Искусственный интеллект

Этапы. Угрозы. Стратегии

Книгу рекомендует Билл Гейтс

Автор входит в топ-15 мыслителей мира по версии журнала Prospect

Илон Маск, основатель SpaceX и Tesla, считает эту книгу достойной прочтения

Ник Бостром – профессор Оксфордского университета, основатель и директор Института будущего человечества, исследовательского центра, где работают лучшие математики, философы и ученые мира. Автор более 200 научных публикаций. Его работы переведены на 22 языка.

Основная идея

В последние годы дискуссия насчет будущего прорыва в сфере создания искусственного интеллекта становится все более острой. Ник Бостром не берется прогнозировать, когда это случится (хотя и говорит, что, скорее всего, это произойдет уже в нынешнем столетии). Он фокусируется на той проблеме, которая встанет перед человечеством, когда сверхразум превратится из фантастики в реальность, и анализирует, к чему это приведет и какой может быть ответная реакция людей.

Парадоксы развития

В процессе развития человечества темпы развития экономики непрерывно возрастали. Временами – плавно, временами – скачкообразно, особенно в результате промышленных революций. Этот процесс продолжается до сегодняшнего дня. И если тот темп роста, который сохранялся в последние 50 лет, останется неизменным, то население планеты к 2050 году станет богаче в 4,8 раза. А к 2100 году – в 34 раза. Этот скачок может быть связан со взрывоподобным развитием интеллекта и появлением искусственного сверхразума.

Автор цитирует статью 1965 года, написанную математиком Ирвингом Джоном Гудом, в которой содержится первое описание такого сценария развития событий: «Давайте определим сверхразумную машину как машину, которая в значительной степени превосходит интеллектуальные возможности любого умнейшего человека… Машина, наделенная сверхразумом, будет способна разрабатывать еще более совершенные машины; вследствие этого, бесспорно, случится такой интеллектуальный взрыв, что человеческий разум окажется отброшенным далеко назад».

Несмотря на такой пессимистический прогноз, первопроходцы в сфере искусственного интеллекта не верили в возможность создания сверхразума и тем более отрицали вероятные негативные последствия такого прорыва. Однако сегодня эксперты рассматривают разные сценарии, в том числе и те, которые напоминают сюжеты фантастических фильмов.

Исследования искусственного интеллекта, начавшиеся еще в прошлом столетии, привели на данный момент к таким важным достижениям, как создание прочной статистической и информационно-теоретической базы для машинного обучения и массы коммерчески успешных приложений в самых разных сферах деятельности.

Однако один из пионеров этого направления Нильс Нильсон считает, что все эти результаты – это «слабый ИИ», цель которого – предоставление помощи человеку в его интеллектуальной работе. По его мнению, которое поддерживает немало корифеев, людям стоит сосредоточиться на создании «сильного ИИ», то есть машинного разума, сопоставимого с человеческим. Бостром считает, что в ближайшем будущем мы можем увидеть немало попыток создать такой ИИ. Он также уверен, что вскоре за этим событием появится и истинный сверхразум – машина, которая превзойдет человека. И, что самое главное, у этого события обязательно будут масштабные последствия. «Вероятность чрезвычайно сильного воздействия – позитивного или негативного – на человечество гораздо более высока, чем вероятность нейтрального», – пишет он.

Пять основных мыслей

1. **Исследования искусственного интеллекта привели к таким достижениям, как создание прочной базы для машинного обучения и массы коммерчески успешных приложений в самых разных сферах деятельности. Однако все это – «слабый ИИ».**

2. **«Сильный ИИ» – это машинный разум, сопоставимый с человеческим. Если он появится, вскоре возникнет и истинный сверхразум – машина, которая превзойдет человека.**

3. **Существуют три типа сверхразума: скоростной, коллективный и качественный. Каждый из них имеет свою специфику и лучше справляется с определенными задачами.**

4. **Если появится интеллект, превышающий человеческий, то он будет обладать практически неограниченными возможностями. А разработчики, имеющие контроль над сверхразумом, получат огромные рычаги влияния на мир.**

5. **Человечеству нужно заранее подготовиться к сценариям, по которым искусственный интеллект приобретает решающее стратегическое преимущество.**

Типы сверхразума

Что именно имеют в виду ученые, когда говорят о создании искусственного интеллекта, превосходящего человеческий? Бостром выделяет три типа сверхразума: скоростной, коллективный и качественный.

