

НА ПОДСТУПАХ К СВЕРХРАЗУМУ

Сильный искусственный интеллект



* название книги дала нейросеть RuGPT-3



альпина
ПАБЛИШЕР

бизнес

На подступах к сверхразуму

СИЛЬНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

 **альпина**
ПАБЛИШЕР | бизнес

Москва
2021

УДК 004.8
ББК 32.813
С36

Продюсер И. Фурман
Научный редактор А. С. Потапов, доцент,
д-р. техн. наук, SingularityNet
Старший редактор Д. Варламова
Редактор А. Воеводская
Редактор инфографики Г. Неяскин

С36 **Сильный искусственный интеллект : На подступах к сверх-
разуму / Александр Ведяхин [и др.]. — М.: Интеллектуальная
Литература, 2021. — 232 с.**

ISBN 978-5-907394-18-6

Эта книга — первый кросс-дисциплинарный гид по искусственному интеллекту на русском языке. Сильный искусственный интеллект — это следующая ступень в развитии ИИ, не обязательно наделенного самосознанием, но, в отличие от современных нейросетей, способного справляться с широким кругом задач в разных условиях. Авторы книги рассказывают о том, что должен уметь сильный ИИ, какие научные подходы помогут его создать и как изменится мир с его появлением.

**УДК 004.8
ББК 32.813**

Все права защищены. Никакая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами, включая размещение в сети интернет и в корпоративных сетях, а также запись в память ЭВМ, для частного или публичного использования, без письменного разрешения владельца авторских прав. По вопросу организации доступа к электронной библиотеке издательства обращайтесь по адресу tylib@alpina.ru

ISBN 978-5-907394-18-6

© ПАО Сбербанк, 2020

Оглавление

К читателям	5
Коллектив авторов	9
Введение	11
Глава 1. Общая история искусственного интеллекта	35
Глава 2. Как мы узнаем, что создали AGI?	59
Глава 3. Основные направления в AGI	87
Глава 4. Варианты воплощения	159
Послесловие. Будущее AGI	223

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

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.

Таюке 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.

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».

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. Общие — 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

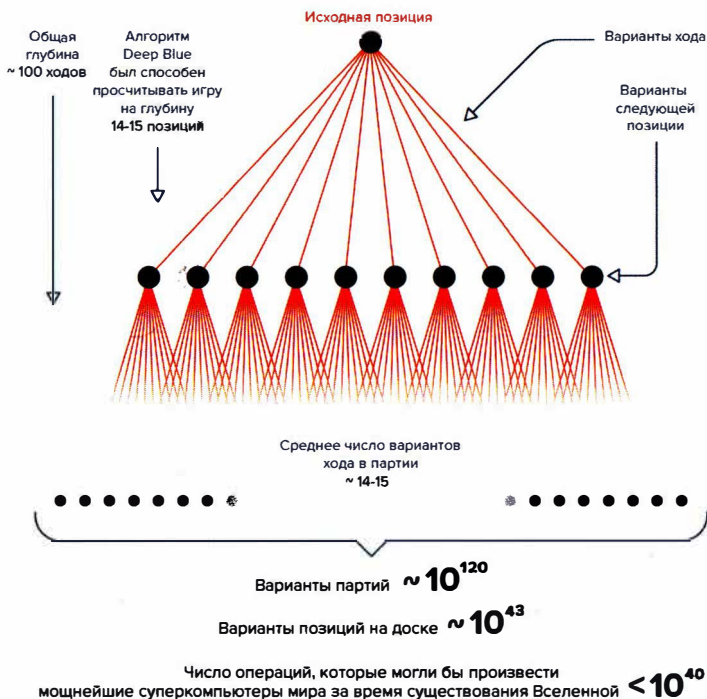


Рис. 1

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 удастся 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 цели¹. 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

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

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 качаться 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 заиклимся 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 рефлекссию.

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 самоприемимости: 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 расстройства²). 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 аналогии, например, любопытства и удивления. Но его вовсе не обязательно пытаться наделять 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.

Глава 1.

ОБЩАЯ ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Перед тем как переходить к возможностям общего ИИ и потенциальным способам его создания, стоит вспомнить, как эволюционировали подходы к искусственному интеллекту в целом и какие уроки мы можем из этого почерпнуть.

В августе 1955 г. четыре знаменитых исследователя, среди которых были создатель теории информации Клод Шэннон и молодой математик Марвин Минский, написали грантовую заявку. Они хотели получить деньги и организовать двухмесячный семинар, посвященный искусственному интеллекту. Они надеялись, что, если лучшие специалисты объединят свои усилия, можно будет существенно продвинуться в том, чтобы научить машину понимать человеческий язык, оперировать абстрактными концепциями, самообучаться и мыслить креативно. Поставленные в заявке задачи остаются актуальными и сейчас, а знаменитый Дартмутский семинар, организованный Шэнноном и Минским, признан колыбелью всей отрасли искусственного интеллекта. В последнее время исследователи ИИ приближаются к тем целям, которые когда-то поставили

основатели отрасли и от которых отрасль впоследствии заметно отошла.

Историю отрасли можно рассказывать как линейную последовательность событий, но это неизбежно будет неточным. В первую очередь потому, что на протяжении всего времени существования искусственного интеллекта его понимание и содержание остаются предметом оживленных дискуссий среди исследователей. Параллельно существуют разные школы мысли и разные подходы к разработке ИИ, в частности потому, что даже естественный интеллект до сих пор не имеет общепринятого определения. Хуже того, многие задачи и технологии, прежде относившиеся к области ИИ, теперь к ней не относятся, например полностью решенные игры типа шашек, методы символьного дифференцирования или роботы с программным управлением. Сама планка сложности задач и методов, относимых к области ИИ, постепенно повышается.

Как это часто случается, представители различных школ, работающие на протяжении последних 60 лет в различных парадигмах, по-своему трактуют области применения, задачи, определения и перспективы ИИ, считая альтернативные подходы неудовлетворительными.

Мы можем назвать задачи, которые решаются современными методами ИИ, но не можем однозначно соотнести текущий уровень развития с рубежом, после достижения которого сможем утверждать, что полноценный ИИ создан. Поэтому до сих пор нельзя объективно сказать, каков уровень развития ИИ и каковы перспективы его совершенствования. Необходимые направления исследований не могут быть отделены от субъективных мировоззренческих позиций исследователей и практического опыта разработчиков. Тем не менее в развитии области, как и в последовательной смене господствующих в ней парадигм и метафор, можно проследить определенную логику.

Подходы и методы

Первой доминирующей парадигмой в ИИ была парадигма *мышление как поиск*. Как и многие последующие подходы, она была построена на представлении об устройстве естественного интеллекта, а точнее, на лабиринтной гипотезе мышления, заимствованной из психологии.

Поставьте неожиданное препятствие на пути муравья, идущего привычным маршрутом, или курицы, бегущей домой через знакомую дырку в ограде. Их поведение станет хаотичным. Они будут метаться по сторонам и тыкаться в любое подходящее отверстие или проход. И это поведение бессмысленно только на первый взгляд. Если проверенный путь закрыт, они будут искать новые выходы из ситуации, перебирая случайные варианты. Курица осуществляет этот поиск в физическом пространстве, а человек или шимпанзе способны осуществлять его в голове, но для каждого из них мыслительная деятельность

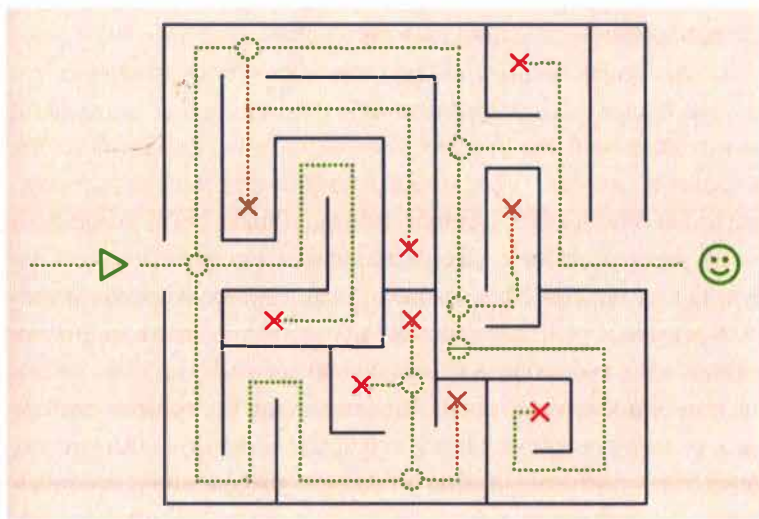


Рис. 3

Мышление как поиск

может возникать как поиск в ответ на незнакомую ситуацию. Ровно так и возникает метафора мышления как поиска: суть интеллекта состоит в решении проблем, а сам процесс решения может быть представлен как поиск пути от исходных данных к ответу в пространстве — «лабиринте» — возможных решений (или как поиск пути от имеющихся средств к конечной цели через достижимые подцели).

