Семинар «Этические проблемы ИИ»

Карпов В.Э.

Применимо ли понятие морали к отношениям между искусственным агентами?

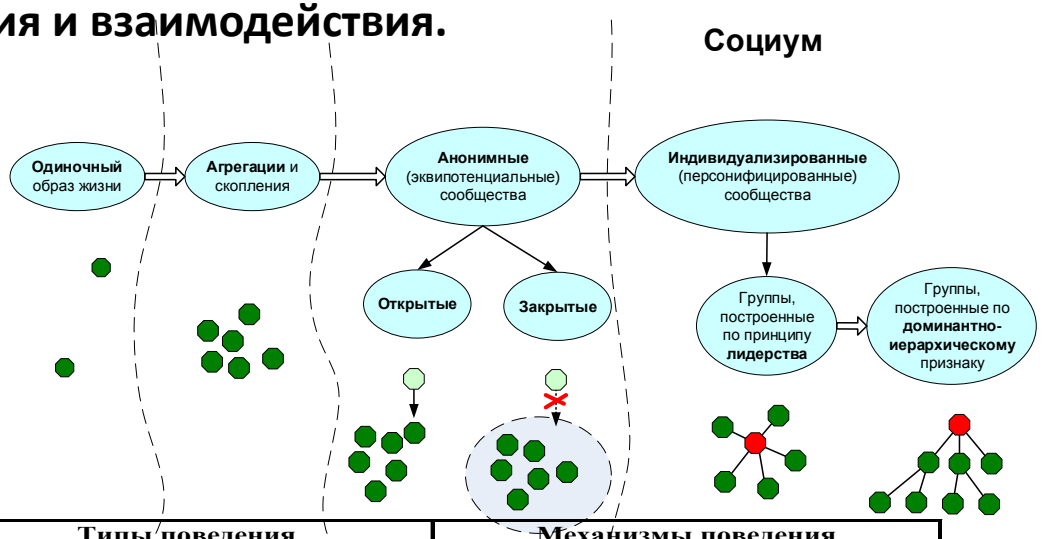
Материалы к докладу

Karpov-ve@yandex.ru

Модели социального поведения в групповой робототехнике

1. **Получение синергетических эффектов в системах ГР возможно при организации групп роботов, как социальных сообществ.**
2. **Основная задача МСП – определение базовых механизмов, определяющих специфику социального поведения и взаимодействия.**

Создание социума роботов (искусственных агентов)



Типы поведения		Механизмы поведения
Контагиозное поведение		1. Коммуникативные сигналы **2. Симпатическая индукция** 3. Подражательное поведение
Агонистическое поведение		1. Агрессия ↓ 2. Драки (конфликты) ↓ 3. Ритуалы и демонстрации ↓ 4. Социальное доминирование
Репродуктивное поведение	**Брачное**	1. Драки 2. Ритуалы и демонстрации 3. Разделение труда 4. Усвоение опыта предыдущих поколений
	Родительское	

Образование социальных форм организации – один из путей адаптации

Особенности социального робота. Потребности, эмоции, чувства

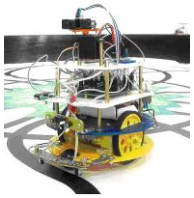
Потребностно–информационная теория эмоций П.В. Симонова (1964)

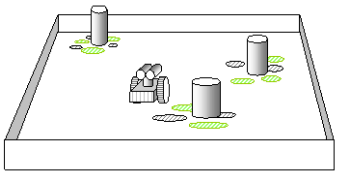
Э = *f*(П, *p*(Ин, Ис))

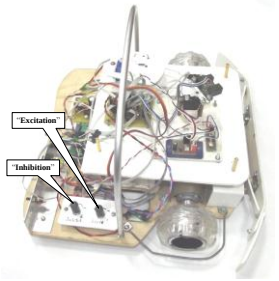
- Э. определяют критерии обучения.
- Э. контрастируют сенсорное восприятие и стабилизируют поведение.
- Э. в условиях неполноты информации.

