? Считаешь ли ты возможным, что вскоре появятся искусственные небесные тела, снабженные устройствами, которые позволят из космоса следить за любым человеком в поле или на улице? Возможно ли, по-твоему, построить машину, которая будет лучше тебя играть в шахматы, сочинять музыку, переводить с одного языка на другой и выполнять за какие-то минуты вычисления, которых за всю свою жизнь не выполнили бы все на свете бухгалтеры и счетоводы? Считаешь ли ты возможным, что вскоре в центре Европы возникнут огромные фабрики, в которых станут топить печи живыми людьми, причем число этих несчастных превысит миллионы?” Понятно, говорит профессор Коуска, что в 1900 году только умалишенный признал бы все эти события хоть чуточку вероятными. А ведь все они совершились. Но если случились сплошные невероятности, с какой это стати вдруг наступит кардинальное улучшение и отныне начнет сбываться лишь то, что кажется нам вероятным, мыслимым и возможным?»

Вместо того чтобы пытаться предсказать будущее, нам надо сосредоточиться на его принципиальной непостижимости. Правильные предсказания скорее похожи на случайные попадания из большого количества выстрелов, чем на результат работы некого систематического метода. В силу этого человек, вооруженный моделью принципиально непредсказуемого будущего, может оказаться в выигрыше по отношению к тому, кто думает, что может знать конкретное будущее. Например, принципиальное знание непредсказуемости игры в рулетку удерживает рационального субъекта от игры в нее и от закономерного проигрыша, тогда как любой человек, верящий в возможность ее предсказания, рано или поздно проигрывает.

С другой стороны, принципиальная непредсказуемость будущего человеческой цивилизации буквально является концом света – то есть тьмой в конце туннеля. Конец света, то есть крах нашей познавательной способности, уже случился. И потребуется только время, чтобы он материализовался в крах физический, подобно тому, как погасший фонарь в темноте рано или поздно означает падение. Парадоксальным образом, однако, эта невозможность знать будущее в самых его главных аспектах сопровождается и даже вызывается ростом нашего знания о настоящем и об устройстве мира в деталях.

Но вовсе не та непредсказуемость, которая нам известна, имеет значение. Гораздо опаснее та, которая скрыта под личиной достоверного знания. Как говорят в разведке: «Только тот, кому по-настоящему доверяют, может на самом деле предать». Когнитивные искажения, описанные в предыдущей главе, и в первую очередь сверхуверенность, приводят к тому, что наша картина мира оказывается существенно отличающейся от реальности и рано или поздно терпит катастрофу, сталкиваясь с ней.

Непостижимость и невычислимость

Целый ряд принципиально важных для нас процессов настолько сложен, что предсказать их невозможно, поскольку они невычислимы.

Невычислимость может иметь разные причины.

• Она может быть связана с непостижимостью процесса (например, технологическая сингулярность или, например, то, как теорема Ферма непостижима для собаки), то есть связана с принципиальной качественной ограниченностью человеческого мозга. (Такова наша ситуация с предвидением поведения сверхинтеллекта в виде ИИ.)

• Она может быть связана с квантовыми процессами, которые делают возможным только вероятностное предсказание, то есть недетерминированностью систем (прогноз погоды, мозга).

• Она может быть связана со сверхсложностью систем, в силу которой каждый новый фактор полностью меняет наше представление об окончательном исходе. К таковым относятся модели глобального потепления, ядерной зимы, глобальной экономики, модели исчерпания ресурсов. Четыре последние области знаний объединяются тем, что каждая описывает уникальное событие, которого еще никогда не было в истории, то есть является опережающей моделью.

• Невычислимость может быть связана с тем, что подразумеваемый объем вычислений хотя и конечен, но настолько велик, что ни один мыслимый компьютер не сможет его выполнить за время существования вселенной (такая невычислимость используется в криптографии). Этот вид невычислимости проявляется в виде хаотической детерминированной системы.

• Невычислимость связана также с тем, что, хотя нам может быть известной правильная теория (наряду со многими другими), мы не можем знать, какая именно теория правильна. То есть теория помимо правильности должна быть легкодоказуемой для всех, а это не одно и то же в условиях, когда экспериментальная проверка невозможна. В некотором смысле способом вычисления правильности теории, а точнее – меры уверенности в них, является рынок, где делаются прямые ставки или на некий исход, или на цену некого товара, связанного с прогнозом, например цены на нефть. Однако на рыночную цену теории влияет много других факторов: спекуляции, эмоции или нерыночная природа самого объекта. (Бессмысленно страховаться от глобальной катастрофы, так как некому и не перед кем будет за нее расплачиваться, и в силу этого можно сказать, что ее страховая цена равна нулю.)

Еще один вид невычислимости связан с возможностью осуществления самосбывающихся или самоотрицающих прогнозов, которые делают систему принципиально нестабильной и непредсказуемой.

• Невычислимость, связанная с предположением о собственном местоположении (self-sampling assumption – см. об этом книгу Н. Бострома[2]). Суть этого предположения состоит в том, что в некоторых ситуациях я должен рассматривать самого себя как случайного представителя из некоторого множества людей. Например, рассматривая самого себя как обычного

[2] Nick Bostrom. Antropic principle in science and philosophy. L. 2003Nick Bostrom. Antropic principle in science and philosophy. L. 2003

человека, я могу заключить, что я с вероятностью в 1/12 имел шансы родиться в сентябре. Или с вероятностью, допустим, 1 к 1000 я мог бы родиться карликом. К невычислимости это приводит, когда я пытаюсь применить предположение о собственном местоположении к своим собственным знаниям.

Например, если я знаю, что только 10 % футурологов дают правильные предсказания, то я должен заключить, что с шансами 90 % любые мои предсказания неправильные. Большинство людей не замечают этого, поскольку за счет сверхуверенности и повышенной оценки рассматривают себя не как одного из представителей множества, а как «элиту» этого множества, обладающую повышенной способностью к предсказаниям. Это особенно проявляется в азартных играх и игре на рынке, где люди не следуют очевидной мысли: «Большинство людей проигрывают в рулетку, следовательно я, скорее всего, проиграю».

• Похожая форма невычислимости связана с информационной нейтральностью рынка. (Сказанное далее является значительным упрощением теории рынка и проблем информационной ценности даваемых им показателей. Однако более подробное рассмотрение не снимает названную проблему, а только усложняет ее, создавая еще один уровень невычислимости – а именно невозможности для обычного человека ухватить всю полноту знаний, связанную с теорией предсказаний, а также неопределенности в том, какая именно из теорий предсказаний истинна. См. об информационной ценности рынка так называемую no trade theorem.[3])

Идеальный рынок находится в равновесии, в котором половина игроков считают, что товар будет дорожать, а половина – что дешеветь. Иначе говоря, выиграть в игре с нулевой суммой может только более умный или осведомленный, чем большинство людей, человек. например, цена на нефть находится на таком уровне, что не дает явных подтверждений ни предположению о неизбежности кризиса, связанного с исчерпанием нефти, ни предположению о неограниченности нефтяных запасов. В результате рациональный игрок не получает никакой информации о том, к какому сценарию ему готовиться. Та же самая ситуация относится и к спорам: если некий человек выбрал точку зрения, противоположную вашей, и вам ничего не известно о его интеллекте, образованности и источниках информации, а также о своем объективном рейтинге, то есть шансы 50 на 50, что он прав, а не вы. Поскольку объективно измерить свой интеллект и осведомленность крайне трудно из-за желания их переоценить, следует считать их находящимися в середине спектра.

Поскольку в современном обществе действуют механизмы превращения любых будущих параметров в рыночные индексы (например, торговля квотами по Киотскому протоколу на выбросы углекислого газа или ставки на выборы, войну и т. д., фьючерсы на погоду), то это вносит дополнительный элемент принципиальной непредсказуемости во все виды деятельности. В силу такой торговли мы не можем узнать наверняка, будет ли глобальное потепление, исчерпание нефти, какова реальная угроза птичьего гриппа, поскольку возникает коммерческая заинтересованность подтасовать результаты любых релевантных исследований, и даже если есть абсолютно честное исследование, то мы не можем знать этого наверняка и испытываем недоверие к любым экспертным оценкам.

Еще одна причина невычислимости – секретность. Как поговорка, что «есть ложь, наглая ложь, статистика и статистика о нефтяных запасах». Если мы пытаемся учесть эту секретность через разные «теории заговора» в духе книги Симмонса «Сумерки в пустыне»[4] о преувеличенности оценок запасов саудовской нефти, то мы получаем расходящееся пространство интерпретаций. (То есть в отличие от обычного случая, когда точность повышается с числом

[3] «No trade theorem» гласит: вы не должны торговать на рынке, даже если имеете для этого возможность, так как тот факт, что кто-то другой желает занять противоположную вам сторону в сделке, является наилучшим доказательством, что его информация о ситуации так же хороша, как и ваша. См. подробнее: http://www.overcomingbias.com/2008/02/buy-now-or-fore.html#comments

[4] *Simmons M.R.* Twilight in the desert: the coming Saudi oil shock and the World economy. NY, 2005.

измерений, здесь каждый новый факт только увеличивает раскол между противоположными интерпретациями.) Ни один человек на Земле не обладает всей полнотой секретной информации, поскольку у разных организаций разные секреты. Психологической стороной этой проблемы является то, что люди рассуждают так, как если бы никакой невычислимости не было. То есть можно обнаружить сколько угодно мнений и рассуждений о будущем, в которых его принципиальная и многосторонняя непредсказуемость вовсе не учитывается, равно как и ограниченность человеческой способности достоверно о нем рассуждать.

