2. «Непредвзятый» универсальный алгоритмический интеллект

2.3. Когнитивное смещение универсального интеллекта

Введение

До настоящего момента нам хватало известных моделей ИМИ, и с не слишком существенными оговорками можно почти согласиться с тем, что «AIXI model is the most intelligent unbiased agent possible» [Hutter, 2007] и что ИМИ в своем поведении будет не более ограниченным, чем человек, но при наличии достаточных вычислительных ресурсов и информации. Последняя оговорка и поясняет основные причины, почему эти модели не привели к созданию реального ИИ и почему их можно рассматривать только в качестве первого маленького шага по направлению к нему. Здесь важно определить, куда двигаться дальше.

Как мы отмечали, интерес представляет реальный, «прагматический», а не «непредвзятый» интеллект, описывающийся моделями ИМИ. Его нельзя построить только на основе сугубо аналитического рассмотрения проблемы выбора оптимальных действий в произвольной среде. Даже максимально эффективная самооптимизация окажется недостаточной, так как потребует не только огромного объема вычислений, сопоставимых с эволюцией, но и соответствующего количества физических взаимодействий агента с окружением. Прагматичность универсального ИИ должна быть обеспечена введением «когнитивного смещения», элементы которого в некотором виде выявлены в классических исследованиях ИИ и человеческого мышления, все результаты которых обладают эмпирической полезностью. Нельзя отвергать накопленные здесь сведения, но необходима их особая интерпретация в рамках теории универсального интеллекта. Покажем, что введение когнитивных функций не расширяет принципиальные возможности ИМИ, но должно повышать его эффективность/прагматичность.

Восприятие

Модели ИМИ не включают такой выделенной когнитивной функции как восприятие, если под восприятием понимать не просто получение на вход сенсорных данных, а именно те специфические когнитивные процессы, которые характерны для естественных систем. При этом естественные системы восприятия обладают выраженной структурой, задающей большое индуктивное смещение, согласованное с регулярностями, встречающимися в реальном мире. Это смещение реализуется в форме представлений информации и делает возможным очень эффективную интерпретацию сенсорных данных без исчерпывающего поиска.

На примере восприятия должно быть абсолютно ясно, что модели ИМИ, требующие прямого поиска алгоритмических моделей, например для изображений, длины которых превышают миллионы бит (т.е. число рассматриваемых моделей $>>10^{100000}$), являются абсолютно нереалистичными. В то же время следует особо подчеркнуть, что человеческая система сенсорного восприятия сохраняет универсальность, что нами уже отмечалось. Она может обнаружить стимул, задаваемый практически произвольной регулярностью. В то же время, весьма легко для любой системы компьютерного зрения найти класс неидентифицируемых ею стимулов. Это хорошо видно на примере попыток промоделировать формирование условного рефлекса с нетривиальными стимулами (см., напр., [Potapov and Rozhkov, 2012] и ссылки там же). Именно в этом контексте высказывается мнение, согласно которому несмотря на заметный прогресс в области робототехники, искусственного интеллекта, машинного восприятия и обучения, имеется недостаток истинно когнитивных систем, которые обладают достаточной общностью для работы в неструктурированной среде [Pavel et al., 2007], что явным образом связано с неуниверсальностью (в смысле алгоритмической полноты пространства моделей) систем восприятия.

В ИМИ процесс построения моделей включает в себя восприятие неявно: в нем оно как специфическая когнитивная функция не отделено от более сложных символьных моделей мира. Само такое разделение (но разделение «мягкое»!) может рассматриваться как эвристика, но ее одной, естественно, недостаточно, и возникает общий вопрос, как сделать процесс построения моделей в ИМИ более эффективным.

Модели

ИМИ не подразумевает выбора и фиксации моделей среды. Для оптимального предсказания перебираются для всех имеющихся данных все возможные модели с разными весами, которые учитываются в предсказании. Естественно, для нас это выглядит абсолютно расточительно. Иногда может даже показаться, что человек склонен искать какую-то одну истинную модель мира. В особенности вся наука представляет собой попытку построения некой единой модели мира, соответствующей однозначным законам. Разумеется, в процессе научного поиска перебираются разные конкурирующие теории, но в конечном итоге между ними осуществляется выбор. При этом множественность теорий во многом поддерживается не внутри отдельно взятого интеллекта, а за счет мультиагентности социума.

В случае восприятия также имеется выраженная склонность к выбору единственной модели или интерпретации. Это особенно хорошо видно на двойственных иллюзиях, когда человеческое зрение выбирает одну интерпретацию из двух равноценных. При этом человек сознательно может заставить зрение переключиться на другую интерпретацию, но не может видеть оба варианта одновременно. Это происходит как на достаточно низких уровнях,

скажем, структурного описания, так и на семантическом уровне. Хорошо известно множество двойственных иллюзий, приводить которые лишний раз не имеет смысла.

Однако здесь нужно особенно подчеркнуть, что выбор единственной непротиворечивой модели для предсказания при выборе действий не только не нужен, но даже вреден. Поэтому и реальный ИИ, вообще говоря, не обязан пытаться строить одну (и единую) модель среды, да еще и в явном виде. Попытки создания ИИ с поддержкой глобальных истинных моделей (большая база непротиворечивых аксиом) наталкивались на значительные трудности. Здесь возникают сложности с системами поддержания истинности при получении новой информации, становится проблематичной реализация индуктивного поведения и т.д.

Человек же вполне естественным образом может руководствоваться противоречивыми не только данными, но и моделями, в одних случаях используя одни модели, в других – другие. Можно даже сказать, что противоречивых данных «в природе» не бывает (по крайней мере, для ИМИ их нет); они появляются как эвристическая характеристика при упрощении моделей универсального предсказания. Также человек можно вполне что-то не разглядеть или не расслышать и выдвигать разные предположения о том, что же там было. То есть теоретически идеальное рассмотрение всех возможных моделей заменяется «умным» анализом результатов перебора моделей.

Выбор и инкрементное уточнение моделей в реальном ИИ, естественно, будет необходим. Каждый раз выполнять индукцию по всем имеющимся данным и перебирать все возможные модели крайне расточительно в условиях ограниченных ресурсов. Но, в то же время, ввод ресурсных ограничений должен быть не настолько жестким, чтобы оставалась единственная «истинная» модель.

В частности, мы уже видели, что необдуманное упрощение при переборе моделей и предсказании ведет к потере важных форм поведения. Также и в науке выбор между теориями осуществляется не просто на основе текущих данных, но разные теории рассматриваются одновременно для того, чтобы определить, какие эксперименты позволят получить новую информацию для уменьшения имеющейся неопределенности. Именно благодаря подобному индуктивному поведению (возможному при рассмотрении многих моделей) накапливается такая информация, которая настолько увеличивает различие качества моделей, что выбор становится почти однозначным. Близкие же по содержанию (и качеству) модели обрабатываются не как независимые разные модели, но как одна «нечеткая» модель. Введение подобных неопределенных/нечетких моделей может позволить получить эффекты индуктивного поведения даже при выборе единственной модели: действительно, некоторые действия будут приводить к большему уменьшению неопределенности, что будет позволять более гарантированно получать то или иное подкрепление при совершении следующих действий. Естественно,

остается требующий теоретического рассмотрения вопрос, как замену множеств моделей на «нечеткие» модели сделать наиболее эффективным образом.