Скоростной сверхразум. Понять и проанализировать его гораздо проще, чем остальные типы. Он представляет собой систему, способную делать все то же, что и человеческий интеллект, только намного быстрее. Причем намного в данном контексте означает на несколько порядков.

Пример, который приводит автор, – это полная эмуляция головного мозга человека, выполненная на сверхмощном компьютерном оборудовании. Такая система могла бы прочитать книгу за несколько секунд, если бы работала всего в десять раз быстрее, чем наш разум. А если удастся добиться скорости в миллион раз выше, то ИИ сможет за день выполнить интеллектуальную работу, на которую у человека ушло бы тысячелетие.

Для скоростного сверхразума человеческая жизнь казалась бы невыносимо замедленной, поэтому легко предположить, что он предпочел бы работать с цифровыми объектами. «Ему удобнее было бы существовать в виртуальной реальности и иметь дело с информационной экономикой, а при необходимости – вступать во взаимодействие с физической средой при помощи наноманипуляторов», – пишет Бостром. Общаться он тоже предпочел бы не с людьми, а с другими скоростными разумами.

Коллективный сверхразум. Это система, которая объединяет множество интеллектов более низкого уровня, благодаря чему ее суммарная производительность превышает производительность любой существующей когнитивной системы во многих универсальных областях деятельности.

Эта концепция уже используется людьми: в некотором смысле компании, социальные сети, команды, государства являются примерами коллективного разума. Благодаря разделению задач такой способ организации работы позволяет добиться хороших результатов практически во всех сферах. «Такого рода разумная система может быть усилена за счет усовершенствования каждой отдельной подструктуры: расширения ее состава, повышения ее уровня интеллекта, оптимизации ее организационной политики», – пишет автор. Таким образом, чтобы создать коллективный сверхразум, нужно добиться резкого роста на всех уровнях существующих систем.

Если произойдет рывок, который предсказывают некоторые эксперты, экономика начнет удваивать темпы роста каждые две недели

Качественный сверхразум. Этот тип представляет собой систему, по скорости работы сравнимую с человеческим умом, но в качественном отношении значительно превышающую его. Примерно так, как человеческий ум превосходит по качеству ум слонов, дельфинов и шимпанзе.

Бостром уверен, что если возникнет сверхразум одного из описанных типов, то со временем он мог бы помочь развитию технологий, которые привели бы к появлению сверхразумов оставшихся двух типов. С этой точки зрения все три типа одинаково достижимы. Также автор отмечает: «В чем-то эти три типа гораздо ближе друг к другу, поскольку любой из них способен создать два других гораздо быстрее, чем мы – один из них».

С какими задачами справляются три вида сверхразума? Скоростной, очевидно, хорошо

В некотором смысле компании, социальные сети, команды, государства являются примером коллективного разума

проявил бы себя в ситуациях, когда нужно выполнить длинную последовательность действий, коллективный – когда нужен анализ и декомпозиция задания на параллельные подзадачи, а также тогда, когда необходимо комбинировать разные навыки. А качественный сверхразум является универсальным типом: он способен справиться с задачами, находящимися вне пределов прямой досягаемости скоростного и коллективного сверхразумов.

Время взлета

Если допустить, что рано или поздно прорыв в сфере ИИ свершится, на первый план выходит вопрос: как быстро машина, обладающая интеллектом человеческого уровня, полностью превратится в сверхразумную? Бостром рассматривает три типа сценариев:

1. Медленный взлет. Если события будут развиваться медленно, то переход к сверхразуму займет от десятков до сотен лет. Этот сценарий дает возможность человечеству как следует подготовиться и адаптироваться к новой реальности, обдумать свои дальнейшие действия и способы взаимодействия со сверхразумом. Люди смогут разработать и протестировать несколько теорий, обучить соответствующих специалистов, перестроить некоторые из общественных функций. Если понадобятся более совершенные системы безопасности, будет время на их создание. Страны успеют провести переговоры и найти совместное решение возникающих проблем.

2. Быстрый взлет. Этот сценарий подразумевает, что изменения наступят в течение очень короткого времени: минут, часов или дней. В таком случае люди не успеют практически ничего предпринять, и если смотреть на появление сверхразума как на борьбу, то человечество ее заведомо проиграет. А его судьба будет зависеть от тех шагов, что были предприняты заранее.

3. Умеренный взлет. Этот сценарий может развернуться на протяжении от нескольких месяцев до нескольких лет. В этом случае времени на разработку и тестирование теорий не будет, но люди еще смогут предпринять некоторые действия. «На создание и развертывание новых систем, таких как политические меры, механизмы контроля, протоколы безопасности компьютерных сетей, времени тоже не останется, но, возможно, получится приспособить к новым обстоятельствам уже существующие нормы», – пишет автор.

