ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ПРОБЛЕМЫ ПСИХОЛОГИИ СОЗНАНИЯ

© Карелов С.В.
кандидат технических наук, экс топ-менеджером компаний IBM и Silicon Graphics Inc. (SGI)
karelovs@gmail.com

© Лебедев А.Н.
доктор психологических наук, главный научный сотрудник лаборатории психологии личности, Институт психологии РАН, Москва, Россия,
lebedev-lubimov@yandex.ru

В интервью гл. ред. журнала «Ученые записки Института психологии РАН» д.пс.н. Лебедева А.Н. с экс топ-менеджером компаний IBM и Silicon Graphics Inc. (SGI), к.т.н. С.В. Кареловым обсуждаются новые технологии искусственного интеллекта, проводится сравнительный анализ языковых и мультимодальных моделей в свете изучения феномена сознания. Рассматриваются возможности и ограничения моделирования сознания средствами искусственного интеллекта, в частности – ошибка антропоморфизма в интерпретации структуры и функций искусственного интеллекта, его эмерджентные свойства и др. Интервью выполнено по материалам выступления С.В. Карелова на заседании Академического дискуссионного клуба ИП РАН на тему «Самый непонятый Х-риск – деградация агентности людей», которое представлено на канале Академического дискуссионного клуба ИП РАН[1].

Интервью выполнено в рамках Государственного задания № 0138-2024-0005

Лебедев А.Н.: Уважаемый Сергей Владимирович, в июне 2023 года в вашем выступлении на заседании Академического дискуссионного клуба ИП РАН, а потом в статье в №3(9) журнала[2], вы говорили о стремительном развитии искусственного интеллекта в мире, в частности, о GPT-4. И вот недавно у вас на канале появился пост под очень интересным названием – «Стохастический попугай умер! Да здравствует близнецы Homo sapiens!». Там речь шла о мультимодальной модели Gemini. Как мы знаем, Gemini может обрабатывать не только текстовую, но и визуальную информацию, которую не может обрабатывать ChatGPT. Могли бы вы рассказать об этом подробнее? И вообще, что нового появилось у разработчиков ИИ?

***Карелов С.В.*:** Когда я говорю о Gemini компании DeepMind, то речь идет о том варианте, который должен появиться в 2024 году. Стратегия OpenAI под руководством Сэма Альтмана, которая разработала ChatGPT, заключалась в том, чтобы привлечь максимум внимания общественности. Она выбрала «хоккейный метод», когда шайба забрасывается на сторону противника и все быстро туда бегут. В свою очередь, компания DeepMind поступила иначе и осторожно сформировала стратегию вывода на рынок своих продуктов.

OpenAI выпустила свой ChatGPT, когда уже была готова четвертая версия, но она решила сначала обнародовать предыдущую, «сырую» версию и позже ее «обновить», чтобы привлечь больше внимания.

DeepMind сделала по-другому. Она сразу заявила о том, что выпустила свой продукт в трех вариантах. Первый – самый маленький – «нано».

[1] https://www.youtube.com/watch?v=0sRiU5mRiuY
[2] *Карелов С.В.* «Ловушка Гудхарта» для AGI: проблема сравнительного анализа искусственного интеллекта и интеллекта человека // Ученые записки Института психологии Российской академии наук. 2023. Т.3. №3. С.5-22.

Искусственный интеллект и проблемы психологии сознания // Ученые записки Института психологии Российской академии наук. 2024. Т.4. №1. С. 76-84.
DOI:10.38098/proceedings_2024_04_01_08

У него изначально упрощенный функционал, который предназначен для того, чтобы запускать на мобильных устройствах. Второй вариант – «ультра», который сейчас все ругают и говорят, что он ничего особенного не делает по сравнению с GPT-4. Есть также и максимальный вариант, который планируется продемонстрировать в начале 2024 года.

