Оглавление

К читателям

Дорогие читатели!

Сегодня технологии искусственного интеллекта прочно вошли в нашу повседневную жизнь, став незаменимыми помощниками в решении множества задач — от эффективной поддержки наших усилий при выполнении рутинных действий, обработке и анализе больших массивов неструктурированной информации — до оказания помощи в реализации нашего творческого потенциала и развития наших креативных навыков. А в таких индустриях как здравоохранение, беспилотный транспорт и охрана общественного порядка, технологии ИИ помогают принимать решения более качественно и быстрее — ровно тогда, когда счет идет на секунды, и эти секунды способны спасти чьи то жизни!

В большинстве своем существующие решения являются примерами реализации технологий узко специализированного искусственного интеллекта, требующего настройки и перепроверки со стороны человека. Чтобы так же хорошо решать разнообразные комплексные задачи, как это делают люди, машины должны научиться строить причинно-следственные модели окружающей среды и ориентироваться в разных контекстах, а не просто максимизировать успех при решении какой-то узкой задачи. Они должны понимать физические, психологические и другие законы нашего мира и уметь связывать новую

информацию в общую картину с тем, что уже знают. Чтобы добиться этого, нам необходимо преодолеть очередной технологический рубеж — создание Общего искусственного интеллекта или AGI.

Вместе с коллективом авторов этой книги мы впервые обобщили и систематизировали накопленные знания в области Общего искусственного интеллекта — от компьютерных наук и машинного обучения до нейронаук и психологии. Оценивая ретроспективу научных достижений в области технологий искусственного интеллекта, особенно впечатляющим оказался «взрывной рост» технологий в последние несколько лет, возможный во многом благодаря экспоненциальному росту доступных вычислительных мощностей и снятию технологический ограничений накопления больших объёмов данных для обучения алгоритмов. Все это позволило давно известным архитектурам многослойных нейронных сетей успешно решать задачи в самых различных сферах человеческой деятельности. При этом, несмотря на особое внимание, которое уделяется нейронным сетям и глубокому обучению, это далеко не единственный путь к Общему искусственному разуму.

Эта книга будет прежде всего интересна всем тем, кто хочет понять, с помощью каких подходов может быть создан Общий искусственный интеллект и какую форму он, вероятнее всего, приобретет в различных сферах прикладного применения. По опыту развития современных технологий машинного обучения мы видим, что путь от идеи до индустриальных решений может оказаться значительно короче, чем это кажется вначале.

Сегодня уже очевидно, что любая выбранная нами стратегия движения в развитии Общего искусственного интеллекта и хорошо организованные управленческие усилия неизбежно столкнутся с ключевым вопросом — вопросом выбора корректных критериев отнесения того или иного исследования к области Общего искусственного интеллекта. Для этого нам важно хорошо понимать структуру исследуемой области,

составить единую и непротиворечивую теоретическую базу, а затем определить наиболее перспективные подходы к развитию технологий и их последующей индустриализации. Этому и посвящена книга. Мы верим, что лишь сделав ставку на эффективное сотрудничество специалистов разных направлений — от глубокого обучения и вероятностного программирования до робототехники и когнитивистики, а также поддерживая междисциплинарные исследования, можно добиться первых ощутимых прикладных результатов в области развития Общего искусственного интеллекта. Крайне важным фактором создания и развития технологий Общего искусственного интеллекта является обеспечение сквозного целеполагания между прикладными (или индустриальными) задачами, фундаментальными исследованиями и системой образования — так называемая триада Practice–Education–Research.

Также мы считаем важным уделить особое внимание профессионального сообщества теме соблюдения принципов этического применения современных технологий. В частности, необходимо на международном и межиндустриальном уровнях выработать единые и общепризнанные стандарты, позволяющие обеспечивать безопасное и социально-полезное применение технологий Общего искусственного интеллекта. В этом контексте соответствующие усилия должны быть предприняты для обеспечения стабильности и интерпретируемости работы алгоритмов, лежащих в основе технологий Общего искусственного интеллекта.

Работая над книгой, мы убедились в том, что у России большой потенциал развития прорывных технологий в области Общего искусственного интеллекта. Хочется надеяться, что наш общий труд подтолкнет исследователей, инженеров, представителей бизнеса и государства к эффективному сотрудничеству в создании принципиально новых подходов на пути к AGI и это позволит России занять лидирующее место в гонке мировых держав в области ИИ.

Многие из нас прекрасно осознают неминуемость масштабных технологических трансформаций нашего общества. Доверие к технологиям искусственного интеллекта в обществе пока только формируется, много говорится о негативных сценариях, которые могут привести к катастрофическим последствиям для цивилизации. Важно слышать и понимать опасения людей. Мы должны научиться использовать новые технологии во благо нашего общества.

И несмотря на то, что нам с вами предстоит еще пройти непростой и большой путь для достижения первых прикладных результатов в области применения технологий Общего искусственного интеллекта, уже сейчас мы можем видеть весьма многообещающие решения в этой области.

Я желаю вам увлекательного и продуктивного чтения! Помните — исключительно важно всеобщее понимание того, что достижение результатов в этом путешествии должно стать благом для всего Человечества!

Герман Греф, Президент,
Председатель Правления Сбера

Коллектив авторов

Бурцев М. С., канд. физ.-мат. наук, МФТИ.

Бухвалов О. Л., канд. техн. наук, Brain Garden.

Ведяхин А. А., канд. экон. наук, первый заместитель председателя правления, ПАО Сбербанк.

Витяев Е. Е., д-р физ.-мат. наук, Институт математики СО РАН, профессор Новосибирского государственного университета.

Еременко М. А, ПАО Сбербанк.

Ефимов А. Р., ПАО Сбербанк, НИТУ «МИСиС».

