1. Введение: основы подхода к построению универсального интеллекта

1.1. От универсального интеллекта к сильному ИИ

Перспективы создания сильного искусственного интеллекта

Область искусственного интеллекта (ИИ) принесла массу замечательных практических результатов в части автоматизации человеческой деятельности в самых разных сферах, что постепенно меняет облик нашей цивилизации. Однако конечная цель – создание по-настоящему разумных машин (сильного ИИ) до сих пор не была достигнута. В то же время, из ученых мало, кто действительно сомневается в том, что такой сильный ИИ в том или ином виде может быть создан. Если какие-то возражения и звучат, то они имеют религиозных характер, апеллирующий к наличию у человека нематериальной души. Но даже при столь радикальных воззрениях на нематериальный мир списывают лишь такие сложные концептуально феномены как свобода воли, творчество или чувства, не отрицая возможности наделения машины почти неотличимым от человека поведением. Гораздо менее однозначными являются ответы на вопросы, когда и как именно может быть создан сильный ИИ?

Искусственный интеллект как область переживал разные периоды. Начальный период, который часто характеризуют как романтический, обещал скорое, в течение пары десятилетий, создание мыслящих машин. Неоправданные ожидания привели к более прагматичному настрою, к ориентации многих исследователей на слабый ИИ – неуниверсальные интеллектуальные системы, способные решать узкие классы практических задач. Пик этой тенденции приходится на экспертные системы (ЭС), которые обещали уже не машинный разум, но эффективные коммерческие решения сложных прикладных задач. Однако и здесь ожидания не оправдались. ЭС, хотя и достаточно успешно применялись на практике, не стали прорывной технологией, которая перевернула бы мировой бизнес, из-за чего инвестиции, хлынувшие было в эту область, заметно уменьшились [McCarthy, 2005]. В США наступила «зима ИИ». Япония потерпела неудачу в проекте компьютеров пятого поколения.

Однако исследования в области ИИ вовсе не затухли. Большое количество подобластей, выделившихся из ИИ, таких как компьютерное зрение, анализ текстов, распознавание речи и т.д., продолжали приносить свои плоды, пусть и не сенсационные, но все более и более значимые. Возродился интерес бизнеса к системам слабого ИИ. Вновь стали повторяться слова о чрезвычайной значимости области ИИ в будущем для всего человечества [Nilsson, 2005a]. И вновь стала озвучиваться мысль, что области ИИ нужно «официально» вернуть ее конечную цель – создание по-настоящему разумных машин [Brachman, 2005].

При этом, однако, в сугубо академических кругах ученые перестали озвучивать сроки возможного создания сильного ИИ. Тем не менее, рядом видных специалистов в этой области снова называются сроки в несколько (а иногда даже и в одно) десятилетий [Hall, 2008]. Причем на этот раз такие экспертные ожидания подкреплены и независимыми свидетельствами. Одно из них связано с тем, что, по крайней мере, по некоторым оценкам вычислительные мощности компьютером, сопоставимые с вычислительными ресурсами человеческого мозга, достижимы к 2030-м годам (а по некоторым – достижимы уже сейчас [Hall, 2008]). С учетом того, что нехватка вычислительных мощностей была (как это понятно сейчас) одной из объективных причин, по которым ранние прогнозы о создании настоящего ИИ были несбыточными, вероятное устранение этой причины в ближайшем будущем внушает оптимизм.

Но вычислительные мощности – лишь необходимое условие для создания сильного ИИ. Помимо этого существует и масса содержательных проблем в теории ИИ, которые до настоящего времени не были решены. Удастся ли их решить за ближайшие десятилетия? Некоторую уверенность в этом дают прогнозы, связанные с технологической сингулярностью (см., напр., [Kurzweil, 2005]). Концепция сингулярности основывается на факте ускоряющегося возрастания сложности технических (а ранее – биологических) систем. Поскольку на каждом этапе глобальной эволюции сложность систем оказывается экспоненциальной (частным примером здесь является закон Мура), а при переходе между этапами показатель экспоненты каждый раз увеличивается, то есть время удвоения сложности уменьшается (так, время удвоения емкости ДНК составляет сотни миллионов лет, а емкости нервной системы – десятки миллионов лет), то следует ожидать выход этого процесса на бесконечность за конечное время. Экстраполяция кривой возрастания сложности не позволяет отнести момент наступления сингулярности позднее 2050 года (а обычно и ранее), и возникновение некоторого сверхчеловеческого разума, вероятно, должно стать одним из последующих этапов усложнения систем. Конечно, возможность достижения истинной сингулярности можно оспаривать: график возрастания сложности объективен, но его экстраполяция может быть различной, но интервалы времени до следующих этапов (метасистемные переходы) не должны начать слишком внезапно и слишком сильно удлиняться. А, значит, данная концепция также подтверждает возможность создания сильного ИИ в течение ближайших десятилетий, что делает данную проблему, хотя и оставляет вопрос о том, как именно к ее решению стоит подходить.