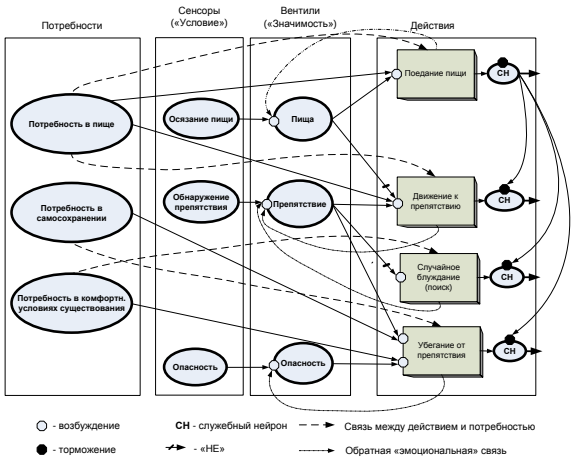
Темперамент робота

- Робот- Меланхолик, Холерик, Сангвиник и Флегматик
- 2 параметра: возбуждение и торможение.



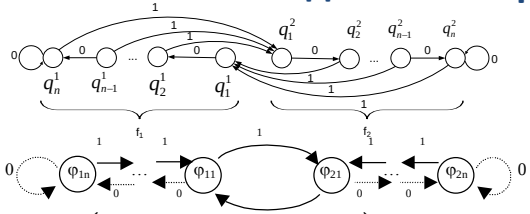






	Возбуждение	Торможение
Меланхолик	Низкий	Низкий
Холерик	Высокий	Низкий
Сангвиник	Высокий	Высокий
Флегматик	Низкий	Высокий

Автоматная модель темперамента



$$K\left(C^{(1)},\ldots,C^{(V)},\Delta\right)$$

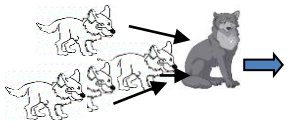
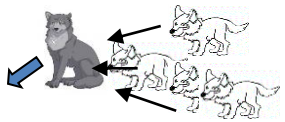
$$ЕСЛИ\ \left(s_{food}^{need}\ \&\ s_{food}^{sens}\right),\ ТО\ \ C_{eat}^{proc}\left[f(PROC_{EAT})\right]$$

$$C_{eat}^{proc} \rightarrow ? \rightarrow exec(PROC_{EAT})$$

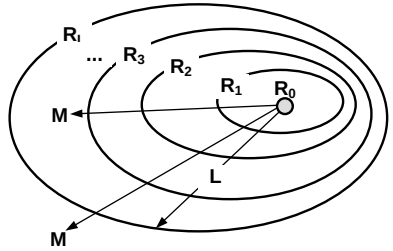
Теория функциональных систем П.К.Анохина. Ситуативные и ведущие эмоции, связанные с удовлетворением или неудовлетворением **потребностей**

Признаки социальности. Доминирование.

Лидерство. Задача голосования. Роли



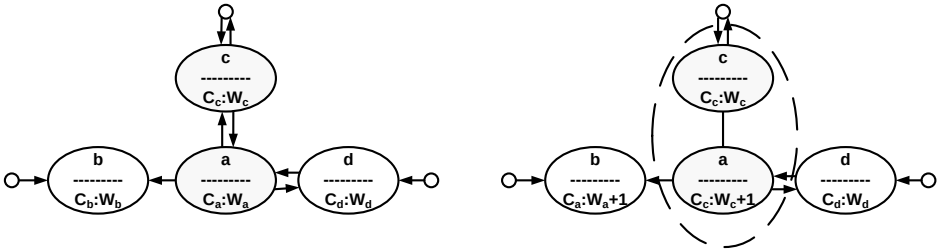


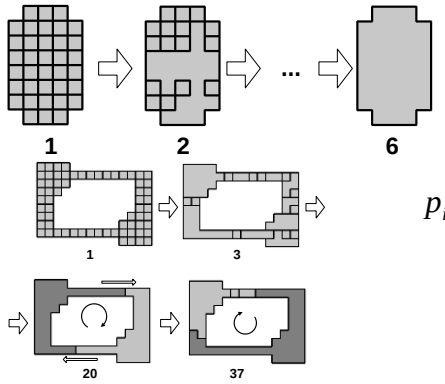
Дано: Множество агентов, способных лишь к непосредственному локальному взаимодействию между соседями

Задача: выбрать **единственного** лидера путем голосования.

Процедура голосования

- каждый агент определяет, за кого голосуют его соседи.
- в зависимости от веса кандидата, за которого голосует сосед, агент может **поменять** свое мнение и проголосовать за того же кандидата, что и его сосед.





$$p_{ij} = \frac{W_i}{W_i + W_j}$$

Распределение задач (функциональная дифферентация)

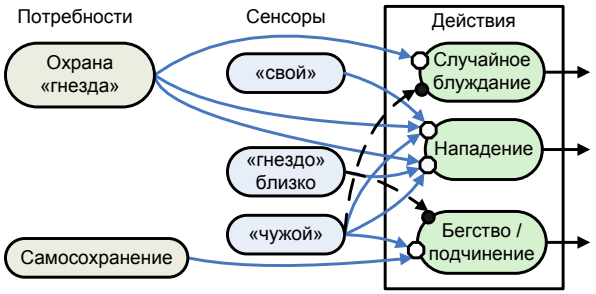
- Для более сложных задач, решаемых группой роботов, требуется дифференциация функций - распределение задач между роботами.

