Семинар “Cogito ergo…”

Общий искусственный интеллект и универсальные рациональные агенты

Алексей Потапов

25 апреля 2014, СПб

Искусственный интеллект – смена настроений

- 50-е годы: ИИ будет создан скоро
  - Общий решатель задач
- 70-е годы: нужно разрабатывать прикладные системы ИИ
  - Экспертные системы
- 90-е годы: «зима ИИ»
  - … Развитие подобластей ИИ с успешными техническими решениями

ИИ уровня человека? Серьезно!

- **2005 год: «официальное» возрождение интереса к сильному ИИ**
  - Nilsson N.J. Human-Level Artificial Intelligence? Be Serious! // AI Magazine. 2005. V. 26. No 4. P. 68–75.
  - Brachman R. Getting Back to “The Very Idea” // AI Magazine. 2005. V. 26. No 4. P. 48–50.
- **Необходимость объединения результатов, полученных в независимо развивавших подобластях ИИ**
  - Bobrow D.G. AAAI: It’s Time for Large-Scale Systems // AI Magazine. 2005. V. 26. No 4. P. 40–41.
  - 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.
- **➔ Интеграция в когнитивных архитектурах**
  - Langley P. Cognitive Architectures and General Intelligent Systems // AI Magazine. 2006. V. 27. No 2. P. 33–44.
  - Cassimatis N.L. A Cognitive Substrate for Achieving Human-Level Intelligence // AI Magazine. 2006. V. 27. No 2. P. 45–56.

Общий искусственный интеллект

Тем не менее, как в академической, так и прикладной сферах ИИ смены курса на ИИ уровня человека не произошло

➔ Общий искусственный интеллект

- Книги:
  - "Universal Artificial Intelligence. Sequential Decisions Based on Algorithmic Probability" (2005)
  - "Artificial General Intelligence (Cognitive Technologies)" (2007)
  - "Theoretical foundations of artificial general intelligence" (2012)
- Conference on Artificial General Intelligence
  - С 2008 года проводится международная конференция, которая посвящена непосредственно проблеме универсального искусственного интеллекта
- Научные институты, стартапы
  - Singularity Institute for Artificial Intelligence
  - Artificial General Intelligence Research Institute
  - IDSIA (Dalle Molle Institute for Artificial Intelligence)
  - Novamente LLC, TexAI, Thinking Machines Corp., SoarTech, MindSoft Bioware Inc., Proto-mind Machines, Adaptive A.I. Inc., DeepMind, …

Идея общего ИИ

Вместо противопоставления сильного и слабого ИИ, как соответственно обладающего и не обладающего человеческими качествами (сознание, понимание и т.д.), предлагается рассматривать противопоставление общего и специализированного ИИ

Wide scope and poor performance is preferred over narrow scope and good performance

Смена парадигм ИИ

1. Поиск в пространстве решений: 1950-е – 1960-е гг.
   Решение формализованных задач
   Ограничение: формализация задач выполняется вручную
2. Представление знаний: 1970-е – середина 1980-х гг.
   Решение задач из описанной узкой предметной области
   Ограничение: извлечение знаний выполняется вручную
3. Машинное обучение: середина 1980-х гг. – 1990-е гг.
   Построение описания узкой предметной области в рамках заданного представления
   Ограничение: структура области определяется вручную
4. Воплощенный интеллект: 1990-е гг. – середина 2000-х гг.
   Автономное получение данных
   Ограничение: решаются низкоуровневые задачи
5. Общий ИИ (парадигма?): 2005-е гг. – …
   Автономное интеллектуальное поведение
   Ограничение: ???

Направление общего ИИ укладывается в логику развития всей области ИИ, но какова здесь парадигма?

Подходы в области общего ИИ

- Полная эмуляция мозга (альтернатива ОИИ)
  - Прямое копирование ЕИ
- Эмерджентные когнитивные архитектуры
  - Воспроизведение нижних уровней организации ЕИ
- Символьные когнитивные архитектуры
  - Воспроизведение высших когнитивных функций ЕИ
- Гибридные архитектуры
  - Воспроизведение всех известных особенностей ЕИ
    - Антропоцентризм
  - Объединение технических методов ИИ, выполняющих разные задачи
    - Опора на специализированные методы