Итак, модели появляются как «кэширование» результатов индукции, полученных в предыдущие моменты времени, и как ограничение перебора всего бесконечного множества моделей в универсальной индукции, но модели эти должны вводиться «мягко», чтобы не терялась универсальность.

Представления

Ограничение числа рассматриваемых моделей необходимо, но далеко не достаточно. Даже использование одной лучшей модели (то есть использование Колмогоровской сложности вместо алгоритмической вероятности) оказывается нереалистично затратным: время поиска модели длины L будет пропорционально 2^L. Мы отмечали, что сложность модели зависит от выбора опорной машины (способа программирования). Но даже при удачном выборе опорной машины длина каких-то моделей (описывающих реальные закономерности) будет оказываться слишком большой, чтобы эти модели можно было построить прямым поиском. Здесь нужны дополнительные метаэвристики.

В рамках проблемы зрительного восприятия нами уже предпринималась попытка сблизить универсальную индукцию и реальные методы анализа изображений [Potapov, 2012]. Речь идет о принципе репрезентационной минимальной длины описания, который проистекает из необходимости декомпозиции задачи построения модели полной истории взаимодействия агента со средой на подзадачи, которые решаются почти независимо. Если некоторую длинную строку сенсорных данных разбить на подстроки, то оценка сложности подстрок окажется гораздо больше, чем алгоритмическая сложность истории в целом. В этой связи такая непосредственная декомпозиция недопустима. Однако если из этих подстрок извлечена взаимная информация, которая используется в качестве априорной при описании каждой подстроки в отдельности, суммарная условная алгоритмическая сложность подстрок будет гораздо ближе к сложности истории. Эта взаимная информация может трактоваться (или быть выражена в форме) как представление (способ описания). Введение представлений сходно с выбором опорной машины, но отличается в двух аспектах: концепция представлений включает дополнительную метаэвристику декомпозиции, и для разных фрагментов данных могут использоваться разные представления, которым могут соответствовать не обязательно алгоритмически полные пространства моделей (тогда как опорная машина задает единое распределение $\xi(q)$).

Действительно, если взять, к примеру, зрение, то описание изображений (как в естественных системах, так и в прикладных автоматических методах) всегда осуществляется в рамках некоторого априорного представления, предназначение которого заключается не только в том, чтобы сместить

распределение вероятностей в пространстве моделей среды, но сделать допустимой их декомпозицию. В частности, благодаря использованию априорных представлений изображений методы компьютерного зрения оказываются применимыми к каждому изображению в отдельности вместо того, чтобы требовать на вход большое количество разнообразных изображений, для совокупности которых сложные закономерности, встречающиеся в каждом из изображений, могут быть выведены.

Понятие представления очень продуктивно. Идея представлений относится и к представлениям сенсорных данных, и представлениям знаний, и ментальным репрезентациям (то есть представления можно назвать общей когнитивной особенностью естественного интеллекта). Даже общая идея иерархических описаний, имеющая самостоятельную ценность, должна рассматриваться как идея о важном, но частном виде представлений. Иерархическая декомпозиция, естественно, потенциально более эффективная. Представления этого типа весьма распространены в машинном восприятии. Однако интенсивная иерархическая декомпозиция ведет к снижению качества моделей, строящихся в рамках соответствующих представлений. Компенсация этого негативного эффекта может быть осуществлена путем введения адаптивного резонанса. Опять же, в некоторых подходах к сильному ИИ значение адаптивного резонанса абсолютизируется (полагается, что он – ключ к СИИ). Хотя значение механизма адаптивного резонанса, безусловно, велико, необходимо понимать, что он является лишь одной из метаэвристик, которая может быть формализована в рамках теории универсальной индукции.

Стоит отметить недостаточность врожденных представлений даже для случая сенсорного восприятия. Имеется множество свидетельств в пользу того, что как у человека, так и у многих других животных, происходит адаптация представлений (даже очень низких уровней восприятия) к конкретному окружению. И для ИИ существует потребность в автоматическом конструировании представлений, которое может также исследоваться в теории универсальной индукции [Potapov et al., 2010] и которое, вероятно, должно быть элементом самооптимизации эффективного универсального ИИ, поскольку обучение представлениям является более конкретным и «прагматичным» способом инкрементного уточнения опорной машины, задающей распределение $\xi(q)$. Но при этом стоит опять подчеркнуть необходимость сохранения универсальности (алгоритмической полноты) множества представлений.

Планирование

Выше мы говорили об инкрементном построении моделей как способе сокращения перебора при постепенном удлинении истории. Однако задача выбора оптимальных действий также обладает высокой вычислительной сложностью, и для этой задачи вполне естественно вводить инкрементные

схемы решения. Такие схемы ведут к концепции планирования, которое также является одной из когнитивных особенностей человека.

Можно отметить, что не только ИМИ, но и методы слабого ИИ, основанные на «грубой силе», не используют планирования. В частности, это относится к успешным шахматным программам, чем они разительно отличаются от игроков-людей, которые практически обязательно опираются на планы, следуя принципу «плохой план лучше, чем никакого» [Bushinsky, 2009]. Планирование, включающее переиспользование результатов поиска, выполненных в предыдущие такты времени, позволяет экономить ресурсы. Действительно, планы строятся заранее (причем когда это позволяют обстоятельства, то есть при наличии свободных вычислительных ресурсов), и они лишь уточняются в процессе выполнения, что подразумевает отсутствие потребности в перестройке всего дерева поиска с нуля в каждый момент времени. Подобная стратегия, естественно, может быть включена в ИМИ, однако ее хорошая реализация может оказаться нетривиальной. В этом смысле шахматная программа является неэффективно интеллектуальной в отличие от человека. Однако то, что для узкого класса сред типа игры в шахматы проще оказывается обойтись неэффективным ИИ, не означает наличия этой же возможности для универсального интеллекта.

Планирование тесно связано с другими способами оптимизации поиска. Так, люди строят планы и выполняют поиск в терминах неких обобщенных действий. Чем более далекими являются планы, тем в терминах более абстрактных действий они описываются. Использование обобщенных действий является очевидно эвристичным. Эти действия также описываются в рамках некоторых представлений, но они напрямую не выводятся в теории универсальной индукции. На практике в методах слабого ИИ такие представления задаются априорно, и для них разрабатываются специфические алгоритмы планирования. Этого явно недостаточно для универсального ИИ.

Кроме самого планирования как инкрементного поиска и представлений для пространства поиска существует много эвристических приемов его сокращения. При этом, с одной стороны, методы поиска и оптимизации, такие как эвристическое программирование, имитация отжига, генетические алгоритмы и т.д. весьма проработаны в классическом ИИ. С другой стороны, общего решения проблемы поиска в настоящее время нет. Весьма вероятно, что какого-то единственного априорно эффективного метода поиска быть не может, и неизбежна необходимость каких-то стратегий самооптимизации, поскольку различные эвристики и специфические методы поиска лучше подходят для разных задач.