При этом ведущими специалистами отмечается невозможность достижения сильного ИИ в рамках краткосрочных проектов [McCarthy, 2005], путем создания узкоспециализированных интеллектуальных систем [Nilsson, 2005b] или даже путем постоянного совершенствования систем, решающих изолированные когнитивные задачи типа обучения или понимания

естественного языка [Brachman, 2005]. Необходимо ставить и решать именно задачу создания сильного ИИ, даже если при этом не ожидается получение каких-либо коммерческих результатов за первые десять или более лет.

В академической среде все ограничивается вполне естественным призывом к объединению подобластей ИИ [Bobrow, 2005; Brachman, 2005; Cassimatis et al., 2006], каждая из которых уже успела приобрести свою глубокую специфику. Достигнутый прогресс в каждой из подобластей дает надежду на то, что объединение полученных результатов позволит построить интеллектуальные системы, существенно более мощные, чем те, что были построены на заре компьютерной эпохи в попытках создания первых мыслящих машин. С другой стороны, такое объединение должно дать многое и самим подобластям: ведь задачи, решаемые в их рамках, зачастую полагаются ИИ-полными. Так, вряд ли можно создать универсальные системы распознавания образов, понимания языка или автоматического доказательства теорем без создания сильного ИИ, поскольку между всеми этими задачами есть принципиальная взаимосвязь [Brachman, 2005].

Использование и исследование когнитивных архитектур как способов объединения в единой системе всех необходимых для полноценного интеллекта функций, таких как обучение, представление знаний, рассуждения и т.д. выделилось в новую господствующую парадигму в области ИИ в целом [Brachman, 2005]. И именно эта парадигма официально связывается с построением систем искусственного интеллекта уровня человека [Cassimatis et al., 2006; Cassimatis, 2006], или универсальных [Langley, 2006].

Подобные интеграционные исследования необходимы, но насколько они достаточны? Общие идеи о том, что сильный ИИ должен создаваться как единая система, которая должна включать некую базовую когнитивную функциональность, достаточно очевидны и высказывались очень давно. Однако до сих пор нет ни минимально необходимого перечня когнитивных функций, ни, тем более, обоснованных деталей их реализации.

Более того, существуют не только многие существенно разные когнитивные архитектуры [Jones and Wray, 2006], но также и архитектурные парадигмы, альтернативные когнитивной [Langley, 2006]. При этом когнитивные архитектуры в основном концентрируются на вопросах интегрирования, взаимодействия отдельных функций. Но можно ли из слабых когнитивных компонент получить сильный ИИ? На наш взгляд, ответ однозначный: нет. Вместо (или, по крайней мере, в дополнение) обсуждения методических вопросов объединения существующих слабых компонент, необходимо разрабатывать теорию сильного ИИ, из которой будет одновременно следовать как структура сильных компонент ИИ, так и необходимая архитектура их объединения.

Как справедливо отмечено в [Cohen 2005]: «Poor performance and universal scope are preferred to good performance and narrow scope». С учетом того, что, как указывалось, создание эффективных узкоспециализированных систем

почти не приближает нас к сильному ИИ, естественно спросить, чего же не хватает современным когнитивным системам в плане универсальности?

Универсальность как алгоритмическая полнота

Исторически в области искусственного интеллекта выделились несколько фундаментальных направлений, таких как поиск или обучение. Эти направления начинают быть четко видны, когда мы ставим интеллектуальные задачи в наиболее упрощенном чистом виде. Так, рассматривая игровые задачи или доказательства теорем, можно предложить для них универсальное решение – полный перебор вариантов в пространстве возможных операций. Конечно, при конечных вычислительных ресурсах полный перебор невозможен, но это не устраняет концепцию поиска как фундаментальной компоненты интеллекта. В случае, когда пространство поиска заранее неизвестно, ставится задача обучения (точнее, предсказания того, как какие-то операции будут влиять на состояния мира и самого агента). Здесь универсальное решение не столь очевидно, но оно также известно практически с самого момента зарождения области ИИ. Это универсальный метод предсказания Р. Соломонова [Solomonoff, 1964] на основе алгоритмической теории информации. Этот метод также на практике не применим, так как требует огромного перебора вариантов (и, вообще говоря, требует решения алгоритмически неразрешимой задачи останова).

Эти идеальные методы – то, к чему нужно приблизиться в условиях ограниченных вычислительных ресурсов, поскольку только ограничение на ресурсы отделяет эти методы от того, чтобы на их основе уже сейчас воплотить сильный ИИ. К примеру, вся проблематика эвристического программирования и метаэвристических методов поиска возникла при попытке решения проблемы поиска при ограниченных ресурсах. Также и проблематика машинного обучения, включающая, например, трансферное обучение, обучение понятиям и многое другое, появляется из-за ограниченности ресурсов. При этом, однако, исследователи, разрабатывающие практические методы, зачастую не оглядываются на тот идеал, к которому необходимо стремиться. Это приводит к созданию методов слабого искусственного интеллекта, обладающих неустранимым дефектом. Этот дефект заключается в том, что эти методы не полны по Тьюрингу, то есть работают в ограниченном пространстве алгоритмов и принципиально не могут выйти за рамки этих ограничений. Хотя для разных частных методов области в пространстве алгоритмов могут отличаться друг от друга, конечное их объединение не может дать алгоритмически полного пространства. В плане методов машинного обучения это означает невозможность выявления произвольной регулярности, которая может иметься в данных, невозможность построить модель мира, которая не была заранее предусмотрена разработчиком.