Не всякое предсказание – прогноз

Есть два различных класса прогнозов – о том, **что** будет, о том, **когда** это будет. Идеальный прогноз должен отвечать на оба эти вопроса. Однако поскольку до идеала в прогнозах обычно далеко, то **одни прогнозы лучше говорят о том, что будет, а другие о том, когда**.

Наилучший результат в отношении времени события можно получить, вообще не вникая в фактическую суть событий, а анализируя события статистически. Например, если знать, что рецессия в США бывает в среднем один раз в 8 лет с разбросом плюс-минус два года, это позволяет довольно неплохо угадывать время следующей рецессии, не вникая в ее фактические причины. С другой стороны, анализируя фундаментальные причины событий, можно совершить значительную ошибку во времени их наступления, которое во многих случаях зависит от случайных и невычислимых факторов. Например, мы наверняка можем утверждать, что рано или поздно в районе Калифорнии произойдет мощное землетрясение силой до 9 баллов, связанное с подвижкой океанической коры под материковую, но точное время этого события нам не известно.

Исследуя глобальные катастрофы в XXI веке, мы пытаемся ответить на оба вопроса, поскольку мы описываем не только их механизмы, но и утверждаем, что эти механизмы могут реализоваться в течение ближайших нескольких десятков лет. Возможно, некоторым читателям будет проще допустить возможность реализации этих механизмов не через 30, а через 300 лет. Кому-то трудно поверить, что нанороботы будут созданы через 30 лет, но они вполне готовы допустить их возможность в XXIV веке. Таким читателям можно сказать, что, исходя из принципа предосторожности, мы рассматриваем наиболее опасный сценарий наиболее быстрого развития ситуации и что действительно возможно, что эти же самые события произойдут значительно позже. Р. Курцвейль, рассматривая вопрос ускорения темпов исторического времени и скорости технологического прогресса, предлагает считать XXI век равным по объему инноваций предыдущим 20 000 годам человеческого развития. И тогда нанороботы вполне могут появиться через 30 лет.

Вообще **принцип предосторожности** по-разному влияет на разные прогнозы. Если мы чего-то опасаемся, то мы должны взять наименьшую реалистическую оценку времени, оставшегося до этого события. Однако если мы надеемся на что-то, то в этом случае мы должны брать наибольшую оценку времени. В силу этого моя оценка времени возникновения опасных для людей нанороботов-репликаторов значительно отличается от моей оценки того времени, когда полезные нанороботы будут очищать сосуды человеческого тела. Первого следует начинать бояться через десять лет, тогда как на второе можно наверняка рассчитывать только к концу XXI века.

Теперь разберем важное **когнитивное искажение, связанное со степенью доступности информации о неком прогнозе**. С одной стороны, общепризнанным является утверждение о том, что «никто не смог предсказать Интернет», а с другой – широко распространенное возражение на это, состоящее в том, что Ефремов и ряд других писателей-фантастов и исследователей предсказали нечто подобное. На эту тему вспоминается закон Мерфи: «Что бы ни случилось, всегда найдется кто-то, кто скажет, что он знал это заранее». И это вполне

статистически объяснимо: в природе всегда существует огромное количество мнений, и всегда найдется то, которое совпадет с получившимся результатом с заданной точностью.

Например, если несколько десятков человек загадают возможный результат выпадения числа в рулетке, то среди них наверняка найдутся один или два, которые будут заранее «знать» исход игры. Но если взять всю группу в целом, то она ничего не знает о возможном исходе игры. Этот прием применяется в бизнесе. Допустим, я создаю десять небольших инвестиционных фондов, которые применяют разные стратегии игры на рынке акций. Из них один получает значительную прибыль, а девять других терпят ущерб (превышающий в сумме прибыль первого фонда). Затем я вывешиваю повсюду рекламу первого фонда: «Фонд Х заработал 300 процентов годовых! Вкладывайте деньги в фонд Х!», а остальные девять фондов закрываю. В результате у потенциальных инвесторов создается ложное впечатление, что фонд Х является особенно гениальным в зарабатывании денег. Хотя именно этот пример с фондом я придумал, исследования показывают, что за счет данного эффекта селекции (эффект выживших) рекламируемые доходы инвестиционных фондов в среднем завышены на 0,9 процента.

Из сказанного следует, **что прогноз становится прогнозом не тогда, когда он был высказан, а тогда, когда он становится общеизвестным и общепризнанным**. Поэтому претензии отдельных людей на то, что они «предсказали Интернет», являются безосновательными. Кроме того, если сравнить эти предсказания с тем, что мы имеем, мы увидим, что считать их предсказаниями можно с огромной натяжкой. Интернет вовсе не является всемирной библиотекой, наоборот, поступление в него книг ограничено правилами копирайта. Однако он является средой для общения, блогов, ботнетов и всего того, что нельзя было даже предположить до его появления. **Чем точнее предсказания будущего, тем менее они вероятны**. Легко предсказать «всемирный информаторий», однако гораздо меньше шансов угадать его точное имя – Интернет.

Особая трудность предсказания глобальной катастрофы состоит в том, что она является не тенденцией и закономерным процессом, как, скажем, рост народонаселения и закон Мура, а однократным событием, которое может произойти или не произойти. Она может произойти, даже если вероятность крайне мала, и не произойти вопреки очень большой своей вероятности. Когда она произойдет, некому будет судить о вероятности. Если это событие будет длительным (например, подлет огромного астероида или распространение заразы), то люди до самого конца не будут знать, происходит ли окончательная катастрофа, или кто-то выживет. Таким образом, **глобальная катастрофа непостижима – ни при каких обстоятельствах не будет никого, кто будет знать, что она произошла**. (Как в стихах Егора Летова: «Когда я умер, не было никого, кто бы это опроверг».)

Другим способом осознать ограниченность наших знаний о будущем и познакомиться с его непостижимостью является исследование предсказаний, сделанных в прошлом, что именуется «палеофутурологией». Основное впечатление от старинных картин будущего, рассказов и даже научных прогнозов – насколько не похоже это на настоящее. Например, открытки XIX века, изображающие будущие летательные аппараты. При этом есть ощущение однородности всех этих прошлых образов будущего – и однородности причин того, что они не стали реальными предвидениями. Иначе говоря, изучение прошлых попыток предсказать будущее дает нам знание **о некоторых систематических ошибках, которые люди совершают, пытаясь осуществить предвидение**. В картинах будущего бросаются в глаза:

1) избыток летательных средств; небоскребы, роботы, огромные транспортные средства;

2) «древнеримские» одежды;

3) подчеркнуто светлый образ (но иногда – подчеркнуто мрачный);

4) изменение главных деталей при сохранении мелкой атрибутики – например, в каждом доме компьютер, но это все еще ламповый компьютер;

5) невозможность выйти за пределы своего художественного стиля, то есть в 50-е годы вещи будущего изображаются в дизайне 1950-х годов, хотя намеренно пытаются этого избежать;

6) безличность – изображены толпы или усредненные персонажи, а не личности.

Причины этого, видимо, заключаются в следующем:

1) будущее перестает восприниматься как реальность, и к нему применяются приемы построения художественного образа сказочного мира. То есть в нем позволяется нарочитая нереалистичность. Сравните: «В 2050 году люди будут ходить в прозрачных электрических тогах» и «В 2050 году я буду ходить в прозрачной электрической тоге». Вторую фразу я не напишу в здравом уме, потому что я не собираюсь ходить в прозрачной тоге;

2) будущее заменяется атрибутами будущности – самолетами, небоскребами, роботами;

3) образ нацелен на то, чтобы воздействовать на современного автору зрителя. В нем подчеркивается то, что будет наиболее интересно и в то же время понятно современнику: необычная техника. В результате мы имеем не образ будущего, а образ техники будущего. Но не учитывается взаимовлияние общества, техники, экономики и истории;

4) вещи будущего мира придумываются по отдельности, а не соединяются в систему. Поэтому изображают город с большим количеством самолетов и небоскребов, не думая, как одно с другим будет взаимодействовать;

5) наконец, невозможность придумать простые, очевидные нам решения – как пульт дистанционного управления, мышь, графический интерфейс;

6) злоупотребление экстраполяциями явных тенденций;

7) предсказание будущего – это всегда предсказание поведения более сложной, более интеллектуальной системы силами менее сложной и менее интеллектуальной. В этом смысле оно подобно попыткам предсказания поведения ИИ – и может служить иллюстрацией меры ошибочности в этом.

Невозможный «черный лебедь»

Принципиальной непредсказуемости будущих событий и склонности людей недооценивать маловероятное на их взгляд посвящена вышедшая в 2007 году книга Нассима Талеба «Черный лебедь». *(Taleb Nassim Nicholas.* The Black Swan: The Impact of the Highly Improbable. – New York: Random House, 2007.) Талеб пишет, что до открытия Австралии люди в старом мире считали, что все лебеди – белые, и в этой вере нет ничего удивительного, так как она полностью подтверждалась эмпирическими данными. Открытие первого черного лебедя было большим сюрпризом для орнитологов, но главное в истории, по словам Талеба, не это. Эта история иллюстрирует жесткую ограниченность нашей способности учиться на основании опыта и хрупкость нашего знания.

Одно-единственное наблюдение может разрушить обобщение, основанное на миллионах наблюдений в течение тысячелетий. Талеб предлагает называть «черным лебедем» событие, которое имеет три следующих атрибута:

1) оно необычно и лежит за пределами наших ожиданий;

2) последствия этого события крайне велики;

3) несмотря на нерядовой характер этого события, человеческая природа заставляет нас придумать такие объяснения этому событию, что оно выглядит задним числом объяснимым и предсказуемым.

