Универсальный алгоритмический интеллект: необходимость и возможность достижения

Алексей Потапов

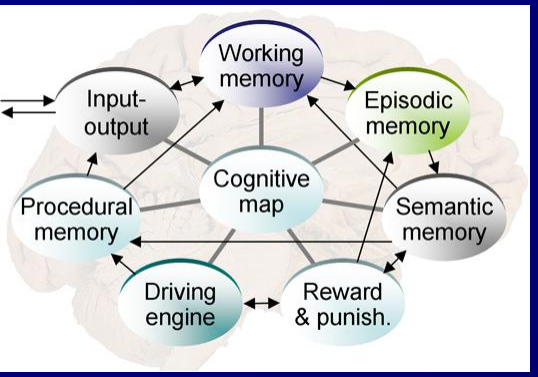
2012

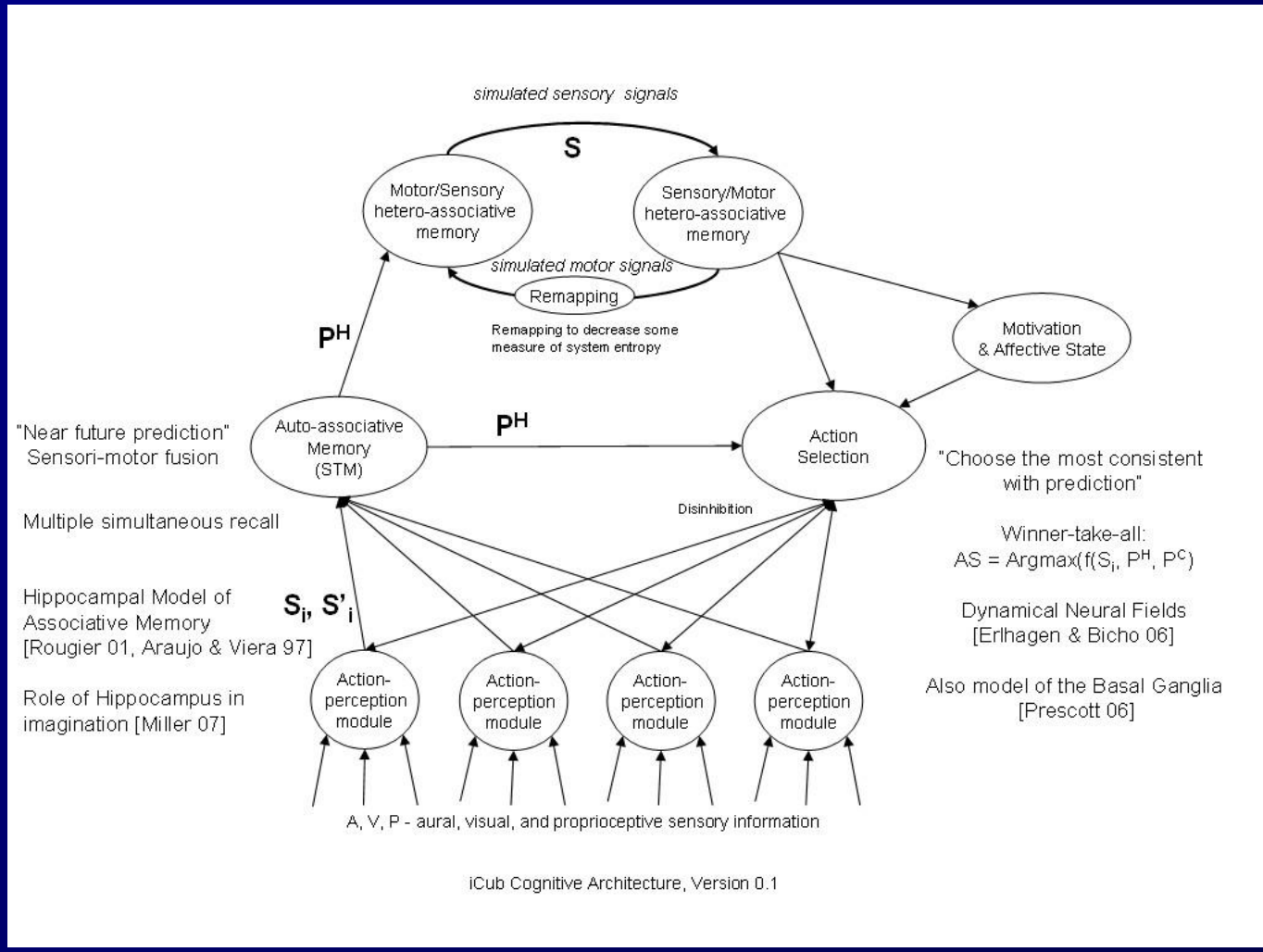
Путь к сильному ИИ?