Здесь кроется ответ, почему работы в области когнитивных архитектур (как подхода к сильному ИИ) не являются достаточными. Они исходят из

посыла, будто современных методов поиска в пространстве решений, представления знаний, машинного обучения достаточно, а не хватает лишь их объединения, при котором возникнет новое качество – сильный ИИ. Мы, однако, полагаем, что свойство универсальности интеллекта кроется в том, что он в принципе может оперировать с любыми моделями из алгоритмически полного пространства (хотя на практике это, естественно, в полной мере не достигается). В этой связи полезно разделить понятие универсального и сильного ИИ. Хотя они фактически могут означать одно и то же, но понятие сильного ИИ неявно подразумевает стремление создавать модели, по внешним признакам напоминающие человеческий интеллект, тогда как понятие универсального ИИ заставляет, в первую очередь, обращать внимание на то, чтобы в строящиеся модели не закладывались непреодолимые ограничения на то, чему ИИ сможет обучиться или в какой среде сможет адекватно действовать.

Для обеспечения этого можно начать с некоторой идеализированной модели сильного ИИ, работающего в условиях бесконечных ресурсов. Поскольку действительно автономный искусственный интеллект должен создаваться как воплощенный интеллектуальный агент, необходимо разработать идеализированную модель такого агента, который бы гипотетически мог решать все те задачи, которые может решать человек.

Попытки создания таких моделей имеются (наиболее известной является AIXI [Hutter, 2005]), и мы их позднее обсудим. Сейчас лишь отметим, что рассмотрение подобных моделей заставляет разных исследователей прийти к выводу, что именно алгоритмической полнотой обеспечивается универсальность интеллекта, и это свойство необходимо пытаться сохранить, по крайней мере, в пределе (см., напр., [Pankov, 2008]).

Таким образом, *первым методологическим принципом является сохранение отсутствия органичений на алгоритмическую полноту можества моделей (закономерностей, понятий, представлений), которые могут быть выведены или использованы системами универсального ИИ*.

Реализуемость как ресурсная ограниченность

Модели универсального алгоритмического интеллекта могут быть хорошим отправным пунктом. Но также очевидно, что необходим учет ограниченности ресурсов, чтобы эти модели были реализуемы. Ведь именно эта ограниченность во многом определяет специфику наших когнитивных процессов.

Действительно, модели универсального интеллекта не имеют с реальным интеллектом почти ничего общего, если судить по их «когнитивным операциям». Такие модели не будут в явном виде строить систему понятий, не будут осуществлять планирования, не будут обладать вниманием и т.д. Крайне сложно сказать, будут ли они обладать функцией «понимания», самосознанием и т.д. Здесь можно провести (неполную) аналогию с шахматной программой,

которая за счет неограниченных ресурсов осуществляет полный перебор. Эта программа крайне проста. Единственная ее фундаментальная операция – это поиск. В ней нет описания шахматных позиций в каких-либо производных терминах, нет ничего похожего на понимание. Но в рамках шахмат она ведет себя идеально. Сходным образом можно попробовать вообразить и идеальный воплощенный интеллект, действующий в реальном мире.

Отсутствие основной части когнитивных функций у такого идеального интеллекта может означать одно из двух. Либо эти функции – следствие ограниченности ресурсов (для ряда из них, например для внимания, это так со всей очевидностью). Либо интеллект – это что-то совсем отличное от того, что под ним обычно подразумевают (а подразумевают средство решения задач, в качестве основной из которых является выживания). Возможно, вторая альтернатива и не столь бессмысленна (и не столь противоречит первой), если интеллектом считать не любой, но некий выделенный способ решения задач (то есть если в интеллекте важна не столько функциональность, сколько способ ее достижения). В то же время, при бесконечных вычислительных ресурсах разумное поведение может достигаться гораздо более простыми средствами. К счастью, обсуждать, следует ли называть разумной систему, реализующую идеальное (по адекватности) поведение за счет «грубой вычислительной силы», а не за счет «интеллектуальности» (некой структурной сложности процессов «мышления»), не обязательно в силу гипотетичности такой системы. Единственное, что нужно обсуждать, – это то, будет ли эта система действительно обладать всеми теми возможностями, что и естественный интеллект. Если в этом будет какое-либо сомнение, то необходимо будет его преодолеть, либо обосновав достижимость соответствующих возможностей, либо уточнив модель.