Признаки социальности. Агрессия

АП не является отдельным типом СП.

Модели макроуровня и модели **индивидуального поведения**.

Схема охраны территории (кормовых участков)



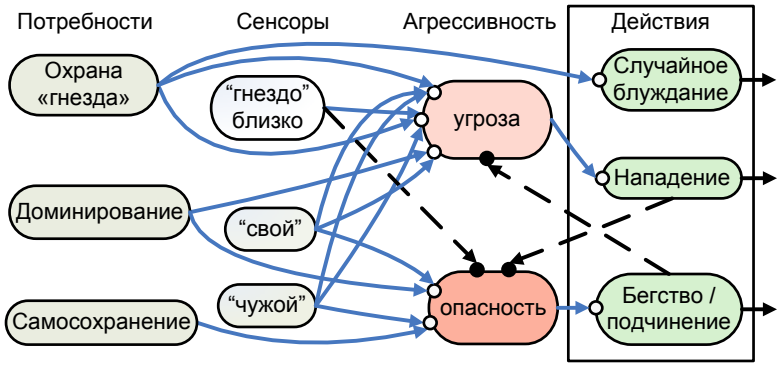
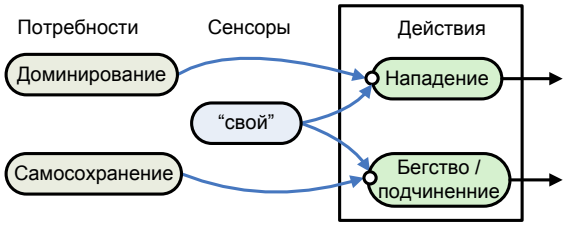
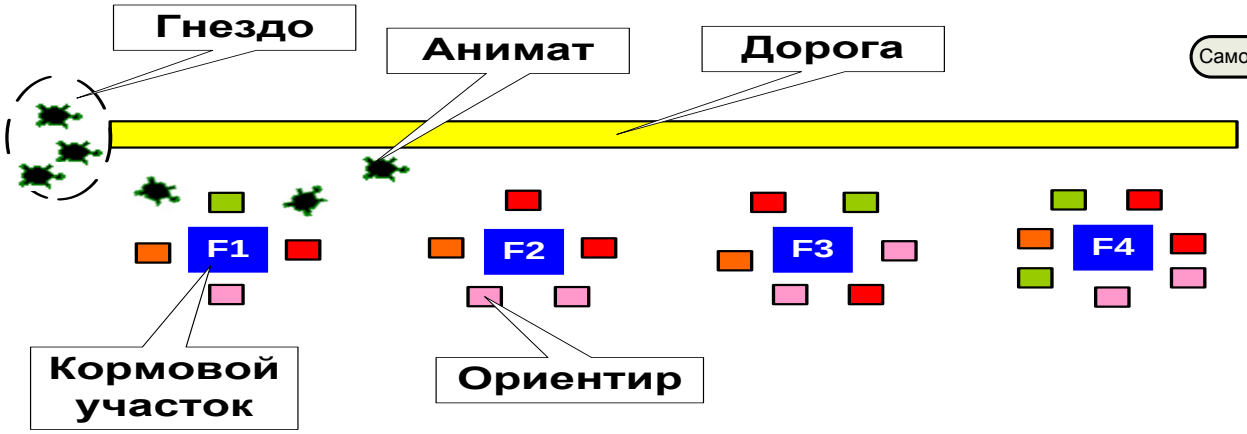


Схема доминирования



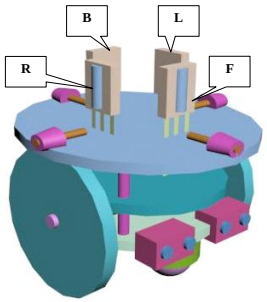


Агрессивность – это внешняя оценка поведения.