Разработка DeepMind основана на идее мультимодальности. До сих пор все феноменальные свойства больших языковых моделей работали сами по себе. Все видели, как появляется тонкая изысканная буква творчества, когда тексты складываются очень красиво и умно, когда есть размышление, когда можно вести развитый диалог. Но это же есть и у Gemini. При этом Gemini одновременно работает с текстом, изображением и видео. Определенные попытки были реализованы и OpenAI, и рядом других фирм, когда изображение генерируется по тексту. Но это был скорее некий трюк. Специалисты удивились, когда вдруг эти модели заговорили на других языках, которым их не обучали (например, обучали английскому, а ИИ ловко заговорил по-русски). Но все это также своеобразный трюк. Просто идет перевод с языка на язык, а потом проводится обратный перевод результатов. То же самое и с изображениями. Что-то описывается словами, переводится в некую промежуточную форму, потом эта промежуточная форма превращается в некое подобие изображения, и только после этого с ним идет работа.

Так вот мультимодальность, которая есть у живых существ, включая человека, работает не так. Дело в том, что мозг обладает своими языковыми возможностями, связанными напрямую с каждым нашим анализатором (зрением, слухом, обонянием). Но он не переводит сенсорную информацию в слова. Мультимодальность Gemini фактически делает нечто похожее, что создает абсолютно иной уровень агентности. Программа не обладает сознанием (в человеческом понимании), агентностью, волей, но за счет того, что разные потоки сенсорной информации у нее естественным образом обрабатываются единой моделью, получается качественный скачок. При этом описать последствия такого качественного скачка мы не можем. Ясно только то, что он качественный, что происходит нечто принципиально отличное от предыдущего функционала.

Самый мой любимый пример – это представить жизнь как очень усложненный химический процесс. Была некая химическая вселенная, где взаимодействовали различные элементы, молекулы становились все более сложными, а потом вдруг появилась жизнь. Вот это тот самый качественный скачок, тот самый переход, который принципиально отличается от предыдущего состояния природы. То же самое происходит с мультимодальностью. Опять подчеркну, что именно возможность обрабатывать видеоизображения, звуки и тексты одновременно превращает этот новый искусственный интеллект, эту мультимодальную модель в некоего эмержентного агента, качества и свойства которого предугадать невозможно. Это так же, как, например, невозможно предугадать свойства развивающейся жизни, то есть, как будут эволюционировать живые существа, если у нас в основе есть лишь предыдущая развитию химическая вселенная. Эволюция неповторима, ее можно увидеть лишь, когда она уже произошла.

Вот, собственно говоря, это и происходит сейчас в связи с появлением первых мультимодальных моделей, которые не просто искусственное сочетание обработки видеоизображения, текста, звука, а всё в одной модели, скажем так, – модели понимания мира. Пусть это совсем другое понимание, не человеческое, но во многом оно уже подобно нашему человеческому, поскольку мы представляем собой, упрощенно говоря, такого же рода мультимодальные модели.

Что в результате всего этого получится, мы увидим в этом году. В лаборатории DeepMind работают с этим уже давно, то есть с мая 2023 года. И по той закрытой информации, которой я располагаю, можно сказать, что то, с чем они сталкиваются, заставляет их все дальше и дальше отодвигать открытый запуск этих моделей для всеобщего доступа.

Лебедев А.Н.: А чем это вызвано, какими-то опасениями?

Artificial intelligence and problems of psychology of consciousness. Proceedings of the Institute of Psychology of the Russian Academy of Sciences. 2024, Vol. 4, No. 1, Pp. 76-84.
DOI:10.38098/proceedings_2024_04_01_08

Карелов С.В.: Это связано с эмерджентностью модели. Начав работать с ней в полном варианте, компания наталкивается еженедельно на какие-то новые проявления ИИ, которые не находят объяснения. И, соответственно, здесь невозможно предвидеть последствия таких способностей. А у разработчиков есть только весьма грубый способ запрета, то есть рамки, определяющие, что модели можно делать, а чего нельзя (так, например, у языковой модели ChatGPT, есть запрет на расистские высказывания, неполиткорректные и т.д.). Но проблема в том, что любую такую возможность можно запретить только после того, как она проявится. А для того, чтобы увидеть ее заранее, нужна предварительная диагностика. Даже когда человек начинает сходить с ума, опытному врачу, прежде чем поставить диагноз, необходимо довольно долго разговаривать с пациентом, чтобы понять, что это – шизофрения, психоз, симуляция, акцентуация личности или что-то другое. Но все эти человеческие свойства и проявления люди изучают как минимум несколько сотен лет, и поэтому уже есть понимание механизмов, классификации этих патологических проявлений. Когда такие проявления появляются у модели, то некоторые из них совершенно не похожи на человеческие и в нашем понимании они не являются никакой акцентуацией.