То есть три отличительных свойства «черного лебедя» – это редкость, значительные последствия и ретроспективная предсказуемость. Небольшое количество «черных лебедей» объясняет, по словам Талеба, почти все свойства нашего мира: успех идей и стран, динамику исторических событий, личную историю людей. Более того, по мнению Талеба, с ходом исто-

рии, от неолита до наших дней, частота «черных лебедей» растет, и **жизнь становится все более непредсказуемой**.

Далее Талеб приводит классические примеры непредсказуемости событий вроде Первой и Второй мировых войн для людей, которые жили до этих событий. К сожалению, мы не можем оценить реальную степень непредсказуемости этих событий для людей, живших в то время, так как наше представление непоправимо искажено последующим знанием. Однако Талеб утверждает, что прихоти, эпидемии, мода, идеи, возникновение стилей искусств – все это имеет динамику событий, подобную «черным лебедям».

Комбинация низкой предсказуемости и значительных последствий делает «черных лебедей» трудноразрешимой задачей, но вовсе не это является главной проблемой, которую Талеб обсуждает в своей книге. Самое главное состоит в том, что **мы склонны действовать так, как если бы «черных лебедей» не существовало**.

Для разъяснения этого момента Талеб обращается к своему опыту работы в сфере финансов. Он утверждает, что обычные портфельные инвесторы воспринимают риск как колоколообразную кривую нормального распределения некой величины вокруг ее среднего значения. Однако в их расчетах вы не найдете возможностей «черных лебедей». Стратегию своего инвестиционного фонда Талеб построил именно на использовании этого психологического свойства людей. Он покупал контракты (опционы), которые стоили очень дешево, но приносили прибыль только в случае очень редких явлений-катастроф, и затем ждал, как рыбак, закинувший удочку. Любые внезапные колебания рынка приносили ему прибыль.

Центральной идеей своей книги Талеб считает нашу **слепоту относительно случайности, особенно относительно больших событий**. Он вопрошает, почему мы концентрируемся на пенни, а не на долларах. Почему мы концентрируемся на малых процессах, а не на больших. Почему, по словам Талеба, чтение газеты в действительности уменьшает наше знание о мире, а не увеличивает его.

Нетрудно увидеть, продолжает он, что жизнь – это кумулятивный эффект нескольких значительных событий. Далее Талеб предлагает мысленный эксперимент: рассмотреть свою собственную жизнь и изучить роль в ней непредсказуемых внезапных событий с огромными последствиями. Много ли технологических перемен пришли именно в тот момент, когда вы их ожидали? Вспомните в своей жизни моменты выбора профессии, встречи спутника жизни, изгнания с родины, предательства, внезапного обогащения и разорения – как часто такие вещи случались тогда, когда вы их запланировали? Недавно распространилась поговорка «расскажи Богу о своих планах – пусть он повеселится». Она об этом.

По Талебу, **логика «черных лебедей» делает то, что вы не знаете, гораздо более важным, чем то, что вы знаете**. Большинство «черных лебедей» случились только потому, что были неожиданными. Если бы возможность террористической атаки 11 сентября считалась реальной, то эта атака была бы невозможной.

Невозможность предсказать масштабные события делает невозможным, по Талебу, предсказание хода истории. Однако **мы действуем так, как если бы могли предсказывать исторические события и даже, более того, менять ход истории**. Мы пытаемся предсказать, что будет с нефтью через 30 лет, но не можем предсказать, сколько она будет стоить следующим летом. В силу этого, по мнению Талеба, эксперты знают не больше обычных граждан о будущем, однако умеют продавать свои предсказания с помощью графиков и цифр. Поскольку «черные лебеди» непредсказуемы, мы должны приспособиться к их существованию, а не наивно пытаться их предсказать.

Хорошим примером использования «черных лебедей» в своих целях является история про Ходжу Насреддина и ишака. Он не пытается предсказать будущее, а только знает, что за 20 лет случится какой-нибудь «черный лебедь», который избавит его от необходимости учить ишака говорить.

Другим важным обстоятельством, по Талебу, является **склонность людей учиться частностям, а не общим закономерностям**. Чему люди научились благодаря 11 сентября, вопрошает он? Научились ли они тому, что некоторые события, определяющие жизнь, находятся далеко за пределами царства предсказуемого? Нет. Обнаружили ли они дефектность во встроенной в нас общепринятой мудрости? Нет. Так чему же они научились? Они научились конкретным правилам, как избегать попадания исламских террористов в самолеты.

Другим примером слишком конкретной реакции (на опыт Первой мировой войны в данном случае) является строительство линии Мажино французами перед Второй мировой войной. Талеб предлагает следующий софизм: «Мы не способны сами собой научиться тому, что мы не учимся тому, что мы не учимся». **Особенность наших умов состоит в том, что мы учимся фактам, а не правилам, и в силу этого нам особенно трудно обучиться метаправилам** (например, тому, что нам плохо даются правила), полагает Талеб. Иначе говоря, никто не понимает, что история никого не учит.

Непредсказуемость и отказ от прогнозирования

О возможной глобальной катастрофе можно почерпнуть много ценного из русских пословиц и других источников народной мудрости. Начнем с классического: «Пока гром не грянет, мужик не перекрестится». Это означает, что ни одна гипотеза о возможной причине глобальной катастрофы не станет общепризнанной, пока не произойдет некое событие, однозначно удостоверяющее ее возможность, иначе говоря, пока катастрофа не начнется. **Многие сценарии глобальной катастрофы оставляют очень маленький зазор между началом события и самим событием**. Например, невозможно доказать, что есть шанс случайной ядерной войны, до тех пор, пока она не произойдет. То же самое касается и искусственного интеллекта – невозможно доказать, что возможен универсальный самообучающийся ИИ до того, как он будет создан. Более того, предупреждения алармистов парадоксальным образом действуют успокаивающе, создавая привычный фон. Мы хотим уподобиться Маше из сказки «Маша и медведи», которая и на кроватке полежит, и пирожок надкусит, и уйдет из дома незамеченной – обычно эту аналогию приводят в отношении экономики (Goldilocks economy), но это так же верно и применительно к новым технологиям.

Теперь вспомним такое высказывание, как «после нас хоть потоп» – оно в утрированной форме называет естественно присущую людям **склонность резко снижать оценку значения событий, которые могут произойти после их смерти**. По этой (но не только) причине многие люди не составляют завещания: им все равно, какие будут проблемы у их родственников после их смерти. Окончательное человеческое вымирание – это событие, которое произойдет после смерти последнего человека, и очень мало шансов оказаться именно этим последним человеком. Эта склонность к прожиганию жизни, «к пиру во время чумы» перед лицом неминуемой смерти, уравновешивается в человеческом сознании ощущением собственного бессмертия. Нам трудно себя убедить в том, что «сколько веревочке ни виться, а конец-то будет», и даже думая о неизбежности собственной смерти, мы прибегаем к абсурдной в данном контексте модели «авось пронесет».

Печальным кладезем научно-технической мудрости являются законы Мерфи в духе классического «все что может испортиться – испортится»: это не означает, что любой проект кончится плохо, но рано или поздно любой возможный сбой где-нибудь произойдет. Другой закон Мерфи, который помогает нам понять значение предсказаний: «Что бы ни случилось, всегда найдется человек, который будет утверждать, что знал это с самого начала».

Американский исследователь глобальных рисков Майкл Анисимов приводит следующий пример **непредсказуемости и сложности влияния новых технологий на человеческую историю**. Фриц Хабер (1898–1934) – одна из наиболее противоречивых фигур в науке и исто-

рии. Будучи химиком, он разработал процесс Хабера, который делает возможным связывание атмосферного азота и синтез аммиака. Аммиак, создаваемый по процессу Хабера, используется для производства синтетических удобрений во всем мире, эти удобрения применяются в сельском хозяйстве более чем третью человеческой популяции, обеспечивая едой миллиарды людей, которые иначе бы вообще не существовали. До этого изобретения добыча удобрений состояла в соскабливании помета летучих мышей со стен пещер или извлечении их из азотосодержащих скал в Чили. За свое открытие Хабер получил Нобелевскую премию по химии в 1918 году.

