Интерфейсы мозг-искусственный интеллект: основания и перспективы

А. Я. Каплан[1]

Мозг человека очевидным образом является «родовым» объектом для определения понятия искусственных «интеллектуальных систем». Однако механизмы собственно интеллектуальной деятельности мозга до сих пор остаются не раскрытыми в своей самой сущностной части. В значительной мере это связано с тем, также очевидным обстоятельством, что естественный интеллект в широком смысле обладает свойством субъективного представления реальности, операциональная архитектоника которого, его форматы, способы и полнота описания реальности вне субъективного «Я» нам неизвестны. В этой связи и с учетом отсутствия «общей теории мозга» становится проблематичным создание нейроморфных систем искусственного интеллекта, построенных на структурно-функциональных аналогиях с мозгом. Уровень такой «нейроморфности» всецело будет определяться не полнотой нейроанатомических описаний, а границами нашего понимания того, как работает мозг человека.

Между тем, свойство нейроморфности систем искусственного интеллекта может быть достигнуто не столько за счет построения подобной мозгу структурно-функциональной архитектоники этих систем (сети искусственных нейронов, алгоритмы глубокого обучения и т. д.), сколько посредством их обучения двусторонней коммуникации с естественным интеллектом. Мы полагаемы, что каналы информационного взаимодействия мозг-искусственный интеллект для запуска процессов обучения «общению» могут быть построены на основе технологий неинвазивных интерфейсов мозг-компьютер (ИМК) нового поколения. Ключевым звеном этих технологий, станут контуры обратной связи, в рамках которых модули искусственного интеллекта, подключенные к анализу ЭЭГ, будут через естественные сенсорные каналы демонстрировать человеку свои гипотезы относительно принадлежности специфических паттернов ЭЭГ тем или иным мысленным актам человека. В интерактивном процессе эти гипотезы будут уточняться до полной сходимости, формируя таким образом в памяти мозга человека и машины траектории взаимодействия до приемлемого уровня «взаимопонимания». В докладе будут рассмотрены нейрофизиологические и нейротехноло-

[1] *Каплан Александр Яковлевич* — д.б.н., профессор, заведующий лабораторией нейрофизиологии и нейрокомпьютерных интерфейсов биологического факультета МГУ им.М.В.Ломоносова, e-mail: akaplan@mail.ru.

Kaplan Alexander Yakovlevich — Doctor of Science, Head of the Laboratory of Neurophysiology and Neurocomputer Interfaces, Lomonosov Moscow State University, Faculty of Biology.

гические основания к созданию интерфейсов «мозг-искусственный интеллект».

Ключевые слова: мозг человека, электроэнцефалограмма, интерфейсы мозг-компьютер искусственный интеллект, нейроморфные системы.

1. Введение

Проблема взаимодействия человека с изобретенными им же самим машинами возникла, наверное, еще со времен строительства великих пирамид. В первую очередь она состояла в том, что даже в самом развитом виде эти машины не являлись продолжением тела и потому не могли управляться напрямую от мозга, не могли в полной мере соответствовать тонкой динамике и направленности ментальных планов. Даже бурное развитие автоматики, кибернетики и робототехники в 20-м веке, в какой-то мере заполнивших зазор между мысленными решениями и исполнительными агрегатами, не смогло снять с повестки дня проблему «человек-машина» и даже наоборот, предъявило особые требования человеку-оператору машин, в частности, по психофизиологическим его параметрам. Однако власть и могущество машин со всей очевидностью проявились в условиях тотальной информатизации и цифровизации жизни человека в 21-м веке. Апофеозом этой взрывной трансформации природной среды человека в кнопочно-виртуальный мир явилось повсеместное и все ускоряющееся распространение высоко производительных элементов искусственного интеллекта (ИИ).

Пусть еще не создан сильный искусственный интеллект, и еще неизвестно будет ли это возможным когда-либо, но уже во всех аспектах своей традиционной деятельности человеку приходится взаимодействовать с высокопроизводительными информационно-аналитическими системами, встроенными не только в стиральные машины и самолеты, но и в сами технологии расчета и проектирования машин, в процессы управления экономическими ресурсами и социальными процессами, наконец, в организацию виртуального мира, опосредующего реальность через масс-медиа, через всевозможные поисковики, интернет-магазины и в скором будущем – через аватары и интернет-вещей. Однако «Это не есть создание нового человека, это есть истребление человека, исчезновение человека, замена его иным существом, с иным, не человеческим уже существованием» [1]. Эти строки написаны философом Н.А.Бердяевым почти век назад, но, как видно, они имели пророческий смысл. В этом «ином существе» явно проглядывает сейчас призрак Искусственного интеллекта, если дать ему свободу выбора и действий.