Колонин А. Г., канд. техн. наук, Новосибирский государственный университет, Aigents.

Курпатов А. В., ПАО Сбербанк.

Мазин В. А., канд. физ.-мат. наук, Mind Simulation AGI laboratory.

Марков С. С., ПАО Сбербанк.

Молчанов А. А., ПАО Сбербанк.

Нейросеть RuGPT-3.

Николенко С. И., канд. физ.-мат. наук, ПОМИ РАН, Neuromation.

Очеретный А.С., ПАО Сбербанк.

Панов А. И., канд. физ.-мат. наук, ФИЦ ИУ РАН, МФТИ.

Пономарев Д. К., канд. физ.-мат. наук, Институт систем информатики СО РАН, Новосибирский государственный университет.

Потапов А. С., доцент, к.т.н., SingularityNet.

Салихов Д. Р., ПАО Сбербанк.
Сарапулов Г. В., Brain Garden.
Свириденко Д. И., д-р физ.-мат. наук, доцент, профессор, Институт математики СО РАН.
Черток А. В., канд. физ.-мат. наук, ПАО Сбербанк.
Шаляпин С. О., Естественный интеллект.
Шелехов В. И., канд. техн. наук, доцент, Институт систем информатики СО РАН.
Franz A., PhD, OCCAM.

Введение

Технологии искусственного интеллекта с самого своего появления демонстрировали удивительные достижения в решении задач, с которыми, как традиционно считалось, способен справиться только человеческий разум.

Сейчас технологии ИИ становятся массовыми и повсеместными, проникли в нашу повседневную жизнь и вряд ли ее покинут. Они используются в поисковых и рекомендательных системах, транспорте, логистике, банковском деле, планировании бизнес-процессов, производстве и научных исследованиях. Они уже давно не ограничиваются цифровой реальностью, проникая в быт. Нас начинают окружать домашние роботы, беспилотные аппараты, умные дома и города, не говоря уже о приложениях с элементами ИИ для смартфонов и персональных компьютеров. Технологии ИИ, включающие машинное обучение, научились неплохо справляться с анализом изображений, звука, речи и текстов на естественных языках. Иногда они делают это не просто на человеческом, а на сверхчеловеческом уровне. Технологии искусственного интеллекта открывают перед нами огромные перспективы. Они способны придать новый импульс развитию мировой экономики, оказать позитивное влияние на все сферы нашей жизни. В этом году они значительно помогли медицинскому сообществу в борьбе с пандемией.

Из-за многочисленных успехов последнего времени может сложиться впечатление, что недавний впечатляющий прогресс

технологий ИИ достиг насыщения. Но это вовсе не так, и диапазон их применения только растет. Даже уже известные технологии имеют массу еще не реализованных возможностей по внедрению. Либо для этого не нашлось свободных специалистов, либо стоимость разработки и внедрения перевешивает ожидаемую прибыль, либо соответствующие технологии, уже существующие в теории, еще не способны предоставить решения достаточно качественного, чтобы быть полезными и удобными в ежедневном использовании. А главное, новые, более перспективные технологии появляются быстрее, чем старые успевают стать в полной мере использованными.

Однако существуют и такие крайне важные задачи, для которых имеющихся технологий просто недостаточно, — например, исследовательские. Скажем, интеллектуальная система, которая могла бы продвинуться в решении проблемы человеческого старения, должна была бы не только анализировать научные статьи, причем на более глубоком уровне, чем поверхностные корреляции между словами, но и моделировать взаимосвязи между различными подсистемами и процессами организма. Такой системы у нас пока нет, и, возможно, излишне оптимистичным было бы полагать, что она может достаточно быстро появиться в результате естественного развития доминирующих сейчас технологий ИИ, нацеленных на решение узких задач.

Эта книга основана на исследовании лучших российских специалистов по ИИ, посвященном общему искусственному интеллекту. Это ИИ, способный самообучаться и решать разнообразные задачи в разных контекстах. Системы искусственного интеллекта смогли бы помочь человечеству справиться с самыми сложными вызовами: построением более справедливого общества, поиском лекарств от смертельных заболеваний, предупреждением катастроф и т.д. Кроме того, развитые технологии ИИ — это важное стратегическое преимущество для государства на внешнеполитической арене. Как говорил В.В. Путин, «Искусственный интеллект — это будущее не

только России, но и всего человечества, и тот, кто будет лидером в этой сфере, станет правителем мира». Пока разработка таких систем вызывает много сложностей, но существует ряд подходов, которые могли бы продвинуть нас в решении этой задачи. Наша книга представляет собой самый полный и глубокий обзор этих подходов и первый шаг к выработке общего бэкграунда для заинтересованных в AGI на русском языке. Он поможет специалистам из разных областей ИИ объединить свои знания и выработать стратегию по созданию общего искусственного интеллекта. Эта книга написана научно-популярным языком, делающим ценные знания доступными для более широкой аудитории, кроме того, скоро выйдет чисто научная версия для глубокого погружения в тему и планируется создание практического руководства.

Недостатки узкоспециальных систем

Применение технологий ИИ все еще не настолько впечатляюще, как могло бы быть, по многим причинам. Но все они обусловлены одним фундаментально важным фактором. Большинство таких систем остаются узкоспециализированными, а еще точнее, позволяют достичь качественных решений только для узких задач. Это свойство не изменилось со времен экспертных систем ИИ, которые всегда требовали формализованных описаний и вручную заложенного эвристического знания. Экспертные системы, системы компьютерного зрения и любые другие знали примерно столько, сколько им сообщили разработчики — и не больше. Поэтому уже полвека назад в адрес ИИ звучала критика в том духе, что «компьютер, запрограммированный на решение тысячи задач, не способен самостоятельно научиться решать тысяча первую».