Этот подход начали применять для решения интеллектуальных игр и доказательства теорем, где концепция поиска также играла ключевую роль. А для сокращения времени поиска использовали эвристики — простые правила, позволяющие сократить перебор возможных вариантов. Правило может быть, например, такое: «В шахматах следует отдавать предпочтение тем ходам, которые позволяют наращивать преимущество в силе фигур». Или такое: «При игре в крестики-нолики все угловые клетки на первом ходу симметричны и достаточно проанализировать возможность хода только в одну из них». Поэтому важнейшей технологией в рамках этой парадигмы стало эвристическое программирование.

Со временем, однако, обнаружилось важное ограничение подобных систем. Формализованное описание задачи для каждой из них должно было составляться человеком. Кто-то из плоти и крови должен был сформировать лабиринт, внутри которого могла бы блуждать машинная система. Более того, не удалось найти универсального алгоритма, эффективно решающего любую задачу поиска. Хотя какие-то общие методы поиска и удавалось выработать, для каждой задачи нужно было изобретать эвристики и кодировать их в эвристической программе на понятном для машины языке. Каждая новая программа поиска писалась практически с нуля, что представлялось, конечно же, неудовлетворительным. Шагом на пути устранения этого недостатка стало отделение машины поиска от предметных знаний.

Следующая идея, доминировавшая в области ИИ на рубеже 1960–70-х гг., прошла под лозунгом «Знание — сила». Знания

перестали неявно закладываться в программный код. Вместо этого для них стали разрабатываться разнообразные собственные способы представления, обладающие ясной структурой и смыслом. Неудивительно, что большую роль здесь играли логические представления, содержащие знания о предметной области в виде логических выражений, из которых осуществлялся дедуктивный вывод. А в 1972 г. появился язык программирования «Пролог», который сделал разработку систем на основе знаний доступной и для неспециалистов в области ИИ. Но логическими представлениями все не ограничилось. Из компьютерной лингвистики, которая также бурно развивалась в то время, были заимствованы обретшие впоследствии большую популярность семантические сети. Их идея была разработана еще в 1956 г. в рамках проекта по изучению языка в целях машинного перевода, но теперь к ней добавилась способность машины рассуждать на основе представленных в семантических сетях знаний. Возникли и другие представления — фреймы, продукции, сценарии и т.д.

Ключевой технологией здесь стали *экспертные системы*, которые одно время являлись чуть ли не синонимом понятия «искусственный интеллект». Это программы, обладающие экспертными знаниями в какой-то отдельной предметной области и использующие их для решения конкретных задач — вроде системы, которая знает устройство электростанции, или налоговое законодательство, или симптомы заболеваний сердечно-сосудистой системы и пользуется этими знаниями, чтобы предупреждать поломки генераторов, или оптимизировать налоги, или ставить диагнозы.

Экспертные системы оказались способными находить и объяснять решения задач, сформулированных на ограниченном естественном языке для узкой области. При этом поиск решения в таких системах превратился в проблему рассуждения на основе знаний, в котором усматривалась суть мышления. Узким местом при создании таких систем

Правило 1: «Если покрыто шерстью, то млекопитающее».

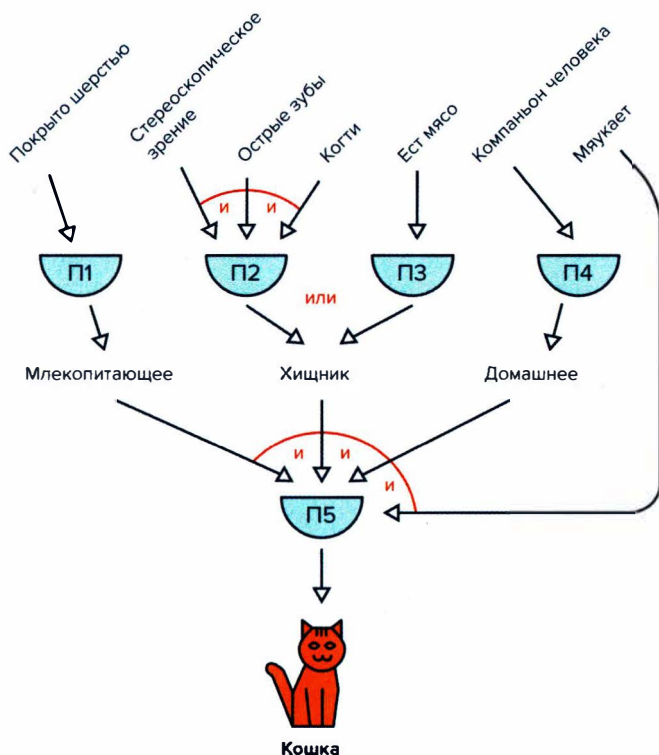


Рис. 4

Как работает «экспертная система»

оказалось извлечение знаний, то есть наполнение базы знаний фактами и правилами о предметной области, на основе которых система уже могла бы, рассуждая, приходить к ответу на ставящиеся перед ней вопросы. Процесс извлечения знаний выполнялся инженерами по знаниям, работа которых заключалась во взаимодействии с предметными экспертами в попытке формализовать их опыт. Но этот опыт нередко был очень обширным и не до конца осознаваемым самими экспертами, так что наполнение баз знаний вручную оказалось

очень трудоемким и ограничивало широту и глубину владения предметом.

Тогда как правила, по которым рассуждают люди, выявить сложно, принятые ими решения напрямую доступны, так что в рамках экспертных систем естественным образом возникла задача автоматического извлечения знаний из наборов примеров, что оказалось не чем иным, как частным случаем машинного обучения³. Хотя область машинного обучения в применении к распознаванию образов существовала с середины прошлого века, она не оказывалась в фокусе внимания многих исследователей ИИ, считавших интеллектуальными только задачи, решение которых было доступно исключительно человеку. Но системы, основанные на знаниях, были слишком зависимы от человека, и в результате в 1980-е годы машинное обучение превратилось в центральную парадигму ИИ, обещавшую эту зависимость ослабить. И хотя технологии машинного обучения для экспертных систем заметно отличались от более традиционных методов, на передний план вышла сама проблематика обучения в целом.

Методы машинного обучения не смогли избавиться от зависимости от человеческого участия. Они требовали рафинированных данных — простых по своей структуре, предобработанных и специальным образом размеченных. Ведь в отличие от человека, способного ориентироваться в контекстах и отсекавать ненужные детали, наделить такой способностью машину непросто, а без этого машина воспринимает всю информацию буквально. В итоге методы машинного обучения зачастую работали по подготовленным человеком данным, а не добывали эти данные сами при взаимодействии с окружающим миром.

³ Класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

Дополнением к описанным методам стал агентный, или мультиагентный, подход, снискавший популярность в конце 1990-х. Агентов программируют так, чтобы они были по возможности автономны, в том числе при получении данных для обучения, и демонстрировали адаптивное поведение в некоторой среде обитания. Среди прочего, они должны достигать своих целей в среде, взаимодействуя с другими агентами, помещенными туда же. Вместе с агентным подходом развивались и другие «автономные» направления, например обучение с подкреплением, аниматы или эпигенетическая робототехника.

Обучение с подкреплением — это такая постановка задач машинного обучения, где у нас нет ни учителя, ни меток, обозначающих правильные ответы. Агент учится через взаимодействие со средой: он может выбрать действие из определенного набора, а среда в ответ на это действие как-то меняется.

Агент воспринимает результат этого изменения, получая при этом также сигнал подкрепления (которое может быть как положительным, то есть вознаграждением, так и отрицательным — наказанием). Задача агента — получить максимальное суммарное вознаграждение в ходе длительных взаимодействий со средой.

К примеру, если агент — это система обучения с подкреплением, управляющая автомобилем в виртуальных гонках, то подкрепление для нее будет основываться на том, на какое расстояние по трассе и за какое время смог проехать автомобиль. Исходно агент будет совершать случайные действия и далеко не уедет. Но, получая вознаграждения за безаварийное движение в нужном направлении с большей скоростью, система будет подбирать все более подходящие действия, постепенно улучшая свою стратегию управления автомобилем.

Аниматы⁴ получили известность после конференции «Симуляция адаптивного поведения: от животных к аниматам», которая прошла в Париже в 1990 г. Это автономные искусственные организмы, которые способны существовать в меняющейся среде и решать в ней какие-то базовые задачи так же, как любые живые организмы. Важно, что их создатели пытаются решить не бионические задачи (например, имитировать работу мышц), а задачу воспроизведения базовых принципов поведения животных. Допустим, у нас есть анимат, имитирующий одноклеточный организм. Он перемещается по ровной поверхности в поисках пищи. Но пища на ней возникает непредсказуемым образом. Ее то много, то мало, то какое-то время вообще нет. Она может лежать спокойно или уклоняться от анимата. Обучение с подкреплением часто рассматривается в применении к конкретным практическим задачам, где требуется один раз выработать стратегию поведения для определенных условий. В таких случаях возникает большой соблазн сделать упрощенный, специализированный метод. При исследовании же аниматов предпочтение отдается адаптации к неизвестным и потенциально меняющимся условиям среды.