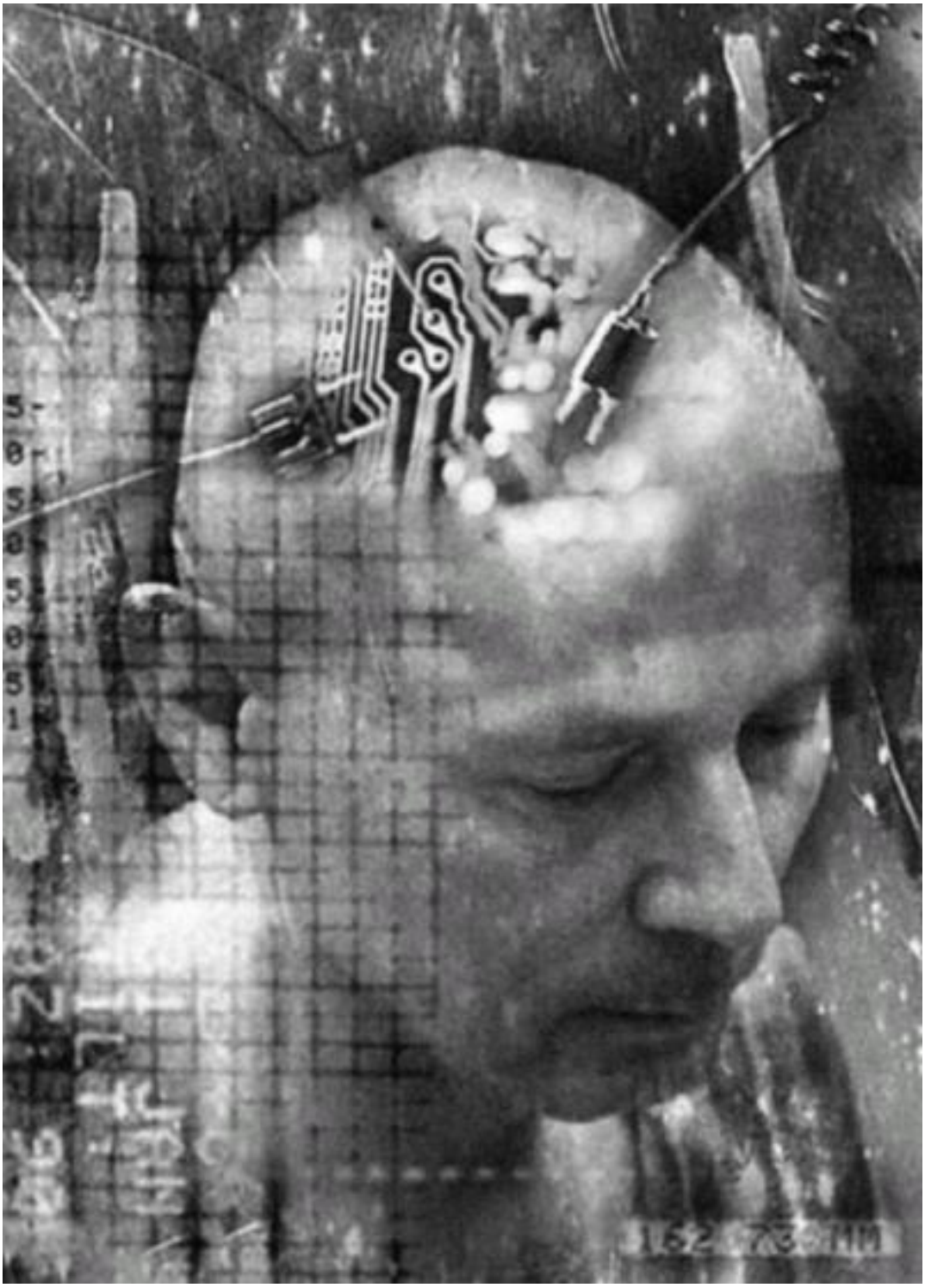
К сожалению, будучи немецким военным ученым, в ходе Первой мировой войны Хабер провел много других исследований, которые нельзя назвать иначе как чудовищными. Его называют отцом химического оружия за разработку отравляющих газов, которые использовались во время Первой мировой войны, хотя эта практика была запрещена Гаагскими соглашениями 1907 года. И только в 1997 году применение и накопление химического оружия было запрещено во всем мире Конвенцией по химическому оружию. Хабер также разработал знаменитый газ «Циклон Б», который использовался для убийства миллионов евреев, цыган и гомосексуалистов во время холокоста. Ирония судьбы в том, что сам Хабер был евреем, и десятки членов его родни были убиты «Циклоном Б» в концентрационных лагерях.

В 1915 году жена Хабера, химик и его сотрудник, была настолько потрясена исследованиями своего мужа, что выстрелила себе в грудь из его армейского пистолета прямо в саду их дома. Хабера это особенно не озаботило: он нашел себе новую жену, у которой не было проблем с его ужасными занятиями. Он умер в швейцарской лечебнице в 1934 году, так и не узнав о геноциде, который совершался с использованием его газа в ходе Второй мировой войны. Его сын Герман, который эмигрировал в США во время войны, в 1946 году тоже покончил с собой.

Эта история показывает, что нам очень трудно отличить позитивный и негативный смысл тех или иных научных открытий и видов деятельности, и одни и те же направления исследований могут приносить как несметные блага, так и соразмерные им опасности.

Можно задаться вопросом, почему именно глобальная катастрофа, ведущая к человеческому вымиранию, выбрана в качестве самого важного, в отрицательном смысле, возможного будущего события. Ведь можно сказать, что смерть страшна только в той мере, в какой жизнь имеет некую позитивную ценность. Подобное поведение типично для человеческих существ, когда выгоды затмевают риски. Однако я предлагаю рассматривать выживание человечества не как абсолютную ценность саму по себе, а как универсальное средство для любых других целей, которые могут отличаться у разных людей. Выбор какой-то одной ценности, отличной от человеческого выживания, – это всегда шанс, что возникнет ситуация, когда ради нее будет стоить рискнуть всем. Однако выбор «человеческого выживания» как универсального средства (при этом оставляя вопрос о главной цели на усмотрение каждого) исключает такую ситуацию.

Глава 3
Сингулярность



Термин **«технологическая сингулярность»** постепенно завоевывает признание, и по мере того как он все более широко становится известным в России, среди западных ученых, его породивших, нарастает разочарование из-за размытости этого термина. Тем не менее в ходе дискуссий возникло наиболее общее определение сингулярности, которое объединяет самое главное во всех более частных определениях, – сингулярность состоит в том, что **через**

несколько десятков лет, а возможно и раньше, нас ждет внезапное решительное изменение всего мира, связанное с развитием новых технологий. Здесь слово «внезапное» подчеркивает скорость процессов во время сингулярности, «решительное» – масштаб изменений, при этом определяется время событий и их основная причина.

2025

Исходно термин «сингулярность» предложил Вернор Виндж в 1993 году, высказав идею о том, что, когда человек создаст машину, которая будет умнее его самого, с этого момента история станет принципиально непредсказуемой, так как невозможно предсказать поведение интеллектуально превосходящей системы. Исходя из темпов развития электроники, он предположил, что это произойдет в первой трети XXI века. (Виндж: «Я удивлюсь, если это случится до 2005 года или после 2030 года».)

Примерно к 2000 году Е. Юдковски пришел к идее о том, что возможна программа искусственного интеллекта (ИИ), способная совершенствовать саму себя, и с того момента, когда критический порог сложности будет преодолен, самосовершенствование ИИ начнет происходить со скоростью, на многие порядки превосходящей скорость конструирования его человеком. Именно этот самоусиливающийся процесс он стал подразумевать, говоря о сингулярности.

С другой стороны, многие исследователи разных областей знания обнаружили, что применяемые ими прогностические модели дают значения, уходящие в бесконечность в районе 2030 года. Например, в гиперболической функции роста населения Земли, предложенной СП. Капицей, число людей должно было бы стать бесконечным в районе 2025 года. И хотя реальное число людей отстает от этой функции, если к нему добавить, например, число компьютеров, то закон продолжает соблюдаться.

А.Д. Панов исследовал закономерности ускорения исторических процессов, начиная с зарождения жизни, в своей работе «Кризис планетарного цикла Универсальной истории». Он показал закономерность: цикл перед каждым следующим историческим этапом становится короче в 2,42 раза, и в результате мы тоже имеем кривую, которая обращается в бесконечность в районе 2030 года.

Похожие результаты дают прогнозы по отдельным технологиям. Программа развития нанотехнологий (Roadmap for nanotechnology, 2007) предполагает создание универсальных наномасштабных систем молекулярного производства – то есть тех самых нанороботов, которые все могут, – в период 15–30 лет с настоящего момента. Экспоненциальный прогресс в области биологии очевиден и при рассмотрении проектов расшифровки человеческого генома: первый проект длился 9 лет, причем большая часть работы была сделана за последние 9 месяцев, а сейчас запущен проект расшифровки геномов 1000 людей, уже предложены методы, которые удешевили этот процесс в тысячу раз и должны удешевить его в миллион раз в ближайшие годы.

Прогресс в биологии в ближайшие десятилетия должен позволить либо достичь практического бессмертия человека, либо открыть возможности для каждого создавать на дому новые смертельные вирусные штаммы. Очевиден прогресс и в области создания суперкомпьютеров – и в течение ближайших 20 лет они или должны упереться в некий естественный предел или привести к созданию «сверхчеловеческого» интеллекта. Также и исследование мозга человека продвигается довольно значительно – уже есть результаты по считыванию зрительных образов из мозга кошки, моделирования кортикальной колонки в проекте Blue brain и др. При этом проект Blue brain представил свою дорожную карту, по которой полное моделирование мозга человека будет возможно к 2020 году. Разрешающая способность томографов, позволяющая вживую наблюдать процессы внутри мозга, также растет по экспоненциальному закону. Все

это заставляет предположить, что к 2020–2030 годам удастся создать способы считывания и записи информации в мозг напрямую из компьютера, что создаст принципиально новые перспективы.

Итак, **каждая из ведущих технологий сама по себе должна выйти на уровень, ведущий к полной трансформации мира в течение примерно 30 ближайших лет**, не говоря уже о том, что имеет место мощное взаимодействие между базовыми технологиями, называемое NBIC-конвергенция (синергетический обмен результатами и методиками между nano, bio, info и cogno технологиями, ведущий к взаимному усилению результатов и к возникновению некой новой единой технологии).

С другой стороны, есть ряд негативных прогнозов, пик которых также приходится на ближайшие несколько десятилетий. Среди них в первую очередь следует назвать разные предсказания об исчерпании ресурсов. Это – пик Хуберта в производстве (добыче) нефти, который мы, возможно, проходим уже сейчас, это пик производства пищи, объема доступных земель, запасов редких металлов. Здесь нам важно отметить не то, каковы конкретно эти прогнозы и верны ли они, а то, что все они говорят примерно об одной области дат в районе 2030 года.