В настоящее время не существует теории эффективной прагматической общей самооптимизации, способной изобретать произвольные эвристики поиска. Однако если бы метод такой самооптимизации и существовал, он бы требовал для своего ускорения неких общих метаэвристик (в противном случае он бы не был прагматическим).

В целом видно, что планирование, как и прочие методы сокращения перебора, – это «всего лишь» элемент оптимизации вычислительных ресурсов. Его можно ввести и не как эвристику – с сохранением точного соответствия с ИМИ, но в таком виде оно будет не слишком эффективно. Более эвристические способы реализации планирования не будут работать для всех возможных сред, но при этом они могут быть весьма эффективными для конкретного, но весьма широкого класса сред. Так возникают такие концепции (которые по сути носят эвристический характер в том смысле, что для определенных классов сред они бессмысленны), как приостановление и возобновление выполнения плана. В то же время, остается открытым вопрос, какие именно механизмы планирования (и различные метаэвристики поиска) следует делать врожденными, а каким у интеллектуального агента будет шанс обучиться за обозримое время.

Знания

Знание играет особую роль в человеческом интеллекте. В то же время, в ИМИ знания в явном виде не используются. Вместо этого в ИМИ строятся холистические модели истории взаимодействия со средой без явного извлечения знаний из них. В принципе, знания нередко рассматриваются просто как верхний уровень иерархических моделей восприятия и управления (например, как верхний уровень зрительной системы). В этом контексте немногое может быть добавлено к тому, что уже обсуждалось в разделах, посвященных, восприятию, планированию и представлениям. Однако системы знаний имеют и свои особенности. В частности, только представления знаний (а не более низкоуровневые представления) являются модально-неспецифичными и описывают «смысл», и знания используются не только для описания внутренних моделей среды, но и используются для передачи между разными агентами (социальные взаимодействия составляют отдельный блок когнитивного смещения, о чем мы скажем ниже).

В целом, представления знаний, вероятно, могут быть обнаружены в процессе самооптимизации ИМИ, но этот процесс требует крайне длительного взаимодействия со средой. Полезные представления среды, абстрагированные от конкретных модальностей, могут дополнительно ускорить расширение ИМИ до прагматического эффективного СИИ. Но, опять же, эти представления не должны ограничивать универсальность ИМИ, как это имеет быть практически во всех существующих когнитивных архитектурах и более частных системах, основанных на знаниях.

Память

Можно сказать, что в ИМИ память есть, но самая примитивная. ИМИ просто хранит все сырые данные, не выполняя какой-либо иной функции. В то же время память является одним из центральных элементов большинства когнитивных архитектур. Также и память человека устроена гораздо сложнее, и ее функции далеко не ограничиваются просто хранением. Как известно, главная

функция памяти человека, составляющая основную проблему для воспроизведения на компьютере, – это извлечение информации по содержанию. Скажем, мы можем вспомнить какое-то событие, место, предмет, человека по их словесному описанию, фрагменту изображения, карандашному наброску и т.д.

В ИМИ ничего похожего как будто нет. Означает ли это, что универсальный агент не сможет проявлять то поведение, которое нам доступно благодаря нашей памяти? Вовсе нет. Память нам нужна, в первую очередь, для предсказания. Мы помним прошлое, чтобы предсказывать будущее (или, по крайней мере, осуществлять в будущем выбор более хороших действий). Сложно придумать какой-то иной биологический смысл памяти. По сути, естественная память настолько тесно интегрирована с функциями индукции и предсказания, что практически невыделима в чистом виде. Особое устройство памяти обусловлено тем, чтобы это делать вычислительно наиболее эффективно (с учетом особенностей нашего мира). Обоснуем второй тезис отдельно от первого. Если бы наша память хранила бы просто сырые данные (например, как один длинный фильм), то для того, чтобы найти в этом «фильме» сцены, удовлетворяющие некоторым критериям поиска, пришлось бы просмотреть весь фильм заново, обработав каждую сцену. Какой смысл это делать, если «фильм» уже один раз был просмотрен и проинтерпретирован? Естественно, экономичнее запоминать уже построенные его описания и проводить поиск сразу среди них.

При неограниченных ресурсах такая экономичность не нужна, и в каждый момент времени ИМИ просто заново обрабатывает всю историю взаимодействия. Уже отмечалось [Goertzel, 2010], что отсутствие памяти как когнитивной структуры в ИМИ связано с предположением неограниченных ресурсов. Но как только мы захотим повысить реалистичность нашего универсального агента с учетом ограниченных ресурсов, мы вынуждены будем усложнять структуру памяти и интегрировать ее с процедурами построения моделей, предсказанием и выбором действий.

Помимо особых функций естественной памяти, она обладает определенной организацией (эпизодическая/семантическая; кратковременная/долговременная и т.д.). Частично эта организация следует из других рассмотренных аспектов. К примеру, в ИМИ модель воспроизводит всю историю взаимодействия. Она одновременно описывает эпизодическое и семантическое содержание. Как только вводятся представления, которые не воспроизводят конкретные данные, но определяют «термины», в которых производится описание этих данных, появляется и соответствующее разделение типов памяти. Нужно рассматривать динамику развертывания представлений во времени, чтобы понять многие особенности устройства памяти.

Есть и другие особенности организации памяти, которые могут дать дополнительные элементы когнитивного смещения или эвристики поиска моделей и действий. К примеру, очевидной эвристикой является наличие

модально-специфической памяти. Отсюда следует банальный (но немаловажный в контексте ИМИ) вывод о том, что для упрощения индукции (процесса построения моделей) данные разных модальностей интерпретируются сравнительно независимо. Такое разделение кажется слишком естественным и само собой разумеющимся, но опять подчеркнем, что оно по сути эвристичное и далеко не полное.

Приведем еще одну весьма показательную особенность устройства памяти человека. Это чанки (chunks), которые берутся даже в качестве основы некоторых когнитивных архитектур [Gobet and Lane, 2010]. Вероятно, они связаны с предельной декомпозицией моделей в памяти (то есть разделение всего запоминаемого множества объектов на минимальные группы, объединяемые индивидуальными моделями). Возможно, чанки являются лишь эпифеноменом процесса декомпозиции задачи индукции, но они явно показывают, насколько сильно реальный интеллект пытается минимизировать затраты при ее решении.

Таким образом, особенности человеческой памяти являются важным источником элементов «когнитивного смещения», но правильное понимание этих особенностей также требует детального анализа в рамках универсального интеллекта.

Символьный и субсимвольный уровни

В методологии классического ИИ имеется достаточно жесткое деление на субсимвольные (например, нейросетевые) и символьные (например, логические) методы. Оно проявляется и в разделении когнитивных архитектур на эмерджентные и символьные. Сейчас возникает тенденция к объединению обоих подходов, в частности, в форме разработки гибридных архитектур. Но примечателен сам факт наличия такого деления. Ведь в ИМИ его нет. Есть ли оно у человека, то есть является ли явное деление на символьный и субсимвольный уровни особенностью естественной когнитивной архитектуры?