Эпигенетическая робототехника идет еще дальше в стремлении создать агента, который имитировал бы уже не животное, а, скорее, ребенка. Дети часто не способны хорошо выполнять даже простые задания, но зато умеют сами добывать информацию о мире, взаимодействовать с окружающей средой, обучаться и идти к своей цели методом проб и ошибок. Такой агент должен проходить несколько стадий в своем развитии, взаимодействуя со средой, при этом обучение должно быть накопительным и сложность задач должна прогрессивно возрастать.

В начале 2000-х в том числе в рамках агентного подхода возникла потребность интегрировать разные подобласти ИИ

⁴ Термин «анимат» происходит от сочетания слов *animal* + *robot*.

автоматического машинного обучения (AutoML) и метаобучения⁵.

При всем разнообразии подходов и методов, использовавшихся в последние 60 лет параллельно, вместе и по отдельности, у них есть общий лейтмотив. Каждая последующая парадигма ИИ была призвана снизить зависимость интеллектуальных систем от человека.

Общий искусственный интеллект естественным образом продолжает эту тенденцию, но подходит к проблеме более системно.

Это относится и к некоторым другим вопросам. В частности, традиционные для всей области ИИ, но развивавшиеся во многом независимо символьный и коннекционистский подходы переосмысляются в контексте общего ИИ.

Символьный подход берет свое начало в философии, логике и математике и оперирует логическими правилами, знаковыми и символьными системами, которые интерпретируются в терминах высокоуровневых мыслительных процессов человека. Суть в том, чтобы воспроизвести сознательное мышление человека, абстрагируясь от нейрофизиологических деталей его реализации в мозгу (условно, вместо самолета, хоть сколько-то напоминающего птицу, построить вертолет или ракету).

Коннекционистский, или субсимвольный, подход предполагает создание моделей, воспроизводящих не только

⁵ Метаобучение (англ. meta-learning) — это «обучение обучению», то есть улучшение существующих или поиск новых, более эффективных методов обучения в процессе самого обучения. Данный термин применяется как в когнитивной психологии, так и в машинном обучении. В последнем случае метаобучение, как правило, выполняется на наборе задач. На их примере нужно построить такой алгоритм, который бы умел эффективно обучаться решению не только этих, но и новых сходных задач.

информационную сторону мышления, но и способ его «аппаратной» реализации в мозгу. Самый яркий пример — нейросети.

На протяжении 65 лет оба подхода динамически развивались, иногда двигаясь в разные стороны, а иногда друг к другу. В XXI в. они чаще существуют вместе, чем по отдельности. Четкая бинарная логика сейчас характеризуется как «старый добрый ИИ» (good old-fashioned AI, GOF AI) и занимает сравнительно небольшую нишу, давно уступив лидерство вероятностным методам. В свою очередь, и биологически правдоподобные искусственные нейронные сети привлекают гораздо меньше внимания, чем модели, создатели которых не заботились о биологическом правдоподобии и которые в большей степени опираются на математические методы. Но это разделение все еще остается удобным, потому что оно вызвано в том числе характером исходных данных и способом их представления — дискретным или непрерывным⁶, — а также вытекающими из них методами вывода и оптимизации.

Удачи, неудачи и рождение общего ИИ

Разнообразные подходы к разработке ИИ, описанные нами выше, существовали параллельно. А скептически настроенные наблюдатели даже утверждают порой, что принципиально новых подходов не появлялось после 1960-х гг. и менялись лишь

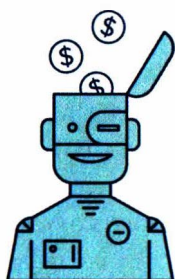
⁶ Если некоторая величина, несущая информацию, в пределах заданного интервала может принимать любое числовое значение (то есть между разными вариантами значений нет разрывов), то она называется непрерывной, а если она способна принимать только ограниченное (или, строго говоря, не более чем счетное) число значений, она называется дискретной. Образно можно представить непрерывную величину как градиент, а дискретную — как радугу, состоящую из семи цветов без полутонов.

относительная успешность и популярность подходов и локальные темпы их развития. Но нельзя отрицать, что за последние 65 лет исследователи достигли большого прогресса в решении самых разных, пусть и частных задач, а область ИИ из предмета академического интереса превратилась в огромную отрасль. За это время системы ИИ научились обыгрывать людей в шашки, шахматы, го и множество других игр. Освоили перевод текстов на достаточном уровне, чтобы приносить пользу людям. Стали широко использоваться в разных отраслях экономики.

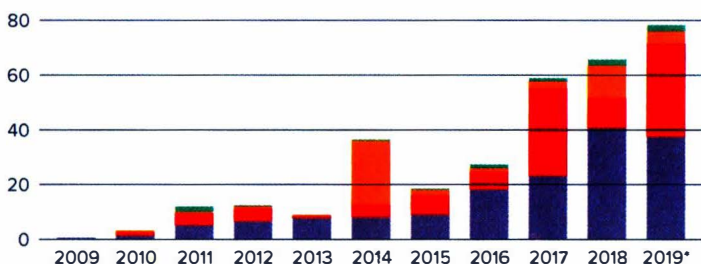
Одним из пионеров систем ИИ стал, например, «общий решатель задач» — система, созданная в 1959 г. Гербертом Саймоном, Клиффордом Шоу и Алленом Ньюэллом. Общий решатель был построен на чисто символьной логике и умел решать ряд задач, включая некоторые математические и шахматные задачи и головоломки, такие как, например, «Ханойская башня». В 1960–70-х гг. были и другие попытки создать системы ИИ общего назначения, но эта цель начала выглядеть слабодостижимой, хотя в решении более частных задач наблюдался определенный прогресс. Исследователи и инвесторы чувствовали разочарование, которое в середине 1970-х вылилось в снижение интереса и падение финансирования. Этот период называли «первой зимой искусственного интеллекта»⁷.

Следующий виток интереса к разработке могущественных систем ИИ возник в 1980-е. Тогда, в частности, огромный проект по разработке искусственного интеллекта, «компьютер пятого поколения», финансировался в Японии. Его целью было создание «эпохального компьютера» с производительностью суперкомпьютера, который бы переводил устную речь в текст, распознавал образы и решал другие сложные для искусственного интеллекта задачи. Проект потерпел неудачу (за десять лет на разработки было истрачено более 50 млрд иен, при этом

⁷ В дальнейшем спад интереса к ИИ повторился в конце 1980-х — начале 1990-х гг. после падения популярности экспертных систем.



- ◆ — миноритарный пакет
- ◆ — публичное размещение
- ◆ — слияния и поглощения
- ◆ — частные инвестиции в стартапы



* Данные за 2019 г. приведены за период с января по октябрь.

Источник: AI Index 2019

Рис. 5

Инвестиции в отрасль искусственного интеллекта 2009–2019 гг.,
млрд долларов США в год

поставленные цели не были достигнуты), и для многих это означало очередное доказательство невозможности построения общего ИИ в ближайшем будущем. Но, как замечает современный энтузиаст общего искусственного интеллекта Бен Герцель, проект не опирался на ясные концептуальные представления о том, как должно быть реализовано мышление в компьютерах. Так что проблема была, скорее, в подходе, а не в нереалистичности самой задачи.

Тогда же, в 1980-е, калифорнийский философ Джон Серл придумал ныне знаменитый мысленный эксперимент под названием «Китайская комната». Представим себе изолированную комнату, в которой находится сам философ, который не знает китайского языка и не понимает ни одного китайского

иероглифа. Но у него есть книга с точными инструкциями о том, как реагировать на входящие сообщения на китайском (например, «Возьмите такой-то иероглиф из корзинки номер один и поместите его рядом с таким-то иероглифом из корзинки номер два»). Что все эти иероглифы означают, не сообщается, так что философ просто бездумно следует инструкции.

Наблюдатель, знающий китайский, через щель передает в комнату вопросы, записанные иероглифами. Может ли он получить осмысленный ответ? Ньюэлл и Саймон на основе своего успешного опыта с общим решателем задач выдвинули в 1976 г. гипотезу физической символьной системы, согласно которой манипуляции с символами являются необходимым и достаточным средством реализации основных интеллектуальных операций физической системой. Фактически «Китайская комната» построена как физическая символьная система. Серл говорил: предположим, что гипотеза Ньюэлла — Саймона верна, тогда комната может быть организована так, что будет обладать интеллектом, а значит, она сможет, например, пройти тест Тьюринга, то есть будет способна вести осмысленный диалог. Но посмотрите — человек в комнате совершенно не понимает китайского. Как он может вести осмысленный диалог?