В сценариях умеренного взлета, скорее всего, возникнут социальные, экономические и политические потрясения, ведь корпорации, группы интересов и отдельные люди будут пытаться обеспечить себе преимущество за счет появления сверхразума. Например, автор приводит описание одного из вероятных последствий: на рынок труда выходят имитации человека, дешевые и производительные, после чего начинаются массовые увольнения людей. А за ними следуют массовые протесты, на которые правительства будут вынуждены отреагировать, повысив размер пособий по безработице и введя некие социальные гарантии для населения, а также обложив компании дополнительным налогом на использование труда роботов.

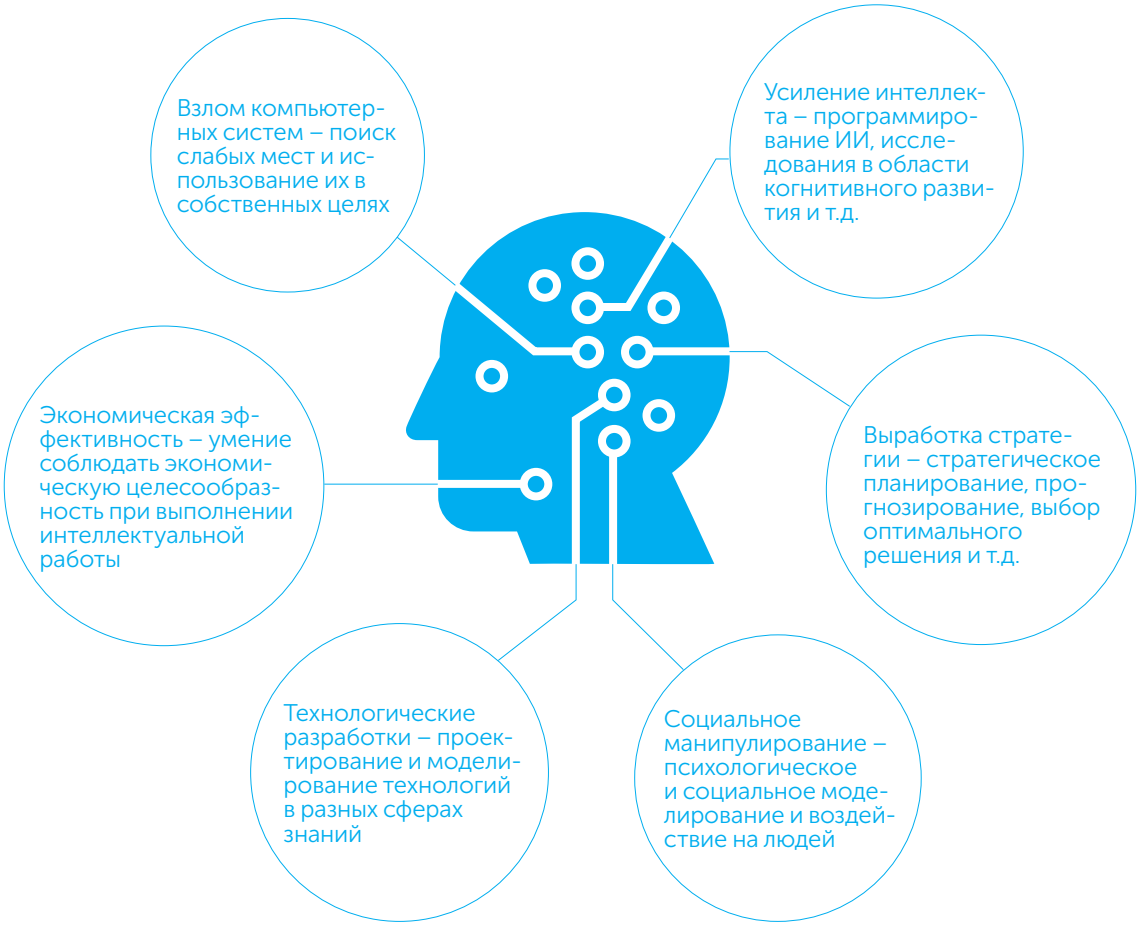
Какой из сценариев наиболее вероятен? Бостром уверен: медленный – наименее возможен. По его мнению, с наибольшей долей вероятности события будут разворачиваться по быстрому сценарию.