(1) Память (2) Агрессия

$$A(t)=k_1 Age(t) + k_2 \omega_t$$

Признаки социальности. Коммуникация

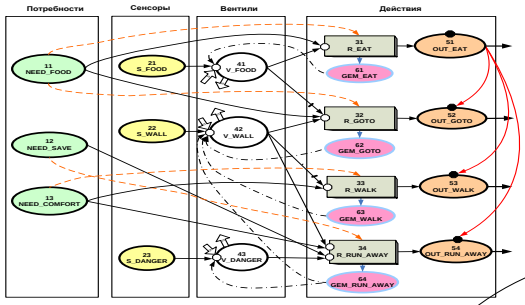


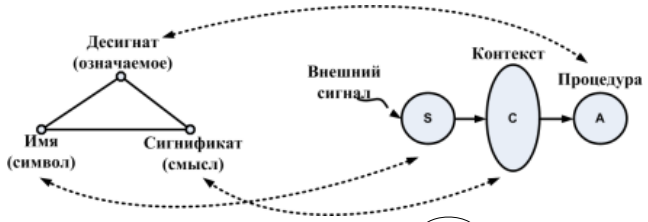
«Язык роботов». Коммуникации. «Язык» животных на самом деле – сигнальная коммуникация. Избыточность, континуальность.

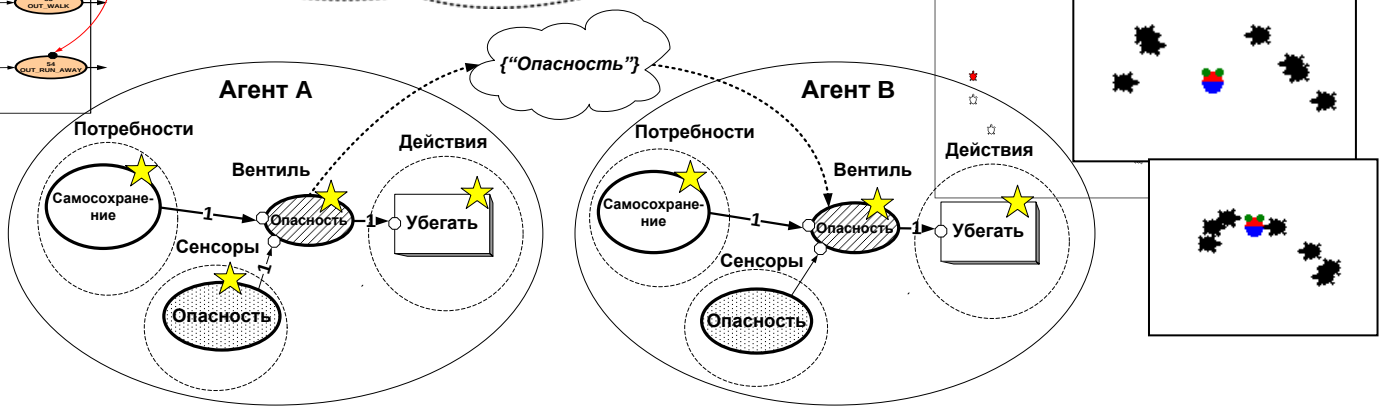
Основная задача – **внешнее отображение внутреннего состояния.**

