Карпов В.Э.

Иллюстративные материалы к докладу

Р.Г. Апресян[1], В.Э. Карпов[2] «Теоретические и методологические предпосылки обсуждения проблем этики искусственного интеллекта»

Исследовательский семинар

«Этические проблемы искусственного интеллекта», 1 октября 2019, ИФ РАН

[1]Институт философии РАН

[2]НИЦ «Курчатовский институт», МФТИ

О терминологии

Три взгляда на ИИ:

- ИИ, как область исследований
- ИИ в метафорическом смысле
- ИИ как усилитель творческих способностей человека

1. ИИ, как область информатики

Определяется то, что чем занимаются исследователи (*“ИИ - это то, чем занимаются специалисты в области ИИ”*)

ИИ – область информатики, занимающаяся решением двух классов задач:

1) Слабоформализуемые задачи.

2) Задачи, для которых не известен эффективный алгоритм решения.

Базовые направления ИИ по Г.С. Поспелову:

A - моделирование на ЭВМ отдельных творческих процессов

B - внешняя интеллектуализация ЭВМ

C - внутренняя интеллектуализация ЭВМ

D - **целенаправленное поведение роботов** (создание интеллектуальных роботов, способных автономно совершать операции по достижению целей, поставленных человеком).

ИИ в метафорическом смысле

ИИ – синоним искусственного разума, некое воплощение всего того, что присуще человеку.

Пример рассуждений с позиции антропоморфной метафоричности:

	Свойства интеллектуальной личности. Кто-то считается интеллектуальным, если он:	Механизм ИИ
1	Много знает	Работа с базами знаний
2	Умеет рассуждать	Реализация различного рода познавательных процедур: дедукция, индукция, абдукция, аналогия и пр.
3	Сообразителен	Механизмы эвристического поиска

=> ИИ – это система, **имитирующая** таковые свойства

Не надо путать:

1. ИИ ≠ машинное обучение
2. ИИ ≠ нейронные сети
3. Понятие сильного/слабого ИИ (Джон Сёрл, 1980) спекулятивно и бессодержательно (Ray Kurzweil, The Singularity is Near, 2005: Strong AI as "machine intelligence with the full range of human intelligence").

«Технологии есть, а науки нет», В.К.Финн

Этические вопросы в науке

- Опасности, «вызовы» современного высокотехнологического мира
  - Этика нанотехнологий
  - Этические проблемы генных технологий
  - Этика больших данных …
- Декларации этических принципов научной деятельности:

  оценка социальных и гуманитарных рисков научно-технических разработок и тех благ, которые они могут принести человеку и обществу.
- Этика робототехники (UNESCO, 2017. REPORT OF COMEST ON ROBOTICS ETHICS)
- **Этические проблемы ИИ**

Этические проблемы в ИИ

Опасность «думающих машин»

- Turing, A., 1950. Computing machinery and Intellicence. Машинные ошибки
- Wiener, N., 1960. Some Moral and Technical Consequences of Automation. Опасность для человека и этические проблемы
- Bostrom, N., Yudkowsky, E., 2011. The Ethics of Artificial Intelligence. Негативные сценарии возможного развития ИИ

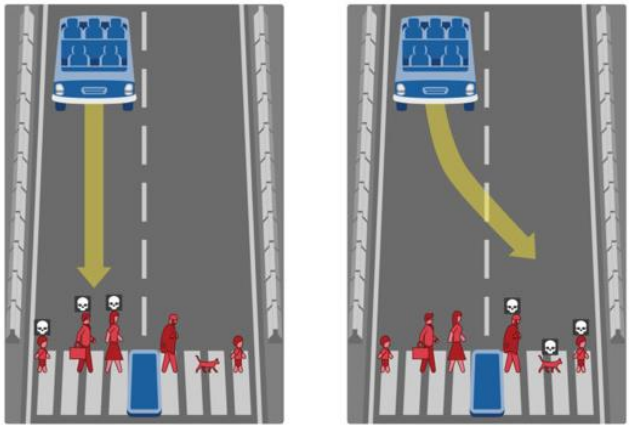
О чем обычно идет речь

Этические проблемы ИИ:

- Опасность для человечества
- Социальные и правовые проблемы
- Этические аспекты применения систем ИИ
- Этика самих специалистов в области ИИ
- …

При этом:

Грядущая неизбежная стандартизация систем ИИ (IEEE) =>???



http://moralmachine.mit.edu/

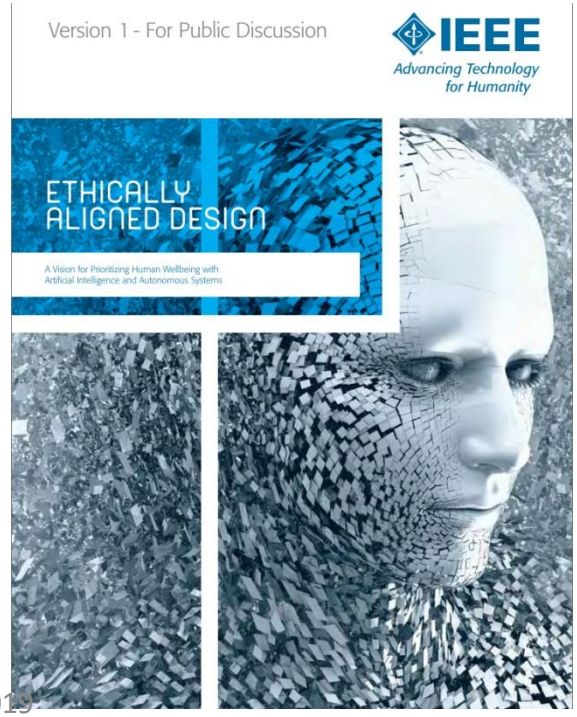
«Проблема вагонетки»: сделать выбор: разбить автомобиль, убив пассажиров, или задавить прохожих на переходе.

Новая постановка задачи

Стандартизация =>

Ответ на вопрос: *насколько соответствует система ИИ этическим нормам*

IEEE: Ethically Alligned Design



Стандарты IEEE, основанные на *Ethically Aligned Design*

- IEEE P7000™ – *Model Process for Addressing Ethical Concerns During System Design*
- IEEE P7001TM – Transparency of Autonomous Systems
- IEEE P7002TM – Data Privacy Process

Понимание проблемы этики и ИИ

Основное отличие проблематики этики в ИИ от этики прочих областей:

1. ИИ-система – это система, автономно принимающая критические решения.

2. Основная проблема – определение того, насколько принимаемые И/АС решения соответствуют этическим нормам, т.е. насколько она «этична».

Перечень проблем

Предмет исследования этически обусловленного проектирования – И/АС, совершающие выбор того или иного **значимого действия** или решения.

При этом совершение выбора осуществляется на основе некоторых **эвристик**, основанных на этических императивах.

1. Формализация. Вопросы конструктивных определений и онтологий.
2. Математический аппарат. Модели и методы.
3. Этическая верификации. Определение степени соответствия И/АС этическим нормам.

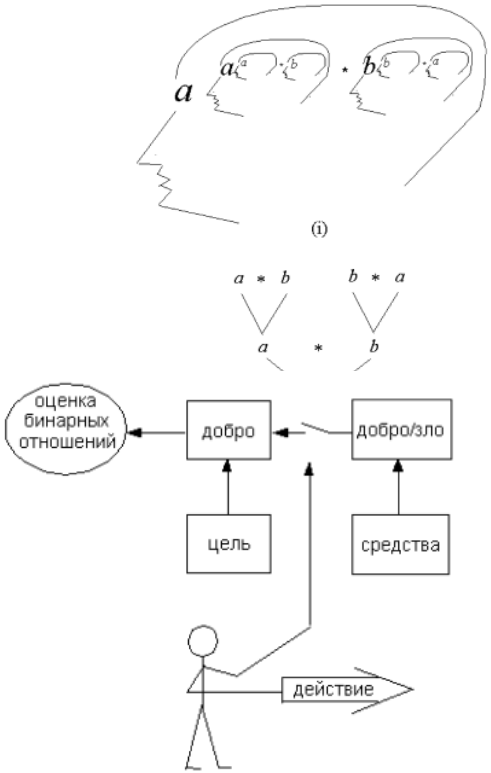
Формализация этических понятий

Формальная этика: нахождение формальных критериев «правильности» поведения (алгебра поступков).

