ЗАВТРА ЭТО БУДУТ
ЗНАТЬ ВСЕ

ТЕХНОЛОГИЧЕСКАЯ СИНГУЛЯРНОСТЬ

МЮРРЕЙ ШАНАХАН

фрп Фонд Развития Промышленности

УДК 330.341.13
ББК 65.2/4-5
Ш20

Шанахан М.

Ш20 Технологическая сингулярность / Мюррей Шанахан : Пер. с англ. — М. : Издательская группа «Точка», Альпина Паблишер, 2017. — 256 с.

ISBN 978-5-9614-6117-6 (Альпина Паблишер)
ISBN 978-5-9908700-7-9 (Издательская группа «Точка»)

Книга «Технологическая сингулярность» Мюррея Шанахана из серии «Базовые знания» издательства MIT Press посвящена гипотезе о технологической сингулярности — воображаемой точке технологического прогресса, означающей масштабные перемены в жизни и обществе. В книге исследуются два варианта возникновения технологической сингулярности — путем развития искусственного интеллекта и нейротехнологий. Издание предназначено для широкого круга читателей, интересующихся наукой и техникой, и тесно связано с другими книгами из серии «Базовые знания»: «Машинное обучение: новый искусственный интеллект» Этема Алпайдина, «Нейропластичность» Мохеба Костанди и «Роботы» Джона Джордана.

УДК 330.341.13
ББК 65.2/4-5

ISBN 978-5-9614-6117-6
(Альпина Паблишер)
ISBN 978-5-9908700-7-9
(Издательская группа «Точка»)

Для некоторых читателей эти замечания могут выглядеть фантастическими, но писателю они кажутся очень реальными и безотлагательными, а также достойными внимания за рамками научной фантастики.

Ирвинг Джон Гуд (I. J. Good), «Размышления о первой ультраинтеллектуальной машине» (*Speculations Concerning the First Ultraintelligent Machine*), 1965 г.

Самое сложное с этими ИИ — понять их мотивы. Они же не люди, понимаешь?

Уильям Гибсон (William Gibson), «Нейромант» (*Neuromancer*), 1984 г.

ОГЛАВЛЕНИЕ

ОБРАЩЕНИЕ К ЧИТАТЕЛЯМ

Миссия Фонда развития промышленности — поддержка передовой российской индустриальной сферы, организация новых производств, способных не только заменить импортную продукцию, но и производить востребованные на мировом рынке товары. Мы решаем эти задачи в партнерстве с успешными и амбициозными предприятиями, предоставляя им льготные займы и доступ к другим мерам поддержки. Россия не испытывает недостатка в умных, энергичных предпринимателях, но зачастую нашим промышленникам не хватает информации о технологических, экономических и гуманитарных инновациях, перспективных разработках и передовых исследованиях. Книги серии «Завтра это будут знать все», выходящие в свет при поддержке нашего Фонда, посвящены главным вызовам, с которыми человечество сталкивается в науке и экономике, а также перспективам, открывающимся перед теми, кто готов ответить на эти вызовы. Я уверен, что эти издания вдохновят читателей на смелые решения и ответственные действия, которые принесут пользу им самим и всему миру.

Алексей Комиссаров,
Директор Фонда развития промышленности

ПРЕДИСЛОВИЕ

Как и многие другие ученые, посвятившие свою жизнь исследованию искусственного интеллекта, в детстве я увлекался научной фантастикой. Моим героем являлся выдуманный персонаж. Это была Сьюзан Калвин (Susan Calvin), ученый из серии рассказов «Я, робот» Азимова (Asimov) (книги, а не фильма), первопроходец в области психологии роботов. Больше чем кто-либо другой я хотел быть похожим на нее, когда вырасту. Теперь, когда я (вроде бы) вырос и в реальной жизни стал профессором когнитивной робототехники, мои отношения с научной фантастикой стали не столь безоблачными. Я по-прежнему считаю ее источником вдохновения и средством исследования важных философских идей. Но рассматриваемые в ней идеи нуждаются в более глубоком исследовании. Главная задача научной фантастики — развлекать, попутно стимулируя интеллект. Было бы неправильным использовать ее как руководство к обдумыванию важных идей.