Достаточно очевидно, что выделение двух столь разных уровней связано с тем, что верхний уровень доступен посредством сознания, а нижний – посредством нейрофизиологических исследований (результаты которых наиболее непосредственным образом могут быть связаны с низшими уровнями сенсомоторных представлений). Промежуточные уровни просто недоступны непосредственному наблюдению, поэтому гораздо хуже исследованы. В этой связи в ИИ иногда выделяется «проблема среднего слоя» (или проблема «семантической пропасти») как одна из наиболее трудных. Однако наличие промежуточных уровней организации, хотя и несколько сглаживает остроту дихотомии символьный/субсимвольный, но не отменяет того факта, что в естественном интеллекте присутствуют разделимые уровни организации.

Такое четкое деление вряд ли возникнет само по себе, если его не заложить архитектурно. В частности, вряд ли его можно было бы выделить в моделях, формируемых ИМИ (в этих моделях даже если и будут

присутствовать какие-то концепты разных уровней, они будут безнадежно перемешаны). В естественном же интеллекте не только строящиеся модели среды имеют достаточно выраженную многоуровневую структуру, но и методы работы на разных уровнях заметно различаются (на себя обращает внимание хотя бы тот факт, что сознание привязано преимущественно к моделям верхних уровней). Так, на субсимвольном уровне преимущественно учитываются типичные закономерности в большом массиве сенсорных данных, тогда как символьный уровень работает с произвольными закономерностями, но в сильно редуцированных данных. Однако это лишь общая характеристика уровней. При их введении все же должна сохраниться универсальность в форме прямых и обратных связей между уровнями, а также в виде возможности построения любых вычислительных предикатов (базовых перцептивных понятий) на субсимвольном (и промежуточных символьных) уровне. Примером могут служить законы гештальта (законы перцептивного группирования), которые типичны для всех людей, но все же у культурных и примитивных людей могут различаться (что может проявляться, например, в (не)восприимчивости к некоторым оптическим иллюзиям). Иными словами, законы перцептивного группирования соответствуют типичным закономерностям в сенсорных данных, но эти закономерности могут быть вынесены или не вынесены в соответствующие представления в зависимости от особенностей онтогенеза.

Все это может трактоваться как общая априорная структура представлений и эвристик построения моделей в их рамках, которые (в дополнение к самой концепции представлений) обеспечивают существенную экономию вычислительных ресурсов (но без фатального нарушения универсальности).

Ассоциации

Существует множество когнитивных особенностей человеческого мышления, которые так или иначе связаны с ассоциированием. Это многоликий феномен, поскольку ассоциирование может выполняться как для представлений, так и для моделей, причем на всех уровнях абстракции. Но во всех случаях имеется нечто общее.

Очевидно, декомпозиция реальных задач как индукции, так и выбора действий всегда оказывается неполной. Процессы, родственные ассоциированию, можно пытаться проинтерпретировать как процессы, устанавливающие возможные связи между элементами данных, моделей, представлений, которые в результате декомпозиции стоящей перед универсальным интеллектом единой задачи рассматривались исходно независимо.

Наиболее очевидной такая интерпретация является для случая декомпозиции задачи индукции. Ассоциация устанавливается между двумя моделями фрагментов сенсорных данных, если между ними имеется взаимная информация, которая может выражаться в статистических терминах (частое

взаимное появление) или в структурных терминах (наличие простого алгоритма, переводящего одну модель в другую). Последнее также является основой аналогий и метафор.

Примером наиболее сложного ассоциирования является трансферное обучение, при котором представления из одних предметных областей переносятся в другие предметные области. Тот факт, что такое в принципе возможно и полезно, является свидетельством особых свойств нашего мира, на использовании которых и основывается трансферное обучение. Хотя само наличие остаточных связей, взаимной информации между разными предметными областями и не является каким-то особым свойством нашей среды (скорее, было бы удивительно, если бы таких взаимосвязей не было вообще), но способность находить и использовать эти связи в условиях ограниченных ресурсов является показательной. Она, во-первых, явно свидетельствует об универсальности человеческого интеллекта, отсутствии у него жестких ограничений на структуру строящихся представлений, на установление связей между фрагментами реальности, а, во-вторых, является при этом и элементом когнитивного смещения.

Сложно сказать, являются ли механизмы трансферного обучения, установления ассоциаций, аналогий, метафор, существенно различными, или это разные приложения одного механизма. Но все эти механизмы (как и упоминавшийся адаптивный резонанс) можно рассматривать и как способы снижения требований к ресурсам, и как способы устранения негативного эффекта от этого снижения в зависимости от того, идем ли мы от прагматического эффективного интеллекта в сторону универсальности или от универсального интеллекта в сторону эффективности/прагматичности. Сейчас трансферное обучение рассматривается отдельно от проблематики универсального ИИ, и не удивительно, что современные модели трансферного обучения чрезмерно специализированы: в них отображение между двумя представлениями (между которыми осуществляется трансфер знаний) всегда задается вручную и работает только для них. Видимая универсальность трансферного обучения у человека говорит о том, что оно должно очень тесно прилегать к ядру универсального ИИ.

Повторимся, что в ИМИ трансферного обучения в отдельности нет, и оно там не нужно (но лишь благодаря нереалистичным неограниченным ресурсам): любая взаимная информация между любыми фрагментами данных там учитывается, а трансфер на уровне эвристик поиска не нужен в силу отсутствия таковых. Поскольку трансфер осуществляется на уровне представлений, то в теории он должен появляться вместе с ними и позволять осуществлять более плавный переход от универсального ИИ к использованию представлений в реальном ИИ.

Трансферное обучение [Senator, 2011] – пример наиболее развитого ассоциирования. Не менее примечательно (своей чрезвычайной распространенностью) и самое низкоуровневое ассоциирование. На

поведенческом уровне это условные рефлексы (а на еще более низком нейросетевом уровне это правило Хебба). Конечно, под самим ассоциированием нередко понимают нечто более сложное, чем просто условные рефлексы (к примеру, по В.Ф. Турчину ассоциирование – это система управления сложными рефлексами, то есть метасистема по отношению к наиболее развитым рефлексам). Однако основа у них одинаковая.

Ассоциирование нередко рассматривают в качестве самостоятельного (иногда – основополагающего) принципа работы естественного мышления, противопоставляя при этом индукции (якобы в индукции обязательно строятся явные модели, тогда как ассоциирование вообще безмодельно и не связано с какой-либо направленной оптимизацией, а связано с уникальными принципами самоорганизации). Конечно, за ассоциированием стоит очень эффективная метаэвристика, отражающая регулярно встречающуюся особенность нашего мира (которая, грубо говоря, сводится к тому, что чем ближе события во времени и пространстве, тем более вероятно они связаны; но, конечно, развитое ассоциирование этим не ограничивается). Естественно, эвристики (включая ассоциирование) не выводятся из теории непредвзятого универсального интеллекта и в этом смысле могут считаться дополнительными принципами.