Идея ограниченных ресурсов как принципиального свойства сильного ИИ, определяющего его архитектуру, уже высказывалась [Wang, 2007]. Но руководствоваться одной только этой идеей также недостаточно, что будет обсуждено ниже. Сейчас лишь отметим, что учет ограниченности ресурсов не должен нарушать (алгоритмической) универсальности интеллекта. Условно говоря, реальный интеллект – это «any-time» метод, который стремится к идеальному интеллекту при неограниченном увеличении вычислительных ресурсов.

С необходимостью ввода ресурсных ограничений согласны и разработчики универсальных моделей алгоритмического интеллекта (см., напр., [Schmidhuber, 2007], [Hutter, 2007]). Попытки ввода ограничений ресурсов в эти модели могут быть рассмотрены как второй шаг в направлении к универсальному ИИ, хотя насколько этот шаг существенный, судить сложно: зачастую эти модели «слишком универсальны» в том смысле, что авторы пытаются заложить в них минимальную предвзятость относительно того, в каком мире предстоит функционировать.

Таким образом, *второй методологический принцип заключается в построении архитектуры реального универсального интеллекта путем ввода ресурсных ограничений в модель идеального универсального интеллекта*.

Априорная информация о мире как основное содержание феномена интеллекта

Воплощенный интеллект ограничен не только по количеству совершаемых вычислительных операций при решении задач индукции и дедукции, но также и по числу действий, совершаемых в физическом мире. Второй тип ограничений принципиально несводим к первому, хотя некоторая взаимосвязь между ними есть: совершение некоторого действия может избавить от необходимости рассуждать, и, наоборот, подумав, можно уменьшить количество пробных действий в физическом мире. Именно этот тип ограничений не учитывается в моделях идеального алгоритмического интеллекта с ограниченными вычислительными ресурсами.

В глобальном плане повышение эффективности совершаемых действий связано, в первую очередь, с накоплением информации о внешнем мире. Можно представить модель идеального интеллекта, обладающего минимумом априорной информацией. Этот интеллект сможет научиться, чему угодно (в том числе, и эффективному использованию своих вычислительных ресурсов) и в пределе будет настолько же эффективен, насколько эффективен специализированный интеллект, но на это уйдет слишком много времени. И, естественно, такой интеллект не сможет автономно выживать в процессе начального обучения.

При этом априорная информация для реального интеллекта может иметь самую разнообразную форму, в частности, иметь форму способностей, таких как подражание. Действительно, от идеального интеллекта необходимо ожидать того, что он сможет выполнять подражание, заранее не обладая этой способностью, однако для этого ему придется сначала накопить слишком много информации. Если же эта способность имеется сразу, то она может существенно ускорить оптимизацию собственных действий в физическом мире. Стоит отметить, что модели обучения роботов путем подражания сейчас широко исследуются (равно как и исследование зеркальных нейронов в нейрофизиологии). Проблема, однако, в том, чтобы данный механизм (как и все прочие дополнительные априорные механизмы) был согласован с универсальностью интеллекта. Аналогично, и лингвистические способности должны быть в какой-то мере заложены априорно. Это должно быть сделано не потому, что универсальный интеллект в принципе не сможет приобрести их самостоятельно, а потому, что это приобретение может занять слишком много времени.

Объяснение ряда когнитивных способностей как априорной информации о внешнем мире (как сугубо физическом, так и социальном), позволяющей ускорить развитие интеллекта (которое, собственно, и сводится к накоплению

информации и ее обработке), достаточно очевидно. Однако это объяснение не использовалось для определения устройства универсального интеллекта. Нас интересует минимальный объем априорной информации и формы ее представления, которые позволят реальному ИИ развиваться не медленнее человека. Принципиальным вопросом при этом является встраивание априорной информации в структуру универсального ИИ.

Важность этого момента видна на примере гибкости архитектуры естественного интеллекта. Например, мозг человека заранее не ориентируется на то, что лингвистическая информация будет передаваться через речь. При формировании протопонятий работают механизмы, родственные условным рефлексам. Если способность к формированию истинных понятий и заложена априорно, то она не привязана к сенсорной модальности. Подобную универсальность необходимо оставлять и при введении каких-то априорных элементов в структуру ИИ. Сейчас же в моделях обучения понятиям не только разделение на семантический и лингвистический каналы выполняется априорно, но делается и привязка к модальности. Аналогичное заключение можно сделать и относительно моделирования прочих когнитивных механизмов, отражающих априорную информацию. Наиболее ярким примером этого служат экспертные системы, в которые априорно закладывается большой объем знаний при отсутствии возможности их автономного расширения, чего, очевидно, следует избежать в случае универсального ИИ.

С другой стороны, именно необходимый для реального интеллекта объем априорной информации и многообразие ее форм (это может быть информация как о самых разнообразных аспектах внешнего мира, так и об эвристиках оптимального использования собственных ресурсов) делает создание ИИ столь сложным. В этом смысле простые модели универсального интеллекта нас мало приближают к его созданию. Практически используемые когнитивные архитектуры могли бы даже оказаться полезнее, если бы не требовали полной переделки при попытке сделать их универсальными. Вместо добавления свойства универсальности в существующие системы, исходно составленные из слабых компонент, продуктивнее будет начинать с универсальной непрактичной системы, добавляя в нее согласованным образом те эвристики, которые были накоплены в области классического ИИ.