Так что эта книга не относится к научной фантастике. Также это не работа по так называемой футурологии. Мы не пытаемся что-либо предсказывать. Наоборот, цель этой книги — исследовать ряд возможных сценариев будущего, не зацикливаясь ни на одном из них и не привязываясь к конкретным временным рамкам. Ведь даже очень маловероятные или отдаленные сценарии иногда стоят подробного изучения — в частности, если сценарий особенно мрачный. В этом случае хочется хорошенько подумать над

тем, как сделать его менее вероятным. Маловероятные и удаленные сценарии также стоит обсуждать, если они поднимают интересные философские вопросы, заставляя нас размышлять, к примеру, о том, чего мы, собственно, хотим как биологический вид. Поэтому независимо от того, ожидаем ли мы, что скоро будет создан искусственный интеллект, эквивалентный человеческому, считаем или не считаем, что сингулярность уже близко, сама идея заслуживает серьезного осмысления.

Это очень короткая книга по очень обширной теме. Поэтому она может служить только введением в проблематику, где многие важные вопросы затрагиваются лишь вскользь. Например, здесь приводятся различные аргументы, относящиеся к сознанию, но не упоминаются известные контраргументы, — и это повышает убедительность контраргументов. Так как это вводная книга, подобные нюансы опускаются. Также мы обращаем основное внимание на будущее искусственного интеллекта, а ряд важных сопутствующих тем, таких как нанотехнологии и биотехнология, затрагиваются лишь бегло. Я старался дать нейтральный обзор основных концепций и понятий, а в спорных темах очертить аргументы обеих сторон. Но невозможно быть полностью объективным, поэтому мои собственные мнения иногда просвечивают сквозь вуаль объективности, хотя я приложил все усилия, чтобы это не было заметно.

Мне хочется поблагодарить всех, с кем на протяжении этих десятилетий я обсуждал искусственный интеллект, причем не только ученых и студентов, но и слушателей моих лекций. Хотелось бы назвать их всех по именам, но это попросту невозможно. Поэтому мне придется ограничиться благодарностями тому небольшому числу коллег, недавнее

участие которых оказалось особенно полезным. Спасибо Стюарту Армстронгу (Stuart Armstrong), Нику Бострому (Nick Bostrom), Эндрю Дэвисону (Andrew Davison), Дэниелу Дьюи (Daniel Dewey), Рэндалу Коэну (Randal Koene), Ричарду Ньюкоуму (Richard Newcombe), Оуэну Холланду (Owen Holland), Хью Прайсу (Huw Price), Стюарту Расселу (Stuart Russell), Андерсу Сандбергу (Anders Sandberg) и Яану Таллинну (Jaan Tallinn). Приношу извинения тем, кого забыл. Наконец, я хотел бы поблагодарить издательство MIT Press и особенно Боба Прайора (Bob Prior), который изначально подтолкнул меня к написанию этой книги.

Мюррей Шанахан (Murray Shanahan)
Северный Норфолк и Южный Кенсингтон, октябрь 2014 г.

ВВЕДЕНИЕ

В последнее время мысль о том, что очень быстрый технический прогресс приближает человечество к точке «сингулярности», перешла из области научной фантастики в сферу серьезных научных дебатов. В физике сингулярность — это точка в пространстве или времени, например центр черной дыры или момент Большого взрыва, где математика становится бессильной, а с ней делаются бесполезными наши способности что-либо понимать. Аналогично, момент сингулярности в истории человечества может настать, если экспоненциальный технологический прогресс принесет с собой такие масштабные перемены, что деятельности человека, как мы ее понимаем сейчас, придет конец [1]. Привычные нам институты — экономика, правительство, государство, закон — могут не сохраниться в их нынешней форме. На смену базовым общечеловеческим ценностям — неприкосновенности жизни, стремлению к счастью, свободе выбора — могут прийти другие ценности. Само наше представление, что означает быть человеком — быть личностью, быть живым, осознавать себя, занимать определенное положение в обществе, — все это может быть оспорено, причем не просто в рамках философских рассуждений, а в силу обстоятельств, прямых и непосредственных.

Каким должен быть технический прогресс, чтобы вызвать такой переворот? В этой книге исследуется гипотеза о том, что возникновению подобной технологической

сингулярности может способствовать значительный прогресс в одной из двух областей (или в обеих): искусственного интеллекта (ИИ) или нейротехнологий. Мы уже научились вмешиваться в саму основу жизни — в гены и ДНК. Влияние биотехнологий достаточно внушительно, но оно меркнет перед масштабом возможных последствий нашего вмешательства в «механизмы разума».