Лебедев А.Н.: Похоже на борьбу с наркотиками. Когда появляется наркотик на рынке, то со временем принимается запрещающий закон. А затем появляется новый наркотик и новый закон. И так по кругу. Получается, что появление новых препаратов опережает методы ограничения их употребления в свободном доступе.

Карелов С.В.: Это так, но здесь ситуация еще хуже, чем в вашем примере. Новый наркотик делают люди, познакомившиеся с этим законом, а что вылезет эмерджентно, даже предугадать невозможно, поскольку неизвестно, как это будет выглядеть и как проявит себя.

Лебедев А.Н.: Сейчас часто идут дискуссии о том, что реализовано в искусственном интеллекте из того, что мы знаем о сознании человека. Как известно, Д. Канеман говорит о двух системах переработки информации. По сути, первая – это бессознательное или подсознание, вторая – это сознание, рациональное мышление. И вот возникает вопрос: нечто подобное есть у моделей ИИ или это принципиально иное, например, по логике и способам решения задач? Можно ли сказать, что у ИИ есть некое сознание и подсознание?

Карелов С.В.: Здесь не только логика иная, здесь базовые предпосылки возникновения самих процессов совсем иные. Пример такой: есть цифровые компьютеры, а есть аналоговые. Аналоговые компьютеры не оперируют никакими битами, это непрерывные потоки информации в аналоговой машине. И в то же время на цифровой машине можно моделировать аналоговую. Так вот наиболее близкий пример того, что такое искусственный интеллект – это цифровая машина, которая имитирует работу аналоговых машин, коими являются люди.

И говорить о том, насколько реальна имитация сознания, которое мы представляем у людей, невозможно. Это моделирование. Здесь бессмысленно говорить, появится у ИИ сознание или нет. Оно может и появиться, и нет. Но в основе лежат абсолютно другие процессы так же, как в аналоговых и цифровых вычислениях. С одной стороны, в цифровой машине это определенный набор триггеров (нули и единицы), усложняя вычисления которых, мы выполняем некие алгоритмы. В аналоговой машине нет никакой программы, записывающей алгоритм, а запускается некий физический процесс, который сам по себе является моделью какого-то явления. Поэтому любая антропоморфизация ИИ и попытка сказать, что сейчас он проявит какие-то человеческие свойства, это будет только моделирование. Но то, что в результате может получиться, мы не способны предугадать из-за его эмерджентных свойств.

Лебедев А.Н.: Когда мы анализируем понятие психики в учебниках по общей психологии, мы говорим, что психика состоит из психических процессов, свойств и состояний, которые включают в себя мышление, память, внимание, эмоции, то есть все то, что можно изучить, измерить и т.д. Но возникает вопрос, с которым психологи часто сталкиваются: может ли выйти

Искусственный интеллект и проблемы психологии сознания // Ученые записки Института психологии Российской академии наук. 2024. Т.4. №1. С. 76-84.
DOI:10.38098/proceedings_2024_04_01_08

машина из-под контроля, начать что-то самостоятельно делать и «восстать против человека»? Дело в том, что у человека есть чувство собственного достоинства, мы переживаем состояние стыда, гордости, чувства вины, справедливости, тщеславия и пр. Нас можно оскорбить, обидеть. Мы иногда хотим поразить мир какими-то новыми идеями или соревнуемся с кем-то за приоритеты, стремимся что-то сделать для того, чтобы пережить состояние некоего превосходства над другими, кого считаем в чем-то хуже нас и пр. Ведь ничего этого нет у ИИ, а значит у него другая мотивация или ее вообще нет?

Карелов С.В.: Александр, я еще раз повторю принципиально важный тезис: чем больше степень антропоморфизации, тем большую ошибку мы допускаем.

Лебедев А.Н.: Но это именно та ошибка, которая пугает людей далеких от этой темы. А таких, похоже, большинство.