Соглашаясь в принципе с проблемой «человек-машина» в трактовке Н.А.Бердяева, в случае с искусственным интеллектом приходится все же смириться с самой необходимостью его появления в технологической эволюции человека. Раз вступив в мир цифровых отношений, человек уже не сможет позволить себе обходиться без помощи информационно-аналитических устройств и технологий, не просто повышающих эффективность его работы, но и определяющих само существование человека в этом дивном во всех отношениях цифровом мире.

В этой связи все большее распространение в среде не только гуманитариев, но и естествоиспытателей, получает идея создания подконтрольного человеку искусственного интеллекта, назовем его «человеко-ориентированным» искусственным интеллектом, обладающего иниициативой и свободой активностей только в пределах, задаваемых собственно потребностями человека. В частном случае подобные операциональные отношения между человеком и ИИ могли бы устанавливаться непосредственно на «линии» между ними, например, посредством технологий интерфейсов мозг-компьютер (ИМК). Задача создания эффективного канала двусторонней связи между мозгом человека и элементами ИИ (искусственными нейронными сетями, к примеру) с возможностью формирования мозг-машинного языка выдвигается сейчас на передний край широкого поля мультидисциплинарных исследований и разработок в области «человеко-ориентированного» ИИ. В первую очередь, это взаимосвязанные задачи от поиска нейрофизиологических маркеров физиологических и ментальных запросов организма до создания программно-алгоритмических решений их выявления, моделирования и классификации. По сути дела, именно эти маркеры становятся тем признаковым пространством, в котором могут быть сформированы элементы символьного взаимодействия между мозгом и ИИ. Рассмотрим основные контуры исследований и разработок в области ИМК.

2. Технологии интерфейсов «мозг-компьютер»: фантазии и реалии

Первые успешные попытки выделения устойчивых командных сигналов в электрической активности мозга человека (ЭЭГ) были продемонстрированы еще в 60-70-х годах прошлого века. Однако настоящий бум исследований и разработок в этой области начался в 21-веке. В последнее время каждый год на эту тему публикуется 2-3 тыс только узко специализированных научных статей. Особенно перспективными с точки зрения планов последующего широкого внедрения ИМК представляются работы с неинвазивной регистрацией биометрической информации,

сопряженной с работой мозга, в первую очередь, с регистрацией ЭЭГ. В силу своей природы этот показатель активности мозга практически без инерции отражает работу больших кооперацией нервных клеток, главным образом поверхностного слоя мозговой ткани, толщиной 3-5 мм, или коры больших полушарий головного мозга. Если не считать мозжечок, то кора головного мозга, несмотря на свое периферийное положение в общем объеме мозга в черепной коробке, содержит около 90% нервных клеток всего мозга. Это почти 20 миллиардов нейронов, которые являются основной сценой, на которой разыгрывается подавляющее большинство ментальных событий нашей психики. Этим обстоятельством и, конечно, практичностью самого метода ЭЭГ, объясняется тот факт, что подавляющее большинство успешных разработок в области технологий интерфейсов мозг-компьютер (ИМК) сделано на основе анализа электрической активности головного мозга, регистрируемой непосредственно с кожной поверхности головы.

К настоящему времени можно выделить три типа технологий надежно работающих ИМК на основе регистрации ЭЭГ. Первый тип нейроинтерфейсов основан на выделении в ЭЭГ спектральных признаков, сопряженных с фокусированием зрительного внимания человека на экранных объектах, ритмически подсвечиваемых каждый на своей частоте, обычно в пределах от 5 до 30 Гц. Таким образом, детектирование внимания оператора к конкретному объекту на экране по ЭЭГ признакам становится равноценным "ментальному"клику на данную иконку. Исследования показали, что, например, для случая из 4 объектов на экране монитора можно добиться 95% правильных срабатываний алгоритма обнаружения фокуса внимания оператора с быстродействием в пределах секундного диапазона. При увеличении числа экранных объектов до 9 такой нейроинтерфейс продолжает «догадываться» о внимании оператора на уровне значительно выше случайного, но число ошибок первого и второго рода при этом может увеличиться до практически неприемлемого уровня. Именно этот тип нейроинтерфейсных контуров часто используется для «мысленного» управления, требующего всего 4-5 команд, например, для управления инвалидными колясками, модулями умных домов и т.д.

