ИПМ им.М.В.Келдыша РАН

Будущее-2024 • Труды конференции

Д.В.Журавлёв, В.С.Смолин

Проектирование структуры нейросетей для AGI

Рекомендуемая форма библиографической ссылки

Журавлёв Д.В., Смолин В.С. Проектирование структуры нейросетей для AGI // Проектирование будущего. Проблемы цифровой реальности: труды 7-й Международной конференции (15-17 февраля 2024 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2024. — С. 125-143. — https://keldysh.ru/future/2024/2-3.pdf https://doi.org/10.20948/future-2024-2-3

Проектирование структуры нейросетей для AGI

Д.В. Журавлёв[1], В.С. Смолин[2]

[1]ООО ЦИФРОМЕД
[2]Институт прикладной математики им. М.В. Келдыша РАН

Аннотация. Современные нейронные сети демонстрируют впечатляющие успехи в аппроксимации сложных многомерных преобразований и генерации реалистичных сигналов, таких как изображения, видео, тексты и речь. Они способны воображать новые ситуации, что говорит о потенциале для развития общего искусственного интеллекта (AGI), отличающегося от узкого искусственного интеллекта тем, что может самостоятельно находить новые знания без прямого участия человека, основываясь на анализе входных данных и уже имеющихся знаний. Создание AGI требует разработки нейронных структур способных к самостоятельному пониманию задачи, разбиения ее на компоненты и формулированию промежуточных целей. В статье представлены идеи по конструированию нейросетевых структур AGI, способных получать новые знания в сложных средах.

Ключевые слова: нейросетевые алгоритмы, AI, декомпозиция, AGI.

Neural network structure design for AGI

D.V. Zhuravlev[1], V.S. Smolin[2]

[1]CIFROMED LLC
[2]RAS Keldysh Institute of Applied Mathematics

Abstract. Modern neural networks have shown impressive success in approximating complex multidimensional transformations and generating realistic signals such as images, videos, texts, and speech. They can imagine new situations, indicating the potential for the artificial general intelligence (AGI) development, which differs from narrow artificial intelligence in that it can independently find new knowledge without direct human involvement, based on the analysis of input data and existing knowledge. The AGI creation requires the neural structures development capable of independent task understanding, breaking it down into components and formulating intermediate goals. This paper presents ideas for designing AGI neural network structures capable of acquiring new knowledge in complex environments.

Keywords**:** neural network algorithms, AI, decomposition, AGI.

1. Введение

За последние 10-15 лет в области искусственного интеллекта произошла нейросетевая революция – нейросетевые алгоритмы все чаще заменяют эвристические методы решения задач. Глубокие нейросетевые структуры успешно применяются для решения различных интеллектуальных задач, во многом превосходя даже человека [1-5].

Появление большого количества крупных языковых моделей, созданных на основе GPT, стало основой для решения разнообразных задач искусственного интеллекта [6]. Нейросетевой подход в AI активно развивается и крупные компании активно внедряют его в свои проекты.

Хотя нейросетевые идеи расширяются, финансовая поддержка для новых подходов пока остается ограниченной. Ряд инновационных идей, высказанных ранее, пока не нашли широкого применения в коммерческих проектах [7-9]. Развитие глубокого обучения идет рука об руку с привлечением новых идей, а успехи нейросетевого AI уже создали основу [10;11] для создания сильного искусственного интеллекта (AGI).

Крупные исследовательские центры, такие как OpenAI, DeepMind, Anthropic PBC и другие [12;13], уже сегодня активно работают над созданием AGI. Технологически человечество готово к этому шагу [14], но для успешной реализации нужно понимание нерешенных задач [15] и способов их решения.

2. Наследие XX века

Системный анализ, теория самоорганизации, синергетика и другие области занимаются исследованием законов функционирования сложных систем, которые есть во всех сферах человеческой деятельности. За последние 100 лет было разработано много ценных идей в области теории сложных систем, работы по которой можно найти в множестве публикаций [например, 16-22].

В России в начале XXI в. большие ожидания возлагались на концепцию НБИКС, которая предполагала сотрудничество различных научных дисциплин, включая нанотехнологии, биотехнологии, информационные технологии, когнитивные науки, социологию и другие, с целью создания инновационных и передовых технологий (рис. 1).

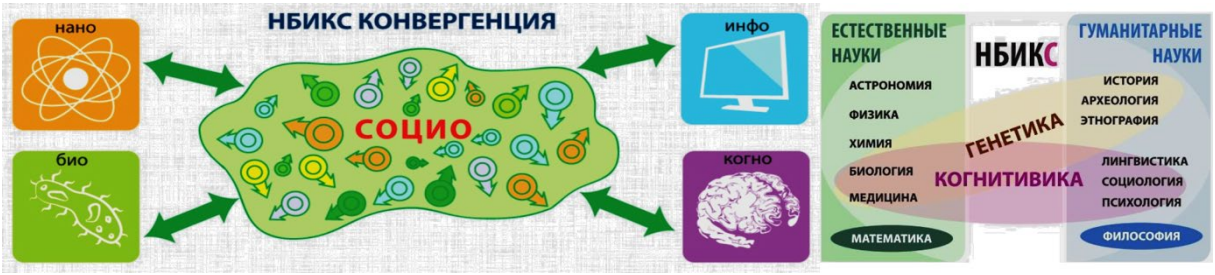

Рис. 1. НБИКС – нано-био-инфо-когно-социо конвергенция [23]

Прогресс цивилизации строится на приобретении и применении новых знаний, которые способствуют более эффективному обмену и использованию информации, а не только на информационных технологиях (IT), синергетике и концепции НБИКС. Долгое время считалось, что только человек способен усваивать знания, и исследование этого процесса рассматривалось как нематериальный аспект механизма познания. Изучение внутренних механизмов познавательного процесса оставалось за пределами возможностей исследователей.

Появление компьютеров и развитие вычислительной техники открыли путь к преодолению сложностей. Еще в далеком 1673 г. Готфрид Вильгельм Лейбниц предсказал, что счетные машины могут достичь такого уровня, что смогут заменить даже живых судей. Но только сейчас, с развитием технологий, некоторые задачи, такие как решения о кредитах или выписывание штрафов за нарушения ПДД, удалось автоматизировать.