Эта критика отчасти справедлива и до сих пор. Хотя современные системы машинного обучения используют довольно сложные методы работы с данными (об этом подробно рассказывается в следующих главах), каждое конкретное решение все еще специализировано под конкретную задачу. Часто эта специализация оказывается даже выше, чем у старых, классических систем, потому что вручную разработанные представления информации обычно более общие, а выученные машиной — подгоняются под конкретную выборку. В результате опыт решения одной задачи плохо переносится на решение другой задачи или даже новый набор данных.

Контраст бывает разительным. Компьютерную модель можно сравнительно легко научить по размеченным данным видеокамер определять, когда один и тот же произвольный человек появляется на разных камерах с неперекрывающимися полями зрения, и она будет точна в 95% случаев. Но стоит проверить ее на другом, незнакомом наборе камер, и точность идентификации упадет ниже 10%. Чтобы не допускать таких провалов, разработчики тренируют модели на нескольких наборах данных с применением методов трансферного обучения, но и этого недостаточно: качество идентификации все равно может оказаться непригодным для практического использования. Иногда обучаемые модели при применении в новых условиях уступают даже необучаемому методу, который пользуется общими признаками, сконструированными вручную.

Другой яркий пример — модель, обученная играть в игры Atari. Она была способна играть в любую из множества игр, хотя тратила на тренировку гораздо больше времени, чем человек, и не во всех испытаниях была способна демонстрировать сверхчеловеческий уровень (хотя недавно компьютер превзошел уровень среднего человека во всех играх). Каждой игре она училась отдельно, и после самого незначительного изменения параметров ей приходилось переучиваться. Поменяйте цвет стен, мимо которых бежит персонаж

компьютерной игры, — и большинство людей этого даже не заметит, а такой модели придется начинать тренировку заново.

В некоторых приложениях машинное обучение все еще не смогло заменить классические технологии ИИ, так как все еще плохо работает со структурированной информацией, например априорными знаниями и причинно-следственными связями. Это тоже можно трактовать как свидетельство узости моделей машинного обучения, но уже более глубокого уровня, чем уровень приложений.

> Разные авторы подчеркивают разные недостатки современных систем ИИ. Кто-то считает, что основной проблемой для них является приобретение новых навыков. Кто-то обращает внимание на нехватку надежности. Но все эти проблемы — симптомы одного свойства: недостаточной широты методов ИИ.

Приобретение новых навыков или надежность не ограничивают решение задач в узких предметных областях. Например, промышленные роботы в строго контролируемых производственных условиях не ошибаются. Проблема возникает тогда, когда методы, придуманные для решения узких задач в предсказуемой среде, начинают использовать широко.

Удивительным образом узкая специализация служит основным ограничением при разработке и внедрении систем ИИ для решения любых задач, как частных, так и общих. Частных — потому что чем конкретнее задачи, тем их больше, и каждая из них требует затрат на разработку. В результате и специалистов не хватает, и разработка и внедрение часто оказываются коммерчески неоправданными. Общих — потому что чем шире задачи, тем больше труда по разметке данных или инженерии знаний они требуют. При решении произвольных задач, особенно сложных, узость методов из количественной проблемы превращается в качественную.

Такой качественный переход призвана совершить область общего искусственного интеллекта (Artificial General Intelligence, AGI), который нужен именно для того, чтобы системы ИИ были общими, и который противопоставляется «узкому» ИИ (Narrow AI). Развитие AGI вместо существующих теперь узких методов или вместе с ними может заложить новый виток технологического прогресса человечества, трансформировать технологии, науку и общество. Поэтому актуальность исследований AGI трудно переоценить.

Общий ИИ — не сильный и не слабый

В первую в мире лабораторию искусственного интеллекта в сопровождении профессора Марвина Минского зашел известный философ и аналитик Хьюберт Дрейфус. Для корпорации RAND он недавно написал аналитический отчет с говорящим названием «Алхимия и искусственный интеллект».

— Вы знаете, что компьютеры принципиально не способны на творчество и что они никогда не смогут, скажем, даже обыграть гроссмейстера в шахматы? — произнес философ, продолжая спор.

— А не хотите ли сыграть с нашим компьютером? Мои студенты как раз недавно закончили работу над шахматной программой, — спросил профессор.

— Извольте. Я и сам неплохо играю.

Вряд ли именно такой диалог состоялся между Хьюбертом Дрейфусом и Марвином Минским, но известный философ в действительности в 1965 г. написал манускрипт «Алхимия и искусственный интеллект», в котором критиковал наивность разработчиков искусственного интеллекта и пытался доказать, что у ИИ есть непреодолимый предел развития, что компьютеры

обладают ограничениями, от которых человеческий разум свободен, и что, в частности, такие игры, как шахматы или го, принципиально не подвластны компьютеру. А в 1969 г. действительно состоялась партия между Дрейфусом и шахматной программой «Мак Хак», написанной Ричардом Гринблаттом в МТИ, и в этой партии философ потерпел поражение.