Содержательная сложность интеллекта, его когнитивная архитектура, – это то, что позволяет действовать нам в имеющемся окружающем мире в условиях ограниченных ресурсов и без чрезмерно длительного обучения. Но это означает, что основная сложность нашего интеллекта связана с его оптимизированностью под окружающий мир. Структура такого интеллекта не может быть выведена теоретически в универсальных моделях интеллекта, а должна быть получена эмпирически либо самим универсальным интеллектом, либо разработчиками. Естественно, мы при этом хотим сделать настолько универсальный интеллект, насколько это возможно. Говоря точнее, такой интеллект может быть настолько же универсальным, насколько являются

упоминавшиеся простейшие модели. Разница между ними будет лишь в смещении предпочтений или предвзятости по направлению к нашему миру. Естественно, повышение эффективности такого интеллекта для нашего мира произойдет за счет снижения его эффективности (но не до нуля, в чем и заключается универсальность) в каких-то других возможных мирах, однако, с учетом того, что ему предстоит действовать в первую очередь в нашем мире, это является вполне допустимым.

Но недопустимым при этом является потеря универсальности, поскольку наш мир сам является «универсальной средой». В этой связи с универсальных «непредвзятых» моделей вполне можно начинать построение реального ИИ. В них могут постепенно вноситься эвристики, связанные с особенностями нашего мира, начиная с самых общих, пока ИИ не сможет самостоятельно действовать (включая самооптимизацию) достаточно эффективно.

Таким образом, *третий методологический принцип – введение в универсальный интеллект априорной информации для уменьшения объема данных, которые в онтогенезе необходимо получить агенту для автономного функционирования в реальном мире, при условии сохранения согласованности последующей универсальной индукции и дедукции с априорной информацией*.

1.2. Краткий анализ существующих подходов к сильному ИИ

Когнитивные архитектуры

При создании сильного ИИ естественно воспроизводить, если не все детали работы человеческого мозга, то, по крайней мере, те функции, которые он выполняет. В противном случае, очень сложно быть уверенным, что создается именно интеллект. Именно такую цель и преследуют когнитивные архитектуры, которые объединяют такие функции, как обучение, память, планирование и т.д., то есть все (или почти все) то, что есть в естественном интеллекте. Это и делает когнитивные архитектуры столь привлекательными и популярными.

Однако само по себе желание наделить компьютер всеми теми же когнитивными функциями, которые есть у человека, не говорит о том, как это правильно сделать. В результате к настоящему моменту разработано множество когнитивных архитектур, ряд из которых нередко позиционируется как путь к построению сильного ИИ. К ним, в частности, относится такие популярные у «строителей сильного ИИ» архитектуры, как Soar и ACT-R.

Многие архитектуры зачастую отталкиваются от феноменологии высших когнитивных функций человеческого разума. Однако из-за отсутствия полного понимания природы этих функций и требований к ним их реализации оказываются во многом произвольными.

Нередко даже построение подобных архитектур ведется в рамках традиционного символьного подхода, моделирующего лишь «вершину айсберга» человеческого мышления. Тем не менее, нередко производится и

попытка построения ахритектур, воспроизводящих не только высокоуровневые, но и низкоуровневые функции (т.н. эмерджентные архитектуры). Более того, исследователи ИИ хорошо понимают необходимость объединения символьных и субсимвольных уровней и разработку гибридных архитектур, а также необходимость построения воплощенных систем (являющихся ключевыми, в частности, для получения семантической основы понятий), которые в сугубо символьных архитектурах реализовывать весьма проблематично (см. [Duch et al., 2008] в качестве обзора).

Тем не менее, отмечается [Duch et al., 2008], что весьма нечасто удается применять для решения реальных задач, не говоря уже о том, чтобы масштабировать до уровня автономного поведения в реальном мире. Так почему же когнитивные архитектуры не привели к существенному прогрессу в области сильного ИИ? Ответ на этот вопрос уже дан выше.

Эти системы, вероятнее всего, обречены на неуниверсальность, поскольку собираются из слабых компонент. Это, видимо, относится и к таким системам, исходно позиционировавшимся в качестве систем общего интеллекта, как Novamente (описание которой дано в [Goertzel and Pennachin, 2007]). Конечно, не исключена возможность внесения свойства универсальности как расширения той или иной архитектуры (в конце концов, универсальность интеллекта вряд ли можно приписывать большинству животных, а, значит, она появилась как эволюционная надстройка над более частными формами интеллекта). Тем не менее, такой путь нам представляется более трудоемким и менее оптимальным.

