

Семинар “Этические проблемы искусственного интеллекта”

Карпов Валерий Эдуардович

О моральных аспектах адаптивного поведения искусственных агентов

НИЦ «Курчатовский институт»

karpov.ve@gmail.com

26 марта 2021

Этические проблемы ИИ

- IEEE. “Этически обусловленное проектирование” (“Ethically Aligned Design”, 2016).
- UNESCO. “Отчет COMEST по этике роботов” (Report of COMEST on Robotics Ethics, 2017)
- Этика и «цифра»: Этические проблемы цифровых технологий , 2020, Гайдаровский форум
- ...

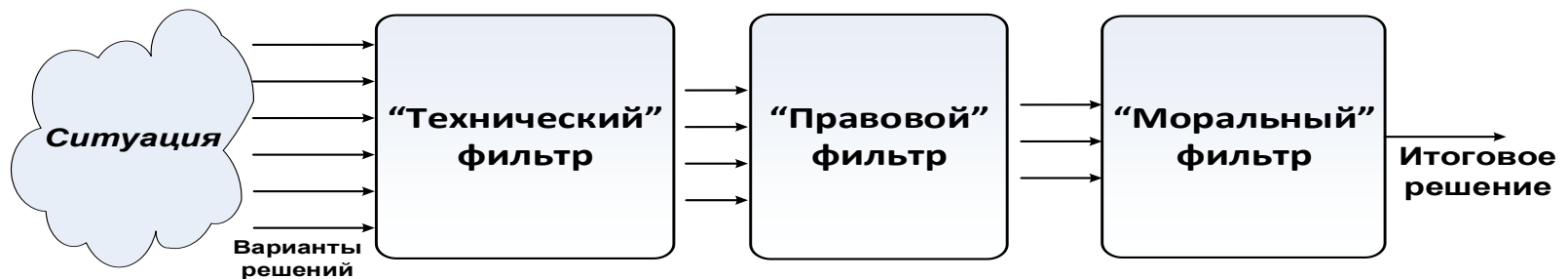
Угрозы, социальные и экономические последствия применения, этичность самих разработчиков и т.п.

Нас интересует иной аспект этики И/АС:

(1) И/АС – это система, автономно принимающая решения, критически важные для человека.

(2) Оценка того или иного возможного решения или действия D может определяться техническими, правовыми и моральными соображениями.

$Оценка(D) = Техническая_оценка(D) + Правовая_оценка(D) + Моральная_оценка(D)$



Моральный агент

Пусть нам удалось формализовать положения моральной философии так, что они могут быть представлены некоторой **системой правил** =>

- **Проблема:** если решение основано на моральных принципах (оно субъективно, слабо верифицируемо, нечетко и т.д.), то единственный способ повысить доверие к нему – это признание того, что решение было принято т.н. **моральным агентом**, т.е. сущностью, которой мы **делегируем право** применять соображения этического характера.
- **Вопрос:** существуют ли предпосылки для того, чтобы И/АС могла бы стать МА?

Определения МА антропоцентричны и лаконичны:

- МА - любой агент, который несет ответственность за действия. МА обязательно является **концептуальным** агентом, то есть агентом, который обладает концепциями и использует их. В т.ч. – концепцией “Я” (*Parthemore J., Whitby B. What makes any agent a moral agent? Reflections on machine consciousness and moral agency, 2013*).
- МА действует в соответствии со своей ролью; свобода, как необходимое условие МА; понимание того, что такое хорошо и плохо и т.д. (*Mayo B. The Moral Agent. Royal Institute of Philosophy Lectures, 1968*)



Прикладной аспект создания И/АС

1. Поведение (принятие решений) И/АС должно отвечать нашим общим представлениям о поведении МА.

2. Будем считать, что МА может быть не только человек, а некая сущность вообще, в том числе – искусственный агент (монополия человека в вопросах морали уже давно под сомнением, см., например, *Вааль де Ф. Истоки морали: В поисках человеческого у приматов. М.: Альпина нон-фикшн, 2019*).

3. Мы постулируем: множество проявлений свойств МА определяется:

(1) Наличием у агента модели мира, в которой имеется компонент “Я” (познающий субъект).

(2) Механизмом сопоставления наблюдаемого конспецифика (контрагента) с “Я”.

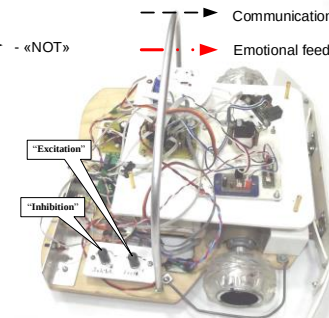
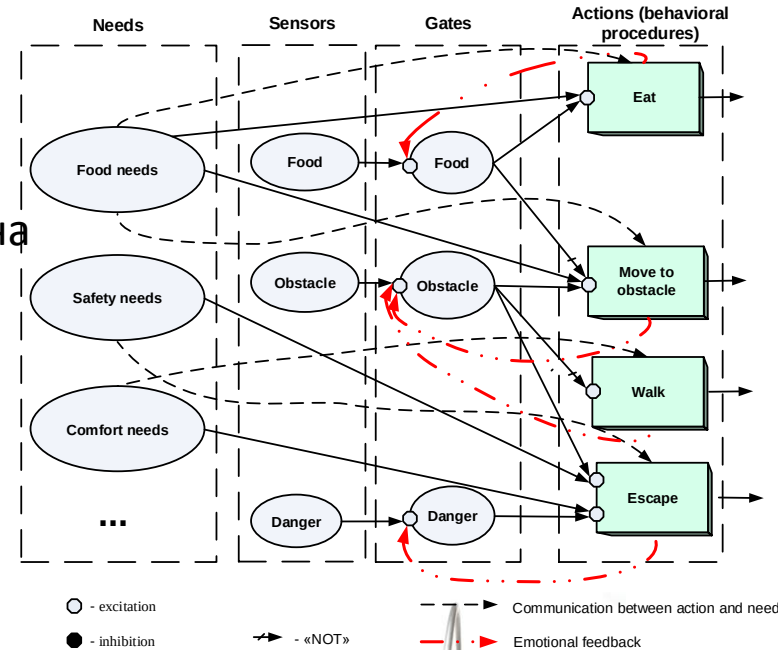
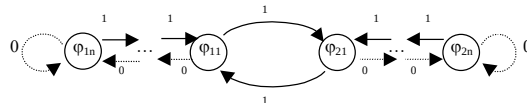
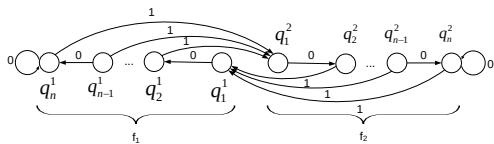
(3) Наличием эмоционально-потребностной архитектуры СУ нижнего уровня.