- **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.
- **когнитивные архитектуры – господствующий подход**
  - 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.

Когнитивные архитектуры

- Soar
- ACT-R
- iCub
- …





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 можно ли на основе слабых компонент создать сильный ИИ?

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. Когнитивные архитектуры: 2000-е гг. – 2015 г.?
   Автономное интеллектуальное поведение
   Ограничение: архитектуры объединяют слабые методы

6. ???: 2015 г. – 2030 г.?
   Сильный ИИ???

Альтернативные? подходы

1. Подход на основе ресурсных ограничений
   - Неаксиоматические логики, NARS (П. Ванг)
2. Бионика
   - Моделирование мозга (напр., Хьюго де Гаррис)
   - Адаптивное поведение (напр., В. Редько)
3. Интегративный подход
   - Novamente engine (Б. Гёрцель и др.), OpenCog
4. Обучение целевым функциям
   - Singularity Institute for Artificial Intelligence, Э. Юдковский
5. Универсальный алгоритмический интеллект
   - Хаттер, Шмидхубер, Aideus

Альтернативные? подходы

1. Подход на основе ресурсных ограничений
   - Неаксиоматические логики, NARS (П. Ванг) символьная когн. арх.
2. Бионика
   - Моделирование мозга (напр., Хьюго де Гаррис)
   - Адаптивное поведение (напр., В. Редько) эмерджентная когн. арх.
3. Интегративный подход
   - Novamente engine (Б. Гёрцель и др.), OpenCog гибридная когн. арх.
4. Обучение целевым функциям
   - Singularity Institute for Artificial Intelligence, Э. Юдковский ~> когн. арх.
5. Универсальный алгоритмический интеллект
   - Хаттер, Шмидхубер, Aideus

Hutter M. Univereal Artificial Intelligence. Sequential Decisions Based on Algorithmic Probability. Springer, 2007. 293 p.

Универсальный интеллект: постановка задачи

- Рациональный агент действует в некотором мире и максимизирует свою целевую функцию (функцию полезности).
- Ему в дискретные моменты времени доступен конечный набор элементарных действий и показаний сенсоров (как внешних, так и внутренних, включающих, в том числе, и целевую функцию).
- $s(t)$ – показания сенсоров
- $q(t)$ – «датчики» целевой функции
- $r(t)$ – действия агента
- задача агента:

$$\arg\max_{r(t)} \sum_{t \geq T} q(t),$$

где $q(t)$ неявно связана с $r(t)$ через среду

Случай известного вычислимого мира

- Путь среда полностью описывается известной машиной Тьюринга (алгоритмом или программой для универсальной машины Тьюринга) ε, входом которой на момент времени *T* являются действия агента *r*(0),…, *r*(*T*–1), а выходом – значения *s*(*T*) и *q*(*T*). Сам агент также управляется некоторой программой *p*, входом которой являются *s*(0), *q*(0),…, *s*(*T*), *q*(*T*), а выходом – значение *r*(*T*).

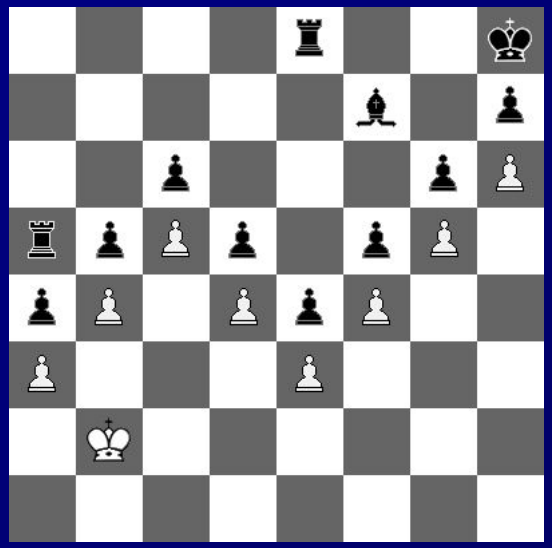
- Текущий результат взаимодействия агента *p* с миром ε может быть вычислен в цикле:
  для *t* от 0 до *T*: $r_p(t)=p(o_\varepsilon(0{:}t))$, $o_\varepsilon(t+1)=\varepsilon(r_p(0{:}t))$.