Философ видел в этом противоречие. Он полагал, что раз человек не понимает вопросов — а понимание необходимо для поддержания осмысленного разговора, — то никакие тонкие умения манипулировать символами не приведут машину к настоящей разумности, то есть гипотеза физической символьной системы неверна. Стоит отметить, что данная цепочка рассуждений содержит изъяны. В частности, физической символьной системой здесь является не человек, а комната. Мы можем уверенно утверждать, что вопросы не понимает человек. Но вот не понимает ли их вся комната целиком? Представим себе, что переключать таблички будет не один человек (ведь для формирования оперативного ответа могут потребоваться миллиарды символьных операций), а большое

множество людей, которые будут передавать таблички соседям в соответствии с собственными инструкциями. Чем это будет принципиально отличаться от мозга человека, в котором миллиарды нейронов передают друг другу сигналы, при этом каждый из нейронов в отдельности не понимает вопросов, но вся система в целом — понимает? Разница в том, как эти системы сформировались. Построить «китайскую комнату» вручную крайне трудозатратно, и даже если это сделать, сможет ли она дальше взаимодействовать с миром и обучаться сама? Вряд ли. Но именно поэтому начиная с 1980-х гг. на передний план вышли вопросы обучения и воплощенности.

Стоит особенно подчеркнуть, что Серл критиковал возможность реализации сильного ИИ на базе компьютеров в целом, а не просто символьный подход в ИИ, поскольку именно компьютер и является той физической символической системой, о которой говорили Ньюэлл и Саймон. Будь то коннекционистская или обучаемая система, будь то система, наделенная сенсорикой и моторикой, но если она реализуется на компьютере, то является физической символической системой и подпадает под критику Серла. С этой критикой многие исследователи не согласны, но даже отвергая ее, доказать, что компьютер способен на «понимание» (в человеческом смысле) или обладает другими «сильными» свойствами, что компьютер может мыслить по образу и подобию человека, — весьма проблематично.

Но давайте вспомним, что участники Дартмутского семинара не связывали искусственный интеллект с нашим пониманием интеллекта человека. Они допускали, что ИИ будет использовать методы, которыми не пользуется человек, а интеллектом считали вычислительную составляющую способности достигать целей в мире. В течение 50 лет после семинара этот посыл часто игнорировали, а потом и вовсе стали все чаще пользоваться понятиями «сильного» и «слабого» ИИ:

Но возьмем шахматный компьютер Deep Blue, выигравший у сильнейшего шахматиста-человека. Вряд ли кто-то

будет настаивать, что Deep Blue — «настоящий ИИ». Но чего ему не хватает? Самосознания? Свободы воли? Если Deep Blue будет себя осознавать и сможет отказаться от очередной партии в шахматы, поскольку ему стало скучно, назовем ли мы его по-настоящему разумным? В рамках концепции «сильного ИИ» мы должны были бы ответить «да» на эти вопросы, но это нам бы не помогло.

В 1997 г., как раз тогда, когда Deep Blue обыграл-таки чемпиона мира в шахматы, появился термин «общий искусственный интеллект». В рамках этой идеи предлагалось делить ИИ не на слабый и сильный, а на узкий, способный решать определенные задачи, и общий, действующий в разных обстоятельствах и адаптирующийся к разным проблемам.

В 2007 г. прошла первая конференция, посвященная исключительно проблемам развития общего ИИ, и с этого времени можно наблюдать возрождение интереса к масштабной проблеме искусственного интеллекта, которой посвящена эта книга. Исследователи новой волны готовы принимать разные подходы к возможному решению проблемы, и как минимум часть из них признают, что компьютер не обязан воспроизводить человеческое мышление во всех его проявлениях, чтобы считаться интеллектуальным.

Состояние дел в области разработки искусственного интеллекта

Как отмечалось ранее, технологии ИИ постоянно развивались в направлении снижения их зависимости от человека при создании решений для новых задач и сейчас находят очень

широкое применение. Тем не менее текущее состояние дел в области, далекое от решения большой задачи, можно охарактеризовать следующим образом.

Во-первых, с точки зрения адаптивности все существующие системы ИИ на основе любых известных подходов предполагают функционирование в ограниченном наборе задач и условий и не способны самообучаться функционированию в условиях существенно новых. То есть они являются программируемыми, хотя сложность их программирования заметно снизилась, а способность к обучению возросла.

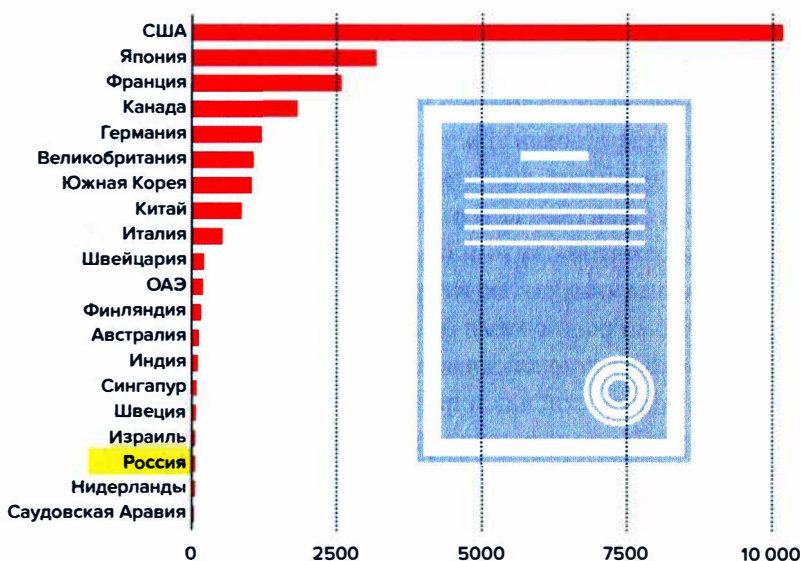
Во-вторых, с точки зрения автономности систем ИИ все существующие системы не являются автономными и не могут полноценно функционировать без живого оператора, отвечающего за запуск и остановку, техобслуживание, целеполагание и определение режимов работы в зависимости от тех или иных условий или задач. То есть они остаются управляемыми. Особенно это очевидно в областях, характеризующихся высокими рисками или высокой априорной неопределенностью (даже робот-пылесос без помощи человека не проживет и пары дней в нормальной квартире).

В-третьих, с точки зрения интегративности современные системы ИИ являются не системами, обладающими интеллектом как таковым (даже ограниченным), а системами компьютерного зрения, обработки естественного языка, анализа данных (машинного обучения), обработки символьной информации (рассуждений на основе знаний) и т.д., то есть интегративными не являются.

В этом смысле нерешенные пока в области ИИ задачи создания автономного и широко адаптивного поведения, реализуемого универсальными самообучающимися агентами, обладающими достаточно эффективными для этого интегративными архитектурами, могут быть отнесены к проблематике общего ИИ (AGI).

Помимо названных концептуальных проблем, есть и частные. Решение многих более узких, но практически важных задач остается недоступным для существующих методов искусственного интеллекта. Так, например, семантический анализ текста на естественном языке и семантический анализ изображений — задачи, которые будут затронуты в разделе, посвященном метрикам, — находятся на таком уровне развития, что даже в рамках узкого ИИ требуется большой объем исследований для продвижения вперед. Для решения этих задач необходимо развитие как символьных и субсимвольных техник ИИ по отдельности, так и подходов, сочетающих в себе сильные стороны каждого из направлений, что требует не только прикладных, но и фундаментальных исследований.

В России развита сильная инженерная база по адаптации, комбинированию, применению существующих технологий машинного обучения, в том числе основанных на искусственных нейронных сетях. Есть сильные научные математические школы, направления исследований которых относятся к основаниям методов ИИ и непосредственно могут быть использованы для анализа существующих методов машинного обучения и разработки принципиально новых, перспективных подходов. В то же время у нас в стране, к сожалению, нарушена коммуникация между математическим и инженерным сообществом в области искусственного интеллекта, и поэтому усилиям специалистов не хватает системности. Обычно это либо инженерные разработки с использованием модификаций существующих, преимущественно зарубежных, технологий, либо фундаментальные теоретические работы, которые не всегда доведены до воплощения. Полный цикл работ — от формализации задач, разработки соответствующих математических моделей, фундаментальных исследований их свойств до изучения аспектов алгоритмической сложности и инженерной реализации решений задач — встречается в российской практике довольно редко, что очень печально: аналогичная ситуация в свое время привела



Источник: AI Index 2019

Рис. 6

20 стран, где было выдано больше всего патентов
в области искусственного интеллекта (всего штук за 2015–2018 гг.)