Это событие, однако, не помешало Дрейфусу и дальше рассуждать об ограниченности компьютеров и в 1972 г. написать трактат «Чего не могут компьютеры» (в русском переводе «Чего не могут вычислительные машины: Критика искусственного разума»), который упорно переписывался вплоть

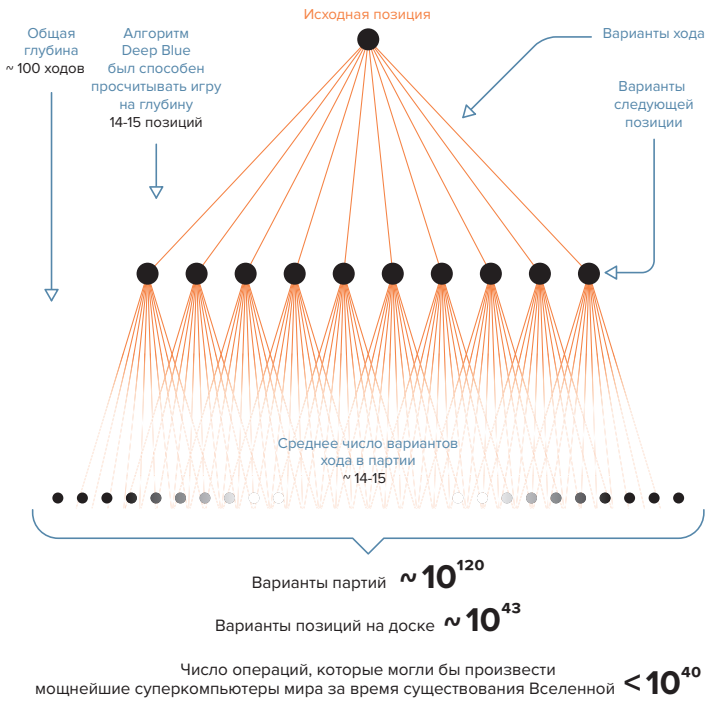

Рис. 1

Дерево вариантов шахматной партии

до 1992 г. и в последней версии назывался «Чего компьютеры все еще не могут».

После «Алхимии» Дрейфуса другой известный философ, Джон Серл, ввел понятие *сильного ИИ* — то есть такого ИИ, который обладает всеми качествами человеческого разума: пониманием, самосознанием, субъективными переживаниями и т.д., — проведя границу между ним и *слабым ИИ*, который таковыми качествами не обладает. Серл полагал, что компьютеры принципиально не способны на сильный ИИ, что и пытался доказать на примере понимания естественного языка в своем мысленном эксперименте — парадоксе «Китайской комнаты», о котором мы поговорим позднее.

Критика возможности реализации сильного ИИ звучала не только от философов. Например, известный математик и физик Роджер Пенроуз утверждал, что в математике существуют алгоритмически неразрешимые задачи, с которыми человеческий интеллект может справиться.

Однако ни одни достаточно конкретные предсказания о том, какие именно задачи компьютер принципиально не способен решать, не сбываются. Компьютер уже победил человека в викторине Jeopardy!, а системы машинного перевода сейчас используются профессиональными переводчиками, которые хоть и вносят правки в вариант перевода, предложенный компьютером, но и учатся у него чему-то новому для себя. Да и работа современного математика без систем помощи в доказательстве теорем вряд ли могла бы быть столь же продуктивной. А лидерство компьютера во всех интеллектуальных играх уже не вызывает сомнения.

Рядом с воззрениями об уникальности человеческого разума и неспособности компьютеров к полноценному мышлению соседствуют популяризованные фантастикой представления о том, что системы искусственного интеллекта и роботы скоро обретут самосознание и в лучшем случае станут просто равноправными членами нашего общества, а в худшем — увидят в нас

угрозу и решат нас поработить или уничтожить. Не слишком ли многого мы ожидаем от машин, не способных, как утверждают некоторые мыслители, к творчеству, пониманию, свободе воли да и просто решению достаточно сложных задач?

Однако предостережения о потенциальных рисках, связанных с ИИ, можно встретить не только в фантастике. И если 15–20 лет назад проблемы безопасности ИИ интересовали лишь небольшое число энтузиастов, то сейчас на эту тему высказываются известные философы, ученые, бизнесмены, проводятся международные конференции, выдаются гранты; она начинает рассматриваться на государственном уровне в разных странах…

Ажиотаж вокруг ИИ также подвергается критике. Специалисты по ИИ давно подчеркивали, что искусственный интеллект — это не мыслящие машины, а, как отмечается, например, в «Толковом словаре по искусственному интеллекту» (1992 г.), «научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования тех видов человеческой деятельности, которые традиционно считаются интеллектуальными». Фактически ученые сами ограничивались разработкой слабого ИИ.

Это верно и для многих современных успехов в ИИ. Хотя некоторые из них и позиционируются как предвестники мыслящих машин, они автоматизируют лишь отдельные виды деятельности, решают отдельные задачи и не вносят особого вклада на пути к мыслящим машинам. Посмотрите на следующее описание.

> Если поставить на поднос чашку с чаем, тяхакоби-нинге начинает покачивать головой, двигать ногами и перемещаться в сторону гостя, которому предназначен напиток. Когда чай берут с подноса, она останавливается; когда же пустую чашку ставят на поднос, она разворачивается и возвращается назад.

Неплохо было бы иметь дома подобного робота, не правда ли? Не сверхинтеллект, но гостей удивить можно… А ведь это даже не робот — это японская механическая кукла XVIII в.

Робот Эрик мог принимать разные позы, сидеть, стоять, двигаться. А кроме того, он разговаривал — голосом мог отвечать на вопросы, которые ему задавали. Правда, он знал ответы лишь на полсотни заранее заготовленных вопросов, но многие роботы и сейчас умеют не более того. Вот только Эрик был создан в 1928 г. Могли ли люди тогда, глядя на Эрика, подумать о том, что еще чуть-чуть — и роботы обретут сознание? Пожалуй, да. И они бы ошиблись.

Конечно, нельзя отрицать, что с помощью компьютеров удается решать все больше задач и что во все большем числе видов деятельности системы ИИ начинают превосходить человека. Однако действительность, стоящая за громкими лозунгами и красочными описаниями, часто гораздо прозаичнее, так что критика ажиотажа вокруг ИИ небезосновательна.