Подход на основе ресурсных ограничений

Данный подход отталкивается от следующего определения, данного П. Вангом [Wang, 2007]:

> *Intelligence is the capacity of a system to adapt to its environment while operating with insufficient knowledge and resources*,

где адаптация (как способность учиться на опыте) является достаточно обычным требованием, тогда как основные особенности подхода выводятся из недостатка ресурсов и знаний (поскольку, когда ресурсов и информации достаточно, могут использоваться и не вполне интеллектуальные методы). Как следствие, в рамках этого подхода строится вариант категориальной логики для учета нечеткости знаний, а также предлагается распределенная система манипулирования знаниями, в которой учитывается ограниченность вычислительных ресурсов.

При этом автор предлагает разделять понятия «интеллектуальный» и «эффективно интеллектуальный». Такое разделение представляется вполне справедливым и отражает то интуитивное впечатление, что, например, шахматная программа, работающая методом «грубой силы», является интеллектуальной не в том же смысле, в котором является интеллектуальным шахматист.

Хотя с самим принципом эффективного интеллекта можно согласиться, данный конкретный подход вряд ли может стать основой для построения СИИ: в нем упускается те аспекты интеллекта, которые выявлены в универсальных алгоритмических моделях и в когнитивных архитектурах. Иными словами, сам тезис о необходимости ресурсных ограничений не говорит о том, как правильно их вводить.

В частности, это видно из того факта, что П. Ванг ввел как основополагающий принцип также недостаток знаний у агента. Недостаток знаний, конечно же, важен, но он вполне учитывается в (критикуемых Вангом) моделях универсального алгоритмического интеллекта, которые включают не только поиск в пространстве действий, но и универсальный индуктивный вывод, для которого учет нечеткости знаний является не основополагающим принципом, а лишь эвристикой для упрощения перебора моделей (что будет продемонстрировано позднее).

В итоге в рамках этого подхода разработана лишь частная когнитивная архитектура, не обладающая принципиальными преимуществами по сравнению с прочими, хотя систематическое следование принципам ограниченности ресурсов и обладает значительной эвристической силой.

Универсальный алгоритмический интеллект

Сама идея данного подхода известна давно, но получил он признание сравнительно недавно в основном через работы [Hutter, 2001], [Schmidhuber, 2003] и другие работы этих авторов. В его рамках основной упор делается на модели универсальной индукции Соломонова, включенные в систему выбора действий в окружающей среде для максимизации некоторой оценивающей функции.

Здесь анализ начинается с простой универсальной модели, на которую не накладываются ресурсные ограничения. Первый шаг нашего подхода аналогичен, так как мы полагаем, что свойство универсальности крайне желательно сразу вводить в модель универсального ИИ и поддерживать сохранение этого свойства при развитии модели, которое осуществляется путем ввода ресурсных ограничений.

В современных версиях рассматриваемых подходов ресурсные ограничения также вводятся, но с сохранением максимальной непредвзятости универсального ИИ, что позволяет строить общие модели самооптимизации.

Такой учет ограничений на ресурсы, однако, не вполне достаточен. Можно сказать, что он требует воспроизводства целиком эволюции, которая также начиналась как универсальный самооптимизирующийся поиск без какой-либо априорной информации. Очевидно, чтобы становление подобного универсального интеллекта могло быть осуществлено за обозримое время, необходимо в него закладывать как достаточно большой объем априорной информации о структуре внешнего мира, так и эвристики для сокращения перебора вариантов моделей и действий. Эти эвристики как раз можно

почерпнуть из феноменологии когнитивных функций естественного интеллекта. С другой стороны, в сильный ИИ нерационально вручную закладывать слишком большой объем специфичных знаний, которые он может почерпнуть самостоятельно (чем грешат такие проекты, как, например, Cyc). Очевидно, необходимо достижение оптимального компромисса между этими двумя крайностями.

Помимо этого, отдельный вопрос для обсуждения заключается в том, а действительно ли представленные модели являются универсальными. Для этого необходимо тщательно сравнить гипотетические возможности этих моделей с возможностями человека. Отчасти такие сравнения проводятся (например, [Hutter, 2005]), хотя их нельзя назвать бесспорными или исчерпывающими. Тем не менее, сомнения в действительной универсальности этих моделей вполне можно выдвинуть, что будет показано при анализе нашей собственной модели универсального алгоритмического интеллекта.

Сейчас отметим лишь одно из таких сомнений, которое заключается в том, что интеллект лишь в нулевом приближении можно свести к максимизации априорно заданной целевой функции. Ведь если, скажем, задача интеллекта заключается в обеспечении выживания, то априорно заданная целевая функция (базирующаяся, скажем, на эмоциональных оценках) может быть лишь грубой эвристической аппроксимацией цели выживания. Это означает необходимость существования специальных механизмов, позволяющих каким-то образом уточнять целевую функцию в онтогенезе. Здесь можно привести следующую аналогию с шахматами. Пусть один интеллектуальный агент может сыграть только одну партию. Имея ограниченные вычислительные ресурсы, он не может осуществить полный перебор вариантов, чтобы предсказать победу или поражение. Рождаясь с минимумом априорных знаний о мире, он не может иметь сложную целевую функцию, которая бы позволяла эффективно отсекать неперспективные варианты на дереве игры. Исходная целевая функция может опираться лишь на какие-то непосредственно воспринимаемые стимулы, скажем на суммарную силу фигур (дающую ощущение боли и удовольствия при потере своей фигуры или съедении фигуры соперника). В процессе взросления (игры) агент может построить более сложные понятия, но самостоятельно (не прожив жизнь целbком) он в принципе не сможет определить, как на основе этих понятий можно улучшить целевую функцию. Эту информацию ему, однако, могут дать другие агенты, но только при условии, что имеется некий хороший механизм модификации целевой функции. Этот аспект имеет отношение и к проблеме дружественного ИИ…