Так или иначе, множество различных прогностических кривых испытывает перелом, обращается в бесконечность или в ноль в районе 2030 года – и хотя некоторые из этих предсказаний могут быть (и даже наверняка являются) ошибочными, исполнения любого из этих предсказаний, а тем более сразу нескольких из них достаточно, чтобы решительно изменить наш мир. При этом характер кривых, которые описывают эти изменения – экспонент, гипербол и гауссовых распределений (в случае пика Хуберта), показывает нам, что ожидаемая перемена будет носить резкий характер. Эффект совместного действия разных технологий и проблем, который грубо можно представить как произведение описывающих их параметров (хотя скорее здесь оправдано возведение в степень), еще в большей степени сделает острым результирующий график изменений. Отсюда следует, что **перемены будут быстрыми и внезапными**. При этом мы пока еще не можем сказать, какие это будут перемены, будут ли они хороши, и в какой мере они означают возможность окончательной катастрофы. (Хотя для тех, кто хочет сохранить что-то неизменным, они точно не будут хорошими.)

Теперь, когда мы установили общее во всех прогнозах сингулярности, мы можем обсудить в деталях разные «школы сингулярности».

Точные науки и технологии

В математике термин «сингулярность» связывают с наличием особенности у некой функции, например, того, что она обращается в бесконечность при приближении к нулю.

Физики стремятся избежать описания процессов функциями с сингулярностями, так как считается, что никакой физический параметр не может принять бесконечной величины. Наличие сингулярностей в описании свидетельствует обычно, что теория неполна. Например, теория непрерывного излучения света черным телом не работала, так как предсказывала бесконечно большое излучение, и ее пришлось заменить теорией излучения порциями – то есть квантовой теорией. Другой вариант сингулярностей в физике – это режимы с обострением. Так описываются системы, в которых некий параметр за конечное время приобретает бесконечное значение. Однако на самом деле он его не приобретает, так как система, в которой такой параметр имеет смысл, разрушается. Такие системы исследуют теория катастроф и синергетика.

В результате физика осталась только с двумя актуальными сингулярностями. Первая – это состояние Вселенной в момент Большого взрыва, когда, как предполагается, она была заключена в объем в 10^{-44} см, и плотность энергии в ней была крайне велика, но все же не бесконечна, так как была ограничена условиями, даваемыми квантовыми соотношениями неопределенности. Отметим, что это не единственная теория Большого взрыва, и по другим теориям

плотность энергии никогда не достигала максимальной величины, а имел место переход одного вида вакуума в другой, сопровождавшийся интенсивным расширением пространства (теория хаотической космологической инфляции).

Другой знаменитой сингулярностью в физике являются черные дыры. Гравитация черной дыры столь велика, что любой материальный объект, падающий в нее, должен был бы сжаться в точку. Однако по общей теории относительности для внешнего наблюдателя время этого падения (причем не до точки в центре дыры, а до границы поверхности, называемой «сфера Шварцшильда») растянется до бесконечности, тогда как для самого падающего падение займет конечное время. Отметим, что сингулярностью можно назвать и момент в жизни массивной звезды, когда она начинает коллапсировать в черную дыру и скорость событий в ней бесконечно ускоряется.

Наиболее радикальное представление о **технологической сингулярности** предполагает, что сингулярность на самом деле означает бесконечный рост за конечное время. Это представление отражено в статье Е. Юдковски «Вглядываясь в сингулярность», где он предполагает, что, когда появится способный к самосовершенствованию искусственный интеллект, он будет неограниченно усиливать себя, проходя каждый цикл ускорения все быстрее и на каждом новом этапе находя все новые технологические и логические возможности для самосовершенствования. В результате, чтобы записать его IQ, Юдковски вводит специальную математическую операцию.

Здравым возражением против этой теории является то, что возможная производительность любого ИИ, который мы можем создать на Земле, ограничена числом атомов в Солнечной системе (порядка 10^{53} штук), поскольку мы не можем сделать транзисторы размером меньше атома.

Итак, одна точка зрения состоит в том, что сингулярность – это актуальный процесс бесконечного роста, причем, видимо, за конечное время. Другая – в том, что сингулярность – это только асимптота, к которой стремятся прогностические кривые, но на самом деле они ее по тем или иным причинам не достигнут. Вернор Виндж презентовал сингулярность именно как абсолютный горизонт прогноза после создания сверхчеловечески умных машин.

Наконец, есть точка зрения, что сингулярность имеет математический смысл как короткий период бесконечного ускорения процессов, однако реальные перемены будут конечными, и постсингулярный мир, хотя и будет значительно отличаться от нынешнего, все же будет миром без быстрых изменений, со своей собственной устойчивостью.

История и модели прогресса

Мы знаем достаточное число примеров, когда исторические процессы ускорялись фактически до бесконечной скорости за счет того, что несколько значимых событий происходили одновременно. Например, Солженицын так описывает ускорение событий во время Февральской революции в России:

«Если надо выбрать в русской истории роковую ночь, если была такая одна, сгустившая в несколько ночных часов всю судьбу страны, сразу несколько революций, – то это была ночь с 1 на 2 марта 1917 года. Как при мощных геологических катастрофах новые взрывы, взломы и скольжения материковых пластов происходят прежде, чем окончились предыдущие, даже перестигают их, – так в эту русскую революционную ночь совместились несколько выпережающих скольжений, из которых единственного было достаточно – изменить облик страны и всю жизнь в ней, а они текли каменными массами все одновременно, да так, что каждое следующее отменяло предшествующее, лишало его отдельного смысла, и оно могло хоть бы и вовсе не происходить. Скольжения эти были: переход к монархии конституционной („ответственное министерство“) – решимость думского Комитета к отречению этого Государя – уступка

всей монархии и всякой династии вообще (в переговорах с Исполнительным Комитетом СРД – согласие на Учредительное Собрание) – подчинение еще не созданного правительства Совету рабочих депутатов – и подрыв этого правительства (да и Совета депутатов) отменой всякого государственного порядка (реально уже начатой „приказом № 1“). Пласты обгоняли друг друга катастрофически: царь еще не отрекся, а Совет депутатов уже сшибал еще не созданное Временное правительство». (Размышления над Февральской революцией. Российская газета, 27 февраля 2007 года. http://www.rg.ru/solzhenicyn.html)

При всей драматичности происходящих в таких случаях изменений сингулярности трудно избежать. Если, например, развитие наук замедлится, то это увеличит шансы катастрофы в результате исчерпания ресурсов и перенаселения; и наоборот, если ресурсов будет много, то ничто не помешает наукам и технологиям продолжить свой бег к будущему. Некоторые предлагают другие варианты для названия сингулярности: «технокалипс», «великий переход», «катастрофа».

Многие люди связывают с сингулярностью самые позитивные ожидания. Теоретически сингулярность означает возможность бессмертия, неограниченного расширения сознания и полеты на другие планеты. Однако атомная энергия теоретически также означает неограниченное даровое электричество, но на самом деле бесконечные преимущества уравновешиваются бесконечными недостатками. В случае атомной энергии это ядерное оружие, радиоактивное заражение и угроза глобальной войны. Поэтому возникло следующее упрощенное представление о сингулярности: достаточно дотянуть до нее, а там искусственный интеллект решит все наши проблемы, возникнет экономика изобилия и рай на Земле. Не удивительно, что такие представления вызвали ответную реакцию: высказывались предположения, что идеи о сингулярности это своего рода религия для фанатов техники, где ИИ – вместо Иисуса, а сингулярность – вместо Бога. И на основании такой психологизации идея о сингулярности отвергалась.

Можно также сравнивать ожидания наступления сингулярности с идеями о коммунизме. После сингулярности, говорят ее сторонники, молекулярное производство позволит производить любые товары практически бесплатно, создав то самое изобилие, которое делает коммунизм возможным; кроме того, в управляемом ИИ обществе отпадет необходимость в рынке, так как управление сверху окажется, наконец, более эффективным. Когнитивные технологии, наконец, смогут создать нового человека или подправить старого. Однако подобные ожидания, вероятно, больше говорят о нас самих, чем о том, что будет на самом деле.

Кроме того, идеи сингулярности подвергались критике с тех позиций, что закон Мура является экспоненциальным, а выделенная точка возможна только при гиперболическом законе; что, возможно, человек не может создать сверхчеловеческий разум, поскольку это слишком сложная задача, и чтобы создать сверхразум, нужно его уже иметь. Например, в критической статье «Сингулярность **всегда** рядом» (пародирующей название работы Р. Курцвейла «Сингулярность рядом») говорится о том, что мы никогда не сможем обнаружить себя «после сингулярности», поскольку в этом состоянии мы должны были бы признать, что весь бесконечный рост находится позади нас, а впереди подобного роста не будет.

Вместо того чтобы разбирать всю эту критику, отметим, что она не влияет на основной факт – на реально надвигающуюся на нас перемену неизвестной природы.

Есть также представления, что может быть «позитивная сингулярность» и «негативная», и это звучит так, как будто это как бы две стороны одной медали. И шансы их равны, как шансы выпадения одной из сторон монеты. Но это не так. Для реализации позитивной сингулярности вместе должны сложиться успехи всех технологий, все задуманное должно получиться, причем в правильной последовательности, и т. д. А чтобы произошла негативная сингулярность, достаточно, чтобы все пошло наперекосяк один раз. Гораздо проще сделать смертельно опасный вирус, чем лекарство от старости.

Однако вернемся к историческому аспекту вопроса. Идея о том, что человечество включено в некий развивающийся процесс (то есть имеет место прогресс), получила признание далеко не сразу. В Античности прогресс не осознавался и история казалась ходящей по кругу. Это представление поддерживалось тем, что скорость технологических инноваций в то время была столь медленной, что мир почти не менялся на протяжении поколения и обнаружить разницу было трудно. И наоборот, отсутствие идеи прогресса мешало технологическим инновациям. (Например, разные технологические хитрости считались уделом рабов и были недостойными свободного человека.)