- Оптимальная стратегия

$$\pi = \arg\max_p \sum_{t\ge 0} q_\varepsilon(r_p(0:t)) \qquad r_\pi(0:T_{\max}) = \arg\max_{r(0:T_{\max})} \sum_{t\ge 0} q_\varepsilon(r(0:t))$$

Адекватность постановки задачи?

1. Диапазон времени суммирования – ?



2. Что такой *q*?

- Телесная функция;
- Внешние сигналы;
- Функция ценностей

Случай известного распределения сред

- Пусть ρ(ε) – (истинные) вероятности сред
- Оптимальные действия определяются, как

$$r_\pi(T:T_{\max}) = \arg\max_{r(T:T_{\max})} \sum_\varepsilon \rho(\varepsilon) \sum_{t \geq T} q_\varepsilon(r(0:t))$$

Строго говоря, суммирование должно проводиться только по тем ε, которые удовлетворяют имеющейся истории, ρ(ε) должно нормироваться соответствующим образов

- Или в явной форме:

$$r_\pi(T) = \arg\max_{r(T)} \sum_{o(T+1)} \max_{r(T+1)} \sum_{o(T+2)} \ldots \max_{r(T_{\max})} \left[\rho(o(T:T_{\max}) \mid r(0:T_{\max}), o(0:T)) \sum_{t \geq T} q(t) \right]$$

Неизвестный мир

- $\rho(\varepsilon) \Rightarrow \rho_{ALP}(\varepsilon)=2^{-l(\varepsilon)}$ универсальное априорное распределение вероятностей программ
- или для явной формы

$$\rho_{ALP}(\alpha) = \sum_{U(\varepsilon)=\alpha} 2^{-l(\varepsilon)}$$

- Универсальность ρ_{ALP}?
- Любая алгоритмическая модель имеет ненулевую вероятность и может быть построена.
- Слабая зависимость от универсальной машины.

Будет ли модель вести себя оптимально?

- Детерминированные задачи
- Парето-оптимальность в произвольных мирах
- Асимптотическая оптимальность в конкретном мире

=> Отсутствие большинства когнитивных функций при видимой оптимальности поведения

Что-то принципиальное не учтено, или данные функции такому интеллекту не нужны?

Ограничения модели

- Универсальность модели?
  - Сверхтьюринговость
  - Целевая функция
- Стохастичность
  - Стохастические модели сред
  - Стохастичность выбора действий
- Отсутствие учета ограниченности ресурсов
  - Вычислительная сложность
  - Время накопления информации

Ввод ограничений на ресурсы

- «Непредвзятый» универсальный алгоритмический интеллект с оптимальной с точностью до мультипликативной константы скоростью работы;
- «Непредвзятый» универсальный алгоритмический интеллект, переводящий мультипликативную замедляющую константу в аддитивную за счет самооптимизации;
- …
- …
- «Предвзятый» субоптимальный квазиуниверсальный интеллект, способный действовать в реальном мире: когнитивные функции как априорная информация и эвристики поиска

Эвристичность структуры ЕИ

- Внимание
- Ассоциирование
- Организация памяти
- Модели и представления
- Адаптивный резонанс
- Логика
- Планирование
- Макиавеллевский интеллект
- Подражание
- Понимание, квалиа, самосознание?

Заключение

- Многообразие когнитивных архитектур связано с тем, что способов слабой реализации функций интеллекта существует неограниченно много;
- Слабость реализации означает ее алгоритмическую неполноту;
- Объединение слабых методов не может позволить достигнуть алгоритмической полноты, то есть сильного ИИ;
- Модели универсального интеллекта практически невычислимы, но они дают понимание причины ограниченности всех существующих когнитивных архитектур;
- Реальный универсальный искусственный интеллект может быть создан путем обоснованного развития модели алгоритмического интеллекта до уровня когнитивных архитектур.

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. Когнитивные архитектуры: 2000-е гг. – 2015 г.?
   Автономное интеллектуальное поведение
   Ограничение: архитектуры объединяют слабые методы

6. Универсальный алгор. интеллект: 2015 г. – 2030 г.
   Сильный ИИ! ☺

Спасибо за внимание!