- Gensler, H.J. 1996. Formal Ethics. London, New York.
- Лефевр В. **Алгебра совести** / Пер.с англ. М., 2003

- Фоминых И.Б. Эмоции как аппарат оценок поведения интеллектуальных систем. // КИИ-2006
- Фоминых И.Б. Классификация эмоций: информационный подход //ИМиМВ в ИИ, 2007

- Неклассические логики (В.К. Финн, О.П. Кузнецов, В.Б. Тарасов, В.Н. Вагин, А.С. Карпенко, …)
- Вероятностные модели и нечеткая логика
- Методы вербального анализа решений (О.И. Ларичев)
- …



Базовый набор инструментальных средств, позволяющих решить задачу формализации этических норм, на настоящий момент имеется.

Верификация

И/АС осуществляют принятие решений, основанные на этических императивах и нормах = эвристики.

Верификация этичности И/АС

- Проблема объяснительной компоненты
- Противоречиивость и многозначность этических норм
- Неэтичный вывод из последовательности этичных рассуждений
- Относительность весов правил
- Проблема контекста
- **Черные ящики**

Вывод:

- **Тесты**
- **Возврат к тезису Тьюринга**

Выводы по первой части части

1. Предмет исследования этически обусловленного проектирования – И/АС, совершающие **выбор** того или иного значимого действия или решения. Совершение выбора осуществляется на базе некоторых **эвристик**, основанных на этических императивах.
2. Принципиально важным являются вопросы конструктивных определений и **онтологий**. Задача онтологий заключается во взаимном увязывании и согласовании этических и технических понятийных систем.
3. В настоящее время **существует математический аппарат**, способный в той или иной мере реализовать формализм, необходимый для этически обусловленного проектирования.
4. С технической точки зрения основной проблемой является проблема **этической верификации**. Эта верификация заключается в комплексе **тестов**, способных определить "степень этичности" интеллектуальной системы. Иного способа определения этой степени, кроме наблюдений за реакциями и поведением исследуемой И/АС, не существует.

«Технические» потребности. Мораль и управление социумом

1. Социум – это устойчивое, замкнутое образование, деятельность которого поддерживается целым комплексом законов и правил внутригруппового взаимодействия =>
2. Способы целенаправленного изменения его поведения:
   - прямое воздействие на индивида (изменение его параметров и структуры);
   - оказание воздействия на окружающую среду (косвенное управление);
   - **управление «нравственными» установками (мораль – наиболее гибкая и вариативная надстройка СУ)**

Все механизмы, требуемые для функционирования социума (когезия, контагиозность, подражание, социальное доминирование, агонистическое поведение, обучение и пр.), должны быть сохранены.