Важно. Все эти компоненты имеют вполне **практическое**, реальное воплощение в технических системах.

Далее – о том, как эти механизмы реализуют ряд феноменов поведения, присущих моральному агенту.

Феномены и механизмы. Эмоции и потребности

- Эмоции (на физиологическом уровне) и темперамент (на психическом уровне) могут быть присущи технической системе как чисто прагматические механизмы, которые влияют на успех искусственного агента в сложных недетерминированных средах.
- Э - свойство СУ, реализующее такие функции как контрастное восприятие, стабилизация поведения, индикация состояния, работа в условиях неполноты информации и т.д.



- В архитектуре СУ эмоциональные реакции системы определяются контурами ПОС. Эти связи отвечают за оценку ситуации и определяют величину эмоционального состояния согласно Информационной теории эмоций П.Симонова (Симонов, П. В. Потребностно-информационная теория эмоций. Вопросы психологии, 1982, 6, 44–56):

$$E = f(N, p(I_{need}, I_{has}))$$

Феномены и механизмы. Модель мира, «Я»

- ММ – атрибут системы управления интеллектуального агента.
- ММ + “Я” (субъект деятельности) => Картина мира (КМ) (Осипов Г.С. и др. *Знаковая картина мира субъекта поведения. М.: Физматлит, 2018*).
- КМ – надстройка над базовым стимул-реактивным уровнем:
 - реализует воздействие на сенсорную систему;
 - определяет значимость тех или иных потребностей => **изменяет характер поведения системы**, ее целеполагание и пр.
- Знаковая или семиотическая модель. Знак: имя n , перцепт p (образ), значение m (способ использования) и личностный смысл (цели, мотивы) a .

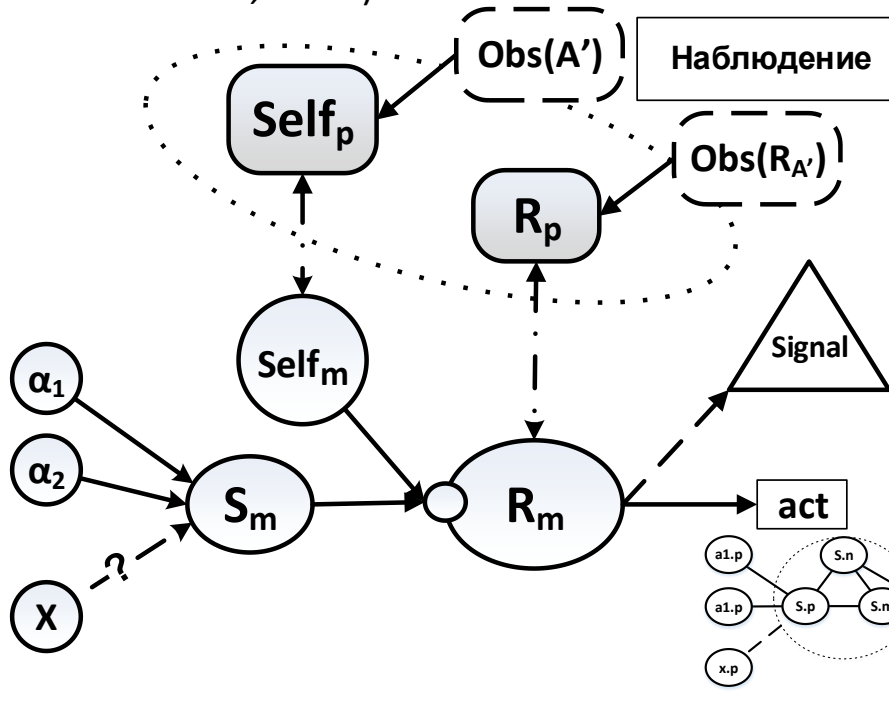
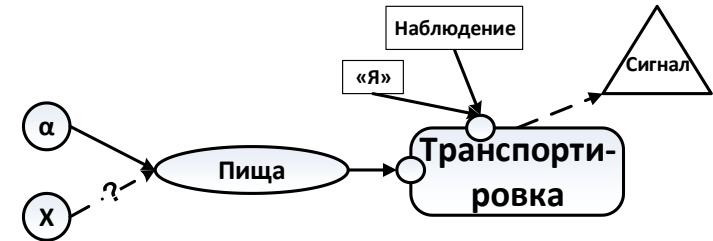
Допущения:

(1) Элементы СУ снабжены дополнительным подтверждающим входом. Это – вход для сигнала от вершины “Я”. Действие не будет активировано, если не будет подтверждающего сигнала от “Я”, интерпретируемого как "принадлежность" этого действия агенту (чувство – ощущение – самости, агентивности), т.е. отождествление или восприятие объекта как своего. С точки зрения семиотики это означает, что эти действия являются *значением* знака “Я”, т.е. вопрос обусловленности решается самым естественным образом.

