


Ордена Ленина и ордена Октябрьской революции

НИЦ «Курчатовский институт»

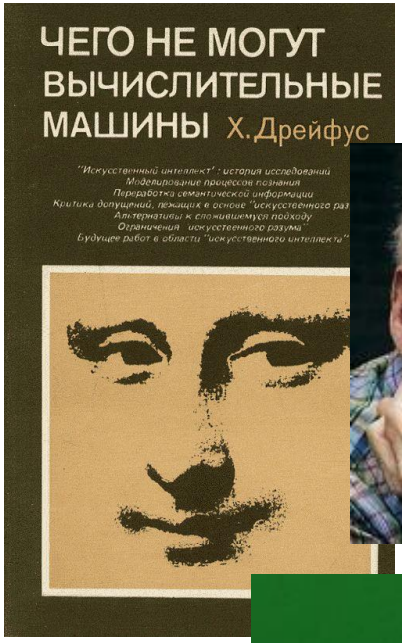
Карпов В.Э.

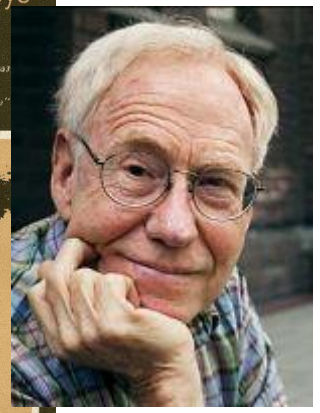
К вопросу о Законах робототехники А. Азимова

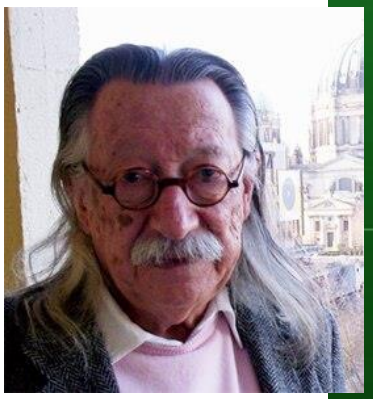
karpov.ve@gmail.com

Принятие «трюкачества»

1. Общество понимает, что ему преподносится некая имитация осмысленности: большие языковые модели (всякие чаты GPT) или хорошо натасканные на каких-то примерах НС и т.д.
2. Сегодня: вместо критического и скептического отношения к подобного рода вещам, разворачиваются обсуждения, дискуссии, принимаются административные решения, строятся планы по использованию таких механизмов в госуправлении, юриспруденции и т.п.
3. Раньше: общество – гуманитарная часть – активно возмущалась, когда вместо чего-то реально подобного человеческому разуму ему предлагались внешние имитации ([Дрейфус, 1978] или [Вейценбаум, 1982] сегодня не воспринимаются).
4. Помимо НС есть еще одна тема, оседланная гуманитарным сообществом – пресловутые Законы робототехники А.Азимова. В профессиональном робототехническом сообществе эти тема почти не затрагивается.







Законы робототехники Айзека Азимова

Одна из самых обсуждаемых тем в изучении взаимоотношений автономных/интеллектуальных систем и человека.

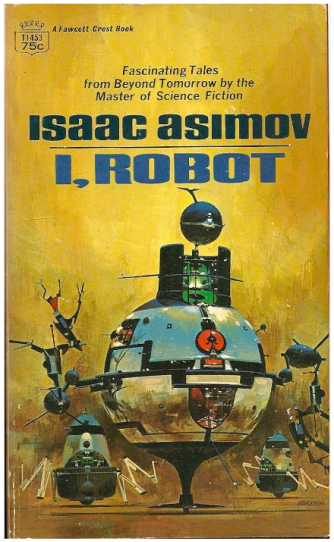
Начиная с 1940-х годов, эти законы неоднократно обсуждались, анализировались и дополнялись.

Рассказ «Хоровод», март 1942 г. (менее явно – рассказе «Лжец!», 1941).

Исаак Юдович Азимов,
- 1919, д.Петровичи, Смоленская губерния, РСФСР
- 1992, Нью-Йорк, США

Законы:

1. *Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинен вред.*
2. *Робот должен повиноваться командам человека, если эти команды не противоречат Первому Закону.*
3. *Робот должен заботиться о своей безопасности, поскольку [в той мере, если] это не противоречит Первому и Второму Законам.*



-М.: Издательство «Знание», 1964

Все не так очевидно, все – о другом

1. Азимов писал как раз о противоречиях этих законов.

2. То, о чем писал Азимов в его «Хороводе» – это вовсе не о Законах робототехники, а об эмоциях.