Второй тип нейроинтерфейсов также основан на выявлении внимания человека к мигающим объектам на экране монитора, но в отличие от первого в ЭЭГ выявляется не «наведенная объектом частота», а специфические реакции ЭЭГ на подсветку каждого из объектов. Нейрофизиологической основой этого метода является хорошо известный феномен появления в зрительной реакции ЭЭГ в области 300 мс после начала подсветки выраженного позитивного компонента – волны П300, если оператор ожидает подсветки конкретного объекта. В таком случае экранных объектов может быть уже несколько десятков, так как решением

статистической задачи является поиск одной реакции, отличающейся от всех остальных по характерному признаку. Таким образом, в контуре нейроинтерфейсов второго типа с большими наборами экранных объектов можно развернуть, в частности, процедуру «мысленного» набора текстов. В авторской технологии «НейроЧат», к примеру, в контуре интерфейса работают несколько матриц экранных объектов, обозначенных символами для набора текстов и пиктограммами для выполнения необходимого пользователю набора «мысленных» команд от звонка другу до управления электронной почтой и социальной сетью [2]. Практика этой процедуры показала, что для гарантированного угадывания задуманной оператором команды, например, на уровне выше 95%, необходимо, чтобы каждый экранный объект был подсвечен минимум 9-10 раз. Это приводит к существенным временным затратам, до 8-10 с на символ. Однако, если речь идет о помощи в замещении коммуникации для пациентов с тяжелыми нарушениями речи и движений, то технология «Нейрочат» оказывается вполне приемлемой. В рамках этого типа нейроинтерфейсов придуманы и более быстрые техники, до 2 секунд на символ, но они отличаются ускоренными и распределенными по экрану подсветками символов, что оказывается слишком нагрузочным для зрительного восприятия.

Наконец, третий тип неинвазивных интерфейсов интересен тем, что для обнаружения мысленной команды он не требует фокусирования внимания пользователя на внешних объектах, наоборот, ему нужно фокусироваться на конкретном мысленном образе, с тем, чтобы в ЭЭГ нашелся специфичный этому образу паттерн изменений. Однако многочисленные попытки обнаружения устойчивых паттернов ЭЭГ в ответ на мысленное представление различных образов, сделанные во многих лабораториях мира с применением самых современных алгоритмов, включая искусственные нейросети, не дали положительных результатов. Кросс-валидационное тестирование обученных алгоритмов для подобных нейроинтерфейсов давало оценки, либо крайне низкие, либо вообще на уровне случайного выбора. И лишь в случае телесных образов, таких, как воображение движений правой/левой руки, ног, лицевой мимики – всего до 5-6 образов, удавалось зафиксировать специфические этим образам паттерны ЭЭГ. При этом надежность распознавания соответствующих паттернов ЭЭГ достаточно высока, но оставляет желать лучшего: при максимально благоприятном выборе «правая-левая рука» ЭЭГ паттерны этого ментального действия в разных лабораториях распознаются с вероятностью всего 0,65-0,8, редко - 0,85. Таким образом, при всех преимуществах интерфейсов на основе мысленного воображения, т.е. без опоры на внешние объекты, шансы с помощью машинных алгоритмов анализа ЭЭГ "отгадать"по ЭЭГ даже самый удобный для

этого телесный образ из набора всего в 5-6 возможных вариантов обычно не достигает даже 65-70 на 100 попыток.

Возвращаясь к исходной задаче создания эффективного канала двусторонней связи между мозгом человека и элементами ИИ возникает вопрос о возможности реализовать такой канал с помощью разработанных к настоящему времени технологий интерфейсов мозг-компьютер. Рассмотрим эти возможности ниже.