Однако ассоциирование нельзя считать ни единственной, ни главной основой мышления. Это хорошо видно и на примере правила Хебба, и на примере рефлексов. Так, само по себе правило Хебба оказывается недостаточным для решения сложных задач обучения, связанных с построением инвариантов. В случае рефлексов также основную сложность представляет не усиление связи между двумя известными стимулами, а выделение классов связываемых стимулов, которые могут описываться произвольными закономерностями (стимулом может быть включение просто лампочки, лампочки определенной яркости или цвета, двойное включение лампочки сначала ярче, а потом тусклее и т.д.). Примечательно, что разные животные имеют разные способности по выявлению закономерностей в стимулах. Так, курицы не способны научиться выбирать более светлую кормушку (из нескольких кормушек, в одной из которых лежит невидимый до момента выбора корм). И даже обезьянам сложно абстрагироваться от локального контекста (например, использовать для доставания плода предметы, в текущий момент не находящиеся в поле зрения). Человеческий же интеллект универсален, и эта универсальность не объясняется ассоциированием, а сочетается с ним.

Рассуждения

Рассуждения – это то, что часто считается собственно мышлением. Есть ли в AIξ рассуждения? В некотором смысле, есть. Некоторые наши рассуждения сводятся к тому, к чему нас приведет то или иное действие (а именно на определение этого и тратятся все ресурсы AIξ). Скажем, обдумывая предстоящий разговор, мы можем предполагать, что нам будет сказано, и что

на это можно будет ответить, а также то, какие эмоции мы при этом испытаем. Однако наши рассуждения далеко не всегда напрямую связаны с предсказанием того, какой сенсорный вход и какое подкрепление мы получим при выполнении тех или иных действий. Нередко мы думаем о вещах, не касающихся нас напрямую. Причем зачастую в наших рассуждениях (которые нам интроспективно доступны) нет никакого намека на индукцию. Ведь с мышлением чаще ассоциируется дедукция. Не случайно во многих экспертных системах рассуждения моделируются с помощью механизмов логического вывода. В ИМИ никаких механизмов логического вывода нет. Хотя в некоторых моделях типа $AI\xi^{tl}$ или машины Гёделя логика вводится для обоснования утверждений относительно алгоритмов, но к обычным рассуждениям это отношения почти не имеет. Не свидетельствует ли это, что в ИМИ чего-то принципиально не хватает?

На самом деле, достаточно очевидно, что не вполне свидетельствует. В методах дедуктивного вывода делается перебор вариантов допустимых цепочек правил вывода, пока не будет получено доказываемое утверждение или его опровержение. Такой перебор сходен с перебором, осуществляемым в ИМИ для одной фиксированной модели среды. Понятное различие заключается в том, что в ИМИ он выполняется в каждый момент времени для целостных моделей сред, которые также перебираются заново с учетом только что полученной информации. А в эффективных прагматических системах это просто невозможно, и приходится рассматривать модели только фрагментов среды, и даже считать эти модели фиксированными. Анализируемый фрагмент среды может быть с нами напрямую не связан, и мы можем относительно него рассматривать те действия, которые совершаем не сами (например, можем думать о том, что произойдет с планетой, если около нее взорвется сверхновая). Склонность анализировать крайне косвенно связанные с нами фрагменты реальности (и даже создавать воображаемые миры) весьма любопытна, но требует отдельного обсуждения и относится, скорее, к вопросам мотивации (целевой функции). Сложно представить, что ИМИ для выбора собственных действий в целях максимизации целевой функции будет (пусть и виртуально) предаваться отвлеченным размышлениям о строении Вселенной (точнее, стремиться получить для этого необходимую информацию), но и противоречия здесь нет, особенно, если за создание хороших космологических теорий он будет получать подкрепление.

Сейчас же нам важно, что дедуктивный анализ моделей фрагментов среды связан с экономией ресурсов. Результаты вычислений алгоритмической модели при разных последовательностях воздействий могут быть запомнены и повторно использованы при неизменности этой модели. Естественно, способы подобной экономии ресурсов тесно связаны с вопросами представлений (причем представлений декларативных, что может иметь отношение к расширению понятия вычислимости) и могут быть крайне нетривиальны. И, конечно, из ИМИ они не выводятся. Так, логика может трактоваться как

метапредставление, полезное для выполнения анализа фрагментов конкретно нашего мира, поскольку возможность выделения объектов и отношений является общим (но нестрогим) его свойством, которое вполне могло бы быть иррелеватно какой-то другой реальности, где наша логика была бы бесполезна. Здесь необходимо проделать большую работу по выявлению принципов осуществления эффективных рассуждений (которые так же далеки от полного перебора элементарных действий, как методы обработки изображений далеки от универсальной индукции на основе алгоритмической вероятности). При этом «кэширование» результатов анализа фиксированных моделей вызовет дополнительные вопросы, связанные с обновлением этих результатов при получении новой информации (в частном виде это хорошо известная проблема «замкнутости мира» или немонотонных рассуждений), что также требует решения.

Здесь опять можно было бы подумать о том, что ИМИ не дает вклада в решение и без него известных проблем (например, логического вывода и поддержания истинности). Однако еще раз отметим, что ИМИ ставит эти известные проблемы в гораздо более общем виде. Так, логика предикатов в рамках универсального интеллекта представляется лишь в качестве метапредставления, которое имеет эвристическую природу и не обязательно должно задаваться априорно: самооптимизирующийся универсальный интеллект (например, человеческий) самой логике и ее эффективному использованию может обучиться; наша задача – создать такой интеллект и сократить время его обучения до приемлемого. Сделать это может оказаться проще, чем вручную создавать множество частных методов, так же, как проще, скажем, реализовать некий метод обучения, чем закладывать все нужные частные факты вручную. При этом при создании ИИ следует попытаться добиться меньших когнитивных искажений, чем у человека. Так, хотя априорное предпочтение в разделении воспринимаемого мира на объекты со свойствами и отношениями, возможно, существенно ускорит обучение, но это предпочтение не должно быть слишком жестким.

Социальные взаимодействия

Взаимодействия с другими интеллектуальными агентами составляет очень существенную часть среды. Эти агенты очень сложны, так что индуктивная реконструкция подходящих моделей других агентов будет требовать очень длительного взаимодействия в реальном мире и огромного количества вычислительных ресурсов. Естественно, некоторая theory of mind (способность моделирования разума, в частности, других агентов) должна быть встроена в эффективный прагматический ИИ. Но в универсальный ИИ она должна добавляться в качестве элемента когнитивного смещения, которое задает смещение моделей, но не накладывает на них непреодолимых ограничений.

Социальные взаимодействия не ограничиваются только предсказанием поведения (или реконструкцией моделей) других агентов как части окружения. Естественно, социальные агенты взаимодействуют друг с другом так же, как и с прочим окружением, – через сенсорику и моторику. Но через них они могут передавать друг другу фрагменты моделей среды, стратегии поведения и даже элементы целевых функций. В действительности именно социум формирует сложные целевые функции, индуктивное смещение и эвристики поиска (в форме этики, науки, искусства и т.д.), благодаря обмену информацией и вычислительными ресурсами между агентами. Непредвзятый универсальный агент может научиться за достаточное время (если в течение этого времени кто-то будет обеспечивать его выживание) правильно интерпретировать сенсорные данные, выявляя из них эту информацию (хотя для обучения целевым функциям потребуются некоторые врожденные механизмы). Но эффективный прагматический интеллект должен иметь эту способность априорно, то есть обладать индуктивным предпочтением социальных сред [Dowe et al., 2011] или иметь «коммуникационные приоры» [Goertzel, 2009]. Конечно, чем более высокоразвитым является животное, тем менее подготовленными к самостоятельной жизни рождаются его детеныши, и универсальному ИИ можно простить длительную «постнатальную» беспомощность, но все же такие априорные навыки, как выделение в сенсорном потоке образов других агентов и подражание, могут весьма заметно сократить период полной беспомощности.