Карелов С.В.: Да, но я сейчас обсуждаю эти вопросы с позиции человека несколько более погруженного в эту тематику. И подчеркиваю, что эту точку зрения разделяют не только всякие там думеры[3], которые боятся рисков, но и такие технооптимисты, как Ян Лекун[4], которых довольно много. Абсолютно все, кто конкретно сталкивался с разработкой этих систем, понимают, что любая антропоморфизация ИИ – это большая ошибка. ИИ гораздо дальше от нас отстоит по всем психологическим параметрам, чем, например, муравей. Муравей гораздо ближе к нам с точки зрения психологии, потому что муравей может быть голодным, у него может быть много стимулов и даже не только с точки зрения его биологической сущности, но и с точки зрения когнитивных возможностей (интересное, опасное, любопытное, странное, новое и т.д.). А есть ли любопытство у ИИ? Есть, но моделируемое. Но поскольку у подобного рода систем нет ни абсолютных аналогов, ни когнитивных свойств, ни эмоциональных, ни связанных со знанием, то здесь все по-другому, есть какие-то другие свойства, но мы их не знаем.

Лебедев А.Н.: Сергей, очевидно, что свойством человеческого сознания является как раз желание найти что-то подобное себе. Так люди придумывают богов, похожих на человека, молятся иконам, считая, что за деревянной панелью есть что-то такое, с чем можно разговаривать, общаться. И вот это стремление искать человеческое и разумное в том, что таковым не является, преследует человечество уже тысячелетия и будет, видимо, преследовать и дальше, потому что искусственный интеллект дает повод к этому. Он имитирует человеческие эмоции, чувства, он способен решать очень сложные логические задачи, но при этом гораздо более эффективно, чем человек. И это вызывает чувство страха.

Карелов С.В.: В первую очередь я подтвержу, что самые серьезные риски, которые реально видны и которых следует опасаться и против которых нужно пытаться искать защиту и спасение, это не риски некоего восстания искусственного интеллекта, где он захочет нанести кому-то вред. Этот вопрос пока отодвигается в неопределенное будущее и застилается куда более серьезными вполне реальными рисками. Например, это риск захлебнуться в фейках и в дезинформации. Но даже это люди начинают понимать в рамках свойственного человеку антрпоморфизма.

Очень глубокие исследования, проводимые в этом году, показали, что лингвистико-семантические, психологические способности порождения текстов большими языковыми моделями сильнее, чем у людей. Они способны порождать более убедительные тексты, направленные на универсальную аудиторию. Если же эти системы знают кое-что о людях, на которых они ориентируются, например, персональные данные, то тут их возможности становятся похожими на ядерное оружие. Если система знает,

[3] Думер, думеризм – термины для описания людей, которые крайне пессимистично относятся к глобальным проблемам, таким как перенаселение или изменение климата. Некоторые из них утверждают, что существует вероятность, что эти проблемы приведут к вымиранию человечества.

[4] Ян Лекун (*Yann LeCun*) – французский учёный, специалист в области информатики, машинного обучения, робототехники и вычислительной нейробиологии. Лауреат премии Тьюринга 2018 года за разработку теории глубокого обучения.

Artificial intelligence and problems of psychology of consciousness. Proceedings of the Institute of Psychology of the Russian Academy of Sciences. 2024, Vol. 4, No. 1, Pp. 76-84.
DOI:10.38098/proceedings_2024_04_01_08

что говорит именно с вами и адресует месседж именно вам, то у человека очень мало шансов противостоять такому влиянию. Настораживает то, что подобные системы уже начали использоваться для создания медийных нарративов. И как они будут развиваться, мы не знаем. Зритель может видеть и слышать настоящего ведущего по телевизору, но текст, который тот произнесет, подготовит и напишет эта система. И его убедительность окажется чрезвычайно мощной.

Такие системы уже искусно оперируют восприятием информации и способны «ввинтить», и ежедневно делают это, через разные медиа каналы, некие нарративы, направляющие и меняющие способы восприятия действительности миллионов людей. И что будет результатом этого влияния, мы не можем даже предположить, потому что не было у человечества ничего подобного. Некоторые люди говорят: «Ну и что? Ведь дезинформация была всегда, менялся только инструментарий. Весь мир всегда строился на обмане». Однако сегодня происходят качественные изменения и бессмысленно говорить, что люди всегда убивали друг друга луками и стрелами, поэтому ничего особенно не произойдет, если в ход пойдут баллистические ракеты с термоядерным оружием.