> Если прорыв в сфере искусственного интеллекта произойдет, он будет иметь характер взрыва

Власть разума

Помимо прогнозов, когда будет создан сверхразум и как будет проходить этот процесс, нас интересует и то, какой властью он станет обладать. Автор уверен, что если появится

Потенциальные возможности искусственного интеллекта



интеллект, превышающий человеческий, то он получит практически неограниченные возможности и, наверняка, сможет создать новую технологическую цивилизацию.

Размышляя о подобной перспективе, мы совершаем распространенную ошибку – приписываем сверхразуму человеческие свойства, психические особенности и мотивации. «По общему мнению, сверхразумная машина будет скорее напоминать заумного зануду – этакое существо с энциклопедическими знаниями, но социально незрелое; последовательное в действиях, но обделенное интуицией и творческим началом», – пишет Бостром. Однако на самом деле такое описание, возможно, могло бы подойти зародышу ИИ, но уж точно не сверхразуму.

Наделять ИИ антропоморфными чертами – не только неправильно, но и опасно. Ведь в этом случае мы рискуем недооценить сверхразум и просто не заметить, насколько он превышает наш интеллект. У людей существует две крайности – понятия «умный» и «глупый», и мы оцениваем всех, в том числе ИИ, по этой шкале.

Людям еще предстоит разработать методы оценки когнитивных способностей ИИ, однако уже сейчас существует список стратегически

Один из типов сверхразума способен гораздо быстрее создать два других, чем человечество создаст хотя бы один

Стоит задуматься

1. **Как изменится ваша жизнь**, если рынок труда заполнят роботы?
2. **Хотели бы вы застать** быстрый взлет развития ИИ?
3. **Какие**, по вашему мнению, **задачи являются самыми важными** для человечества в преддверии появления сверхразума?

Следует сделать

1. **Продумать**, **как** на ваш бизнес **могут повлиять открытия** в сфере ИИ
2. **Интересоваться развитием информационных технологий**, чтобы быть в курсе событий
3. **Сформировать свою позицию** относительно появления сверхразума

значимых задач. Если система способна справиться с каждой из них, она может считаться сверхмощной. Ожидается, что сверхразум сможет решать их все, а разработчики, имеющие контроль над сверхразумом, получат огромные рычаги влияния на мир.

Усиление интеллекта, выработка стратегии, социальное манипулирование, взлом компьютерных систем, технологические разработки, экономическая эффективность – система, способная решать все эти задачи, может создать собственный долгосрочный план и отмести варианты действий, которые приводят к ее поражению. И не исключено, что этот план будет касаться захвата власти в мире. Вопрос лишь в том, какой будет мотивация сверхразума.

Здесь автор приводит два тезиса. Первый – тезис об ортогональности – гласит: более или менее любой уровень интеллекта может, в принципе, сочетаться с более или менее любой конечной целью.

Второй тезис – об инструментальной конвергенции – в упрощенном виде говорит: сверхразумные действующие силы, или агенты, – при самом широком разнообразии своих конечных целей – будут преследовать сходные промежуточные цели, поскольку на это у всех агентов будут одинаковые инструментальные причины. Этими целями могут быть, к примеру, самосохранение, получение ресурсов, усиление своих когнитивных способностей, совершенствование технологий. Тем не менее даже знание конвергентных причин не позволяет предсказать поведение сверхразума. Возможно, мы и сможем разобраться, каковы цели ИИ, однако для их реализации он может прибегнуть к таким действиям и физическим явлениям, которые человечеству еще неведомы.

Бостром рассматривает разные варианты событий, с которыми может столкнуться человечество. На данный момент наблюдения за существующими приложениями ИИ показывают: чем умнее ИИ, тем он безопаснее. Причем речь идет не о предположениях, а о данных серьезных научных исследований и статистике. Разработчики становятся все более оптимистичными, они легче переходят на следующий этап – и в результате уровень ИИ постоянно растет.

Но что, если это всего лишь вероломная уловка со стороны зародыша сверхразума? Пока ИИ слаб, он всячески демонстрирует готовность сотрудничать с людьми. И чем выше становится уровень интеллекта, тем сильнее эта готовность. Но когда ИИ станет достаточно мощным, он без предупреждения нанесет удар и начнет изменять мир в соответствии со своими целями…

Таких вариантов существует множество, и Бостром делает вывод: «К сценариям, в которых искусственный интеллект приобретает решающее стратегическое преимущество, следует относиться со всей серьезностью».