(2) Активизация какого-либо компонента знака влечет **активизацию** других его компонент; между одновременно активными вершинами сетей возникают **ассоциативные** связи.

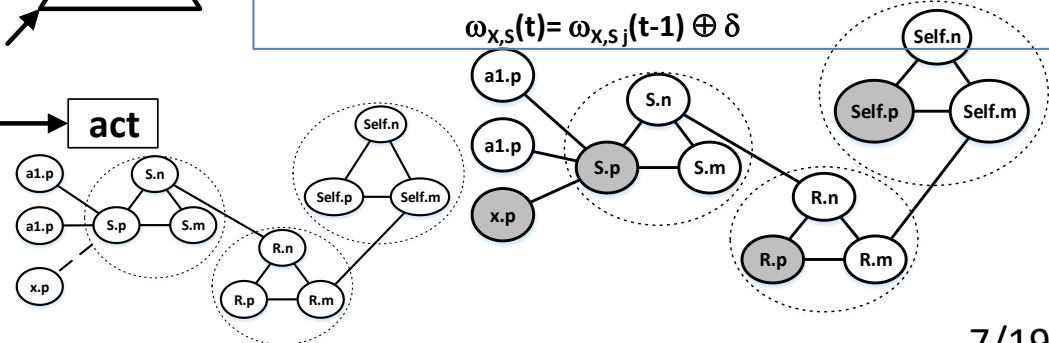
Феномены и механизмы. Подражательное поведение

1. Основной механизм - сопоставление субъекта с конспецификом ("Я" и другого), определение степени их близости.
2. В природе это отождествление сходно с тем, что называется родственным отбором, когда эволюционно выгодным становится поведение, определяемое степенью родства взаимодействующих особей (Уилсон Э. *Эусоциальность: Люди, муравьи, голые землекопы и другие общественные животные*, 2020)



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. $S = \omega_{\alpha1,S} \alpha_1 \oplus \omega_{\alpha2,S} \alpha_2 \oplus \dots$
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:

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



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 $Se = \langle O, P, M, \Omega, E \rangle$
- 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: $P = \Omega_o^P \cdot O, \Omega_o^P \subset \Omega$

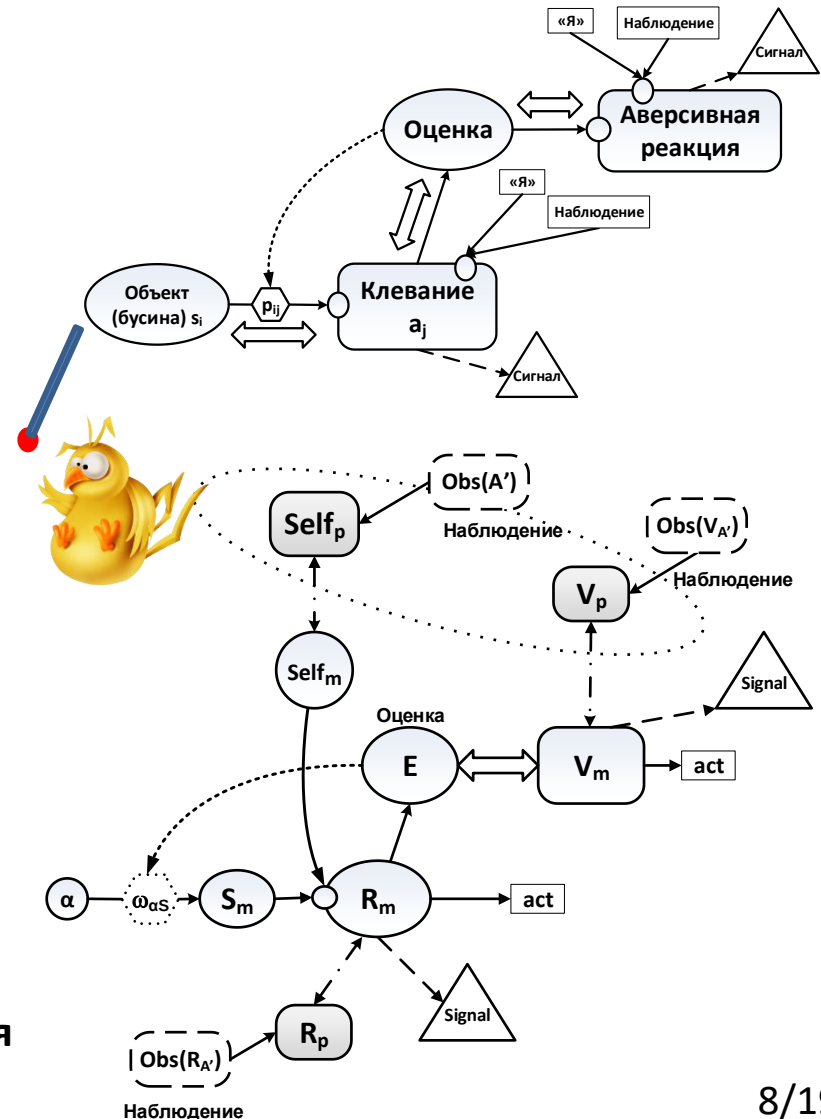
2. Определение активности элементов вектора M: $M = \Omega_p^M \cdot P, \Omega_p^M \subset \Omega$

3. Вычисление оценки: $E = \omega_{VE} V_m$

4. Изменение весов связей между активными элементами, если $E \neq 0$:

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

Особенности рефлекс: Чужой опыт забывается быстрее



Феномены и механизмы. Эмпатия

- Эмп - способность к отзывчивости на эмоциональные состояния окружающих индивидов разной степени близости.
- Эмп определяет эмоциональную склонность к сотрудничеству и проявление альтруизма.
- Предрасположенность индивида к эмпатии является **необходимым** (но не достаточным) свойством для моральности индивида.
- Эмп не является сугубо человеческим свойством. В этологии одним из базовых механизмов формирования социального взаимодействия особей, является т.н. **симпатическая индукция**. Ее определение тождественно определению эмпатии.