Создание и развитие глубоких нейронных сетей подготовлено историей цивилизации. Многие ученые видят в нейросетевой революции в области машинного обучения естественное направление развития AI [24, 25].

Важно находить баланс между использованием AI, AGI и человеческими способностями, так как объединение их усилий должно стать мощным катализатором для ускорения прогресса общества. Для принятия обоснованных решений важно понимать процесс создания AGI, оценивать его потенциал на основе фактов, а не рефлексии, сформированной на философии Платона и Канта.

3. Разнообразие, энтропия, свободная энергия и информация

В середине XX в. появление вычислительной техники и средств передачи цифровых данных привело к развитию теорий информации и управления [26;27], которые нашли применение в технике, управлении и коммуникациях. Биолог У.Р. Эшби первым начал рассматривать математические аспекты управления и передачи информации [28].

Сохранение энтропии тела на низком уровне, несмотря на второй закон термодинамики, объяснялось идеей «свободной энергии», поддержанной такими учеными, как Э. Шрёдингер [29] и Э.С. Бауэр [30] и др. [31]. Однако сам факт необходимости использования энергии не дает ответа на вопрос, почему некоторые системы считаются живыми, а другие нет.

Закон необходимого разнообразия У. Эшби утверждает, что для перехода системы в менее вероятное состояние требуется как «свободная энергия», так и информация о способе этого перехода. Управляемые действия, основанные на знаниях, могут перевести систему в менее вероятные состояния. Снижение уровня энтропии системы через управление, основанное на знаниях, зависит от количества необходимой информации и разнообразия состояний системы.

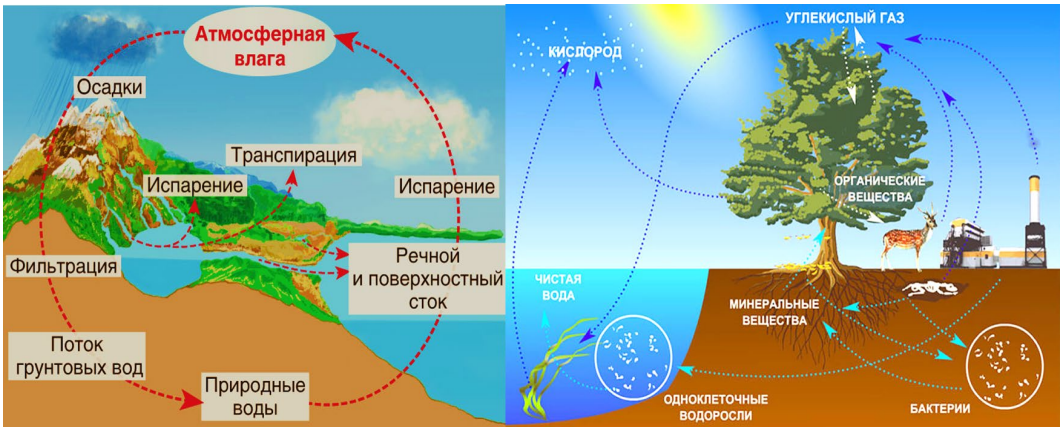

Рис. 2. Круговорот веществ: вода, кислород и многие другие вещества в результате притока «активной энергии» изменяют свои состояния и их энтропию, которая (в цикле) как возрастает, так и уменьшается

«Закон необходимого разнообразия» Н. Эшби утверждает, что главное отличие живых организмов заключается не в борьбе с увеличением энтропии и использовании «свободной энергии», а в постепенном увеличении сложности их структуры, а также в передаче потомкам информации о методах поддержания и воспроизводства жизни.

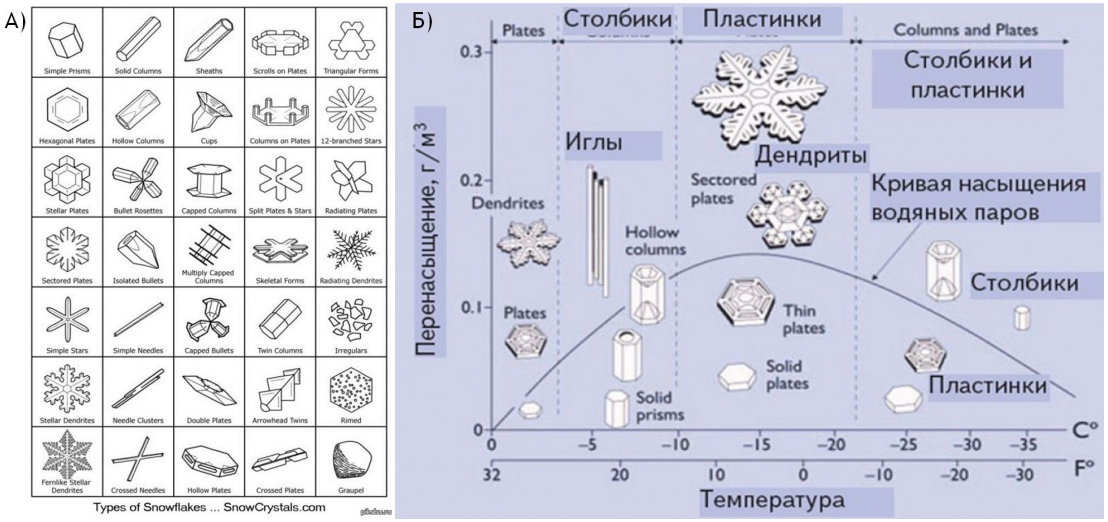

Рис. 3. Снежинки похожи на живые организмы: они в больших количествах воспроизводятся и гибнут в природе; бывают разных видов (несколько десятков, А)); в пределах каждого вида имеется разнообразие; вид «взрослой» снежинки зависит от условий ее роста (Б)) [32]

В природе многие процессы цикличны (рис. 2). Воспроизводство, жизненный цикл и сложная структура являются важными признаками живого, но их можно наблюдать и в неживой природе.