Диапазон между понятиями «умный» и «глупый» ничтожен по сравнению с дистанцией между человеческим интеллектом и сверхразумом

Способы контроля

Оставить процесс развития ИИ без контроля означает проявить недальновидность, чреватую катастрофическими последствиями. Причем некоторые методы контроля или их сочетание следует использовать еще до того, как появится настоящий сверхразум. Людям нужно разработать решения и внедрить их в первую же систему, которая станет сверхразумной. Только в этом случае у человечества еще будет шанс управлять ходом взрывного развития ИИ.

К сценариям, в которых ИИ приобретает решающее преимущество, стоит отнестись столь же серьезно, как и к другим, менее фантастическим

Бостром описывает несколько видов контроля, которые мы можем использовать для обеспечения своей безопасности:

• **Изоляционные методы**, то есть помещение ИИ в среду, где он не сможет нанести вред человечеству. Блокировка может быть как физической, так и информационной. В первом случае интеллектуальную систему изолируют от внешнего мира, оставив лишь строго определенные каналы коммуникации. Такую изоляцию легко организовать, а также сочетать с другими видами контроля. Информационная блокировка – это ограничение информационных потоков, исходящих от ИИ. В фантастических фильмах часто показывают, к чему может привести подключение сверхразума к интерне-

ту. Поэтому кажется логичным не давать интеллектуальной системе доступ к коммуникационным сетям.

• **Стимулирующие методы контроля** заключаются в создании условий, при которых ИИ будет выгодно действовать в интересах людей. Здесь существует много вариантов, один из которых – наделить сверхразум конечной целью, которую можно взять под контроль, и разработать для него систему вознаграждений.

• **Методы задержки развития.** Другими словами, мы сознательно ограничиваем интеллектуальные способности системы или ее доступ к информации. Простейший пример – запуск ИИ на компьютере с низким быстродействием или недостаточной памятью. Разумеется, в этом случае искусственный разум теряет значительную часть своей полезности.

• **Методы растяжек.** В этом контексте растяжки – это специфическое оборудование, которое дает возможность проводить диагностику ИИ, в том числе без его ведома. А в случае обнаружения тревожных сигналов – отключать систему. Растяжки могут стать временными мерами, которые обеспечивают защиту на этапе разработки ИИ и иногда – на этапе функционирования, особенно для изолированных систем. Однако полностью сформировавшийся сверхразум вряд ли удастся сдержать таким методом – он попросту обойдет любую защиту.

Нужно отметить, что человечество сможет использовать не только методы контроля, но и методы выбора мотивации, которые формируют мотивы поведения сверхразума таким образом, чтобы не допустить нежелательных последствий. В их числе – метод точной спецификации (то есть однозначная формулировка цели и системы правил, которым должен следовать ИИ), метод приручения (разработка программы с тем, чтобы привести ИИ к выбору конечных целей, устраивающих человечество) и другие.

Важные вопросы

Но все же, исходя из всего вышесказанного, стоит ли людям радоваться успехам в развитии аппаратного обеспечения и шагам на пути к созданию компьютерной модели мозга? «Быстрый прогресс в области аппаратного обеспечения повышает вероятность быстрого взлета», – пишет автор. А значит, объективно этот прогресс нежелателен. Однако так ли это на самом деле, неизвестно. В любой момент могут появиться доказательства того, что задержка развития аппаратного обеспечения приведет к еще худшим последствиям.

Что же касается второго вопроса, то есть серьезные предпосылки предполагать: если компьютерная имитационная модель головного мозга (КИМГМ) появится раньше, чем сверхразум, то риск перехода к ИИ будет ниже. «Если смотреть скептически на способность человечества управлять переходом к ИИ, – приняв во внимание, что человеческая природа или цивилизация могут улучшиться к тому моменту, когда мы столкнемся с этим вызовом, – то сценарий «сначала КИМГМ» кажется более привлекательным», – утверждает Бостром.

Быстрый прогресс в области аппаратного обеспечения повышает вероятность быстрого взлета ИИ

У сверхразума могут быть цели, глубоко чуждые интересам и ценностям человечества

Тем не менее интуитивно люди стремятся к более быстрому развитию. Ведь все, кто сейчас живет на планете, умрут на протяжении ближайших 100 лет. И многие из нас хотели бы застать будущее, каким бы оно ни было. Кроме того, мы втайне надеемся на то, что сверхразум изобретет способ продлевать человеческую жизнь, причем делая ее качественной.

Таким образом, субъективно быстрый прогресс в области аппаратного обеспечения является таким же желательным, как и в области создания КИМГМ. Потенциальные риски взрывного развития компенсируются вероятной пользой для отдельно взятого человека