3. Технологии интерфейсов «Мозг-ИИ»: пока еще фантазии

Во всех типах современных нейроинтерфейсов элементы ИИ уже используются, но всего лишь в качестве блока алгоритмов для создания классификаторов и для распознавания паттернов ЭЭГ, специфичных ментальным актам. При этом признаковое пространство ЭЭГ для подачи на первый слой нейросетей, как правило, формирует исследователь исходя из своих знаний и предположений о работе мозга. К примеру, наиболее частным методом свертки первичной записи ЭЭГ выбирают ее разложение на ограниченное число спектральных полос, или на компоненты вызванных стимулами реакций. Вся логика работы подобного рода нейроинтерфейсов строится на распознавании в ЭЭГ намерения человека к действию и трансляции его в команду для внешнего исполнительного устройства. В контурах таких ИМК для ИИ отводится лишь техническая роль: по определенному набору признаков научиться различать эпизоды ЭЭГ, сопряженные с заданным исследователем конкретным набором ментальных действий испытуемого.

Между тем совершенно новым и все более актуальным в последние годы становится применение ИИ в контурах ИМК не для построения классификатора паттернов ЭЭГ по известным в науке признакам этого показателя работы мозга, а в качестве активного модуля, нацеленного на оптимизацию и совершенствование самого информационного обмена между мозгом и компьютером. Здесь имеется по крайней мере две возможности, которые могут быть проверены экспериментальным путем.

Первая возможность, это оптимизация с помощью ИИ процесса порождения нового признакового пространства для ЭЭГ, адекватного не столько сумме имеющихся на данный момент нейрофизиологических знаний, сколько задаче эффективного машинного «понимания» мозговых команд. На линии с ИИ эти команды мозга будут формулироваться в контуре ИМК уже не для собственных мотонейронов, а для внешних процессорных устройств. На этом пути ресурсы ИИ должны быть направлены на поиск новых признаковых пространств ЭЭГ с опорой оце-

ночную функцию, характеризующую степень приближения к «пониманию» команды мозга. Здесь можно ожидать, что мозг, реализуя управленческую задачу человека в контуре ИМК, а также в силу своих природных пластических свойств, может пойти «навстречу» ИИ в создании оптимального признакового пространства ЭЭГ, оптимизируя этот сигнал под общую задачу.

Вторая возможность, это создание в рамках ИМК нового контура обратной связи, которая будет информировать оператора ИМК о степени приближении ИИ к распознаванию по ЭЭГ тестируемого ментального действия, например, мысленного представления движения указательного пальца правой руки. Предварительно ИИ может быть обучен, к примеру, распознаванию по ЭЭГ мысленного представления движения правой руки. В новой задаче ИИ предстоит уже без эксплицитного тренинга с учителем распознать по ЭЭГ, какой палец руки представляет в уме испытуемый. Используя генеративно-состязательные сети [3], ИИ может на основе анализа ЭЭГ на фоне ментального акта оператора ЭЭГ синтезировать свои гипотезы в виде визуальных изображений на экране монитора, используя их в качестве обратной связи. Реакцию мозга на обратную связь по шкале «горячо-холодно» тот же ИИ будет детектировать в ЭЭГ уже по итогам предварительного обучения распознавания этих универсальных состояний мозга. Таким образом организуется циклы итераций, которые в простом случае, как в приведенном примере, могут выродится в последовательную демонстрацию на экране каждого пальца руки, и в самообучении, какому пальцу какой паттерн ЭЭГ соответствует. Но даже в этом случае важен будет не достигнутый результат, а опыт ИИ в последовательном приближении к отгадке, какой ментальный акт транслируется в данный момент в ЭЭГ. Очевидным образом в таком ИМК с контуром обратной связи от ИИ репертуар ментальных актов можно расширять, тем самым не просто обогащая опыт ИИ, но в рамках того же самообучения увеличивая обобщающий потенциал ИИ для анализа ЭЭГ с целью распознавания ментальных актов при разных условиях.