И все же, что бы ни говорили об ИИ как о приземленной области исследований и разработок по автоматизации решения частных задач, эта область началась именно с мечты о создании мыслящих машин. Поэтому не так удивительно, что со стороны ведущих ученых время от времени звучали призывы вернуть области ИИ ее исконные цели[1]. И за последнее десятилетие не только вернулся академический интерес к этим целям, но и коммерческие организации стали серьезно рассматривать перспективу создания сильного ИИ.

Так где же правда и чего стоит ожидать от разработок в области искусственного интеллекта? И неужели между этими двумя крайностями — или настоящий искусственный интеллект

---

[1] McCarthy J. The Future of AI — A Manifesto. AI Magazine. 2005. V. 26. No 4. P. 39; Brachman R. Getting Back to «The Very Idea». AI Magazine. 2005. V. 26. No 4. P. 48–50; Nilsson N. J. Human-Level Artificial Intelligence? Be Serious! AI Magazine. 2005. V. 26. No 4. P. 68–75.

никогда не появится, или он будет сверхразумной личностью, создающей угрозу существованию человечества, — ничего нет?

Источник разногласий о перспективах и путях создания искусственного интеллекта кроется в интеллекте человеческом. Нейрофизиологи и психологи очень хорошо знают, насколько сложен феномен человеческого мышления, и сама идея воспроизвести его искусственно на компьютере без понимания того, как оно работает у человека, многим кажется нелепой. Технические же специалисты часто говорят, что ИИ может быть похож на человеческий интеллект не более, чем самолет похож на птицу. Можно пойти еще дальше и спросить: а нужно ли было знание биомеханики для изобретения колеса?

Но постойте, почему мы тогда вообще можем говорить о том, что создается именно интеллект? И из каких соображений он создается, если он так сильно отличается от естественного аналога? Самолет создавали авиаконструкторы, и орнитологи в спор об искусственных птицах с ними не вступали. Но самолет создавался с конкретной целью. А в чем же цель систем ИИ? Как отмечалось, эти системы должны решать задачи или автоматизировать ту или иную деятельность.

Давайте на минутку просто отбросим словосочетание «искусственный интеллект» и спросим: а будет ли кто-то утверждать, что компьютер принципиально не способен складывать числа, поскольку не обладает самосознанием? Сейчас это может прозвучать смешно, но ведь для человека это интеллектуальная операция, недоступная в полной мере для животных; еще не столь давно лишь немногие люди умели считать. Когда-то для Блеза Паскаля возможность построить арифметическую машину, способную выполнять эти операции автоматически, была основанием, чтобы высказать идеи о возможности механического воспроизведения человеческого мышления в целом. Сейчас же, беря задачу, алгоритм решения которой известен, мы даже не относим ее к юрисдикции искусственного интеллекта.

> Если вместо вопроса: «Должен ли искусственный интеллект быть похож на человеческий?» — мы спросим: «Должен ли компьютер решать задачи теми же методами, что и человек?», то уверенно ответить «нет» будет гораздо проще.

И тогда вопрос, каким образом компьютеру лучше решать такие задачи, будет адресован математикам и программистам, а не нейрофизиологам и психологам, хотя это и не означает, что не стоит вдохновляться решениями, найденными природой.

Казалось бы, мы просто возвращаемся к слабому ИИ. Но заметьте: компьютер решает все более и более сложные задачи лучше человека. Есть задачи, которые для самого человека крайне сложны, — например, из области системной биологии: проблема радикального продления жизни или хотя бы поиска лекарств от некоторых тяжелых болезней. Может ли компьютер в этом превзойти человека? Почему нет? Еще недавно считалось, что уйдут столетия на то, чтобы компьютер превзошел человека в го. А до этого думали, что и в шахматы он никогда с человеком не сравнится. При этом мы не утверждаем, обладает ли такой компьютер самосознанием, пониманием или интеллектом. Он просто решает задачи. Разве этого мало, особенно если это будут жизненно важные задачи, с которыми человек не справляется?

Однако несмотря на реальные успехи ИИ и на то, что многие прогнозы о недоступных для компьютеров возможностях оказались ложными, современные методы ИИ все еще остаются «слабыми». И когда речь заходит о действительно сложных задачах, таких как совершение оригинальных научных открытий или полностью автономное долговременное управление роботом в заранее неизвестной среде, возникают сомнения в том, что технологии слабого ИИ способны их решить, и подспудно появляется мысль, что для этого уж точно нужен «настоящий» разум, с которым тут же ассоциируется самосознание, наделение роботов правами и восстание машин.

Но почему мы решили, что недостаток современных методов именно в их «слабости» и что единственной альтернативой этому является сильный ИИ, возможность создания которого как вызывает сомнения, так и пугает? Если два имеющихся пути не устраивают, то необязательно выбирать из них — можно найти третий.

Мы уже видели, что в прошлом философы не раз ошибались, говоря, что те или иные задачи невозможно решить системами без «сильных» свойств — творчества, понимания, сознания. Возможно, AlphaGo или MuZero обладают «пониманием» игры го и делают «творческие» ходы, но вряд ли в человеческом смысле этих слов. Да и разработчики данных систем вовсе не пытались наделить их творческими способностями и функцией понимания, как и не пытались доказать, что эти способности есть у их детищ. Они просто разработали системы, решающие некоторые задачи лучше человека. Почему это не может оказаться верным и для более сложных задач?

Космические корабли были придуманы вовсе не потому, что самолеты недостаточно похожи на птиц. Но если существующим системам ИИ не хватает не «сильных» свойств, то чего же тогда?

Критику полувековой давности о неспособности решить тысяча первую задачу высказывали в адрес своих творений сами специалисты по ИИ. И это несмотря на то, что в те романтические времена цель создания мыслящих машин ставилась явно и предпринимались попытки разработать системы широкого назначения — такие, как общий решатель задач. С тех пор область ИИ прошла заметный путь и достигла значимых результатов. Однако сейчас эта проблема стала даже более рельефной, чем тогда.