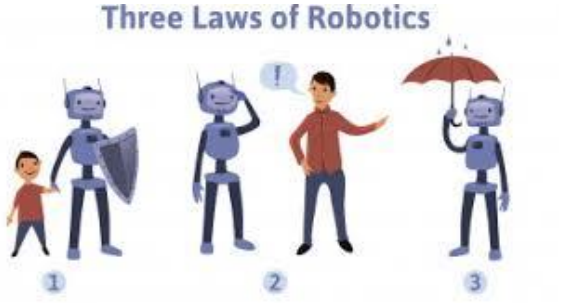
3. В этих законах имеется ряд сугубо технических (робототехнических) проблем, не решенных до сих пор.

4. Законы Азимова имеют тесную связь с вопросами этики автономных/интеллектуальных систем.

- **Внешние противоречия** формулировок Законов – то, что лежит на поверхности – их форма, то, что активно обсуждается, но при этом то, к чему их автор относился весьма критически.
- **Внутренние противоречия** Законов – это про подмену понятий.

Внешние противоречия

1. Все понятно и очевидно, поэтому популярно в гуманитарной сфере.
2. Перечень законов пополнялся и уточнялся (1986, Нулевой: *«Робот не может причинить вред человечеству или своим бездействием допустить, чтобы человечеству был причинён вред»*).
3. Пробуждение интереса гуманитарной общественности к вопросам этики систем ИИ (АИС) привело к тому, что в различного рода обсуждениях и публикациях по этим вопросам упоминание или ссылка на эти три закона является почти обязательным.
4. Но А.Азимов – не только писатель-фантаст, но и ученый (биохимик), популяризатор и историк науки. Он понимал, что пишет, что и как формулирует.
5. Его рассказы про роботов – это истории коллизий этих законов.
6. В такие «очевидные» формулировки заложены сплошные противоречия, неопределенности, конфликты, которые и определяют фабулу и драматизм историй Азимова.



Внешне Законы выглядят хорошо, но в таких формулировках они работать не могут.

Зорькин В.Д. Право и вызовы искусственного интеллекта, РГ, 2024:
«Признать особый правовой режим (правовое положение) роботов, наверное, нужно»