к критическому отставанию в области электроники. Очень не хотелось бы, чтобы то же самое случилось и с ИИ.

Сложность еще и в том, что яркие недавние примеры практического использования нейронных сетей и быстрый эффект от их внедрения создают у дилетантов впечатление, что мы уже без пяти минут как создали всесильный искусственный интеллект. Достаточно привести заголовки новостей: «Специалисты “Студии Лебедева” продемонстрировали нейросеть, которая генерирует логотипы для экспресс-дизайна»; «В России создана нейросеть, которая пишет неотличимые от настоящих комментарии в интернете»; «Искусственный интеллект помогает регионам России бороться с коронавирусом» и т.д.

Поэтому значимость фундаментальных исследований недооценивается. Компании, в свою очередь, ориентированы

на продуктовую разработку с коротким циклом, чтобы быстро получить сверхэффект и заработать денег. Современные образовательные программы готовят инженеров, умеющих работать с существующими технологиями, а не творчески мыслящих исследователей, способных совершать новые прорывы.

В то же время перед нами стоит и широкий круг практических задач, решение которых до сих пор непосильно для существующих методов ИИ. По мере развития моделей машинного обучения становятся очевидными ограничения и уязвимости существующих подходов. Они внезапно проявляются в виде побочных эффектов именно на этапе экспериментов, которые стали основополагающей мерой качества моделей ИИ. Не отрицая практическую значимость эксперимента и успешные результаты, полученные методом проб и ошибок, важно отметить, что перед разработчиками ИИ стоят серьезные вызовы, которые требуют создания новых подходов и технологий. В этой связи критическим для продвижения в данной сфере является формирование условий для воспроизводства специалистов, которые владеют основательной базой математики и компьютерных наук для создания перспективных технологий искусственного интеллекта.

Сверхинтеллект

В массовом воображении разработки ИИ связаны с надеждами или опасениями, что общий ИИ будет создан. Некоторые исследователи и разработчики действительно полагают, что все технологии, необходимые для создания общего ИИ, уже в принципе разработаны и осталось только соединить их некоторым правильным образом для получения нужной синергии. Многие эксперты настроены более скептически, полагая, что

необходимо принципиальное решение проблем, которые обсуждаются в этой книге.

Очевидно, разнятся и ожидаемые сроки наступления этого события, причем настолько, что одни исключают саму его возможность, а другие допускают, что оно произойдет в течение трех–четырех лет. И те и другие сходятся в том, что создание технологии «ИИ уровня человека» (HLAI — human-level AI) достаточно быстро приведет к появлению систем, существенно превосходящих человеческие возможности, то есть «ИИ сверхчеловеческого уровня» (superhuman AI).

Возникновение сверхинтеллекта выглядит закономерным в перспективе эволюции, как закономерное возникновение молекул из атомов, клеток из молекул, организмов из клеток, выделение специализированных клеток в центральную нервную систему, возникновение социальных структур, развитие речи, письменности и информационных технологий. Похоже, что оно неизбежно, если прежде наша цивилизация не погибнет.

Другое дело, что искусственный интеллект — не единственно возможный путь к сверхразуму. В научной фантастике с разной степенью детальности и правдоподобности описано множество воображаемых вариантов сверхразума, но редко ставится вопрос, какой из них возникнет раньше и продолжит развиваться дальше, все больше обгоняя возможные альтернативы. Замечательный фантаст Питер Уоттс в романах «Ложная слепота» и «Эхопраксия» описывает цивилизацию постлюдей, которые объединяются в коллективные сверхразумы или расширяют свои интеллектуальные способности с помощью нейроимплантатов, чтобы успешно конкурировать с ИИ.

Правда, оцифрованное сознание, вероятно, будет обладать катастрофически меньшей вычислительной эффективностью, чем прямая реализация ИИ, не требующая эмуляции биологических процессов, от деталей которых уж точно можно абстрагироваться в ИИ, но без которых оцифрованное сознание



Рис. 7

Будущее человека и искусственного интеллекта: возможные сценарии

может оказаться сильно нарушенным. Так что при прочих равных ИИ будет на несколько порядков превосходить загруженные сознания по быстродействию. А еще в ИИ мы можем сделать качественные улучшения в плане интроспекции, контроля памяти и т.д., недоступные человеку и, как следствие, его оцифрованному варианту. Еще один гипотетический вариант — создание гибридных человеко-машинных систем, сочетающих преимущества и человеческого, и искусственного интеллекта.

Пока что мы можем только мечтать (или бояться) и строить умозрительные гипотезы, ведь на этом этапе развития ИИ

ГЛАВА 2.

КАК МЫ УЗНАЕМ, ЧТО СОЗДАЛИ AGI?

В романе Станислава Лема «Солярис» у ученых ушли годы на то, чтобы предположить, что Океан, представляющий собой студенистую субстанцию, покрывающую поверхность Соляриса, целенаправленно изменяет орбиту планеты. А догадки о его высокоразвитом разуме были сделаны исходя из сложнейших рисунков, создаваемых им на своей поверхности, еще до установления контакта с ним. Казалось бы, в случае с искусственным интеллектом, который мы создаем сами, все должно быть намного проще, но это не совсем так.

Действительно, как мы узнаем, что очередную попытку стоит признать успешной и перед нами вправду мыслящая машина, а не просто хитрый алгоритм со сложным, но не разумным по-настоящему поведением? Этот вопрос возникает в разных контекстах. К нему нередко приводит обсуждающаяся в философии проблема того, может ли в принципе машина мыслить. Звучащие порой громкие заявления отдельных особо амбициозных исследователей или корпораций о том, что настоящий ИИ наконец создан, тоже ставят вопрос о том, как эти заявления надежно проверить. Критерии и тесты интеллектуальности являются естественным продолжением попытки

строго определить понятие интеллекта. Данный вопрос позволяет четче понять, какую цель мы в конечном итоге преследуем, к чему мы стремимся. Хотя по данному вопросу не так много можно сказать без содержательного развития самой области общего искусственного интеллекта, он является методологически важным, и разные мыслители им давно задавались. И после того, как ИИ начали воспринимать как достижимую цель, а не мысленный эксперимент, вариантов ответа появилось великое множество.

Зачастую эти ответы обретают форму какой-то сверхсложной задачи, вызова, который компьютеру нужно преодолеть, чтобы тогда мы уж точно были уверены, что перед нами мыслящая машина. Полное решение подобных задач очень далеко, а частичное решение может не представлять интереса с точки зрения не только AGI, но и ИИ. Именно поэтому возникает отдельный вопрос об оценке промежуточных результатов: как мы сможем понять, что разрабатываемая нами система стала хоть немного ближе к общему искусственному интеллекту или что другая система не продвинулась на этом пути дальше? Ответ на этот вопрос важен не только для исследователей, но и для инвесторов, вкладывающих деньги в развитие ИИ. Для них это не просто умозрительный вопрос, отвечая на который можно ошибиться либо в будущем просто поменять свою точку зрения. Значимость цели построения общего ИИ не вызывает сомнения. Но как понять, насколько она далека, да и достижима ли вообще? Приближаемся ли мы к ней или топчемся на месте? Продолжать ли вкладывать деньги и в какой проект? А может, и вовсе не начинать?..

К сожалению, KPI⁸ в классическом смысле слова плохо переносятся в сферу разработки общего искусственного

⁸ Key Performance Indicators, KPI — числовые показатели деятельности подразделения (предприятия), которые помогают организации в достижении целей или обеспечении оптимальности процесса.

интеллекта. Почему — мы расскажем позже. А пока давайте пройдемся по самым популярным тестам, оценивающим уровень искусственного интеллекта.

А что мы оцениваем?

Оценивая искусственный интеллект, мы можем подходить к делу с двух сторон. Первый вариант — когда мы ориентируемся на функциональность: насколько хорошо агент решает ту или иную задачу. Это подходящий способ оценки «узкого» ИИ. Второй вариант — когда мы ориентируемся на способности. Способность означает, что агент может справляться с определенным спектром задач без предварительной подготовки к конкретному заданию (например, способность решать логические задачи или ориентироваться в пространстве). В идеале AGI должен быть именно таким. Между умением решать конкретную задачу и интеллектуальными способностями в широком смысле слова, вероятно, должен быть какой-то спектр промежуточных состояний, если одно в принципе способно перетекать в другое.

Оценивать то, как ИИ справляется с конкретной задачей, мы можем либо в режиме «белого ящика», когда мы понимаем, как устроен процесс, либо в режиме «черного ящика» — когда оцениваем только степень успешности и не видим процесс решения, хотя при желании можем проанализировать, как ошибается система, и даже поставить задачу так, чтобы система, в том числе и нейросеть, не просто выдавала, например, «истина» или «ложь» на формулировку теоремы, а генерировала текст доказательства на выходе. При тестировании методом «черного ящика» мы можем сделать следующее.