Лебедев А.Н.: Сергей, вот еще вопрос. Сегодня, наверное, можно говорить о развитии не одного какого-то искусственного интеллекта, а о многих, о разных его вариантах, которые будут связанны с культурой и политикой тех или иных стран, которые достигнут в этой области успехов. Сейчас мы понимаем, что США впереди в разработке таких программ. А как быть, например, с Китаем и Россией, и другими странами, в которых эти разработки также ведутся. Насколько мы, например, отстаем от США, и в каком качестве выступает Китай по отношению к уровню развития этих технологий?

Карелов С.В.: Технологии искусственного интеллекта развиваются уже более полувека и тот качественный революционный скачок, который произошел за последние буквально пару лет, связан с технологиями трансформеров. Трансформеры – это особая архитектура, позволяющая масштабировать систему искусственного интеллекта простым наращиванием вычислительной мощности. Берешь и делаешь систему в 100 раз больше, она работает в 100 раз с большим количеством параметров, и ее возможности и ее свойства экспоненциально растут. Так вот это наращивание вычислительной мощности стало определять на 95% потенциал того, чего могут добиться люди в реализации подобного рода систем. И это означает то, что я обозначил в девизе, ставшим когда-то вирусным: «есть железо – участвуй в гонке, нет железа – кури в сторонке».

Поскольку технологии производства интегральных схем для специализированных систем искусственного интеллекта, таких как графические процессоры компании NVIDIA (это наиболее яркий пример), сосредоточены в руках американских фирм, то ситуация получается очень и очень интересная. У американцев есть все для того, чтобы это развивать с той скоростью, с которой позволяют интеллектуальные возможности их лабораторий. У китайцев ситуация сложилась очень тяжелая, потому что последние два года, ровно параллельно тому, как происходила революция трансформеров, американцы зажимали экспортный контроль. И сейчас запрещают не только экспортировать сами графические процессоры, но и технологии, позволяющие самим сделать у себя подобного рода процессоры. А китайцы вкладывают в это дело колоссальные деньги, бросают колоссальные ресурсы на то, чтобы этот разрыв устранить, чтобы самим производить микросхемы у себя. По планам китайцев нужно решить эту проблему к тридцатому году. Решат или не решат – сказать сложно, но до тех пор, пока не решили, они отстают от США во всех областях, связанных с искусственным интеллектом где-то на год, где-то на два, три, пять.

В России ситуация совсем иная. Дело в том, что здесь и не светит ничего. Проблему нельзя решить серыми поставками, как это делается для автомобилей, айфонов или электронных схем, которые используются в вооружении, их все-таки нужны десятки, сотни, тысячи в лучшем случае. Дело в том, что когда речь идет о системах искусственного интеллекта, то этих графи-

Искусственный интеллект и проблемы психологии сознания // Ученые записки Института психологии Российской академии наук. 2024. Т.4. №1. С. 76-84.
DOI:10.38098/proceedings_2024_04_01_08

ческих процессоров только для обучения одной модели требуются десятки тысяч. Наладить производство у себя можно. Для этого нужно несколько десятков, а лучше сотен миллиардов долларов и запас времени в 10-15 лет.

Лебедев А.Н.: Сложная ситуация. Получается, что нужно либо помириться с Америкой и начать сотрудничать, либо забыть об этом навсегда и жить по старинке, то есть заимствовать у конкурентов и платить им большие деньги?

Карелов С.В.: Всегда остается надежда, что вместо одной технологии придумаю другую. Я знаю реально очень много примеров, когда наши разработчики говорят о том, что трансформеры уже «упираются в стену», т.к. невозможно бесконечно масштабировать. У производителя графических процессоров не хватает производственных мощностей, чтобы обеспечить всех желающих. Якобы мы разработали новый метод, который не требует такого масштабирования, то есть не нужно в миллион раз повышать производительность железа, а можно сделать все не менее искусно. Но, к сожалению, все это остается только обещаниями, обогнать американцев пока не удается никому.

Лебедев А.Н.: Не секрет, что фантасты, которые пытаются предположить, что будет в будущем, иногда очень здорово предсказывают то, что происходит на самом деле. Например, в книгах В. Пелевина «Generation П» и «Transhumanism Inc.» было показано, как будут работать фейковые видео программы, и что миром будет управлять не просто телевизор, а телевизор, где роль ведущего возьмет на себя оцифрованный человек, который представит зрителям оцифрованного лидера, а тот, в свою очередь, скажет населению, как нужно правильно жить. Причем В. Пелевин написал это, когда еще не было программ вроде Midjourney, ChatGPT и так далее.