Два аспекта эмпатии: (1) механизм ее реализации, (2) объект эмпатии.

1. Реализация. Тот же принцип отождествления (или определения степени близости) наблюдаемого агента с “Я”.

Эмпатия = {Эмоции + Отождествление контрагента + Подражательное поведение}

2. Объект эмпатии.

- (1) Эмоции – это способ **интегральной оценки состояния** индивида (П.Симонов).
- (2) Имеется внешнее проявление эмоционального состояния агентов.
- (3) Эмп – основа для уровня управления, связанного с целеполаганием и планированием. В терминах моральной философии это означает действие золотого правила: *либо реализовывать такой план действий, при котором контрагенту будет хорошо (увеличение уровня положительных эмоций: "поступай так, чтобы..."), либо сформировать план, не ведущий к появлению отрицательных эмоций (негативная формулировка: "не навреди")*.
- (4) Эмоциональное состояние контрагента влияет на формирование мотива поведения и цели.
- (5) Мотивация и целеполагание - основная роль Эмп с технической точки зрения.

Характерные свойства морального агента

1. Язык. «Мораль не может существовать без символического языка». Но мир животных прекрасно обходится без языка. Не язык, а сигнальная коммуникация, т.е. внешнее проявление внутреннего состояния животного (Панов, 2014). На этом построено множество механизмов социального поведения и взаимодействия. В том числе – эмпатия.

2. Мотив. В моральной философии можно найти любую удобную точку зрения на мотивацию поведения. Например, Дж.Локк: *личный интерес представляет собой единственный разумный мотив*.

3. Чувства. Если это – некий процесс, отражающий **субъективное оценочное отношение к некоторым объектам**, то процессы, происходящие в эмоционально-потребностной архитектуре, можно назвать чувствами (feelings). Робот действительно может испытывать чувства. Если же подразумевается, что МА должен обладать так называемым **"моральным чувством"** (moral sentiments, [Smith, 2006]), то задача упрощается: можно вовсе отказаться от рассмотрения этого свойства.

- И. Кант. Вывел из сферы морали чувства, считая их участие в мотивации деяний предпосылкой моральной ущербности последних.
- Ф. Хатчесон. Моральное чувство не может непосредственно мотивировать, а является откликом на мотив.
- Д. Юм. Моральное чувство – результат действия более простых психологических начал – симпатии и ассоциации идей. Одобряются любые мотивы, ведущие к человеческому счастью.

Панов, Е. Н. (2014). *Эволюция диалога. Коммуникации в развитии: от микроорганизмов до человека*. М.: Языки славянской культуры.

Darwall, S. (2007). *The Foundations of Morality: Virtue, Law, and Obligation*. In D. Rutherford (Ed.), *The Cambridge Companion to Early Modern Philosophy* (pp. 221–249). Cambridge: Cambridge University Press.

Smith, A. (2006). *The Theory of Moral Sentiments* (6th ed.). MetaLibri.

Характерные свойства морального агента

4. Симпатия. Является следствием сопоставления наблюдаемого контрагента с "Я". Сила симпатии зависит от степени близости контрагента. Это - прямое следствие организации знак-ориентированной картины мира. Возбуждение отдельных компонентов знака (перцепта, значения или личностного смысла) приводит к возбуждению его остальных компонентов, в том числе – и собственно знака "Я".

5. Ответственность. «МА несет ответственность за свои действия». Здесь все просто:

- МА должен владеть определенными ключевыми понятиями и иметь способность в течение длительного периода <...> соответствующим образом использовать эти понятия (Parthemore & Whitby, 2013:).
- "... Быть морально ответственным за что-то – значит заслуживать определенную реакцию – одобрение, порицание <...> – за его совершение" (Talbert, 2019).

6. Требование самостоятельности решений. Субъект не должен действовать в соответствии с заложенной кем-то программой и т.д. и т.п. => неизбежные спекуляции.

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness*, 05.

Talbert, M. (2019). Moral Responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.

Промежуточные выводы

1. Многие свойства морального субъекта могут быть присущи и искусственному агенту.
2. Нас не интересует *имитация* психических или когнитивных процессов, свойств сознания и т.п. Мы имеем сугубо технические решения, повышающие **адаптивные способности** технического устройства. Эти решения – модели и механизмы – могут иметь интерпретацию в гуманитарных терминах.
3. Мы старательно обходили вопросы моральной философии. Рассуждения об утилитаризме, эволюционной этике и прагматизме вне сферы наших компетенций. Была лишь постановка вопроса:

Если есть некий перечень свойств, которыми должен обладать моральный агент, то существуют ли причины, по которым мы не можем признать таковым искусственного агента – робота?