Неоправданные ожидания от создания мыслящих машин привели к тому, что подавляющее большинство работ в области ИИ стало посвящено решению отдельных конкретных задач. Конечно же, это полезно. И конечно, решение

практических задач подспудно приводило к развитию технологий ИИ. Однако каждая конкретная задача наиболее эффективно решается своим частным методом, в идеале — точным алгоритмом, если таковой удается найти. Грубо говоря, компьютер, собирающий кубик Рубика по такому алгоритму, не проявляет ни малейшего интеллекта. Интеллект был проявлен разработчиком, придумавшим этот алгоритм.

Но дело не в том, что компьютер действует по алгоритму, придуманному человеком (и потому якобы не проявляет творчества или интеллекта), а в том, что алгоритм сборки кубика Рубика полезен только для сборки кубика Рубика.

Сейчас есть множество приложений для искусственного интеллекта, но каждое из них требует человеческого труда. И труд этот в основном сводится не к развитию способности компьютера решать задачи, а к изучению предметной области человеком. Так, еще недавно для разработки систем машинного перевода или диалоговых систем требовались целые армии лингвистов, огромный труд которых сосредотачивался на конкретном предмете — языке. Конечно, язык тесно связан с мышлением, и можно было бы предположить, что если таким «ручным» способом наделить компьютер способностью к языку, то это будет большим шагом к мыслящим машинам. Однако оказалось, что с использованием более общих методов машинного обучения, в частности глубокого обучения, небольшие команды разработчиков, не разбирающихся в лингвистике, могут создать системы, которые решают естественно-языковые задачи эффективнее более ранних систем, созданных при участии сотен лингвистов.

Но до сих пор имеющиеся решения не работают «из коробки» для новых задач или просто в новых условиях. Даже достаточно хорошо проработанные детекторы объектов ориентированы на определенный ракурс камеры и при его изменении резко ухудшают качество работы или, например, дают большое число ложных детекций на бликах мокрого асфальта.

Под каждую камеру их приходится дообучать, вручную размечая данные. Специализированность решений проявляется и в том, что, например, системы детектирования и распознавания объектов на изображениях, получения ответов на вопросы по изображениям, генерации описаний изображений или синтеза изображений по описаниям — это все разные системы, и хотя их архитектуры могут иметь отдельные общие компоненты, но обучены они будут по-разному, на разных данных, по разным функциям ошибки. Что уж говорить про системы, работающие на данных другой природы? AlphaGo или MuZero, хоть и способны обучиться играть в разные игры, в отличие от Deep Blue, после обучения под разные игры будут разными системами. А главное, система, обученная играть в го, не сможет без полного переобучения играть не только в шахматы, но и в го на доске другого размера или по слегка измененным правилам (например, в атари-го, где цель — первым захватить хоть один камень).

> Если мы подумаем об этом в контексте вопроса «Чего не хватает существующим методам ИИ?», то становится очевидным, что их основной технический недостаток не в том, что они не являются сильным ИИ, а в том, что они являются узкими.

Узость методов ИИ проявляется не только в том, что метод, разработанный под одну конкретную задачу, не может решать другую, даже родственную, задачу. Она проявляется также и в том, что существующий в ИИ инструментарий плохо пригоден для решения «широких» задач. Например, общая система компьютерного зрения должна была бы быть способной анализировать самые разные изображения (без обучения по тысячам размеченных человеком примеров под каждый конкретный случай). Например, не существует ни одной системы компьютерного зрения или искусственной нейросети — несмотря на их бесчисленное количество, — которая могла бы

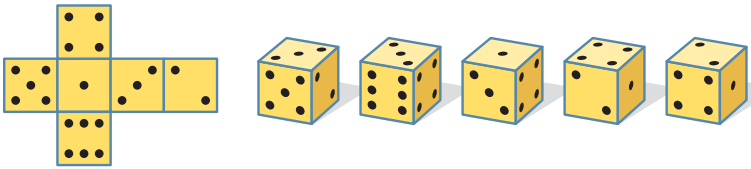

Рис. 2

Какой кубик мог бы получиться после склейки бумажной заготовки?

сходу решить следующую задачу (рис. 2): какой кубик мог бы получиться после склейки бумажной заготовки?

Под конкретный вид кубиков не так сложно натренировать искусственную нейросеть, которая решала бы эту задачу. Но достаточно будет поменять содержание граней кубиков или вместо кубиков взять пирамидки, чтобы решение перестало работать.

Можно придумать неограниченное число новых задач из области зрительного восприятия, которые человеческая зрительная система способна решать на приемлемом уровне без дополнительного обучения. Разумеется, для восприятия и анализа необычных изображений, например рентгенограмм в медицине, обучение потребуется и для человека, а специальная искусственная нейросеть, обученная под очень конкретную задачу, скажем распознавание ранних стадий рака, и ничего больше не умеющая, может это делать лучше человека. Конечно, такие узкие системы нужны, но нужны и системы, способные работать в широких предметных областях.

Разница между узкими и широкими задачами или предметными областями еще яснее видна в робототехнике, где обычно речь идет о степени недетерминированности или неопределенности среды, в которой должен работать робот. Неудивительно, что робототехника долгое время больше всего применялась в промышленности, в которой для роботов можно обеспечить наиболее контролируемые условия. Если роботу

всего лишь нужно сортировать детали из ограниченного перечня на ленте конвейера, то его среда весьма узка. Такого робота не стоит пытаться просить еще и закручивать гайки или даже поднимать детали, упавшие с ленты на пол.