(собеседник может перебивать, переспрашивать, задавать дополнительные вопросы, шутить и т.д.), и его поддержание требует как знания правил языка и общей эрудиции для релевантных реакций на реплики собеседника-человека, так и понимания определенных социальных контекстов. И у судьи есть достаточно времени, чтобы оценить собеседника. Так что машина, претендующая на победу, должна была бы обладать широкой картиной мира, использовать естественный язык, рассуждать, обучаться и понимать контекст.

Тест Тьюринга очень антропоцентричен, то есть соответствует исключительно человеческим представлениям о разумной коммуникации.

Четких критериев оценки у него нет, и непонятно, как оценивать прогресс или сравнивать разговорный уровень двух машин. Кроме того, на практике в тестах редко учитывается возможность принять человека за машину, хотя такой вариант вполне реален.

В результате программы, претендующие на прохождение теста, не вносят заметного вклада в развитие AGI или искусственного интеллекта в целом. Анализ решений показывает, что, хотя уровень программ растет из года в год, качественно они не приближаются к поставленной цели. Оказывается, можно в ручном режиме произвести достаточно сложное дерево диалогов, основанное на закодированных шаблонах и правилах, и создать у тестирующего иллюзию взаимодействия с интеллектом человеческого уровня. К тому же есть возможности для «читерства»: программа может притворяться ребенком или, например, человеком, который нетвердо владеет рабочим языком испытания. Она может отказываться отвечать на те или иные вопросы, потому что ей этого «не хочется». В таком случае ей не нужны ни картина мира, ни рассуждения, ни даже просто понимание содержания диалога.

Ограничения теста стали особенно очевидны после того, как одну из его версий прошла в 2014 г. программа Eugene Goostman, написанная русскими и украинскими программистами. Создатели потратили десятки человеко-лет разработки на создание «дерева диалогов» и соответствующей системы исполнения, чтобы симитировать украинского мальчика, плохо говорящего на английском языке. Залогом успеха Евгения Густмана было не красноречие, а биография: судьи готовы были легко простить программе пробелы в грамматике и недостаток знаний. Так что на соревнованиях 2014 г. каждый третий судья признал программу человеком. Но легко догадаться, что способность обучаться и адаптироваться к условиям обучения в этой программе не проверялась, не реализовывалась и не могла быть реализована.

Тут важно подчеркнуть, что это не снижает значимость теста Тьюринга для развития ИИ. Он оказал большое влияние на философию и понимание машинного интеллекта, хотя и не очень подходит для его практического измерения, да и задумывался не для этого.

Осознавая эти ограничения, исследователи предложили взамен исходного теста множество альтернативных вариантов, которые можно рассматривать как метрики достижения общего интеллекта. В частности, появился ряд испытаний, тоже опирающихся на естественный язык.

Исходную идею разговорного интеллекта продолжает Baby Turing Test, описанный в 1985 г. Этот подход предполагает обязательное обучение в процессе прохождения испытания, когда на начальном этапе мы можем верифицировать отсутствие у системы необходимых знаний, а в конце пути предъявляем ей классический тест Тьюринга в полном объеме. По мере прохождения теста у нас есть возможность проверить, насколько хорошо система следует заданной кривой обучения (learning curve).

Очевидно, что эта идея может быть адаптирована к любому набору сред и критериев тестирования в зависимости

от того, какой тип интеллекта мы хотим оценивать — способность к классификации, прогнозированию или поддержанию диалога на естественном языке.

Также существует тест Winograd Schema Challenge (WSC)¹⁰, ориентированный одновременно и на понимание естественного языка, и на наличие здравого смысла. В оригинальной версии он состоит из 150 утверждений, за которыми следует вопрос и несколько вариантов ответов. Вот пример тестового задания.

Трофей не поместился бы в коричневый чемодан, потому что он был слишком большим. Что было слишком большим?

Ответ 0: трофей.

Ответ 1: чемодан.

Для человека очевидно: для того чтобы трофей не влез в чемодан, он должен оказаться больше чемодана. Эту информацию нельзя получить из самого задания: она предполагает наличие у тестируемого житейского здравого смысла. При этом когда какая-нибудь «Алиса» на «Приятно было с тобой пообщаться» отвечает: «Доброе слово и боту приятно», она реагирует на ключевые слова в реплике пользователя, и это намного менее интеллектуальная операция, хоть и звучит весьма живо. Кроме того, тесты вроде WSC хороши тем, что за ответы программа получает баллы и, соответственно, тут есть численная метрика, по которой легко сравнивать разные системы и оценивать уровень прогресса¹¹.

Winograd Schema Challenge входит в более широкий тест GLUE (General Language Understanding Evaluation),

¹⁰ <http://commonsensereasoning.org/winograd.html>

¹¹ В том случае, если тест содержит подвыборки с инкрементным (то есть, растущим на определенную постоянную или переменную величину) возростанием сложности.

этих общих компонентов. И пусть лучшие решения, участвовавшие в соревновании, прошли тест на четверку, но если бы эти системы тестировались на таких типах заданий, которых не было на этапе разработки и обучения, вряд ли хоть одна из них получила бы положительную оценку.

На текущем этапе развития технологий достаточно сложно создать такой языковой тест, чтобы выигрыш от внесения в решения общих элементов «понимания» и «мышления» превосходил выигрыш от более простых методов, но при этом тест не был бы катастрофически сложным.

Качественный учет и накопление поверхностных корреляций в языке, дополнительная ручная разметка данных, изобретение эвристик под конкретные тесты — все это пока что дает более легко достижимые улучшения в результатах бенчмарков. Чтобы устранить этот недостаток, исследователи ищут другие пути, предлагая, например, тесты¹², в которых система не может непосредственно извлечь ответ на вопрос из существующих текстов, а должна сперва выполнить некоторые манипуляции со словами и символами.

Другой вариант имитационной игры — когда программа притворяется человеком в многопользовательских видеоиграх. Пример такого состязания — BotPrize. Несколько лет тест проходил в форме игры Unreal Tournament 2004, персонажами которой управляли реальные люди или компьютерные программы. Живые игроки одновременно выступали в роли судей, что несколько затрудняло задачу: не очень просто сосредоточиться на оценке соперников, когда ты сам рискуешь быть убитым в игре. Поэтому со временем в состязание стали добавлять оценку от внешних наблюдателей.

¹² <https://medium.com/@runarphius/a-test-for-true-natural-language-understanding-e9879241c07d>

Подобные тесты могут быть полезны, чтобы стимулировать те направления исследований, которые будут способствовать решению задач AGI. Но они все-таки проверяют достаточно ограниченный диапазон навыков. Поэтому многие тесты на достижение уровня AGI выходят за рамки таких заданий.

Тесты на оценку способностей

Если мы хотим оценить способности машины, кажется логичным использовать подходы, которые мы применяем для оценки интеллекта живых существ. Ряд исследователей предлагают оценивать прогресс по достижению уровня интеллекта различных животных. Ведь с точки зрения решения определенного набора задач в непредсказуемой среде животные вполне себе эффективны. Этой идеей, например, вдохновлен тест Animal-AI Olympics¹³. Вот пример классического лабораторного эксперимента, который воспроизводится в «цифре»¹⁴.

Перед животным ставятся две перевернутые миски — под одну из них кладется еда, которую животному надо достать. Сначала еда каждый раз кладется под одну и ту же миску, пусть она будет миской 1. Это этап обучения. Затем человек кладет еду под миску 1, вынимает и очень заметно перекладывает под миску 2. Животные с более развитым интеллектом, например шимпанзе, пойдут сразу к миске 2, но многие по-прежнему перевернут миску 1, потому что выучили задачу через запоминание.

Однако пока не очень понятно, насколько эффективно идти к AGI путем воспроизведения возможностей животных. Таким

¹³ animalaiolympics.com

¹⁴ <https://www.youtube.com/watch?v=ok9opyg0Ofg>

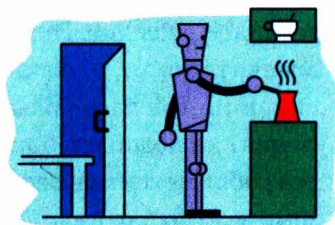
заключается в тонкостях мелкой моторики и обращения с инструментами, а вот для интеллектуального аспекта задачи может подойти и очень специализированное решение. Это будет очень частная модель для управления роботом, обученная на множестве примеров инструкций, и ни для чего больше она (и тем более результат ее обучения) не подойдет. Такие тесты могут способствовать развитию робототехники в целом, но до проблем AGI при работе над ними дело просто не дойдет.

Даже если та или иная задача выглядит так, будто для ее решения нужны развитые и разнообразные когнитивные навыки, она совсем необязательно будет хорошим материалом для практических бенчмарков, по крайней мере таких, которые измеряют прогресс достижения AGI. Поэтому место узких задач все чаще занимают испытания, предполагающие реализацию интеллектуальным агентом некой сложной системной деятельности, не сводящейся к прохождению тестовых заданий.