Сейчас мы часто обсуждаем вопросы трансгуманизма, у которого есть сторонники и ярые противники. Проблема в том, что нас ждет в будущем, будет ли все развиваться так, как нам говорят трансгуманисты и что с этим делать? Нужно ли все это запрещать? У нас есть религиозные ученые, и среди них есть психологи, которые говорят, что с этим надо бороться. Не просто потому, что это вред человечеству приносит, а потому что это те сферы, в которые человек не должен внедряться. Они считают, что не нужны искусственные органы, не нужно улучшать человеческое мышление, что это идет против «божьего замысла», а значит «аморально».

Высказываются и другие крайние мнения, противоположные, например, о том, что хорошо бы разработать некий чип с программой, который можно вставить в голову человеку, и этот модуль, соединяясь с биологическими структурами, расширит возможности человеческого мозга и улучшит способности человека, в частности, в решении сложных интеллектуальных задач. Поскольку известно, что Илон Маск уже добился определенных успехов в данной области, можно, наверное, предположить, что это совсем скоро может стать реальностью, а затем повседневностью. И ведь, наверное, можно такую штуку вставить в голову и животного, тогда оно превратится в оружие или раба, которое будет воевать или трудиться, не заявляя о своих правах. Вот для таких фантастических технологий, как вы считаете, есть какие-то основания или нет?

Карелов С.В.: Начну с того, что фантасты чаще всего предсказывают развитие зачатков уже существующих технологий, а не выдумывают что-то совершенно непостижимое, чего в лабораториях не разрабатывается. В. Пелевин в «Generation П» описывает компьютеры Silicon Graphics Inc. И смею вас заверить, что роман не появился бы в 1998 году, если бы в 1995 году в России не открылось отделение этой компании, которое я тогда и возглавил. И все, что там описано, Deepfake, оцифровка личности, все это уже тогда делалось в лабораториях Silicon Graphics Inc. в 1995 году. Это стоило миллионы долларов, для этого нужны были огромные графические суперкомпьютеры, но все это уже было и, собственно говоря, потребовалось всего тридцать лет для того, чтобы по стоимости подобного рода деятельность стала доступной для массового применения.

Artificial intelligence and problems of psychology of consciousness. Proceedings of the Institute of Psychology of the Russian Academy of Sciences. 2024, Vol. 4, No. 1, Pp. 76-84.
DOI:10.38098/proceedings_2024_04_01_08

Аналогично можно сказать и про нейроинтерфейсы и способы их использования для людей и животных. Все это уже есть в лабораториях, и возможности, которые показывают подобного рода системы, – фантастические. Очень велика вероятность, что всего через 5-7-10 лет проблема понимания языка животных будет успешно решена так же, как решена задача перевода с любого из 200 языков на базе машинного обучения. То, что мы будем понимать язык собак, кошек, енотов и пр., это уже, по сути дела, предрешено. Вопрос только в том, когда такие переводчики подешевеют, чтобы они получили массовое распространение.

С нейроинтерферисами сложнее, но уже сейчас разработаны такие варианты, которые позволяют достаточно четко считывать из мозга информацию и вкладывать в мозг соответствующее управляющее воздействие. Проблема заключается в том, что до сих пор ни один ученый мира не может сказать, что он действительно понимает, каким образом из нейрокоррелятов сознания, которые можно измерить с помощью fMRT, зафиксировать и воспроизвести, рождаются сами феномены – сознание, ощущение боли, голода, тоски, любви и т.д. Нет на это ответа. Летом 2023 года были опубликованы работы Анила Сета[5], где он приводит пятьдесят основных теоретических фреймворков для описания сознания. Но ни один из них не решает этой трудной, возможно непреодолимой, проблемы. Если не будет найден способ объяснить, как из нейрокоррелятов сознания, мышления и прочее рождаются сами эти феномены, то говорить о том, о чем говорят трансгуманисты, о слиянии, переносе и т.д. так и будет оставаться фантастикой.