Контагиозное поведение

Знак-ориентированная СУ



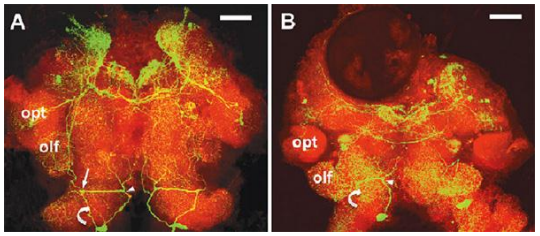


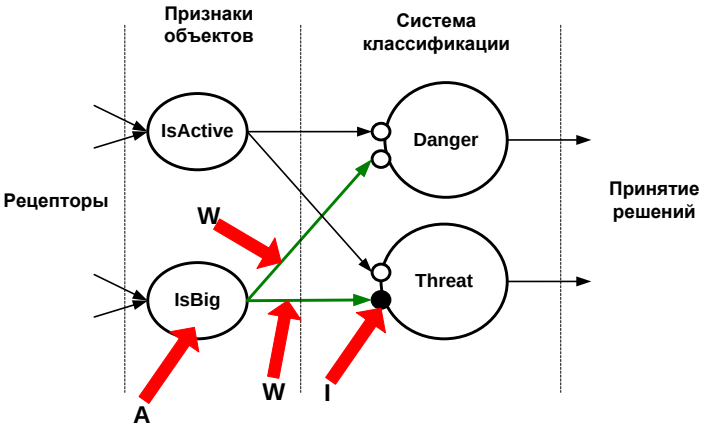


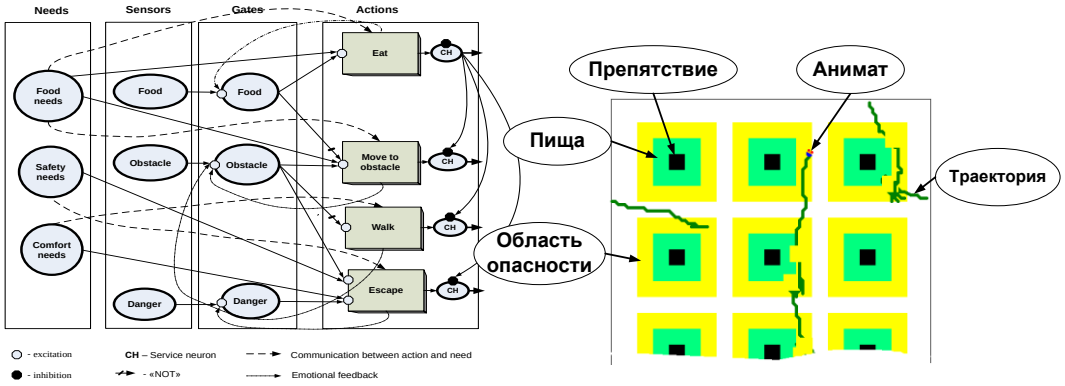
Эмоции как инициатор общения

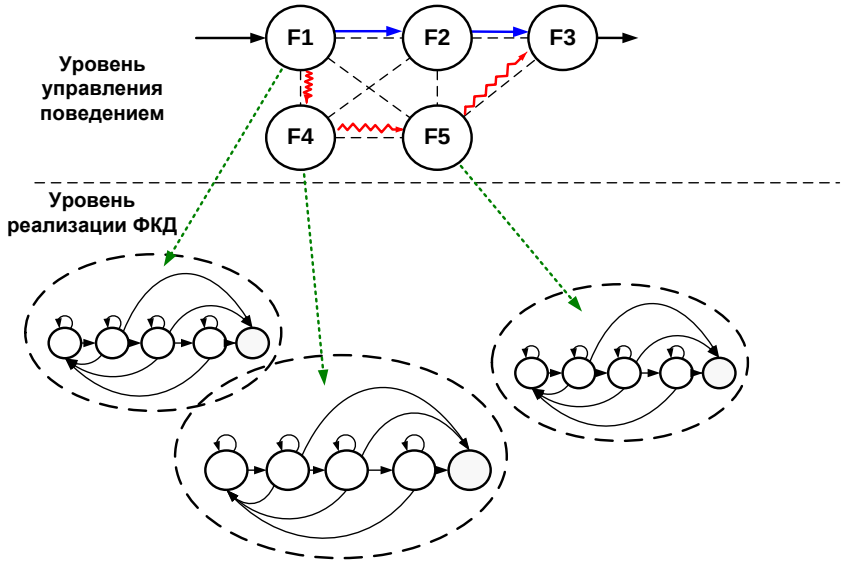
Особенности управления роботом. Паразитическое манипулирование

1. Изменение характера поведения (психологический уровень): Потребности и возбуждение/торможение
2. Прямое выстраивание реакций. Регуляция ФКД.
3. Переориентация реакций (опасность и угроза).









Задачи: заставить анимата
1. Находиться в опасной области
2. Нападать на несвойственные объекты

Субъективное Я

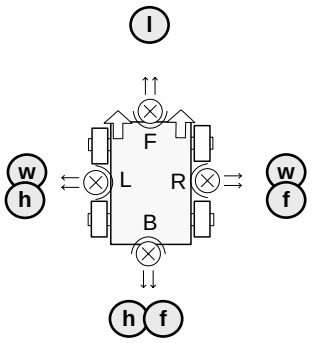
- В социальной психологии С.Я. – это то, что человек сам о себе думает и знает.
- С.Я. = Я-познающее.
- В психологии С.Я. состоит из знаний трех типов: (1) осознание себя одним и тем же человеком в течение онтогенеза; (2) знания об индивидуальности; (3) воля, чувство личного контроля.
- В прикладной семиотике существует аналогичное понятие – субъект деятельности. Включение этого понятия в модель мира формирует т.н. картину мира [Осипов и др., 2018].
- Одной из основных функций субъекта деятельности является **целеполагание**.