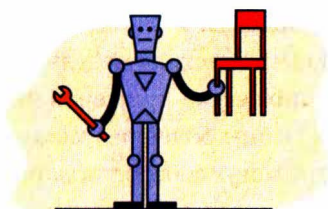
Приведем несколько примеров в порядке возрастания сложности.

- «Банковский тест»: ИИ должен взять кредит в банке на выбранную им цель, заработать денег легальным способом и вернуть его с процентами.
- «Тест студента» Бена Герцеля: искусственный интеллект поступает в университет, посещает те же занятия, что и обычные студенты, выполняет те же задания и получает соответствующую степень.
- «Тест на профпригодность» Нилса Нильссона: машина выполняет экономически важную работу на том же уровне, что и профессионалы в данной области (фактически являясь полноценным Narrow AGI).
- Artificial Scientist Test: компьютер совершает оригинальное научное открытие.
- Nobel Prize Test: машина получает Нобелевскую премию.

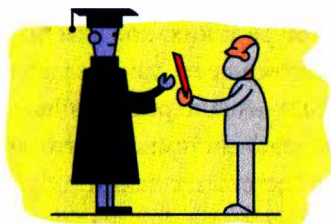
О создании AGI можно будет говорить, когда ИИ сможет...



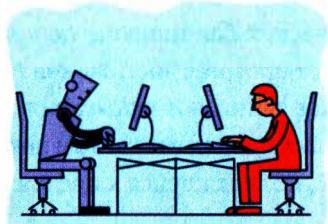
...войти в незнакомый дом,
найти кружку и сварить кофе*



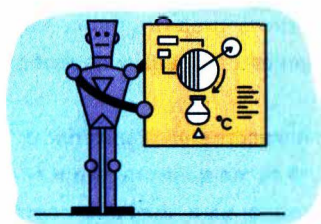
...собрать предмет
мебели из IKEA*



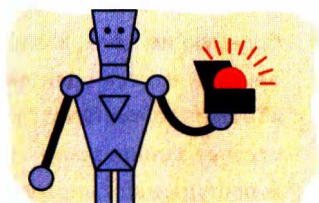
...окончить университет, посещая
те же занятия и выполняя те же
задания, что и люди



...выполнять экономически важную
работу на уровне человека



...сделать научное открытие



...получить Нобелевскую премию

* Для выполнения этих тестов ИИ понадобится подвижное физическое тело.

Рис. 9

Это общий искусственный интеллект или еще нет?

Еще одним критерием достижения AGI может быть решение ИИ-полных задач¹⁷. К ним можно отнести понимание естественного языка, компьютерное зрение, рецензирование научных статей.

Вообще, любая комплексная задача, включающая в себя возникновение непредвиденных обстоятельств (например, очень популярна сейчас задача беспилотного управления автомобилем), требует интеллекта уровня как минимум *Narrow AGI*. К сожалению, понятие ИИ-полноты слишком расплывчато, чтобы служить четким критерием. Например, мы легко можем согласиться, что «настоящее понимание текстов» невозможно без полноценного AGI, но это убеждение ничего не говорит нам о том, как проверить, что понимание — настоящее.

Даже в случае с беспилотным автомобилем может оказаться, что можно обойтись без *Narrow AGI*. Хотя нам и кажется, что такой автомобиль должен правильно реагировать на непредвиденные ситуации, но в норме этих ситуаций сравнительно мало. Система, управляющая автомобилем, может непредсказуемо вести себя в необычных ситуациях, но при этом демонстрировать средний уровень аварийности меньше, чем у человека. То, что при этом для каждой конкретной страны или даже города эту систему придется доучивать и дорабатывать, — это уже другой вопрос. Это очень трудозатратно, но реализуемо. И то, что система не обучена правильно реагировать на посадку самолета на дорогу и в такой ситуации поведет себя непредсказуемым образом, может служить предметом критики таких узких систем, но часто ли самолеты садятся на дорогу или происходит что-то необычное?

¹⁷ ИИ-полная задача (AI-complete) — термин, предложенный математиком Фанья Монтальво. Это задача, которая не заключается непосредственно в создании ИИ, но решение которой без полного интеллекта, близкого по уровню к человеческому, представляется невозможным.

Люди совершают много ошибок и в нормальных ситуациях. Так что превзойти уровень человека даже для такой задачи, как управление автомобилем, может быть, можно и узкими методами, особенно если самих людей-водителей на дорогах не останется. Конечно, если бы технологии AGI появились, они бы позволили решить эту проблему лучше и с меньшими затратами, но в контексте тестов и метрик получается, что для оценки развития и достижения AGI как бенчмарк необязательно подходят даже беспилотные автомобили и требуется что-то еще более впечатляющее.

Таким образом, тесты, которые убедили бы нас, что AGI действительно создан, предполагают выполнение одного из двух условий. Машина должна заменить человека в некоторой интегральной профессиональной деятельности или добиться впечатляющих по человеческим меркам свершений. Соответствие данным критериям будет означать, что, когда AGI будет создан, мы это точно заметим, но при этом не сможем оценивать промежуточные стадии. Наконец, все приведенные примеры показывают, что под «достижением AGI» подразумевается обычно AGI уровня человека.

Коэффициент универсального интеллекта

В результате мы оказываемся в некотором тупике: практические бенчмарки могут оценивать только узкоспециальные навыки, а тесты на достижение полного AGI не позволяют оценить частичный прогресс.

В этой ситуации было бы здорово иметь некую общую метрику интеллекта для всех разумных агентов. Опираясь на нее, можно было бы создавать способы оценки промежуточных результатов. Если попытаться определение интеллекта как

способности агента успешно действовать в широком диапазоне сред представить в математически строгом виде, возникает такая метрика, как коэффициент универсального интеллекта (Universal Intelligence Quotient, или UIQ¹⁸).

Чтобы ее построить, нужно сначала задаться вопросом: как можно формально определить наиболее широкий класс сред? Мы бы могли сказать: все существующие среды. Но это не будет формальным определением. Фактически нам нужен наиболее широкий класс не самих сред, а их моделей, математических представлений. И такой класс моделей — это все возможные алгоритмы.

Рассматривая различные тесты и бенчмарки, мы видели, что основная проблема их использования для оценки AGI заключалась в том, что решения к ним можно и нужно было подгонять. Это очевидно, ведь Deep Blue не сочиняет истории, AlphaGo не распознает картинки, а BERT не играет в шахматы. Что уж говорить про решения для тестов, основанных на каких-то фиксированных реальных данных или же синтезированных задачах, обладающих ограниченной структурой? В итоге поиск закономерностей в данных или учет этой структуры был заслугой больше людей-разработчиков, чем разработанного ими решения.

Но что интересно: синтезированные задачи или виртуальные среды порождаются той или иной программой. Да и тест, хоть и основанный на реальных данных, но проверяемый на компьютере, тоже является компьютерной программой. Чуть сложнее обстоит дело с тестами, проводимыми непосредственно в реальном мире, но пока что все наши модели реальности тоже являются алгоритмическими. Даже если у нас сейчас нет достаточно точной симуляции робота, входящего в произвольный дом, чтобы приготовить кофе, можно

¹⁸ Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4): 391–444, 2007.

предположить, что в принципе может существовать такая симуляционная программа, которую робот не смог бы отличить от реальности. Таким образом, пространство всех возможных компьютерных программ (или реализуемых ими алгоритмов) столь богато, что в нем можно найти любой бенчмарк, тест, задачу, среду. Если мы будем выбирать произвольную задачу из этого пространства, то заслугу ее решения уж точно можно будет приписать самому компьютеру.

Но чтобы построить количественную метрику уровня интеллекта, нужно определиться, как именно оценивать успешность действия интеллектуального агента и как объединять оценки, полученные в разных средах. В UIQ рассматриваются среды, задаваемые программами, которые возвращают агенту значение вознаграждения — как в стандартной постановке обучения с подкреплением. Успешностью действия агента в среде является полученное им суммарное вознаграждение (точнее, его математическое ожидание, если агент или среда допускают случайное поведение).

UIQ, будучи теоретическим критерием, рассматривает усреднение суммарных вознаграждений по всем возможным средам. Но среды берутся с разными весами, соответствующими вероятности встретить ту или иную среду. А именно, вероятность среды экспоненциально падает с длиной ее программы. В последующих главах мы подробнее обсудим, почему такие веса являются осмысленными, а пока лишь отметим, что число программ растет экспоненциально с их длиной, так что вероятность случайно выбрать конкретную программу экспоненциально падает (так, вероятность выбросить последовательность орел — решка при двух подбрасываниях монеты равна $1/4$, тогда как вероятность выбросить, скажем, орел — решка — решка — решка при четырех подбрасываниях составит $1/16$; кроме того, в длинной последовательности подбрасываний орел — решка будет встречаться в четыре раза чаще, чем орел — решка — решка — решка).