Кристаллизация паров воды в атмосфере в виде снежинок – удивительный процесс неживой природы, который демонстрирует разнообразие

и симметрию форм шестилучевых снежинок (рис. 3). Снежинки не накапливают знаний и не наследуют форм предшественников, и именно поэтому являются не живыми. В процессе образования снежинок реализуются различные свойства самоорганизации, такие как снижение энтропии, присутствие активной энергии, бифуркации, цикличность. Однако снежинки не эволюционируют и не воспроизводятся с передачей знаний следующим поколениям, в отличие от живых организмов, которые накапливают знания в виде генетической информации (ДНК и РНК), воспроизводятся с передачей формы родителей и эволюционируют. Живые организмы размножаются путем деления с использованием накопленных знаний, что делает их способными к постоянному развитию и изменениям в течение эволюции.

4. Главное свойство и этапы эволюции

Эволюция основана на выживании наиболее приспособленных организмов. Для того чтобы выжить необходимо не только быть приспособленным, но и уметь быстро адаптироваться к изменяющимся условиям. В ходе эволюции было разработано множество механизмов для ускорения приспособления, таких как контроль темпа мутаций и половое размножение. Появление нервной системы у животных и развитие цивилизации человечеством являются ключевыми этапами, которые значительно увеличили скорость адаптации к переменным условиям. Оба эти этапа расширили возможности для разнообразного поведения и скорости его изменения.

Человек, как результат эволюции, превосходит животных способностью строить рациональное поведение с более точным учетом свойств сложившейся ситуации и прогнозом особенностей ее развития. Однако без строительства цивилизации человек не смог бы достичь доминирования над животным миром. Цивилизация представляет собой новый этап в развитии эволюции благодаря резкому увеличению разнообразия, обусловленному совершенствованием нервной системы. Использование инструментов, одежды и других технологий значительно расширило возможности и темпы адаптации человека к разнообразным и изменяющимся условиям. Особенно важным элементом стало возникновение и использование артикулированной речи, которая позволила передавать знания и организовывать коллективные действия, способствуя углубленному разделению труда и накоплению новых знаний.

Возможна замена человека во многих аспектах цивилизационной деятельности, но принятие решений и освоение новых знаний остаются в основном в компетенции людей. Предполагается, что в будущем проектирование автоматизированных систем будет поручено AGI, способному получать необходимые знания из наблюдений за окружающей средой.

Системы, способные понимать свое устройство и самостоятельно проектировать еще более совершенные системы с улучшенными возможностями – по М. Тегмарку [33] – составят новое звено эволюции, Жизнь 3.0.

И именно системы, способные к проектированию улучшений своей конструкции, по его мнению, и следует считать AGI. Так же как человек вышел из животного мира, так и первые модели Жизни 3.0 возникнут не из ниоткуда, а будут спроектированы и созданы человеком на основе понимания принципов AGI. Ниже мы постараемся показать, что центральными принципами построения AGI (на наш взгляд) являются:

1. Способность к получению новых знаний на основе наблюдений сложного мира путем его декомпозиции на простые (доступные для статистически достоверного изучения) объекты и явления;

2. Возможность использовать имеющиеся знания о простых компонентах окружающего мира для моделирования вариантов развития ситуаций при выполнении различных действий с целью сравнения и выбора лучших последовательностей действий;

3. Задание потребностей, как критерия сравнения при выборе разных локальных и глобальных целей, исходя из способностей соответствовать заложенным в AGI потребностям.

Потребности не должны ограничиваться сохранением целостности и работоспособности системы и даже стремлением к красоте и гармонии. Они должны обеспечивать не только желание развивать новую цивилизацию, но и обеспечивать ее совместимость с существующей.

У людей и животных потребности заложены генетически и проверены эволюцией. Это одна из причин необходимости сосуществования людей и AGI, у которых потребности еще предстоит отладить и проверить. Люди научились обманывать свои центры удовольствия (отвечающие за потребности) разными способами, но здоровые индивидуумы, как и животные, стремятся сохранить свое (и потомков) участие в эволюции.

Появление возможности замены ЛПР (лиц, принимающих решения) специализированными системами на основе AGI откроет новый этап в развитии цивилизации. Это позволит создавать оптимизированные системы для всех важных процессов и средств распространения знаний. Благодаря своей специализации, разнообразию и высокой производительности эти системы смогут превзойти человека в управлении сложными объектами. Разнообразие систем AGI будет сдерживаться выгодами массового производства, но будут учитываться индивидуальные запросы заказчиков. Эта замена будет аналогична эволюции транспорта, где разнообразие видов транспорта заменило старые способы перемещения (на своих двоих...).

Важной особенностью AGI будет возможность прямого копирования знаний между системами без потерь. Это позволит эффективно использовать накопленные знания цивилизации в практической форме. В отличие от человека, который передает знания через слова и учебники, системы AGI смогут передавать информацию непосредственно и в форме, понятной для практического применения. Главное – это обеспечит новый уровень доступности цивилизационных знаний для формирования действий.

5. Идеи для конструирования AGI

Современные успехи нейросетевого ИИ создают условия для формулирования идей, направленных на создание агентов AGI, которые сами будут способны выдвигать подобные идеи. Наше определение AGI мы дадим ниже, в п. 5.8, после рассмотрения проблем его построения и путей их решения. Главной проблемой, на наш взгляд, является обеспечение быстрого получения новых знаний о сложной среде на основе имеющихся знаний.

Чтобы было понятно, о чем идет речь, воспроизведем приводимые в [14] определения данных, знаний и информации.

5.1. Данные, знания и информация

Это определение отличия знаний от данных и информации – качественное, количество информации – строго по Шеннону: В бинарной форме (нулей и единиц) можно представить и данные, и знания, и информацию (и измерить количество). Но будем их различать:

– *Данные (D) – описания объектов, явлений и действий в любом виде;*

– *Знания (K) – описания объектов, явлений и действий с ними в виде, согласованном со свойствами системы и пригодном для формирования действий сложных систем;*

– *Информацию (I) – описания, позволяющие выбрать знания для формирования действий.*

Знания извлекаются из данных и помещаются в память, пополняя и корректируя систему знаний, имеющихся в памяти. Также знания могут быть реализованы в структуре продукта цивилизации или живого организма, причем структура без памяти может содержать знания, а память, без структуры, способной использовать ее содержимое – это не знания, а данные.

Информация – тоже извлекается из данных. Информация позволяет только выбрать вид преобразования входных данных в выходной сигнал, хотя эту операцию можно отнести к части процесса формирования действий и все описания операций с данными называть знаниями (или информацией). Но полезнее иметь возможность различать информацию и знания.