Существенный (но не единственный) аспект социальных взаимодействий – это язык. Анализ языка в контексте универсальных агентов до сих пор мало исследован. К примеру, обсуждалась значимость кодирования в двух частях (в рамках принципа минимальной длины сообщения), которое позволяет агентам эффективно обмениваться регулярными частями моделей, отделенными от шума [Dowe et al., 2011]. Но основная часть важных вопросов до сих пор требуют детального анализа. Сюда относится и семантическое обоснование символов, и та ясная проблема, что для универсальных агентов будет наиболее эффективно (по крайней мере, сначала) перенимать знания, накопленные человечеством, для чего нужно понимать естественные языки, которые связаны с определенными способами представления знаний, а, значит, формирование этих представлений не должно требовать чрезмерных усилий со стороны ИИ.

Один дополнительный немаловажный аспект мультиагентных взаимодействий заключается в том, что окружение оказывается гораздо более сложным и вычислительно мощным, чем сам агент. Этот аспект не является эвристикой или индуктивным смещением, но он также требует учета в моделях ИМИ.

Эмоции

Эмоции нередко рассматривают как компонент когнитивной архитектуры, поэтому их необходимо также обсудить. В то же время, эмоции явно связаны с целевой функцией, поэтому их предназначение (в отличие от других элементов

когнитивной архитектуры) не может полностью сводиться к экономии ресурсов и сокращению времени обучения, для чего служат эвристики поиска и индуктивное смещение.

Проблему целевой функции мы уже кратко обсуждали. «Хорошая» целевая функция (например, хотя бы точно оценивающая выживание) не может быть задана априорно. Врожденная целевая функция является грубой «эвристической» аппроксимацией некоторой «истинной» целевой функции. К примеру, боль и удовольствие являются очень грубой аппроксимацией функции приспособленности – смерть может быть безболезненной, а операция, спасающая жизнь, может сопровождаться сильной болью. Эмоции и другие компоненты оценки качества ситуации позволяют выполнять более точную аппроксимацию. Некоторые из них врожденные. Другие приобретаются в течение жизни.

При этом стоит разделять эвристики аппроксимации истинной целевой функции и оценку качества состояний, учитывающую потенциальные значения целевой функции, связанные с ожидаемыми (предсказываемыми) состояниями. Так, мы можем избегать тех ситуаций, которых боимся, не думая каждый раз о причинах страха. А это уже обычные эвристики, которые не определяют максимизируемую целевую функцию, а сокращают перебор возможных действий при фиксированной целевой функции. К примеру, удовольствие от удовлетворения любопытства и эстетическое удовольствие могут быть введены как отдельные компоненты базовой целевой функции (для чего существуют модели на основе алгоритмической теории информации [Schmidhuber, 2010]). Или же интеллектуальный агент может проявлять любопытство, если он из опыта в состоянии предсказать, что получение новой информации будет полезно для его выживания (точнее, для получения телесного удовольствия и избегания боли). Поскольку это сложная задача предсказания, агент может выработать «чувство любопытства» как элемент беспереборной оценки будущего вознаграждения для экономии вычислительных ресурсов. Стоит подчеркнуть, что эти два варианта принципиально различны, поскольку соответствуют разным максимизируемым целевым функциям, из-за чего соответствующие агенты в некоторой ситуации могут делать разный выбор (что мы уже обсуждали ранее). При этом иметь место в действительности может как каждый из этих вариантов в отдельности, так и их комбинация.

Поскольку человеческий интеллект является эффективным прагматическим, в нем суммарное будущее подкрепление предсказывается преимущественно без явного обращения к базовой целевой функции. Из-за этого обучение самой целевой функции оказывается тесно переплетенным с обучением эвристикам предсказания ее будущих значений. При этом соответствующие механизмы обучения обладают собственным индуктивным смещением, в том числе и в форме неких априорных представлений. К примеру, для некоторых эмоций могут отсутствовать врожденные механизмы «вычисления» их значений, но выделенные типы эмоций могут быть заданы на

уровне представлений. В связи со всем этим относительно человеческих эмоций, чувств и т.д. сложно определить, в какой мере какая из них относится к слагаемым базовой целевой функции, а в какой мере – к эвристикам предсказания ее будущих значений. По этой причине и в психологии до сих пор не удается прийти к единому мнению о механизмах эмоций, их роли и происхождении. Тем не менее, вся эта часть когнитивной системы естественного интеллекта вполне поддается интерпретации в рамках моделей универсального интеллекта, в том числе, и в терминах повышения их эффективности.

Внимание

Такая когнитивная функция, как внимание, является весьма широким феноменом. Однако достаточно очевидно, что его возникновение связано с ограничениями на ресурсы. Например, визуальное внимание направлено на наиболее информативные или значимые (в терминах целевой функции) части сцены, что подразумевает, что эти части подробно анализируются с использованием большего количества ресурсов по сравнению с другими частями. Естественно, распределение ресурсов при решении других когнитивных задач также может быть интерпретировано как внимание.

Этот тезис можно расширить и на мультиагентные архитектуры. Вряд ли можно полагать, что интеллект в принципе не способен решать множество задач параллельно, сохраняя при этом некое единство. По крайней мере, для ИМИ это возможно (и не грозит шизофренией), поскольку ИМИ может работать с любым числом разных источников данных и проводить индукцию и выбор действий одновременно в стольки раздельных «телах», в скольки это нужно. Если разные источники данных будут содержать взаимную информацию, это будет «автоматически» учтено. То есть при неограниченных ресурсах интеллекту, обрабатывающему данные, приходящие от разных тел, нет нужды сосредотачивать внимание на каком-то одном из них. Феномен внимания возникает при введении ограничений на ресурсы, что подразумевает обработку в первую очередь тех фрагментов данных, которые релевантны наиболее насущным задачам с учетом ограниченного набора доступных действий.

Здесь стоит отметить еще одну сторону феномена внимания: его также можно трактовать как направленность действий на какой-то конкретный объект. Такое «внешнее» внимание обусловлено распределением времени между не «внутренними» вычислительными операциями, а внешними действиями. Такое «внешнее» внимание в ИМИ должно реализовываться «автоматически»: универсальный агент должен быть вполне способен, скажем, направить камеру по направлению на резкий звук для получения информации, существенной для избежания сильного снижения значения своей целевой функции (естественно, если у этого агента имеется априорная информация, свидетельствующая о возможной связи громкого звука с опасностью).

Внутреннего же внимания как распределения ограниченных ресурсов в ИМИ нет, так что сведения о том, как работает внимание у человека, может быть полезным для введения данного элемента когнитивного смещения.