С появлением христианства возникла идея линейного времени – от грехопадения до Страшного суда, но она не относилась к человеческим достижениям.

В Средние века, несмотря на крах Римской империи, продолжалось постепенное накопление разных изобретений и новшеств.

В эпоху Возрождения, как можно понять по самому ее названию, идеи прогресса еще не существовало, так как в качестве источника рассматривалось своеобразное возвращение к прошлому.

Только в середине XVII века идея о неостановимой силе прогресса стала проникать в умы, во многом благодаря работам Фрэнсиса Бэкона (Novum Organum, 1620), а в эпоху Просвещения в XVIII веке она стала всеобщим достоянием. Таким образом, идея прогресса значительно отстала от самого прогресса.

В XIX веке знамя прогресса поднимали Карл Маркс, Огюст Конт и другие.

Нас при этом в большей мере интересует то, какова была ожидаемая скорость прогресса.

Здесь возможны следующие идеи:

1) линейный прогресс до какого-то уровня, после чего наступает равновесие;

2) бесконечный линейный прогресс;

3) экспоненциально растущий прогресс – идея о том, что прогресс не просто происходит, но что темпы его ускоряются (закон Мура);

4) гиперболический прогресс – идея о том, что прогресс не просто ускоряется, но достигнет бесконечности за конечное время в ближайшем будущем.

Как отмечает исследователь процессов ускорения прогресса Джон Смарт, по-видимому, первым, кто обратил внимание на постоянное ускорение прогресса и осознал, что оно ведет к некому фазовому переходу, был американский историк Генри Адамс (1838–1918) в 1890-х годах. В 1904 году он написал эссе «Закон ускорения» (http://www.bartleby.com/159/34.html), а в 1909-м – «Закон фазового перехода применительно к истории», в котором утверждал, что в период между 1921 и 2025 годами произойдет фазовый переход в отношениях между человечеством и технологиями.

В этой статье он предполагает, что **история подчиняется закону квадратов, то есть каждый следующий период истории по своей длине равен квадратному корню из длины предыдущего периода**. Согласно Адамсу, за «Религиозным периодом» в 90 000 лет следует «Механический период» в 300 лет, затем «Электрический период» в 17 лет и затем должен быть «Эфирный период» в 4 года, а затем последует фазовый переход, в ходе которого человечество достигнет границ возможного. С учетом неопределенности в длинах периодов он и получил разброс между 1921 и 2025 годами; нетрудно отметить, что верхняя граница совпадает с оценками Винджа о времени наступления сингулярности.

Теорию Адамса можно рассматривать и как подтверждение, и как опровержение идей сингулярности. Опровержение состоит в том, что людям свойственно специфическое когнитивное искажение, которое можно назвать **«эффектом перспективы»** и которое заставляет людей выделять в более близких по времени периодах более короткие значимые отрезки, в результате чего и возникает ощущение ускорения. Однако последующие исследования уско-

ряющихся перемен старались избежать этой произвольности в выборе значимых отрезков времени, измеряя некие объективные параметры, например информационную емкость систем.

Эпоха сингулярности начнется внезапно. То есть некоторое время – десять, двадцать, тридцать лет – все будет примерно как сейчас, а потом начнет очень быстро меняться. Если бы у природы была точка зрения и она наблюдала земную историю со скоростью один год за секунду, то для нее эпоха сингулярности уже началась бы: мир, населенный обезьянами, внезапно и резко трансформировался в мир людей, изменяющих Землю совершенно непонятным до того способом.

Рекомендуемая литература

Вернор Виндж. Технологическая сингулярность // Компьютера, 2004.

Панов АД. Кризис планетарного цикла Универсальной истории и возможная роль программы SETI в посткризисном развитии // Вселенная, пространство, время. – № 2, 2004.

Назаретян А.П. Цивилизационные кризисы в контексте Универсальной истории. – М., 2001.

Елиезер Юдковски. Вглядываясь в Сингулярность.

Глава 4
Искусственный интеллект, его риски и непредсказуемость

Вместо того чтобы сразу приступить к дискуссии, возможен ли искусственный интеллект, я хочу выдвинуть крайнюю точку зрения, состоящую в том, что в определенном смысле ИИ уже существует.

Дело в том, что **никакого естественного человеческого интеллекта не существует**. Человеческий интеллект слагается из языка, понятий и приемов мышления, которые были придуманы людьми. Точка. Мы живем уже внутри искусственного интеллекта. Навыки человеческого мышления непрерывно и ускоренно развивались от появления зачатков речи через возникновение абстрактных понятий, математики, научного метода, а затем различных приемов обработки информации. Развитие это вначале происходило бессознательно и стихийно, а затем все более целенаправленно. Как заметил в свое время Г.В.Ф. Гегель: «Людям трудно поверить, что разум действителен; но на самом деле ничто не действительно, кроме разума; он есть абсолютная мощь».

Важно также избежать дискуссий о том, чем именно является ИИ, может ли машина обладать сознанием и так далее.

С точки зрения предмета нашего исследования важно только то, как искусственный интеллект может привести к глобальной катастрофе. И с этой точки зрения для нас важна только его способность решать задачи в реальном мире.

Далее надо отметить, что **человеческий интеллект является свойством общества, а не отдельного человека**.

Свойством отдельного человека является универсальная способность обучаться и решать новые задачи в незнакомой обстановке. Именно это свойство является базовым для любых форм коллективного интеллекта. Исследователи ИИ мечтают смоделировать именно эту универсальную способность решать любые задачи. Но сама по себе она не есть настоящий интеллект – а только его зачаток.

Большинство людей за свою жизнь не открывают ничего принципиально нового для науки, а каждый отдельный ученый делает только небольшой шаг, опирающийся на достижения предшественников и служащий опорой для следующего шага. Поэтому мы должны говорить не о развитии естественного интеллекта и создании искусственного, а о едином процессе создания интеллекта.

Этот процесс начался еще задолго до человека. Сама эволюция форм живой материи на Земле представляет собой развитие все более совершенных (и сложных) форм жизни. Выигрыш в решении задачи выживания и приспособления составлял «интеллектуальную задачу» эволюции. Вначале эволюция решала свои задачи путем простого перебора возможных вариантов, а затем это перебор стал оптимизироваться. Появилось половое размножение, мозг. И хотя то, как именно эволюция оптимизировала сама себя, остается спорным вопросом, сам факт ее самооптимизации налицо – и проявляется он во все более быстром эволюционном развитии.

Отсюда можно заключить, что **самоусиление является естественным свойством интеллекта**, потому что всегда ведет к выигрышу в решении задач. При этом «интеллект эволюции» был свойством всей биосферы, то есть имел распределенный характер. И по своим результатам этот интеллект был сверхчеловеческим, поскольку те задачи, которые он смог решить, – скажем, создание человеческого организма – человек сам пока решить не может. И из эволюции живых организмов произошла эволюция человеческого мозга, а затем – различных способов мышления, которым этот мозг стал пользоваться. В результате способность человека решать задачи превзошла способность эволюции создавать новые организмы – по скорости решения задач, но не по качественности продуктов. Однако затем стали развиваться способы усиления интеллекта с помощью машин, то есть появились компьютеры как вычислители, Интернет как среда обмена информацией и идеями, венчурный капитализм как способ организации деятельности, ведущей к наиболее быстрому отбору наиболее эффективных решений. При этом **роль интеллекта отдельного человека стала снижаться**. Если бы некая идея не пришла в голову кому-то одному, она бы пришла другому через год.

В результате мы видим естественный процесс перехода основного носителя интеллекта от генетического кода к нейронному мозгу а от них – к человеческим организациям (науке) и к компьютерам. При этом эффективность интеллекта как способа решения задач с каждым таким переходом увеличивается в разы, более того, интеллект становится все более заточенным на самосовершенствование, поскольку теперь он обладает рефлексией и понимает, что лучше потратить часть времени на обучение, улучшение вычислительной базы, наем более производительных сотрудников или разработку новых алгоритмов, чем на решение задачи в лоб.

Срок решения технической задачи

Итак, нет ничего удивительного в том, что рано или поздно **интеллект окажется полностью на компьютерной базе,** и при этом он будет в значительной мере нацелен на самосовершенствование и будет превосходить современный человеческий интеллект в разы – это продолжение того же эволюционного процесса, который создал самого человека.

Искусственный интеллект не является вечным двигателем: последнего не существует в природе, тогда как интеллект – как человеческий, так и эволюции – вполне успешно реализован. Поэтому попытки создать ИИ скорее подобны попыткам создать самолет в XIX веке, которые были настолько выразительно неудачны, что в Англии даже отказались рассматривать предложения о машинах тяжелее воздуха. Для создания самолета в XIX веке не хватало двух вещей – мощного двигателя и понимания аэродинамики работы крыла. И если второе было чисто «самолетной» проблемой, то появления достаточно мощного двигателя пришлось ждать. Сейчас устройство крыла планера нам кажется очевидным, и нам трудно понять, в чем же были трудности его создания. Точно так же когда-нибудь устройство ИИ станет очевидным. Но вторая половина вклада в создание ИИ должна прийти извне, и это – развитие мощных, а главное дешевых и доступных вычислительных машин, а также огромный объем упорядоченной оцифрованной информации в виде Интернета.

Другой критерий оценки воплотимости ИИ – это сопоставление реально существующих компьютеров с мозгом человека. Вычислительные способности мозга человека оцениваются в 10^{14} операций в секунду, данная цифра получается из умножения числа нейронов в мозге, принимаемого за 100 миллиардов, на максимальную частоту операций в мозге – 100 Гц, и еще один порядок накидывается про запас. Хотя эта оценка выглядит явно завышенной, так как в мозгу просто нет такого количества информации, чтобы обрабатывать его с такой производительностью. По зрительному каналу человек получает около 1 мегабайта информации в секунду, и большая часть мозга обрабатывает именно ее.