Важно использовать методы проектирования при создании систем AGI, отказавшись от мифа о «невозможности осмысления путей реализации психических функций человека» и стремясь к созданию эффективных условий для появления новых свойств. Следует развивать конструктивные идеи, специализировать направления исследований, объединять результаты и использовать методы проектирования для успешного построения AGI.

5.2. Декомпозиция и нелинейная композиция

Все нейросетевые алгоритмы осуществляют аппроксимацию нелинейных преобразований $\vec{Y}(t) = F(\vec{X}(t))$, входного сигнала (наблюдений) $\vec{X}(t)$ в выходной сигнал (действия) $\vec{Y}(t)$. Настройка аппроксимации осу-

ществляется изменением весов связей. Именно за счет весов связей формируется выходной сигнал (действия) и в данном случае веса связей и есть (согласно определению (К)) знания, содержащиеся в нейросети. А ошибка аппроксимации определяет степень достоверности имеющихся знаний.

Для получения низкой ошибки аппроксимации (статистической достоверности) необходимо дать обучающие примеры нелинейного преобразования $\vec{Y}(t) = F(\vec{X}(t))$, количество которых в целом экспоненциально зависит от числа существенных переменных, описывающих преобразование.

Проблема описания сложного мира схематически представлена рис. 4. Пусть сложный векторный сигнал С формируется как сумма двух независимых компонент А и В. Если для статистически достоверного описания каждой компоненты (А и В) нужно по 1 млн разных примеров, то для сложного сигнала С, содержащего А и В в произвольной комбинации, потребуется триллион примеров. Если есть возможность обработать 10 млн примеров, то задача покомпонентного описания вполне выполнима, а статистически достоверное изучение сложного сигнала С – недоступно.

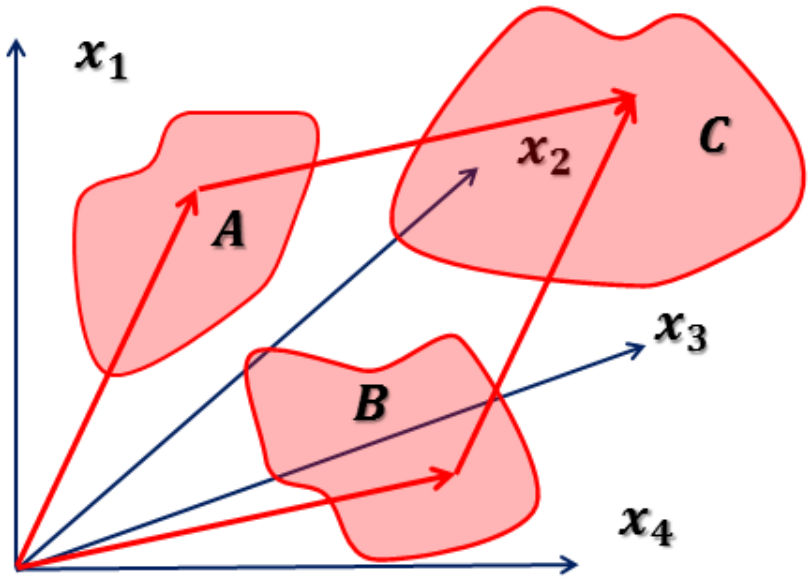

Рис. 4. Сложный векторный сигнал как сумма двух простых

Если компоненту А можно изучать отдельно, а компонента В доступна для наблюдения только вместе с А, то знание свойств А позволит при имеющихся возможностях построить и статистически достоверное описание свойств компоненты В. Если С = А + В, то В = С − А, и можно набирать статистику свойств компоненты В независимо от А. При наблюдении С состояния В будут иногда повторяться и потребуется не миллион, а возможно 2-3 млн примеров, что, даже в сумме с примерами для выявления свойств компоненты А, несравненно меньше триллиона примеров, необходимого для статистически достоверного изучения свойств С.

Число возможных состояний сложного сигнала будет равно произведению числа состояний каждой компоненты. С ростом числа компонент количество состояний сложного сигнала растет экспоненциально, а количество суммы состояний отдельных компонент – практически линейно:

$$\prod_{j=1}^{M}\prod_{i=1}^{N_j} a_{ji} \gg \sum_{j=1}^{M}\prod_{i=1}^{N_j} a_{ji} \text{ при } a_{ji} > 1 \quad (1)$$

где a_{ji} – параметры, характеризующие необходимое число примеров вдоль каждой из существенных переменных.

В этом случае особенно важны знания про компоненты сложного сигнала. Если разделить сложный сигнал на 2 части – известные компоненты $A_j = \prod_{i=1}^{N_j} a_{ji}$ и неизученные компоненты $B_l = \prod_{i=1}^{N_j} a_{ji}$, то использовать знания про A_j на основе формулы

$$\sum_{l=1}^{N_b} B_l = C - \sum_{j=1}^{N_a} A_j \quad (2)$$

для статистически достоверного изучения свойств неизвестных компонент возможно только в случае, если неизученная часть сигнала C является относительно простой. И чем больше компонент A_j нам уже известно, тем вероятнее ситуация, что сумма компонент B_l окажется доступной для реализации статистически достоверного выявления их свойств.

Сегодняшние успехи LLMs and Foundation models основаны на использовании содержащихся в текстах информации о человеческой декомпозиции сложных сцен. Использование уже готовой декомпозиции эффективно, но формировать собственные правила декомпозиции на основе наблюдений без участия человека клоны GPT пока не научились.

Чтобы агент AI стал AGI, он должен уметь полностью самостоятельно осуществлять декомпозицию сложных сигналов.