Подход на основе обучения целевым функциям

Проблема обучения целевым функциям иногда рассматривается в качестве основополагающей при построении сильного ИИ (или, точнее, дружественного ИИ [Yudkowsky, 2011]). В рамках этого подхода совершенно

справедливо замечается, что максимизация априорной целевой функции недостаточна для того, чтобы искусственный интеллект оказался универсальным, особенно, в части эффективного (и желаемого) взаимодействия с социальным окружением, которое является таким же элементом объективной реальности, как и физическое окружение.

Проблема наделения ИИ способностью к модификации собственной целевой функции нетривиальна в силу того, что не ясно, как целевая функция может оптимизироваться, если не под управлением другой целевой функции (или каких-то других априорных механизмов). Важность возможности модификации целевой функции связана не только с тем, что это необходимо для полноценной универсальности агента, но и с тем, что ИИ, стремящийся к максимизации априорной целевой функции вполне может найти такие действия, оптимальные с точки зрения этой функции, которые окажутся крайне нежелательными для людей [Yudkowsky, 2011]. Хотя важность этих аспектов бесспорна, их рассмотрение вне конкретных моделей универсального интеллекта не позволяет наметить путь создания сильного ИИ (а, скорее, задает некоторые ограничения на пути его создания), поэтому данный подход следует считать комплементарным другим подходам. Возможность модификации целевой функции необходимо предусмотреть в архитектуре универсального интеллектуального агента, хотя в целом это можно рассматривать на том же уровне, что и другие когнитивные функции, а именно, как специфическую эвристику повышения эффективности развития «младенческого» ИИ до уровня «взрослого» ИИ.

Адаптивное поведение, самоорганизация и бионика в целом

Существует большое направление исследований в области сильного ИИ, связанное с бионическим подходом. Здесь выделяются попытки (см., напр., [Garis, 2007] [Red'ko, 2007]) моделирования мозга на разных уровнях детальности, воспроизведения адаптивного поведения, начиная с простейших его форм к более сложным, моделирования эволюции, самоорганизации в целом. Зачастую этот подход носит имитационный характер и достаточно жестко противопоставляется алгоритмическому подходу, из-за чего оказывается недостаточно глубоким. В частности, разные имитационные модели эволюции и самоорганизации не приводят к неограниченному развитию по той простой причине, что их авторы даже не пытаются рассматривать вопросы, связанные с вычислительной сложностью решаемых оптимизационных проблем и алгоритмической полнотой тех форм поведения, которые в принципе могут получиться в ходе этого моделирования. Из-за этого весьма сомнительно, что бионический подход сам по себе может привести к созданию сильного ИИ. Однако в то же время он может служить важным источником продуктивных идей, пренебрегать которым было бы слишком расточительно.

Выводы

Как видно, разные существующие подходы к сильному ИИ не столько противоречат друг другу, сколько рассматривают разные аспекты проблемы универсального ИИ, в связи с чем необходимо осуществлять их объединение. Естественно, существует и множество интеграционных подходов, пытающихся выполнить синтез разных имеющихся систем и методов, поэтому идея интеграции в целом не нова. Однако зачастую эта интеграция ограничивается объединением слабых методов, либо же частичным расширением универсальных алгоритмических моделей ИИ. Недостаточная «глубина» интеграции видна по тому факту, что сторонники перечисленных подходов предпочитают их противопоставлять друг другу, критикуя недостатки конкурентных подходов. Здесь же речь идет, скорее, о разработке нового подхода, осуществляющего учет основных ранее полученных результатов и идей на гораздо более глубоком концептуальном уровне (при этом, правда, далеко не всегда легко установить связь между разными подходами).

Необходимо начать с простейших моделей в случае неограниченных ресурсов; убедиться в их универсальности или установить, чего не хватает для ее достижения, что может быть учтено впоследствии. Далее следует рассмотреть универсальные модели с ограничением на вычислительные ресурсы. Такие модели могут быть также относительно просты, но должны включать самооптимизацию. Далее должна вводиться априорная информация о свойствах мира (наиболее общие из которых обусловят особенности когнитивной архитектуры) для сокращения времени становления ИИ, то есть приобретения им автономности.