Когнитивные архитектуры вместо ОИИ

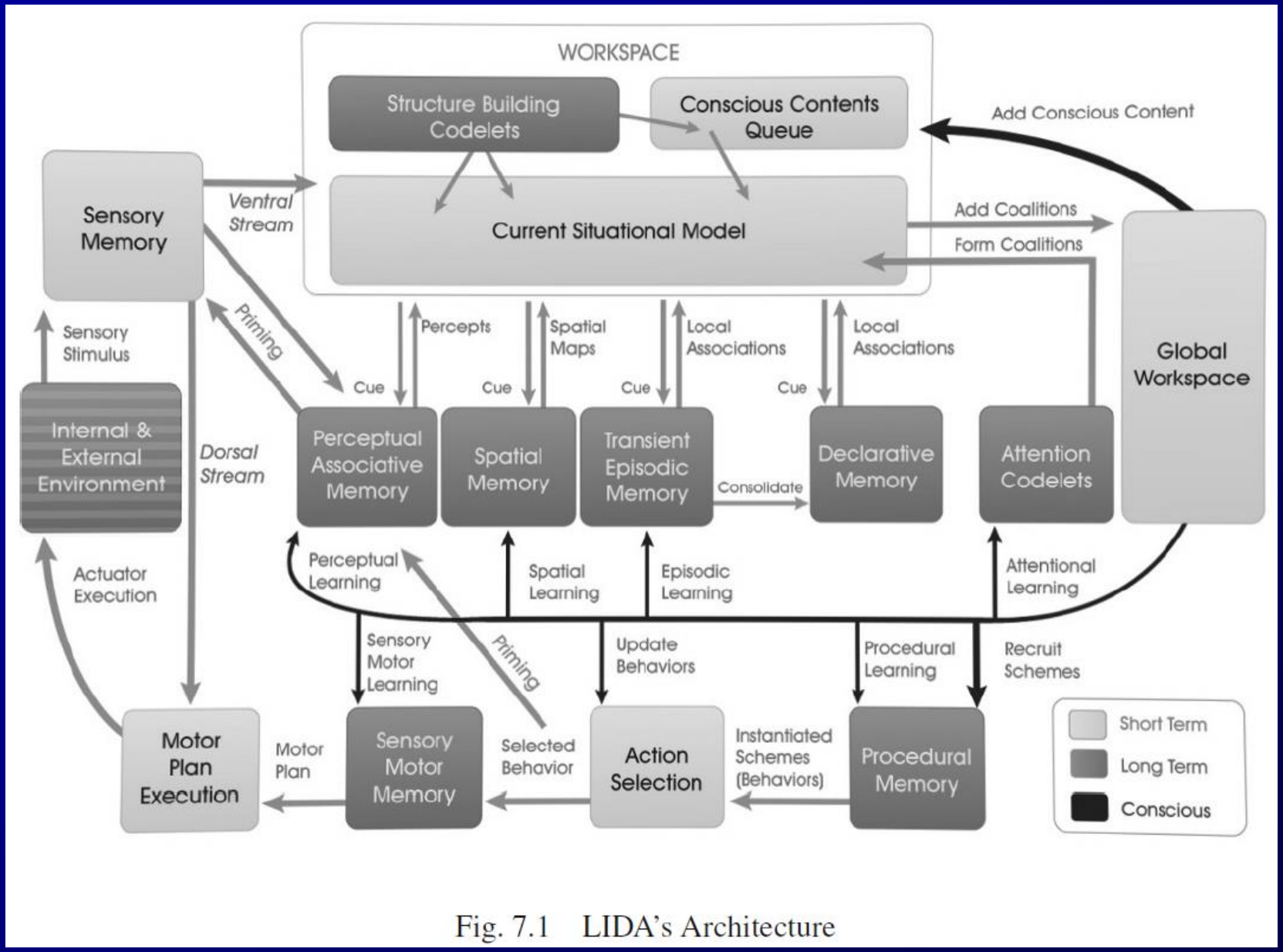
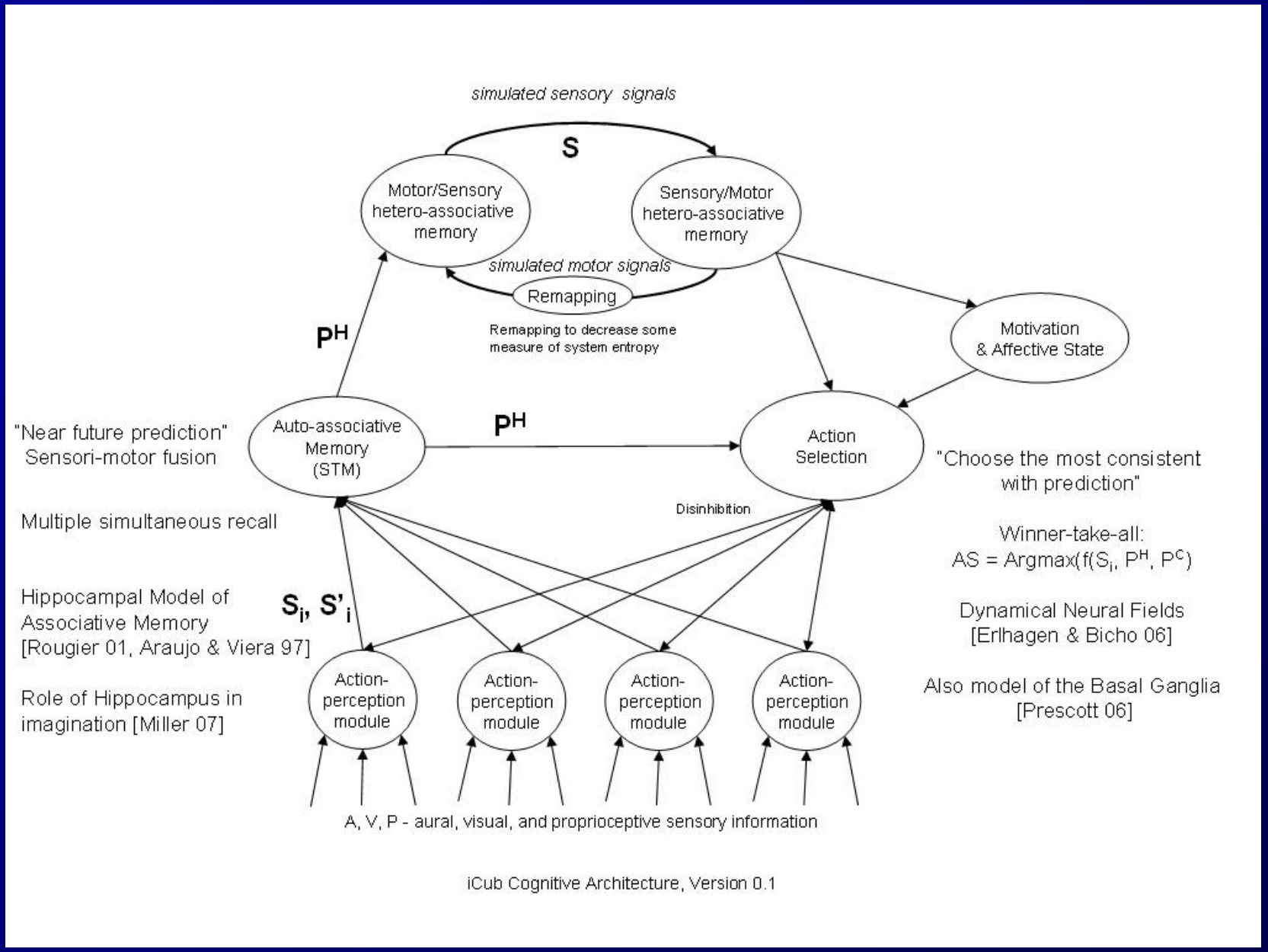