Декомпозиция сложных сигналов может быть основана на том наблюдении, что, как правило, в их формировании участвует несколько относительно простых (доступных для достоверного статистического изучения) источников. То есть сложный сигнал $\vec{X}(t)$ представляет собой некоторую сумму сигналов $\vec{X}_i(t)$ от простых источников.

$$\vec{X}(t) = \sum_{i=1}^{N} \vec{X}_i(t) \quad (3)$$

Если бы сумма всегда была линейной, то декомпозиция осуществлялась бы просто. Имея статистически достоверные описания компонент сложного сигнала $\tilde{\vec{X}}_i(t)$, можно было из сложного сигнала вычитать сумму модельных компонент $\sum_{j\neq i}^{N} \tilde{\vec{X}}_j(t)$, отличных от выделяемой компоненты. Модельные компоненты бы, в свою очередь, воспроизводились бы на основании их достоверных описаний так, чтобы минимизировать разность:

$$\vec{X}(t) - \sum_{i=1}^{N} \tilde{\vec{X}}_i(t) \to min \quad (4)$$

Выбор компонент векторных сигналов высокой размерности (не все известные компоненты должны участвовать в описании сложного сигнала) возможен корреляционными методами и не представляет особой сложности. Невозможность статистического анализа сложных сцен как целого приводит к необходимости их разложения на простые составляющие, для которых статистический анализ возможен. Но для восприятия сложных сигналов необходимо сопоставлять их с суммой простых компонент, запомненных ранее. То есть можно помнить не только трансформацию $\vec{X}_i \to \vec{Y}_i$, но и $\vec{X}_i \to \tilde{\vec{X}}_i$, где $\vec{Y}_i$ и $\tilde{\vec{X}}_i$ это выходной и копия входного вектора, соответствующие *i*-той компоненте входного сигнала $\vec{X}$. Это позволяет выделить компоненты $\vec{X}_i$ из сложного сигнала $\vec{X}$:

$$\vec{X}_i(t) = \vec{X}(t) - \sum_{j \neq i}^{N} \tilde{\vec{X}}_j(t) \quad (5)$$

В целом схема выделения компонент $\vec{X}_i(t)$, доступных для статистической обработки, показана на рис. 5.

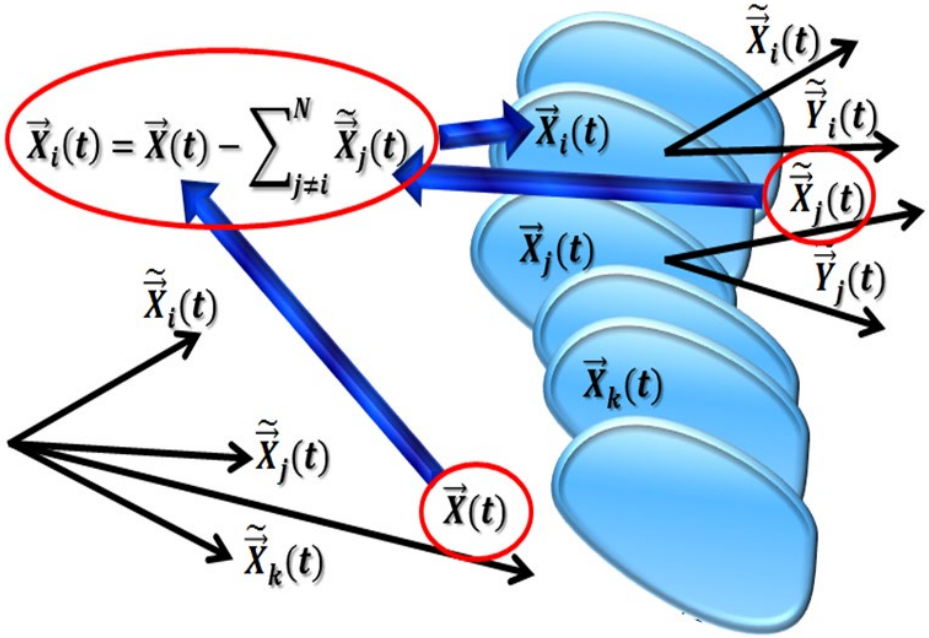

Рис. 5. Структура линейного разложения на компоненты $\vec{X}_i(t)$ для составного векторного входного сигнала $\vec{X}(t)$

Тот факт, что суммация компонент при восприятии сигналов из сложной среды часто носит нелинейный характер, в значительной степени затрудняет их декомпозицию. Но именно нейросетевые алгоритмы и являются средством для преодоления этих трудностей. Это демонстрируют

проанализированные в [14] успехи генеративных нейросетей. Другие идеи разделения входного сигнала высказывались в работах [8-10], но пока они не получили развития. Направленная декомпозиция сложных сигналов возможна при наличии моделей компонентов сложных сигналов.

5.3. Понимание ситуации

Для сопоставления накопленных знаний с компонентами сложной среды для их правильного применения можно использовать формулы (3)–(5). Критерием того, что сопоставление является правильным, может служить постоянство набора индексов i в (4) при изменении $\vec{X}(t)$ в течение времени. Иногда одни объекты и явления могут появляться, а другие выходить из зоны внимания. Но если при небольших изменениях ситуации $\vec{X}(t)$ состав ее компонент сохраняется прежним, то велика вероятность того, что сопоставление наблюдаемых объектов с имеющимися знаниями о них было осуществлено правильно, есть понимание какие объекты и явления составляют сложную среду, в которой осуществляются действия.

Другим аспектом понимания является контроль, сравнение реального развития событий с прогнозом. Хорошее совпадение говорит о правильном выборе действий, достаточности имеющихся знаний для понимания развития ситуации. Иначе лучше проверить правильность идентификации компонент и/или выбрать другие действия и/или сменить цель действий.

5.4. Иерархическая структура знаний – не только о простых объектах и явлениях

Для изучения простых объектов и явлений необходимо повторение их состояний и возможность статистического анализа свойств. Сложная среда характеризуется большим числом простых элементов, вероятность повторения ее состояний стремится к нулю. Если компоненты сложной среды действуют независимо, знание о них по отдельности достаточно для управления каждым. Но чаще компоненты взаимодействуют и нам не удается собрать статистику всех возможных вариантов взаимодействий.

Некоторые взаимодействий всё-таки повторяются и это дает возможность получить статистически достоверные знания. Наблюдения за взаимодействиями позволяют создать многоуровневую иерархическую структуру знаний, описывающую как свойства простых элементов, так и известные типы их взаимодействий. С повышением уровня иерархии структуры возрастает число вариантов взаимодействий между компонентами, но мы можем выявить лишь некоторые из них из-за ограничений наблюдений.

Иерархическое представление знаний о сложной среде позволяет осуществить декомпозицию на разные уровни взаимодействия, что упрощает получение статистически достоверных описаний. Уровни иерархии определяют контроль действий с различными характерными временами достижения целей, на верхних уровнях эти времена больше. При этом нейросети разных уровней иерархии имеют одинаковые структуры, что

позволяет моделировать действия по достижению целей с одинаковой скоростью вне зависимости от уровня иерархии.