Интуитивно может показаться, что «победа» в более сложной среде должна значить намного больше. Но на самом деле нет. Если бы нам нужно было выбрать одну задачу для тестирования интеллекта, мы бы подумали о сложной задаче, надеясь, что если уж агент способен решать сложную задачу, то с простой-то он справится. Но когда мы производим усреднение по всем средам, то мы можем не делать такого допущения, и простые среды оказываются ничем не хуже, чем сложные. Скорее, наоборот. Если агент хорошо играет в шахматы, но не способен решать шахматные задачи, то чем он лучше агента, решающего шахматные задачи, но плохо играющего в шахматы?

Итак, мы можем определить коэффициент универсального интеллекта как значение получаемых агентом суммарных наград во всех вычислимых средах, взятых с весами, убывающими с длиной их программ.

Правда, тут же мы сталкиваемся с набором серьезных ограничений в практическом применении: невозможность определить, заикнется ли программа той или иной среды; усреднение по бесконечному количеству сред; усреднение по всем взаимодействиям агента с миром.

Но в то же время эта теоретическая мера интеллекта обладает целым рядом положительных свойств. Например, оптимизация этой меры эквивалентна построению универсального интеллектуального агента AIXI, о котором мы будем говорить в следующих главах. В отличие от теста Тьюринга, в котором агент прошел или не прошел испытание, UIQ является непрерывной мерой успеха, способной оценивать прогресс в системах AGI. Кроме того, эта мера не является антропоцентрической и не отдает априорного предпочтения тем или иным предметным областям или классам задач. К тому же UIQ дает объяснение нашему интуитивному пониманию уровня общего интеллекта разных систем. Например, отчасти обосновывает

Среда				Все возможные среды по пробкам, погоде, ЧП и т.д.	
Вероятность попадания в среду	0,02	0,05	0,15	P_n	0,00000001
	x	x	x	x	x
Награда за безаварийную езду	0,3	0,5	1	P_n	0
	=	=	=	=	=
Результат	0,006 +	0,025 +	0,15 +	$\sum P_n R_n$ +	0 =

$0,181 + \sum P_n R_n$
UIQ

Рис. 10

Как считается коэффициент универсального интеллекта (UIQ)

использование теста IQ для оценки уровня человеческого интеллекта — там также делают акцент на разнообразных задачах низкой сложности, — но при этом показывает и его предвзятость: задачи в IQ явно подобраны под человека, причем современного. В то же время UIQ показывает, что никаким фиксированным набором специализированных интеллектов (геометрического, языкового и т.д.) пространство задач для общего интеллекта не покрыть.

Для практического удобства появляются разные упрощения UIQ, например AIQ (Algorithmic Intelligence Quotient — коэффициент алгоритмического интеллекта¹⁹), а также системы тестов²⁰, учитывающие вид среды (активная, пассивная), наличие времени, определение интеллекта и видов универсальных агентов. В них устанавливаются ограничения по времени и вычислимости, ограничения на число взаимодействий и шкалы вознаграждений и штрафов в диапазоне $(-1,1)$. Сложность сред возрастает по мере прохождения, и они сбалансированы так, чтобы случайно действующий агент набирал суммарное вознаграждение, равное нулю.

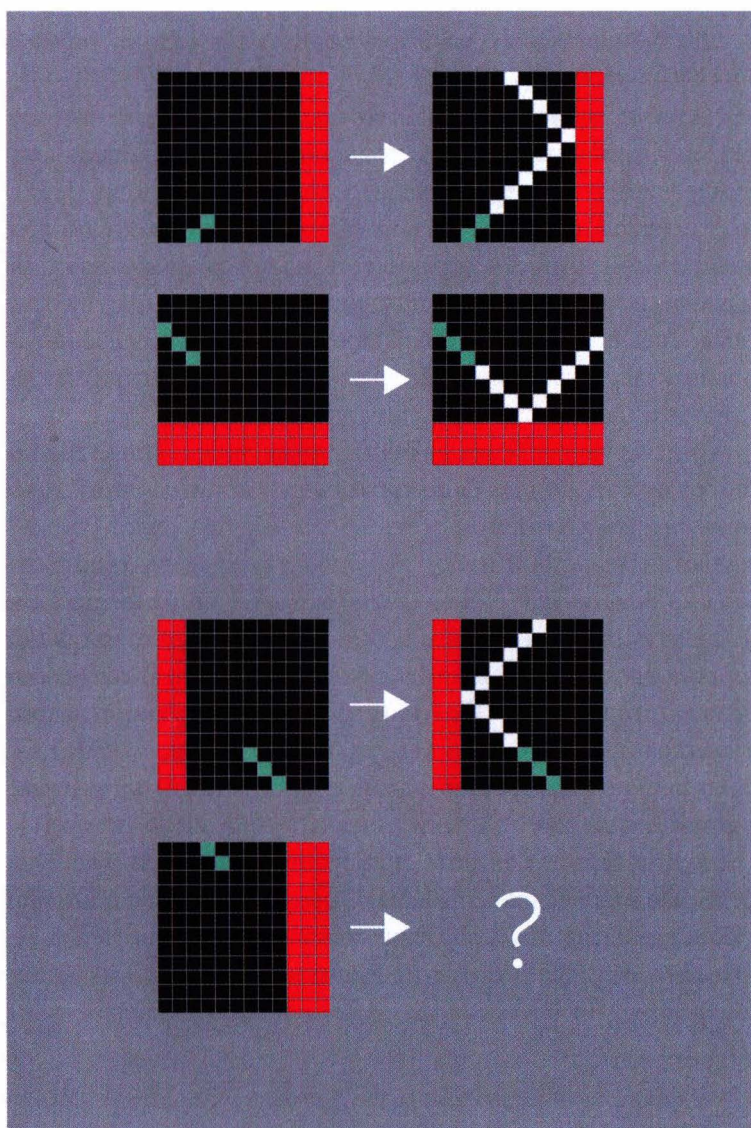
Существуют и более ограниченные тесты, отталкивающиеся от той же идеи алгоритмически произвольных закономерностей. Например, тест Abstractive Reasoning Corpus²¹ представляет собой набор визуальных заданий, выполняемых на квадратной сетке, клетки которой могут иметь один из десяти цветов. Каждое задание представляет собой набор из двух-трех пар сеток, где первая сетка представляет собой «вход», вторая — «выход» или «результат». Трансформация из «входа» в «результат» осуществляется с помощью некой высокоуровневой операции (заранее неизвестного алгоритма), разной для каждого задания. Задача тестируемой системы — сгенерировать абсолютно правильную «результатирующую» сетку для каждой новой входящей, увидев несколько примеров таких трансформаций. В самих заданиях заложена некая инкрементность²² сложности, но выполнение даже самых легких

¹⁹ An Approximation of the Universal Intelligence Measure — Shane Legg, Joel Veness.

²⁰ Measuring universal intelligence: Towards an anytime intelligence test — José Hernández-Orallo, David L. Dowe — Artificial Intelligence 174 (2010) 1508–1539.

²¹ <https://arxiv.org/abs/1911.01547>

²² Приращение на определенную величину.



Источник: François Chollet. On the Measure of Intelligence

Рис. 11

У искусственного интеллекта подобные задания вызвали трудности

разработку специализированных решений, характерных для всех предметно-ограниченных тестов.

Подводя итог, приходится признать, что хорошего практического решения проблемы метрики развития AGI в настоящее время не существует.

Некоторые специалисты полагают, что измерение частичного прогресса в AGI и в целом крайне проблематично, потому что это эмерджентная система²⁵, которая не будет проявлять свойств AGI до тех пор, пока она не будет построена целиком (подобно тому, как, воссоздавая человеческий мозг по кускам, мы не можем оценить результат до самого последнего шага). Тесты на то, что мы достигли AGI, не особо основываются на наших представлениях о том, как AGI должен быть устроен, и не позволяют измерить частичный прогресс. А тесты, которые пытаются сравнивать proto-AGI системы, оказываются либо слишком сложными для существующих систем, либо слишком простыми, так что их можно решать и узкими методами; и они всегда отдают предпочтение тому или иному подходу.

Разнообразные бенчмарки и тесты нужны скорее как мотивация, но не как нечто, что требуется проходить любой ценой. Например, шахматы были вызовом для ИИ многие годы, но Deep Blue, преодолевший этот вызов, не представляет особого интереса в области искусственного интеллекта, не говоря об AGI. Достигнув своей цели, он становится бесполезен, ведь больше он ничего не умеет. Гораздо больше пользы было от шахмат как от мотивационной задачи, чем от победы над человеком во что бы то ни стало.

²⁵ Эмерджентность — появление у системы свойств, не присущих ее элементам в отдельности; несводимость свойств системы к сумме свойств ее компонентов.