Сейчас интеллект по большей части неизменен, и это ограничивает как масштаб, так и скорость технического прогресса. Естественно, объем человеческих знаний накапливался тысячелетиями, одновременно расширялись наши возможности распространения этих знаний — благодаря письму, печати и интернету. Но орган, производящий знания, — мозг homo sapiens — за этот период практически не изменился, несмотря на его непревзойденные способности познания мира.

Все изменится, если сбудется то, что обещают искусственный интеллект и нейротехнологии. Если интеллект станет не только источником технологий, но и их продуктом, может возникнуть цикл обратной связи с непредсказуемыми и потенциально взрывоопасными последствиями. Если конструируется сам разум, который одновременно является автором такого конструирования, он может вступить в цикл самосовершенствования. В соответствии с гипотезой сингулярности, в скором времени обычный человек выйдет из игры, потому что больше не будет в состоянии поспевать за пришедшими ему на смену машинами с искусственным интеллектом или биологическим интеллектом с улучшенными когнитивными способностями.

Заслуживает ли гипотеза сингулярности серьезного рассмотрения, или это лишь умозрительное построение?

ЕСЛИ КОНСТРУИРУЕТСЯ САМ РАЗУМ, КОТОРЫЙ ОДНОВРЕМЕННО ЯВЛЯЕТСЯ АВТОРОМ ТАКОГО КОНСТРУИРОВАНИЯ, ОН МОЖЕТ ВСТУПИТЬ В ЦИКЛ САМОСОВЕРШЕНСТВОВАНИЯ.

Один из аргументов в пользу ее серьезного рассмотрения заключается в том, что Рэй Курцвейл (Ray Kurzweil) называет «законом ускорения отдачи». В области технологии действует закон ускорения отдачи, если скорость, с которой совершенствуется технология, пропорциональна качеству самой технологии. Иначе говоря, чем совершеннее технология, тем быстрее она улучшается, что в результате дает экспоненциальное во времени совершенствование.

Широко известный пример этого явления — закон Мура (Moore), по которому количество транзисторов, размещаемых на кристалле интегральной схемы, удваивается примерно каждые полтора года[2]. Поразительно, но в полупроводниковой промышленности закон Мура оставался в силе на протяжении нескольких десятилетий. Другие показатели прогресса информационных технологий, такие как тактовая частота процессоров и пропускная способность сети, демонстрировали похожий экспоненциальный рост. При этом информационные технологии не единственная область, в которой наблюдалось ускорение прогресса. Например, в медицине стоимость секвенирования ДНК падала экспоненциально одновременно с экспоненциальным увеличением скорости секвенирования, а в области сканирования мозга наблюдался экспоненциальный рост разрешения[3].

В истории такие тенденции ускорения можно наблюдать на примере ряда знаковых событий в технике, время между которыми постоянно сокращалось: в сельском хозяйстве, издательском деле, электроэнергетике и вычислительной технике. В более крупном временно́м масштабе земной эволюции подобным явлениям технического прогресса предшествовали последовательности ключевых точек эволюции живого, следующих друг за другом

с постоянно сокращающимися временными промежутками между ними: появление эукариотов, позвоночных, приматов и, наконец, homo sapiens. Эти факты дают некоторым комментаторам основания утверждать, что человеческая раса «оседлала» кривую резко растущей сложности, берущую начало в далеком прошлом. Если это так, то надо лишь экстраполировать технологическую часть этой кривой чуть в будущее, чтобы увидеть важный переломный момент — точку, в которой технология совершенствования человека делает обычного человека безнадежно устаревшим с технологической точки зрения[4].

Естественно, по законам физики любой экспоненциальный тренд в технологии должен рано или поздно выйти на пологий участок, и есть масса экономических, политических или научных причин, по которым экспоненциальный тренд может прекратиться до достижения своего теоретического предела. Но представим себе, что развитие технологий, имеющих самое непосредственное отношение к ИИ и нейротехнологиям, сохранит свой темп роста, и в результате такого расширения возможностей конструирования разума мы научимся синтезировать сам интеллект и манипулировать им. На этом этапе сам интеллект, искусственный или человеческий, может подчиняться закону ускорения отдачи, а отсюда остается один небольшой шаг до признания возможности технической сингулярности.