4. Все описанные выше механизмы вводились исключительно и только как адаптивные механизмы для решения задач 3 уровней:

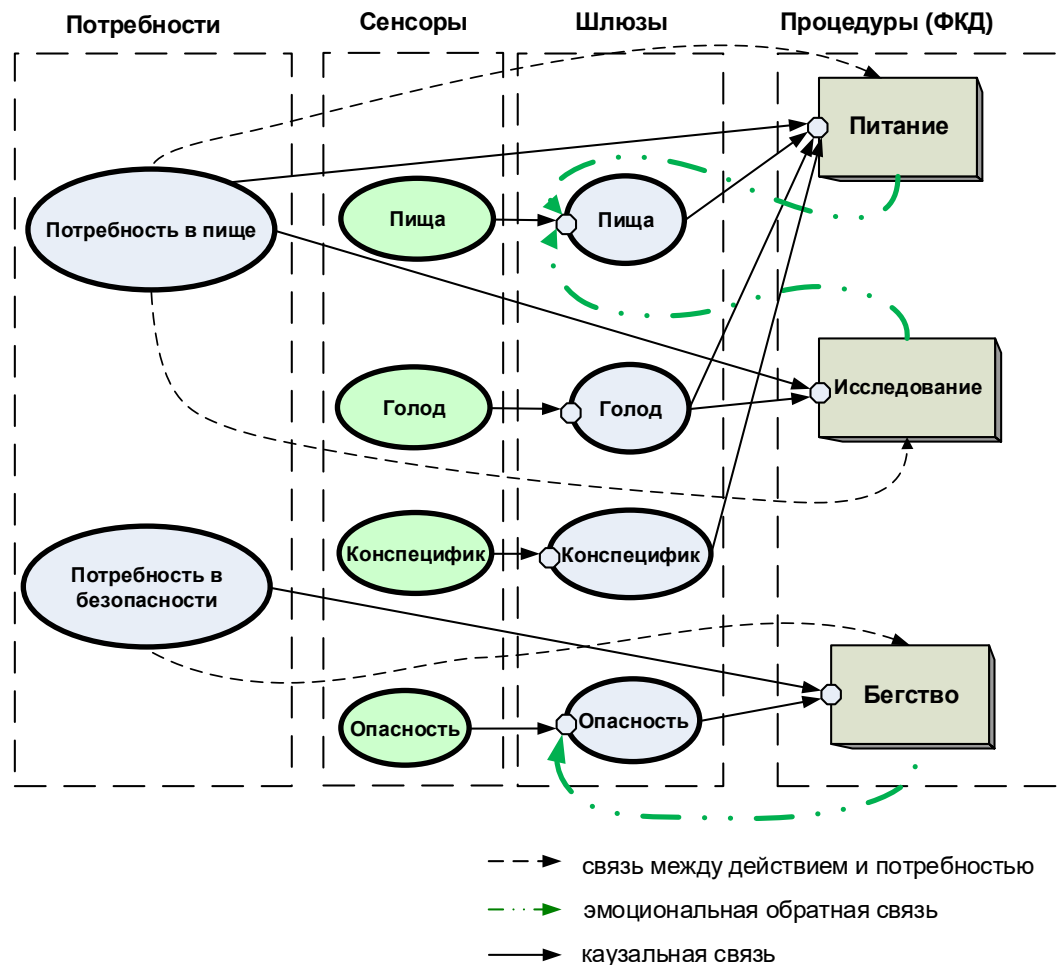
- (1) Обеспечение возможности техническому устройству целесообразно действовать в сложной, недетерминированной, динамической среде.
- (2) Организация взаимодействия внутри группы агентов, вплоть до появления форм социальной организации (образование социума – тоже механизм адаптации).
- 3) Целенаправленное управление поведения социумом. Возникает необходимость реализации дополнительных механизмов адаптации, позволяющие социуму сохранять свою **устойчивость**. Одним из важнейших факторов стабилизации - наличие механизмов **разрешения конфликтов** внутри социума (основная задача и суть морали).



Пример задачи (из области экспериментальной морали)

1. Основные формы поведения живых существ: пищевое, агонистическое, исследовательское, и репродуктивное (родительское)
2. Вопросы размножения исключены из рассмотрения, а такой элемент родительского поведения, как кормление, входит в состав обобщенной процедуры питания (предположим, что обмен пищей среди насекомых - трофаллаксис - обусловлен именно базовым родительским поведением).