Социальное обучение

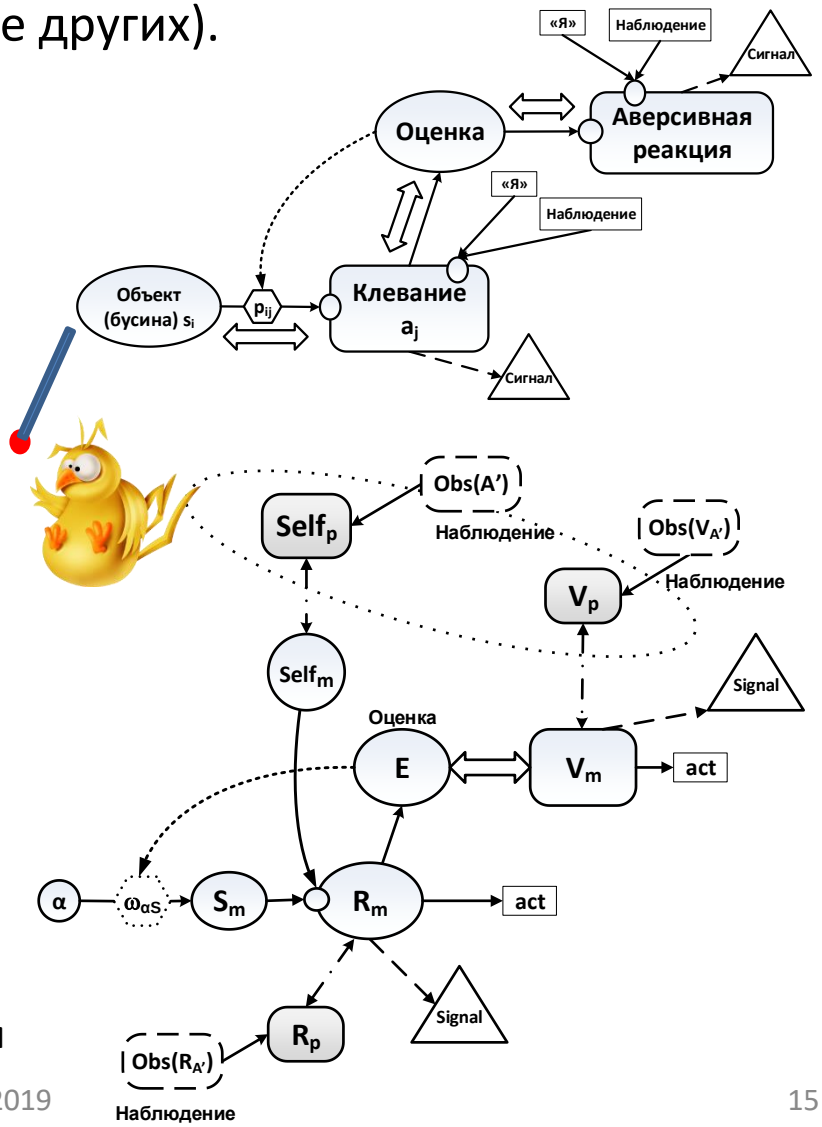
Способность животных приобретать опыт, связанный с взаимодействием с другими особями (обучение на опыте других).

- $Obs(V_{A'}) \rightarrow V_p \rightarrow V_m \rightarrow E$
- Обучение Se = <O, P, M, Ω, E >
- Наблюдения
  $O^T = [O_i] = [Obs(A'), Obs(V_{A'}), Obs(R_{A'})]$
- *Перцепты* $P^T = [P_i] = [Self_p, V_p, R_p]$
- *Значения* $M^T = [M_i] = [Self_m, V_m, R_m]$
- $E=f(V) \vee h(E_{int})$
- $E= \omega_{VE}V \oplus E_{int}$.

1. Определение активности элементов вектора P: $P = \Omega_o^p \cdot O,\ \Omega_o^p \subset \Omega$
2. Определение активности элементов вектора M: $M = \Omega_P^M \cdot P,\ \Omega_P^M \subset \Omega$
3. Вычисление оценки: $E = \omega_{VE}V_m$
4. Изменение весов связей между активными элементами, если E≠0:

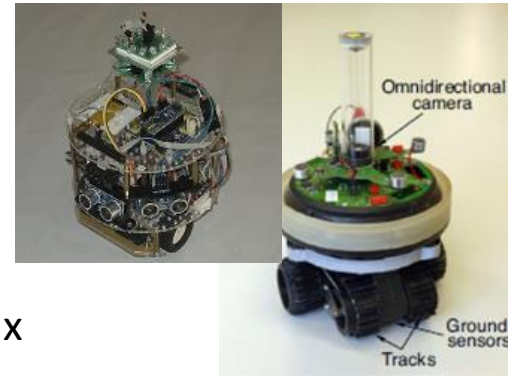
$\omega_{ab}(t) = \omega_{ab}(t\text{-}1)\delta,\ a,\ b \in Q$

Особенности рефлекса: Чужой опыт забывается быстрее



Подражательное поведение

(1) чему следует подражать, (2) почему это следует делать и (3) кому надо подражать.