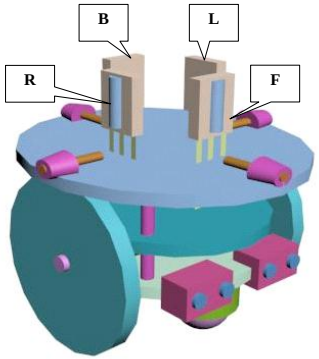
С.Я. как конструктивный элемент

«Я» - элемент ММ для ***прогнозирования*** ситуации. Это прогнозирование сводится к ***моделированию*** развития ситуации.

Описание ММ. ММ - это способ отображения в памяти ИС знаний о внешней среде. Описание отношений между объектами мира, в котором живет робот:

1. Всякое действие робота должно быть отображено в ММ.
2. Действие изменяет отношения между объектами.
3. Одним из объектов является сам робот (**С.Я.)**





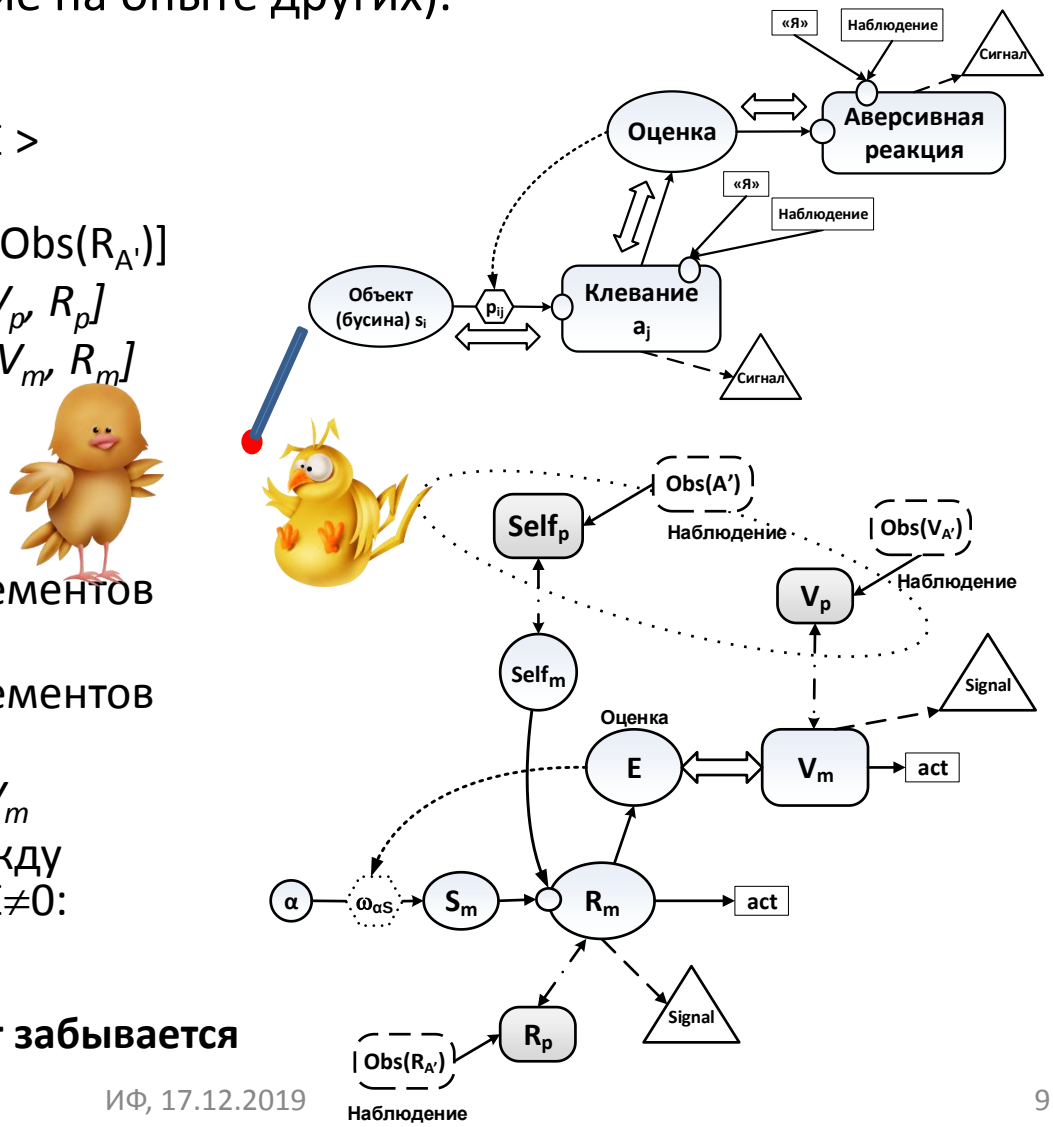
$$R_{r+f} = M_f^{\ T}(M_r^{\ T} \cdot R)$$

$$R = \begin{array}{c} \\ F \\ B \\ L \\ R \end{array} \begin{array}{c} \begin{array}{cccc} w & l & h & f \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