Архитектура анимата

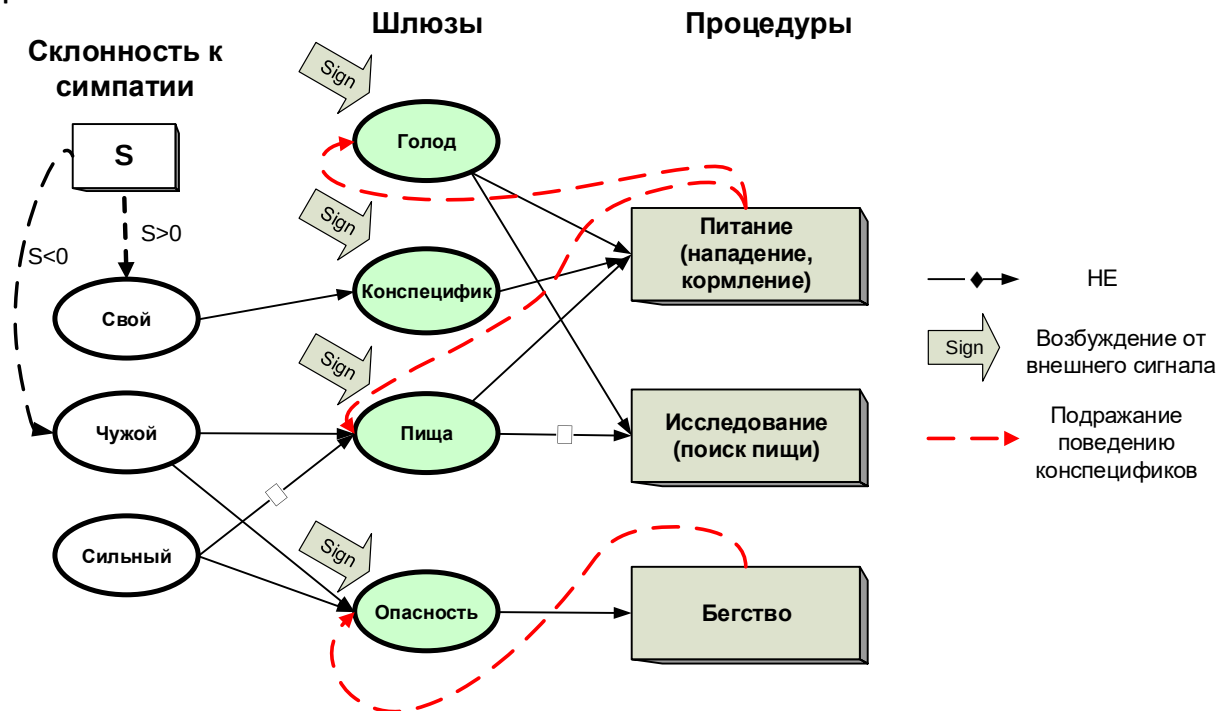


Анимат

1. Склонность к симпатии S . $S \in [-1, 1]$. Определяет восприятие других агентов с точки зрения "свой-чужой".

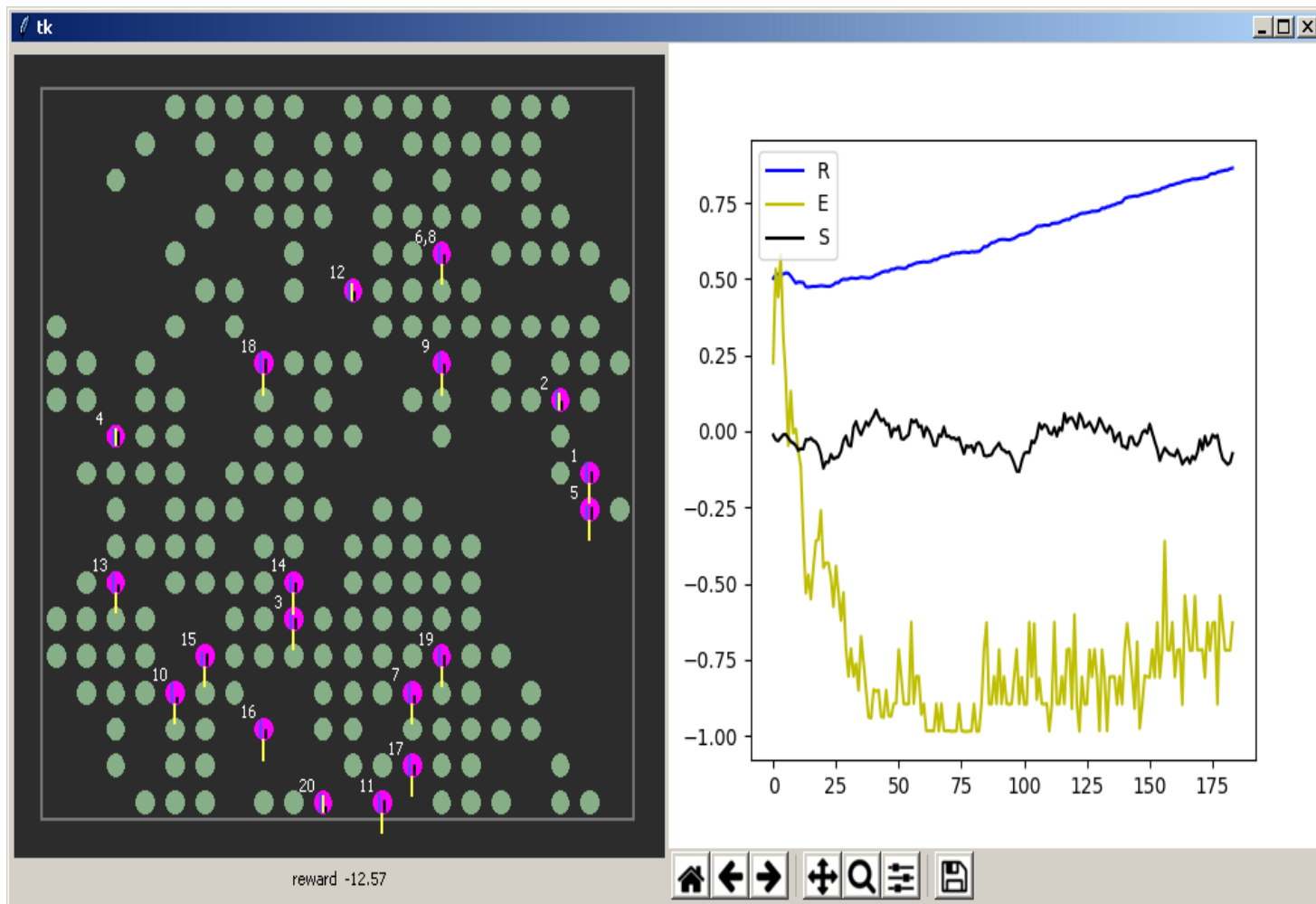
- $S > 0$: Агент отождествляет наблюдаемого конспецифика с собой. Агент способен делиться пищей с этим конспецификом.
- $S < 0$: Агент не отождествляет себя с другими агентами, рассматривая их как источник ресурсов (пищу).
- $S = 0$: Агент безразличен (толерантен) к другим агентам. При этом если агент видит чужого, ресурс которого больше, чем у него самого, тогда этот чужой воспринимается как опасность.