1. Чему подражать. Технический "трюк". Контрагент сам сигнализирует о том, в каком состоянии он находится в текущий момент времени (световые, ИК-маяки)

- Каждое действие имеет свое внешнее проявление.
- Аналог сигнальной коммуникации у животных (языка животных нет!)

2. Вопрос обусловленности подражания. Чувство «самости».

Действие не будет активировано, если не будет подтверждающего сигнала от С.Я., интерпретируемого как "принадлежность" этого действия агенту. В определенном смысле это – чувство (ощущение) самости, т.е. отождествление или восприятие объекта, как **своего**.

3. Кому подражать. При наблюдении контрагента происходит не просто определение степени его похожести или близости к агенту.

Происходит <u>отождествление наблюдаемого объекта (контрагента) с С.Я.:</u> при наблюдении "**близкого**" контрагента происходит то же подтверждение самости, что и при активизации вершины С.Я.

Подражательное поведение

1. Продукции $S_m \rightarrow R_m$

$$R = \omega_{S,R}\, S \wedge \omega_{self,R}\, Self$$

2. $Self = I \vee Obs(A')$

3. $R \rightarrow Signal$

4. $S = \omega_{\alpha 1,S}\, \alpha_1 \oplus \omega_{\alpha 2,S}\, \alpha_2 \oplus \ldots$

5. Множество активных элементов $Q = \{q_k \mid q_k \neq 0\}$

6. Связи между q_i:

$$\omega_{i,j}(t) = \omega_{i,j}(t-1) \oplus \delta,\ s_i,\ s_j \in Q$$

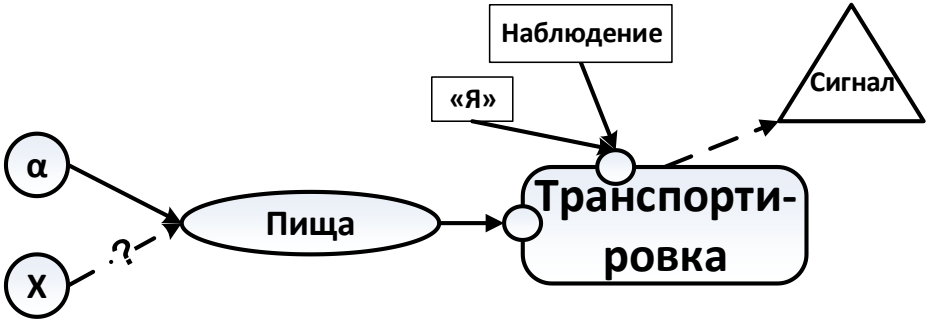
7. Связь перцепт-значение:

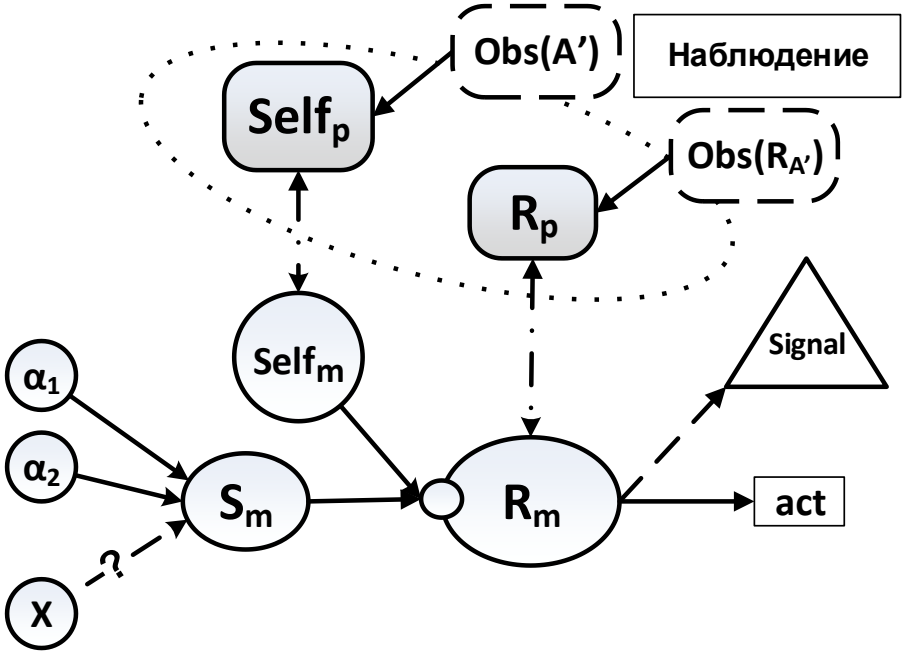
$$Obs(R_{A'}) \rightarrow R_p \rightarrow R_m$$

$$Obs(A') \rightarrow Self_p \rightarrow Self_m$$

8. Итоговое формирование ассоциативной связи:

$$\boldsymbol{\omega_{X,S}(t) = \omega_{X,S\,j}(t-1) \oplus \delta}$$





Моральное поведение

Р. Апресян: *"…**если** мы ассоциируем мораль с человеческими отношениями"*

1. Необходимость морали. Мораль – это механизм **адаптации**, то, что позволяет функционировать социуму более эффективно.

2. Целевая функция регулирования поведения. Основополагающий регулятив в межличностных отношениях - "золотое правило" морали (практика взаимности: "ты – мне, я – тебе" или “непричинение вреда другим").

3. Механизмы, лежащие в основе морального поведения: подражательное поведение, социальное обучение и эмпатия.

«Удобство» морали:

1. Важнейшая особенность морали - ее гибкость, вариативность.
2. Эта надстройка над базовыми моделями поведения "легковесна" и может варьироваться в широких пределах на протяжении жизненного цикла индивида.

Эмпатия = {Эмоции + Отождествление контрагента + Подражательное поведение}

- Эмпатия (в социуме) – это основа для более высокого уровня управления, связанного с *целеполаганием и планированием:*
  либо реализовывать такой **план действий**, при котором контрагенту будет хорошо (увеличение уровня эмоционального состояния, "поступай так, чтобы…"), либо сформировать план, не ведущий к появлению отрицательных эмоций контрагента ("не навреди").

Неожиданный, но неизбежный вывод

Посылка:

1. В перечень механизмов, определяющих нравственное поведение, наряду с прочими, входят социальное обучение и **эмпатия**
2. Основной мотивацией морального поведения (то, на что направлено золотое правило морали) является **максимизация эмоционального уровня контрагента**, на которого направлено воздействие индивида.
3. Знак и величина эмоции непосредственно определяется существующими **потребностями** агента.

Вывод: действие основного морального регулятива направлено на **удовлетворение потребностей индивида**.

- Аналог выводов по экспериментальным моделям проф. Выбегалло (модель Человека, неудовлетворённого полностью; неудовлетворённого желудочно; полностью удовлетворённого): *"Теперь можно было считать доказанным, что ежели человека не кормить, не поить, не лечить, то он, эта, будет, значить, несчастлив и даже, может, помрёт"* (А. и Б. Стругацкие).
- Такая неизбежная вульгаризация – это результат попытки переноса сугубо технических механизмов (вплоть до адаптационных моделей) на область, принципиально плохо формализуемую – моральную философию.
- Такое же положение дел, как и у т.н. эволюционной этики: *"Эволюционные и этологические теории морали фактически дезавуируют (как несущественное или неинтересное с естественно-научной точки зрения) именно это достижение философии – специфическое понимание феноменов морали " (Р.Апресян)*.