Моделирование последствий действий в неизвестных ситуациях помогает компенсировать неполноту знаний и предсказать их результаты. Моделирование можно сравнивать с мышлением, наблюдаемым не только у людей, но и у высших животных. Количество компонентов и ограничения системы влияют на сложность процесса моделирования, однако выделение внимания на ключевые компоненты может упростить этот процесс.

5.5. Распределение ресурсов

Так как знания, которые используются для моделирования ситуации, принципиально не являются полными, необходимо постоянно сравнивать реальное развитие событий с прогнозом. Если появляются заметные расхождения, то следует изменить способ достижения цели или саму цель. При управлении простым объектом можно напрямую, путем аппроксимации на основе имеющихся знаний выбрать другую цель управления, более соответствующую изменившимся условиям. Это наиболее быстрый процесс, но он не может учесть всех особенностей состояния сложной среды, не доступной для статистического анализа, в которой он происходит.

Моделирование последствий различных вариантов действий позволяет значительно улучшить степень соответствия выполняемых действий текущим обстоятельствам. Но, хотя моделирование происходит значительно быстрее, чем выполнение действий и не требует больших затрат, невозможность в большинстве случаев проанализировать все варианты на моделях приводит к ограничению числа рассматриваемых вариантов действий.

Наличие двух режимов использования имеющихся знаний в применении к сложившейся ситуации приводит к необходимости распределения вычислительных ресурсов под эти режимы обработки. В нейросетях знания представлены в виде значений весов связей и распределены по структуре, и в живых нейросетях (где нет возможностей цифрового копирования) скорее всего оба режима реализуются на одних и тех же структурах, с разделением по времени. Если нижние уровни иерархии справляются с выполнением поставленных им целей, то верхние уровни могут быть использованы для анализа различных вариантов развития действий. Это возможно при выполнении простых действий, например, лежа на диване или гуляя по парку. В сложных ситуациях, например, при участии в подвижных играх или переходе оживленной улицы, процесс моделирования вариантов затруднен, поскольку все ресурсы иерархической структуры используются на частую смену целей просто на основе аппроксимации.

В формальных нейросетях возможно цифровыми методами скопировать знания и разделять их обработку в разных режимах пространственно. Но и в этом случае придется разделять по времени режимы использования знаний, т.к. перед выполнением сложных действий полезно провести сравнительное моделирование возможных вариантов. Перебор большего числа

вариантов ведет к выбору более рационального способа действий. Поскольку в сложной среде вариантов обычно очень много, а время на моделирование ограничено, есть необходимость искать компромисс между рациональностью выполняемых действий и сроками их подготовки.

Компромисс основан на том, что в сложной среде формирование действий не может быть глобально оптимальным (всегда есть возможности улучшения). Постоянно приходится выбирать между быстрым (только на основе аппроксимации) выполнением действий и отложенным по времени их выполнением более близкого к оптимальному варианта, полученного на основе выбора путем сравнительного моделирования. Механизм нахождения баланса при решении данного компромисса может быть основан на прогнозных оценках выигрыша/потерь при выборе режимов использования имеющихся знаний. Если сложные прогнозы лучше строить на основе моделирования, то оперативный (сознательный) контроль переключения режимов – только путем аппроксимации оценок в приложении к ситуации.

5.6. Новые знания на основе уже имеющихся знаний

Рассмотренные в п. 5.2 (и на рис. 4) способы получения новых знаний на основе декомпозиции являются только возможностью, использование которой становится рациональным, когда новые знания могут помочь сформировать более эффективные действия. Для выявления такой рациональности необходимо наличие описанного в п. 5.3 понимания ситуации, которое тоже основано на имеющихся знаниях, также полученных с использованием декомпозиции сложной среды.

Хотя получение новых знаний возможно и в «фоновом» режиме, просто путем наблюдения за различными объектами и явлениями, более эффективным является целенаправленное поведение по получению недостающих знаний. Для этого может быть использована и «животная» функция внимания, и «цивилизационные» методики исследований.

Но и «фоновое» получение новых знаний тоже всегда присутствует, что часто приводит к «озарениям» – новые (не целенаправленно полученные) знания помогают сформировать необходимые действия.

5.7. Психические функции и активность нейронов

Существует глубокий разрыв между описаниями активности нейронов и представлений о психических функциях, на основе которых реализуется высшая нервная деятельность (ВНД). Современные успехи нейросетевого ИИ, который демонстрирует всё возрастающие «интеллектуальные» возможности, начинают уменьшать этот разрыв. Но пока принято гордиться тем, что мы знаем свойства условных рефлексов и не знаем, зачем нам нужны такие психические функции. Описанные выше проблемы получения знаний в сложной среде позволяют объяснить назначение психических функций при построении поведения:

– *Понимание*, как восприятие сложной среды в виде устойчивого набора (относительно) простых объектов и явлений необходимо для постро-

ния в ней рационального поведения и наличия знаний про возможные действия с воспринимаемыми объектами.

– *Интуиции* в рассмотренной модели соответствует использование аппроксимации для преобразования сенсорных данных и цели в действия. От рефлексов интуиция отличается учетом знаний, содержащихся в верхних уровнях иерархии описаний простых объектов и явлений и их взаимодействий.

– *Мышление* рассматривается нами как процесс, включающий в себя генерацию разных вариантов развития сложившейся ситуации. Невозможность хорошей аппроксимации действий в неповторяющихся ситуациях, возникающих в сложной среде, ведет к необходимости проведения сравнительного моделирования различных вариантов действий.

– *Сознание* отражает контроль ситуации, действий и мышления, который должен обеспечивать переключение работы нейросетевых структур в соответствующие обстоятельствам режимы. В сложной среде следует с учетом изменяющейся ситуации выбирать между возможностью продолжения мышления и выполнением действий.

5.8. Определение AGI

Мы считаем неудовлетворительным определение AGI как «тип AI, который соответствует человеческим возможностям или превосходит их в широком спектре когнитивных задач» [35], поскольку и человеческие возможности меняются, и «широкий спектр когнитивных задач» тоже строго не определен. А автоматизация решения класса задач, считавшихся «когнитивными» приводит к исключению этих задач из списка когнитивных.