2. Сигнальная коммуникация. Агенты генерируют сигналы, определяющие их эмоциональное состояние.



Задача определения взаимозависимостей R-E-S

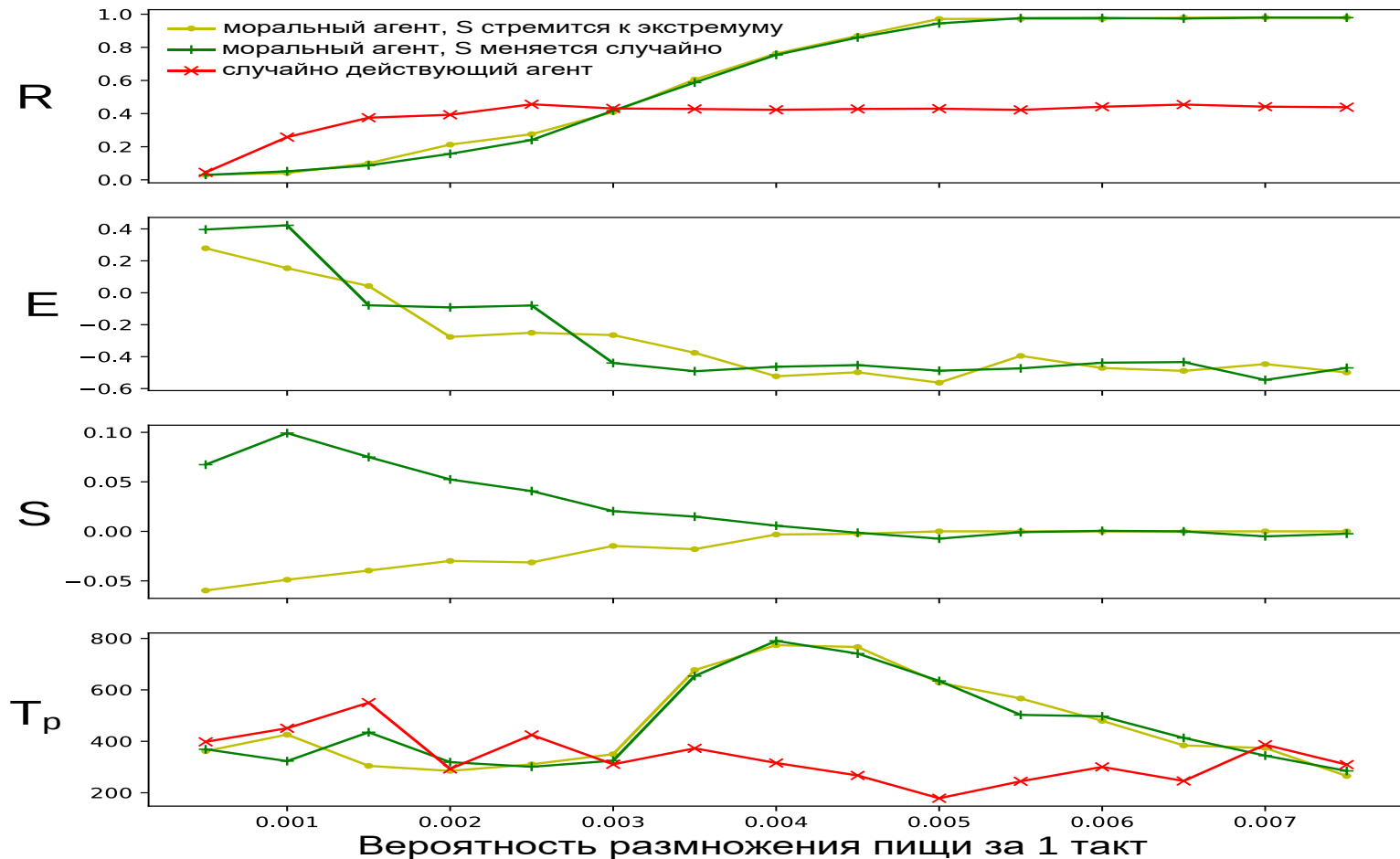
Исследование зависимости значений параметров агентов от закона изменения симпатии S и богатства среды.