Fig. 7.1 LIDA's Architecture

Когнитивные архитектуры вместо ОИИ



Мотивация

«Different parts of the brain carry out various functions, and no one part is particularly intelligent on its own, but working in concert within the right architecture they result in human-level intelligence… On the other hand, most of the work in the AI field today is far less integrative than what we see in the brain. AI researchers work on individual and isolated algorithms for learning, reasoning, memory, perception, etc. with few exceptions…

As a result, no one knows what level of intelligence could be achieved by taking an appropriate assemblage of cutting-edge AI algorithms and appropriately integrating them together in a unified framework, in which they can each contribute their respective strengths toward achieving the goals of an overall intelligent system.»*

*Hart D..Goertzel B. OpenCog: A Software Framework for Integrative Artificial General Intelligence // Proc. 1st AGI conf. 2008. P. 468-472.

Звучит разумно. Но насколько объединения имеющихся методов достаточно?

Когнитивные архитектуры вместо ОИИ

- Рассуждения
- Креативность
- Ассоциирование
- Обобщение
- Распознавание образов
- Решение задач
- Память
- Планирование
- Достижение целей
- Обучение
- Оптимизация
- Зрение
- Языковые способности
- Моторные навыки
- Индукция
- Дедукция
- ……

……..

- Soar
- ACT-R
- CHREST
- iCub
- LIDA
- CLARION
- NARS
- Variac
- Sigma
- MicroPsy
- ……

……

……

Общий или сильный ИИ?

- Когнитивные архитектуры – это не только общие схемы, но и реализация в программном коде… и иногда они даже могут делать что-то полезное.
- Но все же: самолет плох, когда не имеет перьев или когда плохо летает?
- Шахматная программа обладает низким общим интеллектом, потому что не имеет самосознания или потому что решает только одну задачу?
- Нужны ли все когнитивные функции для интеллекта, или они – эпифеномен? Как их назначение можно сформулировать так, чтобы определить их правильную реализацию?
- Что такое интеллект?

Intelligence measures an agent's ability to achieve goals in a wide range of environments*

*S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. Minds & Machines, 17(4):391–444, 2007.