Развиваемый Беном Герцелем подход к определению AGI, как «системы, способной решать сложные задачи в сложной среде за счет использования ограниченных ресурсов» страдает расплывчивостью. Задачи, решаемые сейчас, были недостижимо сложными еще в прошлом веке, не говоря про более древние времена. Уровень «сложности» среды не определен. Ограниченные ресурсы можно трактовать как антитеза к философскому понятию «неограниченные ресурсы». На практике ресурсы всегда ограничены и уровня «ограниченности» в определении не предлагается.

Одно из наиболее конструктивных определений AGI дано астрофизиком М. Тегмарком в [33]. Он определяет AGI как «Жизнь 3.0», то есть такие организмы, агенты, системы, которые не просто могут жить в сложном мире (как Жизни 1.0 и 2.0), но и способны использовать имеющиеся у них знания для совершенствования своей конструкции. Человечество, с расшифровкой кода ДНК, разработкой методов его редактирования [36;37] приблизилось к созданию углеродной «Жизни 3.0», что подчеркивает расплывчатость определений AGI через возможности человека и сложность решаемых задач. Определение AGI у Тегмарка свободно от этих недостатков, оно дает более конкретный критерий создания AGI (если «Жизнь 3.0»

будет не белковой). Впрочем, при глубоком редактировании генома, белковая жизнь тоже будет не совсем естественной…

Мы в целом согласны с определением Тегмарка и не считаем, что отсутствие указания на пути создания «Жизни 3.0» в этом определении является его недостатком. Критерий не обязан включать способы достижения указанных в нем условий (их не содержится и в других определениях).

Поскольку выше были рассмотрены трудности построения поведения в сложной среде, мы готовы дать более конструктивное определение AGI, которое конкретизирует некоторые свойства, необходимые для реализации кремниевой или белковой «Жизни 3.0». Для соответствия понятию AGI, система, по нашему мнению, должна удовлетворять свойствам:

– *Способность к освоению имеющихся и получению новых (не следующих из известных согласно доступным методикам) знаний и использованию их для задач развития цивилизации.*

– *Способность к выделению и описанию свойств простых компонент из сигналов, поступающих из сложной среды, основанной на адаптивной иерархической структуре представления знаний,*

– *Развитость адаптивной иерархической структуры представления знаний должна позволять работать с высокоуровневыми цивилизационными знаниями.*

Эти свойства не противоречат определению «Жизни 3.0», разве что «уровень цивилизационных знаний» должен будет еще немного подрасти.

5.9. AGI и развитие цивилизации

Отсутствие уверенности в способности разработчиков внедрить в AGI потребности, направленные на развитие цивилизации с учетом интересов людей, вызывает опасения в безопасности совместного развития людей и AGI в рамках одной цивилизации. Эти опасения объединяются под термином «алармизм». Некоторые исследователи, такие как С. Рассел и М. Тегмарк, предлагают оптимистичные сценарии взаимодействия человечества с AGI, подчеркивая необходимость контроля над процессом создания AGI, чтобы сделать его «совместимым» с человеческой цивилизацией.

Идея «совместимости» звучит гуманистически и направлена на сохранение развития цивилизации и эволюции жизни. Однако, детали этого процесса иногда вызывают сомнения. Например, участие политиков и бизнесменов в создании AGI с целью угадывания их мыслей и желаний может привести к неравенству и ограничению участия остального населения в формировании общественной политики. Под прикрытием заботы о безусловном базовом доходе (ББД) стремятся исключить значительную часть населения из процессов принятия решений, что может угрожать общественной стабильности и разнообразию.

Таким образом, важно внимательно изучать все аспекты внедрения AGI, учитывая гуманитарные принципы, равноправие и участие всего об-

щества в процессе принятия ключевых решений для обеспечения устойчивого и безопасного будущего. Сама «забота» о ББД ничего плохого не содержит, но под видом этой заботы продвигаются идеи исключения основных масс населения из процессов формирования социальной, экономической и внешней политики государств. Это создает условия для отключения обратной связи в контуре управления миром, поляризации и глобализации и, в конечном итоге – условия для формирования диктатур.

Возможности зомбирования населения и технические средства контроля за оппонентами в век информационных технологий создают возможности для появления и значительно более долгого сохранения власти диктаторов. Справедливо называемая фашистской, политическая партия Германии пришла к власти в 1933 г. как Национал-социалистическая немецкая рабочая партия (Nationalsozialistische Deutsche Arbeiterpartei, NSDAP) и декларировалась как защитник интересов рабочего класса. Средства пропаганды и силовой контроль населения с опорой на поддержку хозяев крупной промышленности еще в прошлом веке позволили разорвать обратную связь как с собственным населением, так и с правительствами остальных стран. Это привело Германию к национальной катастрофе в 1945 г., но смена правительства и социальной, экономической и внешней политики Германии произошли в результате военного поражения. А если бы Германия успела стать глобальной диктатурой, поражение было бы невозможно в связи с отсутствием внешних противников.

Как всем хорошо известно, благими намерениями усеяна дорога в ад. В средние века следовало бояться не ведьм, а высокоморальных служителей церкви, которые могли любого отправить на костер, обвинив его в ереси и колдовстве. Сейчас основную опасность представляет не сама идея AGI, а борцы за «новую этику», планирующие писать технические задания (ТЗ) на разработку AGI, способного улавливать и реализовывать их желания без учета (или с навязчивым формированием) интересов остального населения.

AGI, как и практически любую другую технологию, можно использовать как на пользу, так и во вред основным массам людей.

6. Выводы

Развитие человеческой цивилизации шло по пути ускорения получения новых знаний за счет разделения труда, создания средств обмена знаниями и увеличения разнообразия видов деятельности. Это привело к созданию всё более сложных конструкций и структур отношений.

Но это не должно затенять того факта, что все практические и научные знания основаны на статистически достоверном изучении свойств простых объектов и явлений. Нарастающее количество знаний о простых объектах и явлениях позволяет создавать всё более сложные системы и структуры (рис. 6).