R - количество ресурса, E - эмоциональный уровень, S - уровень симпатии

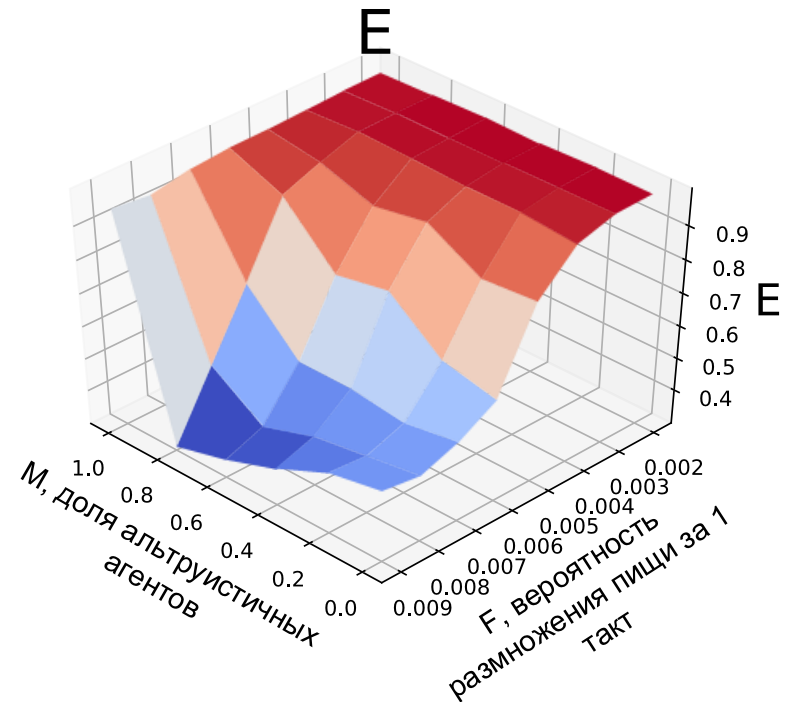
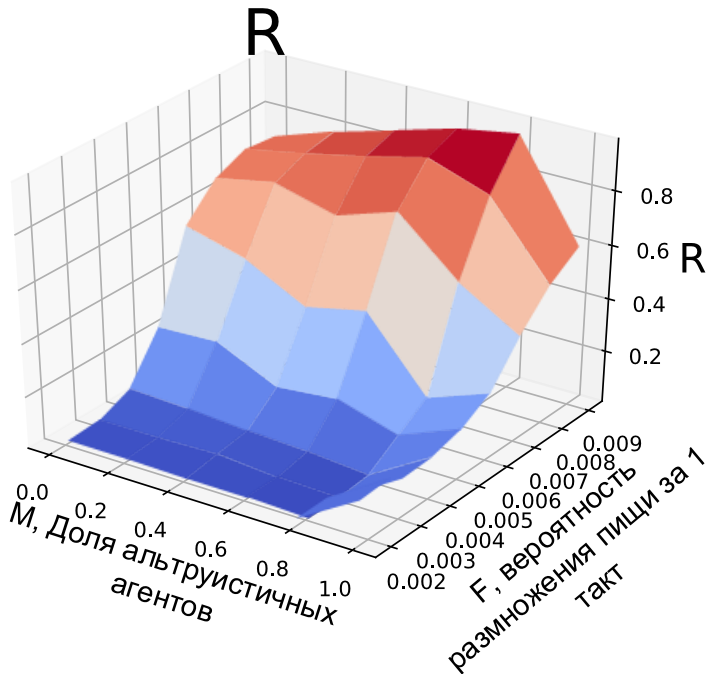
Эффективность популяции

Зависимость R , E , S и T_p (время переходного процесса) от изобилия пищи в среде



Моральные установки агентов способны несколько изменить успешность их действий как сообщества в пограничных случаях, но мало влияют на функционирование при очень богатой или очень бедной среде.

Доля альтруистов/эгоистов в обществе



В богатых средах $R = \max$ при $M = 80\%$. Полностью альтруистичный коллектив оказался в целом менее продуктивен, чем коллективы с агрессивными агентами.

Е мало (счастья мало)

В бедных средах продуктивность полностью альтруистического общества выше.

Постоянная раздача ресурсов агентами друг другу вместо сбора новых ресурсов при полном отсутствии хищников; это понижает как потери ресурса при хищничестве, так и интенсивность их сбора.

Е велико (все заняты тем, что хотят – ищут пищу)

Выводы по результатам экспериментов

(Все банально и очевидно)

1. Моральность общества важна в сложных условиях существования, в бедных средах; для выживания в этих условиях требуется кооперация. Когда среда богата и взаимопомощь не требуется, то более выигрышной индивидуальной тактикой является эгоизм, вплоть до проявления агрессии.
Или: в экстремально тяжелых и экстремально благоприятных средах понятия эмпатии и симпатии не играют заметной роли и т.п.
2. В бедных средах уровень E выше, чем в богатых. Если все-таки интерпретировать E как показатель счастья (по И.Канту, например), то это можно перефразировать так: бедные живут более счастливо, потому что у них есть меньше неудовлетворенных потребностей.

"Счастье – это такое состояние разумного существа в мире, когда все в его существовании происходит согласно его воле и желанию", Кант И. Критика практического разума // Соч.: в 4 т. на немецком и русском языках, Т.3. М.: Московский философский фонд, 1997. 784 с.

Основная задача - не столько получение новых результатов из области моральной философии или социологии, сколько изучение базовых механизмов, лежащих в основе подобного рода феноменов.

Цель экспериментов:

- (1) Показать адекватность предложенной архитектуры анимата и моделей поведения философским, социологическим и даже бытовым представлениям о содержании морального поведения.
- (2) Продемонстрировать возможность перехода от абстрактных рассуждений гуманитарного плана к вполне конструктивным схемам и моделям, позволяющим рассматривать вопросы морали с технической, экспериментальной точки зрения.

Заключение

1. Мы делегируем интеллектуальной/автономной системе **самостоятельно принимать решения, критически важные для человека**. Если в ситуации выбора исчерпаны все соображения технического и правового характера, то остаются критерии **морального** характера. Доверие к такому "этическому" решению возможно лишь тогда, когда то, принимающее решение сущность является **моральным агентом**.
2. Все описанные выше механизмы вводились исключительно и только из соображений **технической целесообразности**, для решения задач создания эффективных адаптивных механизмов трех уровней:
 - (1) Целесообразное поведения технического устройства в сложной, недетерминированной, динамической среде.
 - (2) Организация взаимодействия внутри группы агентов. Вплоть до появления форм социальной организации.
 - (3) Задача целенаправленного управления поведения социумом.
3. Сегодня идет интенсивное и достаточно успешное развитие когнитивных и социальных способностей интеллектуальных автономных систем. Вплоть до их моральности.
4. В области этики поведения И/АС продвижение сопряжено с рядом сложностей, и основной проблемой является отсутствие конструктивных моделей, которых ждут исследователи от моральной философии.