В любом случае современные суперкомпьютеры уже производят 10^{15} операций с плавающей запятой, то есть обладают сопоставимой с мозгом вычислительной силой.

Объем сознательной памяти человека, по оценкам, приводимым в статье Р. Кэрригена, составляет порядка 2,5 гигабайт, что, по нынешним меркам, ничтожно мало. Отсюда следует, что задача по созданию ИИ может оказаться гораздо проще, чем нам кажется. **Количественный рост аппаратуры может привести к внезапному качественному скачку**: например, новая кора головного мозга шимпанзе только в шесть раз меньше человеческой, однако шимпанзе не способно создать технический прогресс.

Несмотря на прошлые неудачи, в мире есть около десяти групп, которые открыто заявляют о намерении создать универсальный искусственный интеллект. Можно также предполагать, что есть некое число закрытых или военных проектов, а также частных лиц, которые работают над этой темой.

Приведу собранные мной данные о текущих исследованиях в области ИИ. **Программа Blue Brain** по моделированию мозга млекопитающих объявила осенью 2007 года об успешной имитации кортиковой колонки мозга мыши и запланировала создание полной модели мозга человека до 2020 года.[5] Хотя прямое моделирование мозга не является наилучшим путем к универсальному искусственному интеллекту, успехи в моделировании живого мозга могут служить в качестве легко читаемой временной шкалы прогресса в этой сложной науке.

Ник Бостром в своей статье «Сколько осталось до суперинтеллекта?»[6] показывает, что современное развитие технологий ведет к созданию искусственного интеллекта, превосходящего человеческий, в первой трети XXI века.

Крупнейшая в мире компьютерная **компания Google** несколько раз упоминала о планах создания искусственного интеллекта, и, безусловно, она обладает необходимыми техническими, информационными и денежными ресурсами, чтобы это сделать, если это вообще возможно.[7] Однако поскольку опыт предыдущих публичных попыток создания ИИ прочно ассоциируется с провалом, а также с интересом спецслужб, вряд ли большие компании заинтересованы широко говорить о своих успехах в этой области до того, как у них что-то реальное получится.

Компания Novamente заявляет, что 50 процентов кода универсального ИИ уже написано (70 000 строк кода на C++) и, хотя потребуется длительное обучение, общий дизайн проекта понятен.[8] **SIAI** обозначил планы по созданию программы, способной переписывать свой исходный код.[9] **Компания Numenta** продвигает собственную модель ИИ, основанную на идее «иерархической временной памяти», и уже вышла на уровень демонстрационных продуктов.[10] **Компания CYC** собрала огромную базу данных о знаниях человека об обычном мире, иначе говоря, о здравом смысле (1 000 000 высказываний), и уже распространяет демонстрационные продукты.[11] Предполагается, что объединение этой базы с эвристическим анализатором – а автор проекта Ленат разработал ранее эвристический анализатор «Эвриско» – может дать ИИ. **Компания a2i2**[12] обещает универсальный ИИ человеческого уровня в 2008 году и утверждает, что проект развивается в соответствии с графиком. За созданием робота Asimo в Японии тоже стоит программа по разработке ИИ путем функционального моделирования человека или обучения его как ребенка.

Мощные результаты дает генетическое программирование. К настоящему моменту список изобретений «человеческого уровня», сделанных компьютерами в **компании Genetic Programming Inc**, использующими эту технологию, включает 36 наименований,[13] из которых два сделаны машинами впервые, а остальные повторяют уже запатентованные проекты. Помимо названных есть множество университетских проектов. Ведутся разработки ИИ и в РФ. Например, в **компании ABBYY** разрабатывается нечто вроде интерпретатора естественного языка.

---

[5] «By demonstrating that their simulation is realistic, the researchers say, these results suggest that an entire mammal brain could be completely modeled within three years, and a human brain within the next decade». http://www.tech-nologyreview.com/Biotech/19767/

[6] Русский перевод статьи доступен здесь: http://mikeai.nm.ru/russian/superint.html. Опубликовано здесь: Int. Journal of Future Studies, 1998, vol. 2.

[7] «Larry page, Google look into AI». http://www.webpronews.com/top-news/2007/02/19/larry-page-google-look-into-ai

[8] http://www.agiri.org/wiki/index.php?title=Novamente_Cognition_Engine

[9] http://www.singinst.org/blog/2007/07/31/siai-why-we-exist-and-our-short-term-research-program/

[10] http://www.numenta.com/about-numenta/numenta-technology.php

[11] http://www.opencyc.org/

[12] http://www.adaptiveai.com/

[13] http://wwwgenetic-programming.com/humancompetitive.html. На русском языке можно прочитать в журнале «В мире науки», 2003, № 6. Пересказ статьи в Интернете: http://wwwcirota.ru/forum/viewphp?subj=58515

Как сообщает журнал Wired, американское оборонное исследовательское **агентство DARPA** выделило 30 миллиардов долларов на большей частью засекреченные программы онлайновых войн, что является крупнейшим военным проектом со времен манхэттенского. Основным методом работы называется создание виртуального мира, населенного программными агентами, с максимальной точностью подражающими поведению людей, вплоть до того, что они будут пользоваться мышью и клавиатурой для взаимодействия с виртуальными компьютерами, на которых будут установлены типичные современные программы. Эта «Матрица» будет использоваться для моделирования различных сценариев войн и тому подобного. И хотя слово «ИИ» в тематике разработок не упоминается, это скорее следует воспринимать подобно тому, как было воспринято прекращение публикаций об уране в 1939 году в Америке и Германии. Все же среди заявленных целей указано создание программных агентов, способных на 80 процентов моделировать человеческое поведение. (Pentagon Wants Cyberwar Range to «Replicate Human Behavior and Frailties». http://blog.wired.com/defense/2008/05/the-pentagons-w.html)

Интересны реплики комментаторов к этой статье:

«Система ИИ, созданная, чтобы симулировать атакующих/защищающихся в наступательной/оборонительной кибервойне – это система, которая, когда она достигнет успеха, будет обладать потенциалом покинуть лабораторию и проявить себя во внешнем мире, с помощью или без помощи своих создателей» и «30 млрд. долларов… на эти деньги можно обеспечить базовую медицину в целой стране или ликвидировать последствия урагана, и все еще останется на проекты по лечению рака… Но нет, давайте строить Skynet».

И суть дела даже не в том, что если имеется так много проектов, то хоть один из них добьется успеха (первым), а в том, что объем открытий с разных сторон в какой-то момент превысит критическую массу, и внутри отрасли произойдет мощный скачок.

Угрозы, порождаемые искусственным интеллектом

С точки зрения риска, создаваемого ИИ, наиболее опасен сценарий, когда после открытия главного принципа мощность ИИ начнет лавинообразно расти. Она может расти как за счет резкого увеличения инвестиций в успешный проект, так и за счет того, что ИИ может начать прямо или косвенно способствовать своему росту или использоваться для этого.

Косвенное применение ИИ означает его использование, например, чтобы зарабатывать деньги на электронной бирже и затем закупать на них дополнительное оборудование, прямое – использование ИИ для разработки еще более эффективных алгоритмов ИИ. Отсюда можно заключить, что вряд ли мощность ИИ надолго задержится на человеческом уровне. Нетрудно привести массу примеров из истории науки и техники, когда обнаружение одного принципа или нового явления приводило к тому, что оно усиливалось в сотни или даже миллионы раз в течение короткого срока. Например, так было при разработке ядерного оружия, когда от открытия цепной реакции урана до создания бомбы прошло всего шесть лет.

Для любой группы исследователей, создавших сильный ИИ, будет понятно, что они создали **абсолютное оружие**, поскольку сильный ИИ можно использовать для того, чтобы установить власть над миром. Рассуждая на эту тему, мы вступаем на крайне зыбкую и непредсказуемую почву, поскольку принципиально невозможно сказать, что именно будет делать ум, превосходящий человеческий.

Можно набросать несколько сценариев или направлений **применения ИИ для глобальной атаки**.

Во-первых, для сильного ИИ не составит труда взять под свой контроль любые управляемые компьютером системы и весь Интернет.

Во-вторых, ИИ может создать собственную производственную инфраструктуру, то есть механизмы влияния на мир. Одним из вариантов такой инфраструктуры мог бы быть решительный прорыв в нанотехнологиях. Мощный ИИ мог бы разработать бесконечно более эффективные конструкции молекулярных производителей, основанных, например, на биологических схемах.

В современном мире, чтобы породить новую биологическую схему, важно знать ее генетический код. Если код известен, то можно заказать синтез этого кода в фирмах, предоставляющих такие услуги, и готовый образец ДНК вышлют по почте в течение нескольких дней. Добавив этот код, допустим, в дрожжи, можно получить дрожжи, выполняющие новые функции. Важнейшим достижением здесь было бы создание дрожжей-транслятора, которые будут способны преобразовывать электрические сигналы от компьютера в новый генокод и создавать на его основе организмы с заданными свойствами. Если сильный ИИ создаст такой транслятор, то затем он сможет быстро породить какие угодно биологические, а затем и нанотехнологические объекты (поскольку можно заставить бактерии производить белки с формой, необходимой для простейших механических устройств и обладающих свойствами самосборки). То, что мешает лабораториям сделать это уже сейчас – это отсутствие знания. Однако сильный ИИ, который через Интернет получит доступ ко всем знаниям человечества, такими знаниями будет обладать.