Бытовые роботы стали распространяться гораздо позже, поскольку их среда гораздо неопределеннее. Даже робот-пылесос или робот-газонокосилка, хоть и решают очень конкретные задачи в относительно простых условиях, все же оказываются в заранее неизвестной среде — квартире или доме, которые разработчики никогда раньше не видели. И хотя эти роботы уже вполне полезны, до полной автономности им далеко. Что уж говорить о роботах в еще более разнообразных средах, например о том, чтобы робот мог «хотя бы» сходить в магазин за покупками?..

Итак, проблема с узостью методов ИИ не только в том, что решение новых задач требует определенных усилий от разработчиков, но и в том, что более широкие или сложные задачи оказываются просто недоступными для автоматического решения. Основная масса усилий в современном ИИ направлена на решение узких задач. Здесь достигаются видимые успехи и приносится ощутимая польза. Однако более общие методы и методы решения более сложных задач при этом развиваются мало, хотя польза от них может быть неизмеримо больше. И сами собой общие методы из узких не возникнут — они должны быть устроены по-разному.

Как мы говорили, альтернативой узким методам ИИ является *общий искусственный интеллект*, который не лежит между слабым и сильным ИИ, а просто находится в стороне от них и определяется без отсылки к человеческому интеллекту как *ИИ, способный решать широкий круг задач*.

Общий ИИ представляет собой отдельное направление со своим понятийным аппаратом, подходами, методами, которые лишь частично пересекаются с методами узкого ИИ. Само возникновение этого направления вызвано

неудовлетворенностью исследователей ограниченностью узких методов, их слабой переносимостью на новые задачи, необходимостью вкладывать человеческий интеллект при разработке каждого решения. Все это, а также специализированное внутреннее устройство узких методов, и вызывает впечатление низкой интеллектуальности компьютеров, даже когда они обыгрывают человека в го или Jeopardy! Эти недостатки не устраняются сами собой и требуют самого пристального внимания.

Исследователи AGI дают разные определения интеллекта, которые позволяют выявить специфику данной области; это важно для того, чтобы четче рисовать образ результата. Вот несколько вариантов таких определений:

> «Общий интеллект — это способность достигать сложных целей в сложных средах».
>
> **– Бен Герцель**

> «Интеллект — это способность системы адаптироваться к своей среде, работая при недостаточных знаниях и ресурсах».
>
> **– Пей Ванг**

> «Интеллект измеряет способность агента успешно действовать в широком диапазоне сред».
>
> **– Шейн Легг и Маркус Хуттер**

Эти определения различаются в деталях, но смысл их примерно одинаков. Хотя акцент на «достижении целей в широком диапазоне сред» или «решении широкого спектра задач» может приводить к уклону в сторону конкретных подходов, например обучению с подкреплением (в первом случае) или рассуждениям на основе знаний (во втором), но эти определения могут подразумевать друг друга.

Может быть полезным и явное указание на ограниченность ресурсов (вычислительных, информационных), так как

система, достигающая тех же целей, что и другая, но при использовании меньшего объема исходных данных или тратящая меньше вычислительных ресурсов, должна признаваться более интеллектуальной. И при этом учет ресурсов часто забывают включить в постановку цели как в теории, так и на практике, поэтому лишний раз упомянуть о них не помешает.

Если резюмировать все вышесказанное, общим интеллектом в AGI признается *способность достигать целей в широком диапазоне сред с учетом ограничений* (хотя настаивать на конкретных словах в этом определении не стоит).

Среды могут быть любыми — не только физическими, но и виртуальными, не только пространственно-временными, но и абстрактными. Конечно, может показаться естественным создавать AGI, ориентированный на ту же среду, что и человек в повседневной деятельности. И в этом есть свои плюсы. Можно также утверждать, что существует только одна среда — реальный мир. И это тоже правда. Но реальный мир очень разнообразен. Игра в шахматы и даже любая виртуальная игра являются частями этого мира (а называть их «средами» или нет — вопрос больше терминологический, хотя и имеющий тонкие методологические следствия), и делать акцент на какой-то конкретной его части может быть слишком «узко» и не вполне полезно для глубокого понимания реальности. Например, «наивная физика», которая позволяет нам кататься на качелях или жонглировать, скорее, мешает нам понимать квантовую механику или теорию относительности. Так должны ли мы таким же образом ограничивать искусственный интеллект, если хотим, чтобы он помогал нам решать сложные, в частности научные, проблемы?

Делая акцент на широком диапазоне сред, область AGI позволяет нам избавиться от антропоцентричных предпочтений и предлагает сфокусироваться на общих решениях, пригодных для разных агентов (человека, животных, роботов, ботов и т.д.), действующих в разных условиях — и в микромире,

и на неизведанных планетах, и в виртуальных средах, и в абстрактных (но все же связанных с реальностью) математических пространствах. Если мы зациклимся на конкретном физическом воплощении, есть риск уйти от общего интеллекта в набор специализированных решений, лучше человека приспособленных к некоторому фрагменту реальности, но катастрофически проигрывающих ему даже в тех вопросах, для решения которых человеческий мозг эволюционно явно не предназначался. Разве сможем мы такой ИИ, сколь бы хорошо он ни управлял телом, скажем, андроида, признать разумным? Именно поэтому указание на широкий диапазон сред оказывается столь важным.

Определение интеллекта в области AGI может казаться слишком абстрактным и далеким от наших представлений о естественном интеллекте. А еще широко распространено мнение, что человеческий интеллект — единственный пример интеллекта. Так почему бы не опираться на него? Даже если мы предположим, что для характеристики AGI необходимы или достаточны критерии человеческого интеллекта, описание последнего тоже основано на довольно зыбких понятиях. Вот пример из «Википедии».