Существует множество моделей внимания для когнитивных архитектур (например, [Ikle' et al., 2009], [Harati Zadeh et al., 2008]). Можно сказать, что механизмы внимания присутствуют даже в простых универсальных решателях (например, [Hutter, 2002]), которые принимают в расчет вычислительную сложность и пытаются оптимально аллоцировать ресурсы между разными рассматриваемыми гипотезами. Естественно, более развитые механизмы внимания должны присутствовать в эффективном прагматическом ИИ. Но детали этих механизмов существенно зависят от других частей когнитивной архитектуры. Таким образом, содержательные модели внимания должны разрабатываться совместно с ресурсно-ограниченными расширениями моделей ИМИ.

Метакогнитивные функции

Нередко считают, что главное, что отделяет компьютер от человека, – это отсутствие у первого таких функций как самосознание, понимание и т.д. Это мнение свойственно не только людям, далеким от ИИ, но и людям, которые им занимаются (по крайней мере, в философском плане). Даже сильный ИИ Сёрлом был определен как ИИ, обладающий всеми такими функциями. И невозможность истинного понимания – это то, что приписывается компьютерам Пенроузом и прочими сторонниками взглядов о невозможности сильного ИИ.

Многие специалисты, не ограничивающиеся общими рассуждениями об ИИ, а занимающиеся разработкой конкретных решений в этой области, видят гораздо более серьезные трудности, например, в проблемах поиска, обучения, представления знаний и т.д., тогда как указанные «человеческие» функции считают не столь сложными. Так, самосознание интерпретируется просто как модуль управления верхнего уровня, получающий и обрабатывающий информацию о работе других блоков когнитивной архитектуры. Понятно, что подобные *метакогнитивные функции* в полной мере не реализовать без самого интеллекта. Тогда компьютер не наделяется самосознанием не потому, что это что-то загадочное и присущее только человеку, а потому, что не реализованы более базовые функции. Из-за этого технические специалисты нередко сторонятся этих аспектов мышления, считая их «гуманитарными» и в противовес философам интерпретируя их слишком упрощенно. Тем не менее, метакогнитивные функции начинают привлекать все большее внимание [Anderson and Oates, 2007] и даже в каком-то виде реализуются в некоторых когнитивных архитектурах [Shapiro et al., 2007] (хотя эти реализации достаточно интересны и познавательны, на наш взгляд, они являются «слабыми»). Нельзя полностью обойти их обсуждение и в контексте разговора об универсальном интеллекте.

Действительно, в моделях ИМИ в явном виде ни самосознание, ни понимание не реализуются, что вызывает естественный вопрос о том, не упущено ли в этих моделях что-то важное. Из анализа ряда метакогнитивных функций (метаобучение, метарассуждения) ясно [Anderson and Oates, 2007], что их предназначение связано с компенсацией неоптимальности работы базовых когнитивных функций. При этом причина таких ошибок, скажем, обучения, которые могут быть исправлены самим агентом, может быть связана только с тем, что на решение соответствующей задачи обучения было выделено недостаточно ресурсов. Ведь при использовании универсальной индукции при неограниченных ресурсах результат в принципе не может быть улучшен на тех же данных, и метаобучение бессмысленно. Конечно, метакогнитивные функции не сводятся просто к перераспределению ресурсов (это лишь частный прием, являющийся прерогативой внимания). Так, в случае обучения экономия ресурсов может проявляться в использовании только части данных, игнорировании контекста, использовании упрощенных представлений и т.д. И метаобучение должно заниматься не тем, что оттягивать больше ресурсов на метод универсального обучения, а тем, что оценивать успешность работы блока обучения и привлекать, к примеру, более общие методы тогда, когда более простые методы потерпели неудачу. Вводится даже понятие т.н. метакогнитивного цикла, в котором должно определяться «что пошло не так и почему» [Shapiro and Göker, 2008].

Такая интерпретация метакогнитивных функций является слишком общей. Относительно же конкретных функций возникают вопросы. Так, понимание (которое, правда, не всегда трактуют как метакогнитивную функцию, но которое все же, на наш взгляд, обладает несомненными атрибутами таких функций) не так просто связать с «когнитивным смещением». Существует много примеров, показывающих, что конкретные системы (слабого) ИИ не реализуют понимания. Но эти примеры не свидетельствуют о принципиальной невозможности машинного понимания, а как раз позволяют определить роль понимания в экономии ресурсов. Мы уже рассматривали классический пример шахматной позиции, в которой компьютерная программа, способная обыграть гроссмейстера, играет неправильно из-за отсутствия понимания этой позиции. При неограниченных вычислительных ресурсах в результате глубокого перебора программа могла бы избегать ошибочного хода. Более того, возможно такое (алгоритмическое) описание данной ситуации (например, в форме оценивающей функции), которое позволяло бы без перебора определить ошибочность соответствующего хода. То есть понимание ситуации связано с использованием такого ее представления, которое позволяет выбирать эффективные действия без больших затрат вычислительных ресурсов.

Аналогичный вывод можно сделать и при использовании других примеров. Так, показательна следующая классическая задача. Имеется доска 8x8 клеток, из которой вырезаны две угловые клетки, находящиеся на одной

диагонали. Требуется замостить доску с помощью костяшек домино 1x2 клетки. Неэффективный интеллект (не обладающий пониманием, но имеющий неограниченные ресурсы) мог бы перебрать все варианты замощения. Человек же испытывает эффект понимания, когда представляет, что эта доска имеет шахматную раскраску, так что на ней оказывается 32 клетки одного цвета и 30 клеток другого цвета при том, что каждая костяшка занимает обязательно по одной клетке разных цветов. Выбор подходящего представления делает задачу элементарной. Возможно, еще более показательны такие задачи на «творческое мышление», как задача о построении четырех равносторонних треугольников с помощью шести спичек. Здесь выбор представления ситуации также имеет принципиальное значение.

И понимание изображений связано с построением их описаний в рамках определенных представлений (которые, как правило, должны облегчать совершение адекватных действий). Видимо, то же можно сказать и о понимании естественного языка, хотя оно включает и дополнительную проблематику.

Возможно, понимание – это не само использование эффективных представлений, а метакогнитивная функция, которая дает (доступную для сознания) оценку эффективности представлений. Если человек чего-то не может понять или понимает недостаточно хорошо, он зачастую (хотя и не всегда) отдает себе в этом отчет; равно как человеку доступно и ощущение достижения ясного понимания, что, вероятно, должно быть связано с проблематикой самооптимизации.

Доступ к внутреннему содержанию процессов мышления характерен для всех метакогнитивных функций, что интегрально выражается в феномене самосознания. Ничего похожего в ИМИ в явном виде нет (за ненадобностью контроля собственных мыслей – они и так идеальны), но это не означает, что он не сможет вести себя как самоосознающий агент. Но сможет ли он корректно пользоваться такими выражениями, как «я думаю», «я полагаю», «я знаю», «я умею», «я хочу», «я помню» и т.д., если в нем методы выбора действий не получают никакой информации о собственной работе (а использование таких выражений может быть важно для выживания в существующей мультиагентной среде)? Однозначно ответить на этот вопрос непросто. Возможно, ИМИ сможет правильно (в прагматическом плане) использовать эти выражения без понимания их смысла, но для этого потребуется крайне обширный опыт взаимодействия с социальной средой и, естественно, неограниченные вычислительные ресурсы. Ведь произнесение слов принципиально не отличается от какого-либо моторного выхода, и если существует вычислимое отображение между входными воздействиями и требуемыми выходными воздействиями, то ИМИ может его реконструировать по подходящей истории взаимодействий. Тем не менее, возможность «неосознанного» использования действий, требующих отсутствующей интроспективной информации, продолжает вызывать сомнения. К счастью,

развеивать эти сомнения не нужно, так как для создания эффективного прагматического ИИ доступ к этой информации пригодится не только для общения с другими агентами, но и для самооптимизации.