Третьей возможностью реализации прямого канала интерактивной коммуникации между мозгом и ИИ является обобщение первых двух возможностей при длительном обучении ИИ тактикам и стратегиям распознавания все большего числа ментальных актов в контуре ИМК с обратной связью. Это будет путь к появлению собственного опыта ИИ для все более эффективной идентификации ментальных образов конкретного человека, выраженных на языке специфических паттернов ЭЭГ. На этом пути можно ждать взрывного расширения канала интерактивной коммуникации «Мозг-ИИ», поскольку каждый новый образ будет расшифровываться по ЭЭГ уже с учетом накапливающегося потенциала

предсказательной функции машинного опыта. Этот машинный опыт, по сути дела, может стать той истинной нейроморфной составляющей ИИ, которая будет нести в себе не столько структурное и функциональное сходство с нервной тканью, сколько саму семантику отношений ментальных объектов, проявляющихся в ЭЭГ конкретного человека. ИИ с семантическим нейроморфным модулем, сформированным в результате длительного «общения» с данным человеком в контуре ИМК с обратными связями, станет в буквальном смысле человеко-ориентированным ИИ.

Удачная реализация предложенных возможностей может привести созданию комплексов «Мозг-ИИ», способных автоматизированным образом формировать мозг-машинный язык, по крайней мере, в границах какого-то пула ментальных активностей. Насколько развиваемым с помощью технологии «Мозг-ИИ» может оказаться этот мозг-машинный язык и как далеко он заведет ИИ на пути формирования человекоподобного машинного опыта покажут мультидисциплинарные исследования в этой области.

Список литературы

[1] Бердяев М.А., "Человек и машина (Проблема социологии и метафизики техники)", *Путь*, **38** (1933), 3–38.

[2] Ганин И.П. и др., "Набор текста пациентами с постинсультной афазией в комплексе "НейроЧат" на основе технологии интерфейсов мозг-компьютер на волне P300", *Журнал высшей нервной деятельности им. И. П. Павлова*, **70**:4 (2020), 435-445.

[3] Goodfellow I.J., Pouget-Abadie J., Mirza Mehdi, Xu Bing, Warde-Farley D., Ozair Sh., Courville A., Bengio Yo., "Generative Adversarial Networks", 2014, arXiv: 1406.2661.

Brain-artificial intelligence interfaces: foundations and perspectives

Kaplan A.Y.

The human brain is obviously a "generic" object for determining, including artificial "intelligent systems". However, the mechanisms of the actual intellectual activity of the brain still remain undiscovered in their most essential part. To a large extent, this is due to the also obvious fact that natural intelligence in a broad sense has the property of a subjective representation of reality, the operational architectonics of which, its formats and completeness of description are unknown to us. In this regard, and taking into account the absence of a "general theory of the brain", it becomes problematic to create neuromorphic artificial intelligence systems based on analogies. The

level of such “neuromorphism“ will be entirely determined by the boundaries of our understanding of how the human brain works. Meanwhile, the neuromorphic property of artificial intelligence systems can be achieved not so much by building brain-like structural and functional architectonics of these systems (artificial neuron networks, deep learning algorithms, etc. D.), how much through their training of two-way communication with natural intelligence. We believe that the channels of brain-artificial intelligence information interaction for launching the learning processes of “communication“ can be built on the basis of technologies of non-invasive brain-computer interfaces (BCI) of a new generation. The key link of these technologies will be feedback loops, within which artificial intelligence modules connected to EEG analysis will demonstrate to a person through natural sensory channels their hypotheses regarding the belonging of specific EEG patterns to certain mental acts of a person. In the interactive process, these hypotheses will be refined to full convergence, thus forming interaction trajectories in the memory of the human and machine brains to an acceptable level of “mutual understanding“. The report will discuss in detail the neurophysiological and neurotechnological grounds for creating “brain-artificial intelligence“ interfaces.

Keywords: human brain, electroencephalogram, brain-computer interfaces artificial intelligence, neuromorphic systems.

References

[1] Berdyayev M.A., “Man and Machine (The Problem of the Sociology and Metaphysics of Technology)”, *Way*, **38** (1933), 3–38 (In Russian).

[2] Ganin I.P. et al, “Typing by patients with post-stroke aphasia in the NeuroChat complex based on the technology of brain-computer interfaces on the P300 wave”, *Journal of Higher Nervous Activity named after I. P. Pavlov*, **70**:4 (2020), 435-445 (In Russian).

[3] Goodfellow I.J., Pouget-Abadie J., Mirza Mehdi, Xu Bing, Warde-Farley D., Ozair Sh., Courville A., Bengio Yo., “Generative Adversarial Networks”, 2014, arXiv: 1406.2661.