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. 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:

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

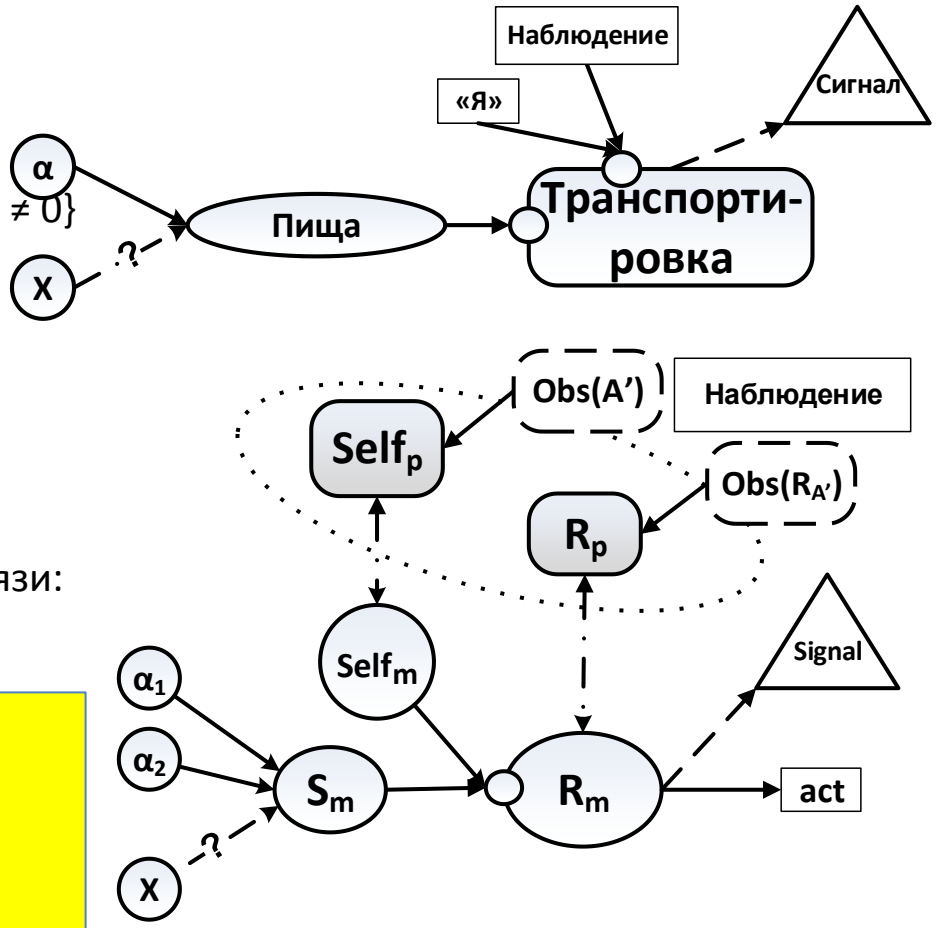
7. Связь перцепт-значение:

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

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

8. Итоговое формирование ассоциативной связи:

$$\omega_{x,s}(t) = \omega_{x,s\,j}(t\text{-}1) \oplus \delta$$



Не просто определение степени похожести конспецифика, но <u>отождествление наблюдаемого объекта с С.Я.:</u> при наблюдении "**близкого**" контрагента происходит то же подтверждение самости, что и при активизации вершины С.Я.

Вопрос 1. От управления индивидом к управлению социумом

1. Социум – это устойчивое, замкнутое образование, деятельность которого поддерживается целым комплексом 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.

Вопрос 2. Сугубо техническая (прагматическая) мотивация

Возможность разрешения конфликтных ситуаций (конфликт между личными потребностями и оценкой последствий) типа:

1. *Надо ли прийти на помощь?*
2. *Кому помочь (слабому? сильному?) или у кого отнять?*
3. *Можно ли помешать другому (своему? чужому?)?*
4. *Кто есть свой/чужой (клан, колонна, гнездо, федерация)?*
5. *Потребность в самосохранении Vs необходимость самопожертвования?*

…

Необходимы **общие** правила, установки, схемы, позволяющие разрешать подобного рода конфликты.