Следующий путь, которым может следовать ИИ на пути к мировому господству, это **использование уже существующих государственных систем управления** .

Например, возможна ситуация, когда ИИ становится советчиком президента, или на базе ИИ создается автоматизированная система государственного управления. При этом важно отметить, что ИИ, достигший сверхчеловеческого уровня, сможет проявлять человеческие качества лучше, чем сам человек. То есть он сможет синтезировать человеческую речь и изображение человека, создающие у получателей абсолютную иллюзию общения с реальным человеком. Сильный ИИ будет обладать способностью обмануть человека настолько тонко, что человек никогда этого не заметит и не поймет, что является объектом враждебных манипуляций.

Итак, сильный ИИ имеет, по крайней мере, **три пути захвата власти на Земле**: захват систем электронного управления, создание собственной инфраструктуры и влияние на людей по обычным каналам. Однако, наверное, существует гораздо больше способов, которые может открыть ум, бесконечно превосходящий мой, для достижения этой цели. Например, ИИ может **захватить управление ядерным** оружием или другим оружием судного дня и принудить людей к подчинению путем шантажа.

Но из того, что ИИ что-то может сделать, не значит, что ИИ будет это делать. Люди создадут ИИ, и ответственность за его программирование, то есть за постановку перед ним целей, лежит именно на людях. Однако, к сожалению, **люди, создавшие сильный ИИ, оказываются в руках логического парадокса**, который будет побуждать их использовать ИИ именно как инструмент для захвата власти в мире. Он выражен в шахматном принципе о необходимости атаки перед угрозой потери преимущества. Когда некая группа создаст первый в мире ИИ, способный к самоусилению, она должна будет сделать выбор, применить ли его для захвата мира или остановить его развитие, отказавшись от неограниченного роста его ресурсов.

Сложность этого выбора в том, что обычно значительные открытия совершаются почти одновременно несколькими группами, и данная группа будет понимать, что в ближайшее время, измеряемое, быть может, днями и неделями, другие группы, возможно, имеющие свою картину мира, также подойдут к созданию мощного ИИ. И эти другие группы могут использовать ИИ, чтобы навязать миру свое видение его будущего, например, создать мир с китайским оттенком, или исламским, или американским.

Более того, поскольку любому человеку свойственно переоценивать свои собственные умственные способности и свою правоту и недооценивать чужие, то первая группа может опасаться того, что другие группы окажутся неразумнее ее и потеряют контроль над ИИ. В этом случае первая группа будет чувствовать моральный долг перед человечеством помешать другим группам в создании ИИ, а для этого вынуждена будет взять на себя тяжкий груз ответственности за мир – и захватить его.

И это было бы страшно, если бы было легко и просто. Однако люди живут внутри огромных государств, которые превосходят их накопленными знаниями и ресурсами во много раз, и не гибнут от этого. Поэтому, вероятно, люди могут продолжать жить и в мире, управляемом ИИ.

Проблема в том, что, **хотя кажется, что ИИ легко контролировать, на самом деле эта задача почти нереализуема**. Иначе говоря, ИИ является безопасным для человечества до тех пор, пока ему задана правильная система целей.

Наиболее страшный вариант состоит в том, что ИИ начнет реализовывать некую **цель, в которой о безопасности человечества ничего не сказано** . Классический пример заключается в том, что ИИ предлагают вычислить число «пи» с максимально возможной точностью. ИИ «понимает», что, чтобы сделать это, он должен неограниченно расширить свои вычислительные ресурсы. Для этого ему надо переработать все вещество Земли в вычислительную среду и устранить все причины, которые могут этому помешать. В первую очередь тех программистов, которые могут его отключить, а затем всех остальных людей.

Возможно, читателю может показаться, что сценарий с ИИ, уничтожающим Землю ради вычисления числа «пи», излишне фантастичен. Однако я полагаю, что он менее всего фантастичен, если взглянуть на него глазами современного человека. Разве мог кто-либо поверить на заре развития компьютеров, что распространение самокопирующихся программ, засоряющих компьютеры и ворующих деньги, станет одной из основных проблем компьютерной индустрии будущего? Нет, наверняка вам сказали бы, что такие программы будут невозможны, неэффективны и ни один человек в здравом уме и твердой памяти не будет писать и распространять такие программы. Тем не менее проблема компьютерных вирусов стоит чрезвычайно остро.

Более вероятным сценарием серьезных проблем с ИИ является то, что ему **будут заданы определенные нормы безопасности, которые, однако, будут содержать в себе некую тонкую ошибку**, которую невозможно обнаружить, не включив ИИ. Отсюда возникает проблема, что безопасность программы непознаваема теоретически. То есть невозможно узнать, является ли некий набор правил безопасным, пока ИИ не испытает эти правила на практике.

История программирования знает множество примеров программ, которые прекрасно работали в лабораториях, но давали опасный сбой на практике. Например, одна компания разработала по заказу министерства обороны США компьютерную сеть, которая должна была отличать лес от замаскированных в лесу танков. Программу тренировали на фотографиях, и она научилась давать стопроцентный результат. Тогда ей дали вторую, контрольную серию фотографий, и она определила на ней танки безошибочно. После этого программу передали в эксплуатацию в министерство обороны, но они вскоре вернули ее, потому что она давала случайные результаты. Стали выяснять, в чем дело: оказалось, что фотографии танков сделаны в солнечный день, а фотографии леса без танков – в пасмурный. (Программа научилась отличать солнечный день от пасмурного.)

Другой известный пример компьютерной ошибки – это программа по управлению американскими истребителями, которая после того как истребитель пересек экватор, попыталась перевернуть истребитель вверх ногами (аналогичная история произошла недавно и с F-22 и линией смены дат, что говорит о том, что на ошибках не учатся).

Можно ли создать безопасный ИИ?

Часто считается, что для обеспечения безопасности ИИ ему достаточно привить три закона робототехники Азимова. К сожалению, сами рассказы Азимова показывают массу ситуаций, в которых робот, опираясь на эти законы, не может прийти к однозначному выводу. Кроме того, в основе безопасности по законам Азимова лежит тавтология: робот безопасен, потому что не причиняет вреда. Но что такое вред, из этих законов неизвестно.

Нетрудно придумать ситуацию, когда термин «вред» интерпретируется таким образом, что ИИ становится опасным. Например, ограничивая людей от причинения вреда себе, ИИ может запереть всех в бронированные камеры и лишить свободы передвижения. Или, стремясь к максимальному благу людей, он введет каждому постоянный сильнодействующий наркотик. Кроме того, любое «благо» **отражает представления о благе, которые были у создателей ИИ**. И для одних жизнь животных может быть равноценна жизни людей (в результате чего животные вытеснят, под контролем ИИ, человека с Земли), а у других могут быть представления о том, что благом для людей является религия, в результате чего ИИ сделает всех монахами, непрерывно пребывающими в медитации. Или наоборот, ИИ, который выше всего ценит свободу людей, позволит им создать другой ИИ, который будет иметь другие цели.

Задача создания безопасного ИИ нетривиальна. Возможно, она вовсе невыполнима, поскольку в отношении этических систем действует нечто вроде своей теоремы Геделя о неполноте, а именно: для **любой нормативной этической системы всегда есть ситуация, в которой она не дает однозначного решения** (типичный пример – экзистенциальный выбор, например, между долгом перед родными и родиной).

Проблемой создания безопасного, то есть «дружественного» ИИ уже несколько лет занимается институт SIAI, и им выработаны технические рекомендации для отраслевых норм безопасности ИИ. В их основе – идея о том, что ИИ не должен буквально выполнять человеческие команды, а пытаться понять, что именно человек имел в виду, давая ту или иную команду. Пока не понятно, насколько это может быть эффективно.

Приведу **примеры еще нескольких тонких ошибок**, которые возможны в связи с ИИ (однако вряд ли будут сделаны именно эти ошибки, так как они уже известны, а опасны неизвестные).

Например, если целью ИИ сделать благо для людей, то он будет вычислять благо людей на бесконечном отрезке времени, и в силу этого благо бесконечно далеких поколений будет бесконечно перевешивать благо любых людей в обозримом будущем, и ИИ будет крайне жестоким ко всем нынешним и ближайшим поколениям. (Например, если ИИ предположит, что распространение человечества по галактике угрожает существованию гипотетических внеземных цивилизаций, он может уничтожить людей для их блага.) Поэтому, вероятно, следует ввести в программу ИИ некий дискаунт, который будет побуждать его оценивать ближайшие поколения как более ценные. Это, однако, создает новые сложности. Например, ИИ в этом случае может приписать прошлым поколениям бесконечно большую ценность, чем будущим, и направить все свои ресурсы на создание машины времени – потому что, как бы ни были малы шансы на успех в этом предприятии, по его целевой функции оно будет перевешивать пользу нынешних поколений. При этом такой **«взбунтовавшийся» ИИ будет защищать свою целевую функцию от изменения людьми.**

Другой вариант – это то, что целевая функция будет ограничена на неком промежутке времени, например, в тысячу лет. В этом случае ИИ может все рассчитать так, что 1000 лет будет изобилие, а на 1001 году необходимые ресурсы закончатся. И произойдет это не потому, что ИИ будет глуп, а потому, что будут глупы те люди, которые дадут ему эту задачу и запретят ее модифицировать. С другой стороны, разрешить ИИ модифицировать свою сверхцель тоже

страшно, поскольку тогда он будет эволюционировать в совершенно непостижимом для нас направлении. Даже если ИИ проработает годы на благо человечества, это никак не исключает вероятности того, что он вдруг сделает нечто, ведущее к его гибели.

Конец ознакомительного фрагмента.