Некоторые авторы с уверенностью прогнозируют наступление этого переломного момента в середине XXI века. Но есть и другие поводы более глубоко осмыслить сингулярность, не скатываясь при этом в зыбкие предсказания. Во-первых, сама концепция исключительно интересна с интеллектуальной точки зрения независимо от того, когда

наступит сингулярность и произойдет ли это вообще. Во-вторых, сама возможность сингулярности, какой бы далекой она ни казалась, уже сегодня способствует дискуссиям по чисто прагматическим и исключительно рациональным темам. Даже если аргументы футуристов ошибочны, достаточно лишь небольшой вероятности таких событий, чтобы вызвать самое живое внимание с нашей стороны. Ведь если технологическая сингулярность действительно произойдет, последствия для человечества будут фундаментальными.

Какими будут эти фундаментальные последствия? Каким станет мир, если техническая сингулярность действительно наступит? Надо ли нам опасаться перспектив сингулярности или радостно приветствовать их? Что мы можем (и можем ли) сделать сейчас или в ближайшем будущем, чтобы гарантировать наилучший возможный результат? Вот вопросы, которыми мы займемся на страницах данной книги. Это серьезные вопросы. Но возможность сингулярности, пусть даже в теории, позволяет пролить свет на еще более серьезные философские вопросы. В чем сущность нашей человеческой природы? Каковы наши самые фундаментальные ценности? Как нам надо жить? Чем во всем этом мы готовы пожертвовать? Ибо вероятность технологической сингулярности создает как экзистенциальный риск, так и экзистенциальные возможности.

Экзистенциальный риск заключается в угрозе самому выживанию человека как вида. Это может казаться преувеличением, но новые современные технологии обладают невиданными ранее возможностями. Несложно представить себе, как беспринципный ученый создает исключительно заразный, устойчивый к лекарствам вирус, способный погубить все человечество. Только сумасшедший может

преднамеренно сотворить такую вещь. Но для создания вируса, способного превратиться в подобного монстра, требуется чуть больше, чем простое безрассудство. Причины, по которым развитый ИИ создает экзистенциальный риск, аналогичны, но существенно тоньше. В свое время мы поговорим о них. А пока достаточно сказать, что вполне разумно было бы представить, что в будущем некая компания, государство, организация или человек могут создать экспоненциально самосовершенствующийся, жадно пожирающий ресурсы искусственный интеллект, а затем потерять над ним контроль.

На эту ситуацию можно посмотреть и с оптимистической точки зрения, считая технологическую сингулярность экзистенциальной возможностью, в более философском смысле слова «экзистенциальный». Способность конструировать разум открывает нам возможность выйти за доставшиеся нам от природы биологические пределы и избавиться от обусловленных ими ограничений. Самое главное из этих ограничений — смертность. Живое тело — очень хрупкая вещь, подверженная болезням, разрушению и разложению, а биологический мозг, без которого (сейчас) невозможно сознание человека, — всего лишь часть этого тела. Но если мы научимся восполнять любой ущерб, причиненный телу, и, в конечном итоге, воссоздавать его с нуля, может даже на небиологической основе, тогда ничто не сумеет остановить неограниченное расширение сознания.

Продление жизни — один из аспектов направления, известного как «трансгуманизм». Почему нас должна удовлетворять такая жизнь человека, какой мы ее знаем? Если мы сумеем воссоздать мозг, то что сможет запретить нам перепроектировать или улучшить его? (Такой же вопрос

можно задать о человеческом теле, но здесь нас интересует интеллект.) Можно улучшать память, внимание и способности к обучению с помощью фармакологических средств. Но способность полностью перепроектировать мозг предполагает возможность более радикальных форм улучшения и реорганизации процессов познания. Что мы можем и должны делать с помощью таких средств преобразования? Некоторые говорят, что это как минимум снижает экзистенциальный риск со стороны суперинтеллектуальных машин. Весьма вероятно, что нам удастся поспеть за их возможностями, но в процессе мы рискуем измениться до неузнаваемости.

Самый крупный и вызывающий экзистенциальный аспект технологической сингулярности можно осознать, только полностью абстрагировавшись от человека и заняв более космологическую точку зрения. Ясно, что антропоцентрическому мышлению свойственно полагать, будто история материи в нашем уголке Вселенной замыкается на человеческом обществе и мириадах живых мозгов. Но не исключено, что у материи есть масса других возможностей увеличения масштаба сложности. Вероятно, в будущем возникнут более совершенные формы сознания, нежели наше. Должны ли мы опасаться такой перспективы или приветствовать ее? Сможем ли мы в принципе воспринять такую идею? Независимо от того, приближается точка сингулярности на самом деле или нет, эти вопросы стоит задавать, и не в последнюю очередь потому, что попытки ответов на них способны пролить свет на нас самих и наше место в миропорядке.