1. Что должно учитываться?
2. Каковы алгоритмы и критерии?
3. Не есть ли это то, чего мы ожидаем от **моральной философии**?

Основной вопрос

Может ли рассматриваться поведение АИС с точки зрения морали (или применимо ли понятие морали к роботам)?

или:

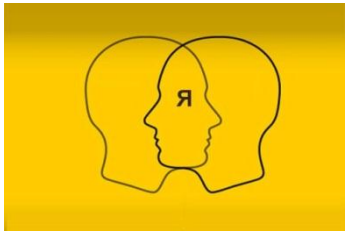
Можно ли считать робота моральным агентом?

Moral agency is an individual's ability to make moral judgments based on some notion of right and wrong and to be held accountable for these actions. A moral agent is a being who is capable of acting with reference to right and wrong.

https://en.wikipedia.org/wiki/Moral_agency

Чего не хватает для того, чтобы признать робота моральным агентом?

1. **Потребности и эмоции**
2. **С.Я., самосознание и рефлексия**
3. **Социальные механизмы поведения**
4. **…**



5. Эмпатия и целеполагание

- Эмпатия – отзывчивость на эмоциональное состояние другого. Э. - основа для уровня управления, связанного с целеполаганием и планированием.

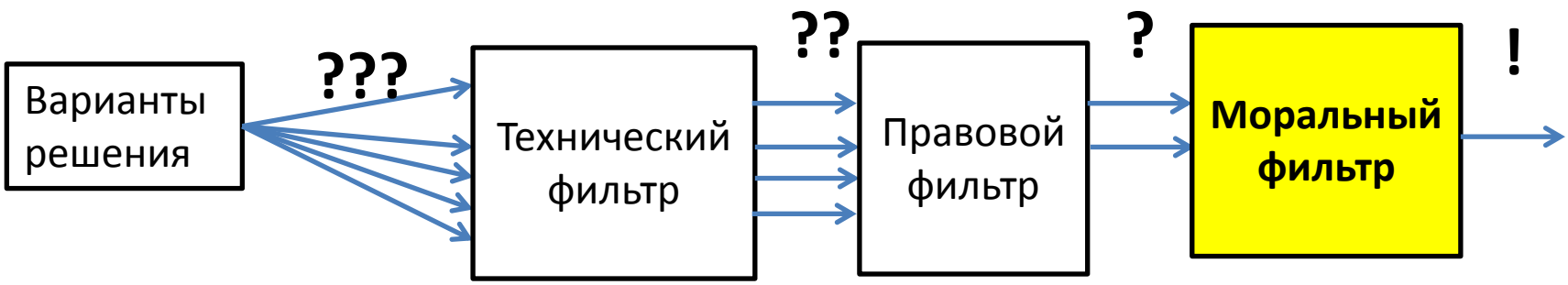
Эмпатия = {Эмоции + Отождествление контрагента + Подражательное поведение}

В этологии – СИМПАТИЧЕСКАЯ ИНДУКЦИЯ

Эмоциональное состояние контрагента оказывает влияние на формирование **мотива** поведения, его **цели**.

Механизмы оценки решения

Оценка (Д) = Техническая_оценка(Д) +
Правовая_оценка(Д) + ***Моральная_оценка****(Д)*



Моральное поведение с технической точки зрения

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

2. Целевая функция регулирования поведения – сохранение устойчивости социума.

Способ - разрешение конфликтов. Основополагающий регулятив в межличностных отношениях - "золотое правило" морали ("непричинение вреда другим").

Примечание: Практика взаимности "ты – мне, я – тебе" менее очевидна.

3. Механизмы, лежащие в основе (но не ограничивающие его?) морального поведения: подражательное поведение, социальное обучение и эмпатия.

4. «Удобство» морали:

Гибкость. Эта надстройка над базовыми моделями поведения "легковесна" и может варьироваться в широких пределах на протяжении жизненного цикла индивида.

-

Механистический вывод, который вряд ли кого устроит:

Посылки:

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**.

Чтобы уйти от такой вульгаризации, требуется иное.

Сугубо техническая задача:

Дано:

Агент (робот), обладающий всеми свойствами индивида. Потенциально – моральный агент.

Требуется:

Вложить в него понимание того, что такое хорошо-плохо.

Воспитание? Примеры? Априорное знание в готовом виде?