Лебедев А.Н.: Да, в психологии мнения по поводу этой проблемы, которая у нас благодаря Г. Фехнеру получила название психофизиологической, очень разные. Многие считают, что она вообще не имеет решения и для этого находят разные основания, не только философские, но и научные. Парадоксально, но сегодня ученые часто спорят не о том, как решить психофизиологическую проблему, а о том, как лучше доказать, что решить ее невозможно в принципе. Мы на эту тему тоже много спорим в нашем Академическом дискуссионном клубе. На заседаниях часто такие вопросы возникают, но пока, действительно, все это очень сложно. Может быть, как раз работа над искусственным интеллектом позволит нам в нашей науке продвинуться и понять, как же все-таки нейроны работают и создают психические переживания и движение мысли? Как они формируют этот спектр индивидуальных субъективных психологических ощущений, которые каждый знает, но никто не знает, как это все можно изучить и воспроизвести в экспериментальных условиях?

Есть еще одна проблема, связанная с тем, о чем мы говорим. Интересно ваше мнение и как вы к этому отнесетесь. На заседании нашего клуба выступал философ Д.И. Дубровский, который считает, что поскольку все субъективные переживания возникают на основе работы нейронов, из которых состоят сложные структуры мозга, то их работа, разумеется, гипотетически, воспроизводима и не на биологической основе. И если мы поймем, как работает мозг, то сможем воспроизвести структуры мозга таким образом, чтобы полученная некая «интеллектуальная активность» смогла переживать что-то похожее на эмоциональные состояния, например, такие, как страх, боль и т.д. Одни этого боятся, другие говорят, что субъективные переживания – это продукт определенного этапа развития биологии и ни на какой технической основе их воспроизвести нельзя. Есть еще третья группа людей, которые, следуя Декарту и Лейбницу, полагают, что, помимо нейронов мозга, существует некая духовная субстанция, обладающая волей, нечто «внетелесное», что вообще невоспроизводимо путем какого-либо моделирования. Что вы можете добавить по этому вопросу?

[5] Анил Сет (*Anil Seth*) – британский нейробиолог, специалист в области когнитивной и вычислительной нейробиологии. Сторонник материалистического объяснения сознания, который входит в число наиболее цитируемых в мире ученых по темам нейробиологии и когнитивной науки.

Искусственный интеллект и проблемы психологии сознания // Ученые записки Института психологии Российской академии наук. 2024. Т.4. №1. С. 76-84.
DOI:10.38098/proceedings_2024_04_01_08

Карелов С.В.: Я исхожу из того, что мне известно о работах, которые проводятся в самых лучших передовых лабораториях мира. Многие из этих работ еще не опубликованы, но тем не менее в подавляющем большинстве результаты исследований говорят о том, что все феномены, о которых мы говорим применительно к человеческому сознанию, разуму, мышлению, все они, являются вычислительными, т.е. в их основе вычислительный механизм. Это первое. Многие копья ломались по вопросу: является ли это все-таки результатами вычислений, обработки информации или нет? Ответ на этот вопрос с вероятностью 90% – да, это результат вычислительных действий.

Второе это то, что материальной основой реализации этих вычислений может быть что угодно. На сегодняшний день уже известны разные механизмы реализации вычислений. Например, возьмите то, как умеют вычислять растения, насекомые и люди, они просто принци-пиально по-разному реализованы. То, что процессы саморегуляции происходят через тот же самый вычислительный механизм, показывает, что материальный носитель вычислений может быть любым. Это также экспериментально подтверждено.

Наконец, третье и последнее заключается в том, что физика, то есть материальные процессы, реализующие эти вычисления, могут быть разными. До тех пор, пока мы их не изучим и не поймем, как вычисляет дерево или звезда, мы не поймем, за счет каких процессов реализуется вычисление в мозге. В этом смысле попытка смоделировать отдельный нейрон или же сети нейронов является аналогичной, например, желанию папуасов сделать соломенный самолет в надежде, что он взлетит и потом принесет подарки от наших умерших родственников. Речь не о том, чтобы смоделировать элемент и соединить элементы, а в том, чтобы понять физику процесса, то есть за счет чего самолет поднимается в воздух. И до тех пор, пока мы это не поймем, говорить о реализации чего-то подобного бессмысленно.