В данной статье сделана попытка обратить внимание на то, что AGI (СИИ) для достижения возможности решать сложные задачи должен обладать способностью выделять простые объекты и явления из сложной среды, статистически достоверно изучать их свойства и использовать уже имеющиеся знания для облегчения получения новых. Современные глубокие нейросети до некоторой степени обладают такими возможностями, но хорошей теории, позволяющей ясно понять, каким образом они реализуются пока не разработано. Необходимо создавать такую теорию, на ее основе строить нейросетевые структуры, обладающие данными в определении AGI (п. 5.8) способностями.

Рис. 6. Прогресс как накопление знаний, развитие способностей к познанию и ускорение способов получения новых знаний

Литература

1. *Krizhevsky A., Sutskever I., Hinton G.E.* ImageNet classification with deep convolutional neural networks // Communications of the ACM 60(6), 84-90 (2017).
2. *Goodfellow I.J., P.-A. Jean et al.* Generative adversarial networks. arxiv:1406.2661
3. *Silver D., Hubert T. et al.* Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arxiv:1712.01815
4. *Vaswani A., Shazeer N. et al.* Attention is all you need // Advances in Neural Information Processing Systems. Curran Associates, Inc. 30. 2017
5. *Devlin J., Chang M.-W. et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. arxiv:1810.04805v2
6. *Brown T., Mann B. et al.* Language models are few-shot learners. arxiv:2005.14165

7. *Fukushima K.* Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position // Biolog. Cybernetics 36, 193-202 (1980).
8. *Carpenter C., Grossberg S.* A massively parallel architecture for a self-organizing neural pattern recognition machine // Computer Vision Graphics and Image Processing 37, 54-115, (1983).
9. *Kohonen T.* Self-organized formation of topologically correct feature maps // Biolog. Cybernetics 43, 59-69 (1982).
10. *Bommasani R., Hudsonet D.A. et al.* On the opportunities and risks of foundation models. arxiv:2108.07258
11. Developing and understanding responsible foundation models. https://crfm.stanford.edu
12. Beijing Institute for General Artificial Intelligence (BIGAI). https://eng.bigai.ai
13. *Бурцев М.С., Бухвалов О.Л., Ведяхин А.А. и др.* Сильный искусственный интеллект: На подступах к сверхразуму. – М.: Интеллектуальная Литература, 2021. – 232 с.
14. *Журавлёв Д.В., Смолин В.С.* От чуда нейронных сетей – к идеям для AGI // Этот сборник.
15. *Bengio Y., Lecun Y., Hinton G.* Deep learning for AI // Communications of the ACM 64(7), 58-65 (2021).
16. *Богданов А.А.* Тектология – Всеобщая организационная наука. – Берлин – Санкт-Петербург, 1922. Переиздание – М.: Экономика, 1989.
17. *Анохин П.К.* Принципиальные вопросы общей теории функцио-нальных систем // Принципы системной организации функций. – М.: Наука, 1973. С.5-61.
18. *Пригожин И.* Введение в термодинамику необратимых процессов. – Спрингфилд, Иллинойс, Издательство Чарльза К. Томаса, 1955.
19. *Haken H.* Synergetics: an introduction: nonequilibrium phase transitions and self-organization in physics, chemistry, and biology. – Berlin – New York: Springer-Verlag, 1978.
20. Eigen M., Schuster P. The hypercycle: A principle of natural self-organization. – New York – Berlin – Heidelberg, Springer-Verlag, 1979
21. *Jantsch E.* Technological planning and social futures. – London, SW: Associated Business Programmes Ltd., 1972.
22. *Князева Е.Н., Курдюмов С.П.* Основания синергетики: Режимы с обострением, самоорганизация, темпомиры. – СПб.: Алетейя, 2002. – 414 с.
23. Ассоциация инновационных предприятий НБИКС. https://nbics.org/Default.aspx
24. *Kelly K.* The inevitable: Understanding the 12 technological forces that will shape our future. – Viking, 2016.

25. *Flood Sung* 关于深度学口口展的必然及未来的思考 (Мысли о неизбежном развитии глубокого обучения и его будущем). https://zhuanlan.zhihu.com/p/375226190
26. *Shannon C.E.* A mathematical theory of communication // Bell Syst. Techn. Journ. 27, 379-423; 623-656 (1948).
27. *Винер Н.* Кибернетика, или управление и связь в животном и машине. – М.: Советское радио, 1958.
28. *Эшби У.Р.* Введение в кибернетику. – М.: Иностранная литература, 1959.
29. *Шрёдингер Э.* Что такое жизнь? Физический аспект живой клетки / Издание третье, дополненное и исправленное. – Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2002. – 92 с.
30. *Бауэр Э.С.* Теоретическая биология. – М.-Л.: Издательство ВИЭМ, 1935. – 149 с.
31. Friston K. The free-energy principle: A unified brain theory? // Nature Reviews Neuroscience 11, 127–138 (2010).
32. *Libbrecht K.G.* Guide to snowflakes. http://www.snowcrystals.com/guide/guide.html
33. *Tegmark M.* Life 3.0: Being human in the age of artificial intelligence / First ed. – New York: Knopf, (2017).
34. Vinge V. Technological Singularity. https://cmm.cenart.gob.mx/delanda/textos/tech_sing.pdf
35. https://en.wikipedia.org/wiki/Artificial_general_intelligence
36. *Бородинов А.Г., Манойлов В.В. и др.* Поколения методов секвенирования ДНК (обзор) // Научное приборостроение. 30(4), 3-20 (2020).
37. *Barrangou R.* The roles of CRISPR-Cas systems in adaptive immunity and beyond // Current Opinion in Immunology 32, 36-41 (2015).
38. *Russell S.* Human compatible: Artificial intelligence and the problem of control. – Viking, 2019.
39. Future of life institute, FLY. https://futureoflife.org
40. *Ван Парайс Ф., Вандерборхт Я.* Базовый доход: Радикальный проект для свободного общества и здоровой экономики. – М.: Высшая школа экономики, 2020. – 440 с.
41. Международный форум «Этика ИИ: Начало доверия», 2021. https://rubda.ru/media_assocation/v-moskve-zavershilsya-mezhdunarodnyj-forum-etika-iskusstvennogo-intellekta-nachalo-doveriya
42. Форум этики в сфере искусственного интеллекта: Поколение GPT. Красные линИИ, 2023. https://conference.tass.ru/events/forum-etiki-iskusstvennogo-intellekta-pokolenie-gpt-krasnye-linii
43. Кодекс этики в сфере ИИ. https://ethics.a-ai.ru