Литература

(Bobrow, 2005) Bobrow D.G. *AAAI: It's Time for Large-Scale Systems* // AI Magazine. 2005. V. 26. No 4. P. 40–41.

(Brachman, 2005) Brachman R. *Getting Back to "The Very Idea"* // AI Magazine. 2005. V. 26. No 4. P. 48–50.

(Cassimatis et al., 2006) Cassimatis N., Mueller E.T., Winston P.H. *Achieving Human-Level Intelligence through Integrated Systems and Research* // AI Magazine. 2006. V. 27. No 2. P. 12–14.

(Cassimatis, 2006) Cassimatis N.L. *A Cognitive Substrate for Achieving Human-Level Intelligence* // AI Magazine. 2006. V. 27. No 2. P. 45–56.

(Cohen, 2005) Cohen P.R. *If Not Turing's Test, Then What?* // AI Magazine. 2005. V. 26. No 4. P. 61–67.

(Duch et al., 2008) Duch W., Oentaryo R.J., Pasquier M. *Cognitive Architectures: Where Do We Go from Here* // Frontiers in Artificial Intelligence and Applications (Proc. 1st AGI Conference). 2008. V. 171. P. 122–136.

(Garis, 2007) Hugo de Garis. *Artificial Brains* // in Artificial General Intelligence. Cognitive Technologies, B. Goertzel and C. Pennachin (Eds.). Springer. 2007. P. 159–174.

(Goertzel and Pennachin, 2007) Goertzel B., Pennachin C. *The Novamente Artificial Intelligence Engine* // in Artificial General Intelligence. Cognitive Technologies, B. Goertzel and C. Pennachin (Eds.). Springer. 2007. P. 63–130.

(Hall, 2008) J Storrs Hall. *Engineering Utopia* // Frontiers in Artificial Intelligence and Applications (Proc. 1st AGI Conference). 2008. V. 171. P. 460–467.

(Hutter, 2001) Hutter M. *Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions* // In Proc. 12th European Conf. on Machine Learning (ECML-2001). 2001. V. 2167 of LNAI, Springer, Berlin.

(Hutter, 2005) Hutter M. *Universal Artificial Intelligence. Sequential Decisions Based on Algorithmic Probability* / Springer. 2005. 278 p.

(Hutter, 2007) Hutter M. *Universal Algorithmic Intelligence: A Mathematical Top→Down Approach* // in Artificial General Intelligence. Cognitive Technologies, B. Goertzel and C. Pennachin (Eds.). Springer. 2007. P. 227–290.

(Jones and Wray, 2006) Jones R.M., Wray R.E. *Comparative Analysis of Frameworks for Knowledge-Intensive Intelligent Agents* // AI Magazine. 2006. V. 27. No 2. P. 57–70.

(Kurzweil, 2005) Kurzweil R. *The Singularity is Near*. Viking, 2005.

(Langley, 2006) Langley P. *Cognitive Architectures and General Intelligent Systems* // AI Magazine. 2006. V. 27. No 2. P. 33–44.

(McCarthy, 2005) McCarthy J. *The Future of AI—A Manifesto* // AI Magazine. 2005. V. 26. No 4. P. 39.

(Nilsson, 2005a) Nilsson N.J. *Reconsiderations* // AI Magazine. 2005. V. 26. No 4. P. 36–38.

(Nilsson, 2005b) Nilsson N.J. *Human-Level Artificial Intelligence? Be Serious!* // AI Magazine. 2005. V. 26. No 4. P. 68–75.

(Pankov, 2008) Pankov S. *A computational approximation to the AIXI model* // Frontiers in Artificial Intelligence and Applications (Proc. 1st AGI Conference). 2008. V. 171. P. 256–267.

(Red'ko, 2007) Red'ko V.G. *The Natural Way to Artificial Intelligence* // in Artificial General Intelligence. Cognitive Technologies, B. Goertzel and C. Pennachin (Eds.). Springer. 2007. P. 327–352.

(Schmidhuber, 2003) Schmidhuber J. *The new AI: General & sound & relevant for physics*. Technical Report TR IDSIA-04-03, Version 1.0, cs.AI/0302012 v1, IDSIA. 2003.

(Schmidhuber, 2007) Schmidhuber J. *Gödel Machines: Fully Self-Referential Optimal Universal Self-improvers* // in Artificial General Intelligence. Cognitive Technologies, B. Goertzel and C. Pennachin (Eds.). Springer. 2007. P. 199–226.

(Solomonoff, 1964) Solomonoff R.J. *A formal theory of inductive inference: parts 1 and 2. Information and Control*. 1964. V. 7. P. 1–22, 224–254.

(Wang, 2007) Wang P. *The Logic of Intelligence* // in Artificial General Intelligence. Cognitive Technologies, B. Goertzel and C. Pennachin (Eds.). Springer. 2007. P. 31–62.

(Yudkowsky, 2011) Yudkowsky E. *Complex Value Systems in Friendly AI* // Proc. Artificial General Intelligence – 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Lecture Notes in Computer Science 6830. Springer. 2011. P. 388–393.