Человечество и так продвинулось очень далеко с таким способом вычислений, который реализован в наших компьютерах. Компьютеры были созданы всего-то в середине прошлого века, и с тех пор прошли колоссальный путь и масштабировались в миллионы и миллиарды раз по своей вычислительной мощности. В результате на этой вычислительной основе (кремниевой, цифровой) мы смогли достичь такого уровня реализации, что уже делаем вещи, которые очень похожи на мышление, где-то даже на сознание. Но представьте себе, что реализовать все это можно и абсолютно на другой основе, абсолютно на других физических принципах.

Таким образом, подводя итог, можно сказать, что существует бесконечное количество способов реализации тех феноменов, о которых мы говорим. Один из этих способов – биологический – нам показала сама природа, второй способ люди сделали самостоятельно, но он еще довольно основательно отстает по своей вычислительной мощности от того, что есть у человека. Удастся ли на основе вычислительных цифровых машин, реализованных в кремниевых микросхемах, достичь уровня, сопоставимого с человеком, пока считается теоретически возможным. Есть ли альтернативные способы? Наверняка есть, и их огромное количество. Основной вопрос – их найти.

Лебедев А.Н.: Но ведь задавшись такой целью и с генетикой можно поэкспериментировать. Взять какую-нибудь обезьянку и улучшить ей мозг путем генетической инженерии, сделать из нее более разумное существо. Такая идея тоже может появиться. Разве нет?

Карелов С.В.: В связи с этим я вам назову лишь два имени. Это Майкл Левин и Джош Бонгард. Они создали ксеноботы, экзоботы, то есть живые программируемые организмы, которые сделаны из адаптированных стволовых клеток африканского вида лягушек (*xenopus laevis*). Разработчики уже близки к решению проблемы регенерации органов. Это тот самый стык вычислений и биологии, цель которого понять, как выполнять вычисления на уровне биологического субстрата, а не на уровне кремниего. Они страшно перспективны и лично я думаю, что это будет магистральной линией, которая позволит совершить следующий качественный рывок. И

Artificial intelligence and problems of psychology of consciousness. Proceedings of the Institute of Psychology of the Russian Academy of Sciences. 2024, Vol. 4, No. 1, Pp. 76-84.
DOI:10.38098/proceedings_2024_04_01_08

это будет уже не наращивание масштабов по числу процессоров, которое сейчас характерно для реализации трансформеров, а нечто более глобальное.

Лебедев А.Н.: Сергей, для психики и сознания характерно саморазвитие. Мозг как саморегулирующаяся система себя строит, меняет и так далее. А языковые модели, похоже, пока еще не очень?

Карелов С.В.: Языковые модели решают свои задачи. Эволюция создала живые организмы, в том числе и мыслящие, для решения задач выживания. Компьютеры строились совсем для другого, они не предназначены для собственного выживания. Я еще раз повторю, антропоморфизм – это самая большая ошибка, которая может быть допущена в разговорах об искусственном интеллекте!

Лебедев А.Н.: Сергей, спасибо вам за участие в наших дискуссиях и за очень интересное интервью.

ARTIFICIAL INTELLIGENCE AND PROBLEMS OF PSYCHOLOGY OF CONSCIOUSNESS

© Sergey V. Karelov
PhD in Engineering, Extop Manager of IBM and Silicon Graphics Inc. (SGI)
karelovs@gmail.com

© Aleksandr N. Lebedev
Sc.D. (psychology), Chief Scientific Officer, laboratory of psychology of personality,
Institute of psychology, Russian Academy of Sciences, Moscow, Russia
lebedev-lubimov@yandex.ru

Interview of the editor-in-chief of the journal "Proceedings of the Institute of Psychology of the Russian Academy of Sciences", Doctor of psychology science Lebedev A.N. with ex-top manager of IBM and Silicon Graphics Inc (SGI), Ph.D. S.V. Karelov. New artificial intelligence technologies are discussed, and a comparative analysis of linguistic and multimodal models is carried out in the light of the study of the phenomenon of consciousness. The possibilities and limitations of modeling consciousness by means of artificial intelligence are considered, in particular, the error of anthropomorphism in interpreting the structure and functions of artificial intelligence, its emergent properties, etc. The interview was conducted based on the materials of the article by S.V. Karelov and his speech at the meeting of the Academic Discussion Club of the IP RAS on the topic "The most misunderstood X-risk is the degradation of human agency", which is presented on the channel of the Academic Discussion Club of the IP RAS.