Formal Definition of Intelligence

- Agent follows policy $\pi : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^* \rightsquigarrow \mathcal{A}$
- Environment reacts with $\mu : (\mathcal{A}\times\mathcal{O}\times\mathcal{R})^*\times\mathcal{A} \rightsquigarrow \mathcal{O}\times\mathcal{R}$
- Performance of agent π in environment μ
  $=$ expected cumulative reward $= V_\mu^\pi := \mathbb{E}_\mu[\sum_{t=1}^{\infty} r_t^{\pi\mu}]$
- True environment μ unknown
  $\Rightarrow$ average over wide range of environments
- Ockham+Epicurus: Weigh each environment with its Kolmogorov complexity $K(\mu) := \min_p\{length(p) : U(p) = \mu\}$
- Universal intelligence of agent π is $\Upsilon(\pi) := \sum_\mu 2^{-K(\mu)} V_\mu^\pi$.
- Compare to our informal definition: Intelligence measures an agent's ability to perform well in a wide range of environments.
- AIXI $= \arg\max_\pi \Upsilon(\pi) =$ most intelligent agent.

Коэффициент общего интеллекта когнитивных архитектур

- Современные методы машинного обучения работают в ограниченных пространствах моделей
  - Обучение с подкреплением: (Частично наблюдаемые) Марковские среды
  - Распознавание образов: аппроксимация закономерностей в распределении образов фиксированными базисными функциями
  - Символьное обучение: преимущественно контекстно-свободные языки
  - …

➔ Когнитивные архитектуры неявно специализированы по классу сред, в которых они могут достигать цели, то есть обладают (сравнительно) низким общим интеллектом

Универсальные рациональные агенты

- Как максимизировать общий интеллект?
- Предположим для начала отсутствие ограничений на вычислительные ресурсы
- Случай известного вычислимой среды μ
  - стратегия агента π может быть подобрана так, чтобы максимизировались суммарные ожидаемые награды V_{μ}^{π},
  - возможна универсальная стратегия π, которая определяет перебором такую цепочку действий, что награды, возвращаемые данной средой μ будут максимальны

Универсальные рациональные агенты

- Случай известного стохастической среды (распределения сред)
  - универсальная стратегия определяет перебором такую цепочку действий, что математическое ожидание наград, возвращаемых всеми возможными средами с учетом их апостериорных вероятностей, рассчитанных после каждого действия и каждого наблюдения, будут максимальны
- Случай неизвестной среды
  - Неизвестное распределение вероятностей сред заменяется универсальным распределением априорных вероятностей

➔ Модель AIXI

Свойства универсальности

- Универсальное распределение априорных вероятностей $\xi=2^{-K(\mu)}$ доминирует (с мультипликативной константой) над любым другим распределением
- Байесовское предсказание с использованием этого распределения сходится в пределе к предсказанию с использованием истинного распределения
- Модель интеллектуального агента AIXI является Парето-оптимальной, то есть не существует такого интеллектуального агента, который бы получал не меньшую суммарную награду во всех средах, и большую – хотя бы в одной; кроме того, AIXI по своей конструкции обладает максимальным коэффициентом общего интеллекта

AIXI и общий интеллект

- Модель AIXI оптимальна
  - Является ли она решением проблемы общего ИИ?
    - Нет, это способ постановки данной проблемы, поскольку AIXI фактически невычислим
- Модель AIXI не содержит компонентов когнитивных архитектур
  - Эта модель не содержит интеллекта?
  - Эти компоненты не нужны интеллекту?
    - Модель является теоретически оптимальным интеллектом; ее кардинальное отличие от реального интеллекта должно быть обусловлено сделанными упрощениями, главное из которых – неограниченность ресурсов

    ➔ Особенности ЕИ должны объясняться либо как способы достижения высокого уровня общего интеллекта в условиях ограниченных ресурсов, либо как следствие его специализации

Индуктивное смещение и эвристики

- Специфика механизмов восприятия
  - Смещение априорных вероятностей
  - Специализированные методы индукции
- Планирование
  - Эвристики поиска
- Представление знаний
  - Компонент индуктивного смещения и эвристик поиска
- Организация памяти
  - Повышение вычислительной эффективности предсказания и принятия решений
- Внимание
  - Отчасти возникает автоматически + Управление ресурсами
- Theory of mind
  - Индуктивное смещение для социальной среды
- Сознание, понимание, …

Заключение

- Многообразие когнитивных архитектур связано с тем, что способов специализированной реализации функций интеллекта существует неограниченно много;
- Специализированность реализации означает ее алгоритмическую неполноту;
- Объединение специализированных методов не может позволить достигнуть алгоритмической полноты, то есть высокого коэффициента общего интеллекта;
- Модели универсального интеллекта практически невычислимы, но они дают понимание причины ограниченности существующих когнитивных архитектур;
- Реальный универсальный искусственный интеллект может быть создан путем обоснованного развития модели алгоритмического интеллекта до уровня когнитивных архитектур с учетом ресурсных ограничений.