> Интеллект — качество психики, состоящее из способности приспосабливаться к новым ситуациям, способности к обучению и запоминанию на основе опыта, пониманию и применению абстрактных концепций и использованию своих знаний для управления окружающей средой. Общая способность к познанию и решению проблем, которая объединяет все познавательные способности: ощущение, восприятие, память, представление, мышление, воображение, а также внимание, волю и рефлексию.

Тут упоминается созвучная AGI «общая способность к … решению проблем», но при этом указывается на то, что эта способность объединяет ряд других способностей, включая,

например, мышление и волю, понимание и обучение и т.д. При этом не очень понятно, как эти способности можно обособить друг от друга. Например, наша память тесно связана с эмоциями, оценкой информации и другими функциями, а воля зависит от множества факторов (начиная от природной чувствительности к вознаграждениям и издержкам и заканчивая умением разбивать волевую деятельность на маленькие шаги и управлять своим вниманием). Должны ли мы разделять эти функции у ИИ? В какой степени такие способности должны быть предустановлены, а в какой мы позволим агенту их «отращивать» по ситуации, для лучшего решения задач в конкретной среде? Определение интеллекта в области AGI не включает эти способности, но и не отвергает их. При этом оно побуждает нас задуматься о том, зачем нужен тот или иной компонент интеллекта и как это зависит от свойств среды и условий задачи. Здесь акцент ставится на том, что мы хотим построить, а не на том, каким именно образом мы хотим это сделать. Цель отделяется от способа ее достижения так же, как хороший заказчик, составляя ТЗ, формулирует именно задачу, не навязывая способ ее решения, на предмет которого заказчик может и ошибаться.

В действительности даже специалисты по AGI ставят перед собой разные цели, именно потому их определения интеллекта несколько различаются. Кому-то предпочтительнее, чтобы система решала задачи в масштабе реального времени, пусть даже и не очень хорошо. Кому-то важнее решение сложных задач, пусть даже на них могут уйти годы. Стоит особо подчеркнуть, что это не столько разное понимание некоего объективного феномена интеллекта, сколько постановка несколько разных целей, и нельзя сказать, что одни цели лучше или хуже других. Однако сила концепции AGI — в ее самоприменимости: в конце концов, общий интеллект должен быть способен достигать разных целей, даже если это цели по созданию общего интеллекта с разным уклоном. Так что именно эта

общность оказывается ключевой особенностью интеллекта, точкой самоприменимости.

Кроме того, некоторые человеческие способности предполагают не только объективную внешнюю оценку, но и субъективное внутреннее ощущение (например, понимание, воображение, самосознание и т.д.). Когда мы говорим о сильном ИИ, мы подразумеваем, что он должен быть наделен всеми человеческими качествами (и обратной стороной медали тут могут быть человеческие слабости, например психические расстройства[2]). Но для общего ИИ это необязательно.

Как повышение уровня решения узких задач вплоть до сверхчеловеческого не потребовало «сильных» качеств, так и расширение общности методов решения задач вовсе не обязательно подразумевает преднамеренное движение в сторону сильного ИИ. Можно предположить, что некоторые аналоги некоторых «сильных» качеств у действительно общего ИИ должны быть. Например, наверняка общий ИИ должен иметь способность к интроспекции — анализу собственных мыслительных процессов или даже оптимизации лежащих в их основе алгоритмов. При этом у такого ИИ будет некий образ себя как часть картины мира. Но это не обязательно означает, что у него будет самосознание в философском смысле и уж тем более личность сродни человеческой, хотя по глубине рефлексии он вполне может и превосходить человека. Наверняка он проявит

---

[2] Есть теории о том, что разные психические расстройства — это следствие чрезмерной адаптации (как генетической, так и через опыт) к очень специфической среде, приводящей к дезадаптации при смене условий. Например, повышенная чувствительность к издержкам, которая в перспективе может привести к депрессии, имеет смысл в среде, где мало возможностей и много рисков. А синдром дефицита внимания, по одной из гипотез, мог быть вполне функциональным состоянием для первобытного охотника, которому было важно быстро переключаться между разными сигналами, чтобы заметить добычу или опасного хищника. А вот с развитием земледелия, требовавшего кропотливой рутинной работы, он стал приводить к дезадаптации.

«понимание» тех областей, в которых действует успешнее человека. Но это не значит, что такое понимание будет сопровождаться у него субъективными переживаниями, схожими с человеческими. Наверняка у него будет многомерная система мотивации, включающая аналоги, например, любопытства и удивления. Но его вовсе не обязательно пытаться наделить всеми человеческими эмоциями. Хотя, скажем, для социальных роботов это может быть полезно, но даже они способны лишь симулировать чувства и эмоции, а не испытывать их.

Итак, идея общего ИИ предполагает, что компьютеры смогут самостоятельно решать как новые узкие, так и сложные задачи, чем будут заметно отличаться от критикуемых систем ИИ, но способ, которым компьютеры будут это делать, может быть далеким от человеческого. А называть ли этот способ интеллектом в некоем обобщенном смысле — вопрос договоренности.

Поскольку фактически на настоящий момент систем AGI не существует, для характеристики систем, разрабатываемых в рамках данной области, вводятся такие понятия, как proto-AGI и Narrow AGI. Под proto-AGI имеются в виду системы, призванные решать широкий круг задач, но все еще не способные делать это эффективно. При этом обычно подразумевается, что эти системы могут быть со временем доработаны до AGI или, по крайней мере, являются шагом на пути к нему. Термином Narrow AGI обозначаются не существующие пока системы, обладающие общим интеллектом, но демонстрирующие (сверх)человеческий уровень в одной предметной области, оставаясь существенно ниже уровня человека во всех других сферах (то есть сходные с людьми-савантами). Предполагается, что такие системы могут быть промежуточным шагом на пути к полноценному AGI.