Использование этой информации нетривиально. Мы уже отмечали, что помещение ИМИ в среду, включающую другие ИМИ, вызывает противоречие (один агент моделирует другого агента, который, в свою очередь, моделирует первого агента, и так далее до бесконечности). Полная интроспекция вызвала бы аналогичное противоречие. Данное противоречие снимается вместе с вводом ресурсных ограничений, которые, однако, нарушают и абстрактную идеальную интеллектуальность ИМИ. Значит, проблема интроспекции (и в целом проблема «theory of mind») не решается в рамках ИМИ и требует разработки дополнительных принципов. И хотя проблематика «theory of mind» (и метакогнитивные функции в целом) связана с «когнитивным смещением» (особенно в части эвристик самооптимизации), она также может иметь отношение к недостаточной универсальности базовых моделей ИМИ.

Выводы

Мы разобрали некоторые когнитивные особенности человеческого мышления, которые достаточно естественно могут трактоваться как эвристики и индуктивное смещение, обеспечивающие эффективную прагматичность естественного интеллекта, то есть приемлемую его работу в определенном классе сред в условиях ограниченных ресурсов и времени обучения.

В целом требования к ограниченным ресурсам – это не новость; и достаточно очевидно, что многие когнитивные особенности берутся именно отсюда. Однако до сих пор нетривиального рассмотрения связи математической теории универсального ИИ и сложных когнитивных архитектур не проводилось [Goertzel and Iklé, 2011]. Для установления такой связи нужно не просто поверхностно описывать когнитивные функции, а строго вводить их в качестве расширения моделей ИМИ с сохранением той универсальности, которой эти модели обладают. Именно эта задача нами далее будет решаться.

Литература

(Anderson and Oates, 2007) Anderson M.L., Oates T. *A Review of Recent Research in Metareasoning and Metalearning* // AI Magazine. 2007. V. 28. No. 1. P. 7–16.

(Bushinsky, 2009) Bushinsky Sh. *Deus Ex Machina – A Higher Creative Species in the Game of Chess* // AI Magazine. 2009. V. 30. No. 3. P. 63–70.

(Dowe et al., 2011) Dowe D., Hernández-Orallo J., Das P. *Compression and Intelligence: Social Environments and Communication* // Lecture Notes in Computer Science 6830 (proc. Artificial General Intelligence – 4th Int'l Conference). 2011. P. 204–211.

(Gobet and Lane, 2010) Gobet F., Lane P.C.R. *The CHREST Architecture of Cognition. The Role of Perception in General Intelligence* // E.Baum, M.Hutter, E.Kitzelmann (Eds), Advances in Intelligent Systems Research. 2010. V. 10 (Proc. 3rd Conf. on Artificial General Intelligence, Lugano, Switzerland, March 5-8, 2010.). P. 7–12.

(Goertzel, 2009) Goertzel B. *The Embodied Communication Prior* // In: Yingxu Wang and George Baciu (Eds.). Proc. of ICCI-09, Hong Kong. 2009.

(Goertzel, 2010) Goertzel B. *Toward a Formal Characterization of Real-World General Intelligence* // E.Baum, M.Hutter, E.Kitzelmann (Eds), Advances in Intelligent Systems Research. 2010. V. 10 (Proc. 3rd Conf. on Artificial General Intelligence, Lugano, Switzerland, March 5-8, 2010.). P. 19–24.

(Goertzel and Iklé, 2011) Goertzel B., Iklé M. *Three Hypotheses About the Geometry of Mind* // Lecture Notes in Computer Science 6830 (proc. Artificial General Intelligence – 4th Int'l Conference). 2011. P. 340–345.

(Harati Zadeh et al., 2008) Harati Zadeh, S., Bagheri Shouraki, S., Halavati, R.: *Using Decision Trees to Model an Emotional Attention Mechanism* // Frontiers in Artificial Intelligence and Applications (Proc. 1st AGI Conference). 2008. V. 171. P. 374–385.

(Hutter, 2002) Hutter M. *The Fastest and Shortest Algorithm for All Well-Defined Problems* // International Journal of Foundations of Computer Science. 2002. V. 13. No. 3. P. 431–443.

(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.

(Ikle' et al., 2009) Ikle' M., Pitt J., Goertzel B., Sellman G. *Economic Attention Networks: Associative Memory and Resource Allocation for General Intelligence* // In: B. Goertzel, P. Hitzler, M. Hutter (Eds), Advances in Intelligent Systems Research. 2009. V. 8 (Proc. 2nd Conf. on Artificial General Intelligence, Arlington, USA, March 6-9, 2009). P. 73–78.

(Pavel et al., 2007) Pavel A., Vasile C., Buiu C. *Cognitive vision system for an ecological mobile robot* // Proc. 13 Int'l Symp. on System Theory, Automation, Robotics, Computers, Informatics, Electronics and Instrumentation. 2007. V. 1. P. 267–272.

(Potapov and Rozhkov, 2012) Potapov A.S., Rozhkov A.S. *Cognitive Robotic System for Learning of Complex Visual Stimuli*. 2012. (in print)

(Potapov et al., 2010) Potapov A.S., Malyshev I.A., Puysha A.E., Averkin A.N. *New paradigm of learnable computer vision algorithms based on the representational MDL principle* // Proc. SPIE. 2010. V. 7696. P. 769606.

(Potapov, 2012) Potapov A.S. *Principle of Representational Minimum Description Length in Image Analysis and Pattern Recognition* // Pattern Recognition and Image Analysis. 2012. V. 22. No. 1. P. 82–91.

(Schmidhuber, 2010) Schmidhuber J. *Artificial Scientists & Artists Based on the Formal Theory of Creativity* // In: E.Baum, M.Hutter, E.Kitzelmann (Eds), Advances in Intelligent Systems Research. 2010. V. 10 (Proc. 3rd Conf. on Artificial General Intelligence, Lugano, Switzerland, March 5-8, 2010). P. 145–150.

(Senator, 2011) Senator T.E. *Transfer Learning Progress and Potential* // AI Magazine. 2011. Vol. 32. No. 1. P. 84–86.

(Shapiro et al., 2007) Shapiro S.C., Rapaport W.J., Kandefer M., Johnson F.L., Goldfain A. *Metacognition in SNePS* // AI Magazine. 2007. Vol. 28. No. 1. P. 17–31.

(Shapiro and Göker, 2008) Shapiro D. and Göker M.H. *Advancing AI Research and Applications by Learning from What Went Wrong and Why* // AI Magazine. 2008. V. 29. No. 2. P. 9–10.