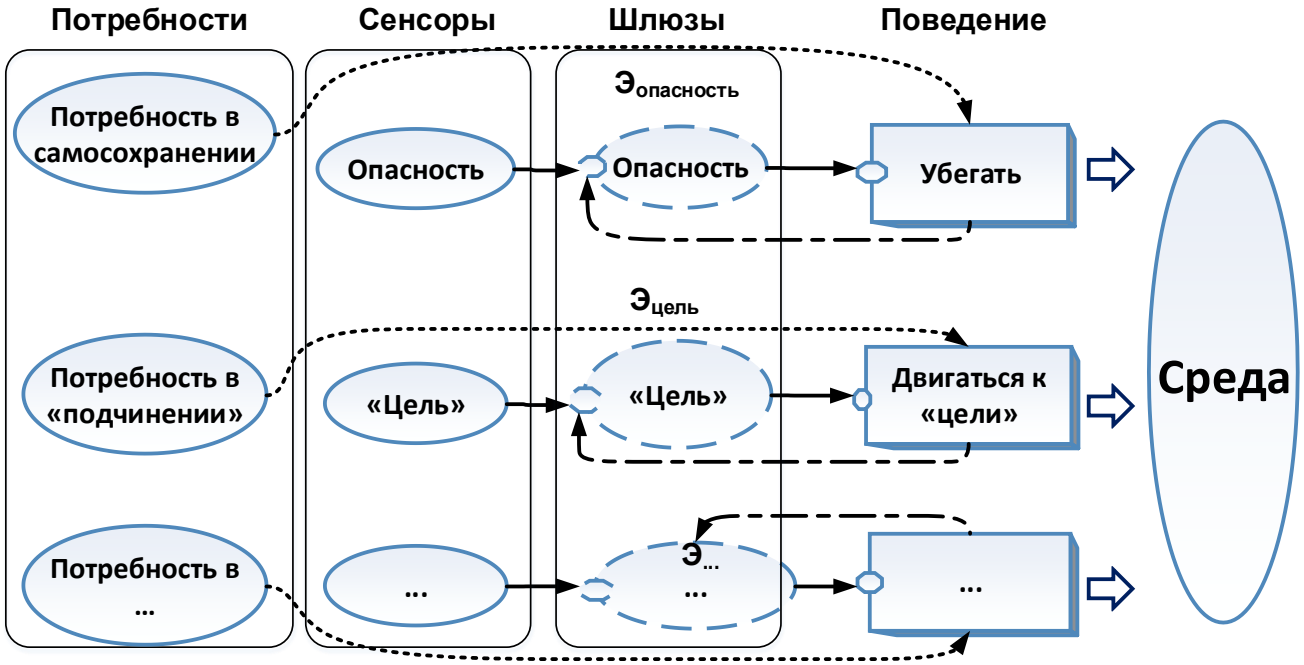
Эмоции и поведение

1. В «Хороводе» поведение робота имеет весьма косвенное отношение к трем законам. Причина неустойчивого поведения робота – его **эмоциональная несбалансированность**.
2. Эмоции, как комплекс процессов, присущи не только человеку или высшим млекопитающим.
3. Э. – это нижний, психофизиологический уровень управления животного, обладающего более или менее сложным поведением.
4. Основой эмоциональных процессов является интегральная оценка ситуации, когда определяется баланс между теми средствами и ресурсами, которые нужны животному для удовлетворения своих актуальных потребностей и теми, которые имеются в наличии. Согласно В.П. Симонову:

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

- Различие между потребностями и возможностями определяет эмоциональную оценку текущей ситуации.
- Эмоции отвечают за стабилизацию поведения, контрастирование восприятия, даже за кратковременную память.
- Механизм реализации эмоций - контуры ОС в СУ, которые отвечают за устойчивость поведения.
- Нарушение эмоционального механизма влечет серьезные функциональные нарушения, известные в нейропсихологии, подобные тем, что были у робота Спиди.

Фрагмент эмоционально-потребностной схемы системы управления робота Спиди



Здесь одной из базовых потребностей робота является потребность в подчинении человеку, а целевой объект (селен, за которым посылали робота) обозначено как «Цель».

- Шлюз – узел, который формирует процессы, которые мы называем эмоциями.
- С каждым шлюзом связана своя частная эмоция, которая зависит от того, какова сила соответствующей потребности, что наблюдает робот (сенсорика) и чем он реально занят.
- Шлюз ответственен за сенсорику и фантомные сигналы (раздражителя уже нет, опасность не видна, а агент продолжает убегать), именно там накапливается отрицательная эмоция, связанная с тем, что агент не выполняет сейчас ту поведенческую процедуру, которая должна быть актуальной и т.д.

Схема поведения и характер

Поведенческая процедура, которую выполняет робот, зависит напрямую от эмоций, определяемыми степенью удовлетворения потребностей.

1. Роботу отдан приказ доставить селен, потребность в этом актуальна и велика.
2. Известно, куда двигаться и что с ним делать.
3. Вместо этого робот вынужден удаляться от опасного места (потребность в самосохранении).
4. Через какое-то время накапливается отрицательная эмоция, формируемая шлюзом «Цель».
5. Ослабевает активность процедуры убегания (робот удаляется от опасности). В результате робот начнет выполнять процедуру движения к цели, приближаясь при этом к опасному участку.
6. Goto 1.

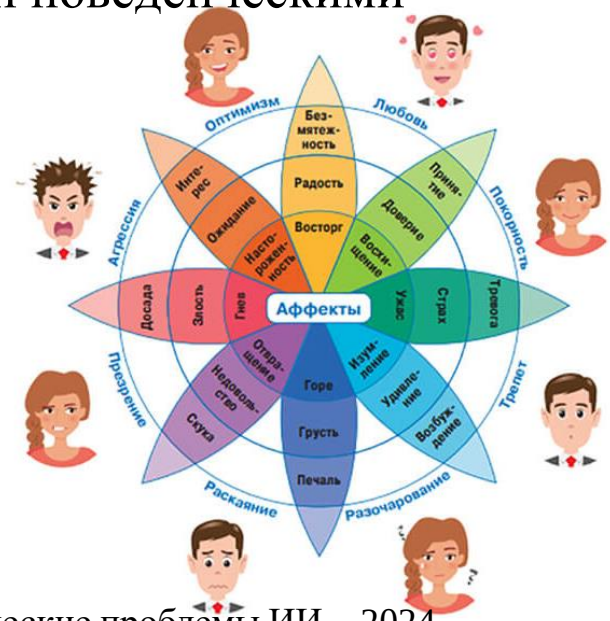
«Регулировка» поведения: изменение силы связей между элементами схемы (сделать его более эмоционально устойчивым).

«Подкручивая» параметры этой схемы, получаем различный характер поведения робота (меланхолик, флегматик, сангвиник, холерик). Спиди был типичным холериком.

«Разнообразие» эмоций

- Часто эмоции путают с чувствами, а сами эмоциональные процессы – с внешним проявлением эмоций.
- Чувства направлены на что-то конкретное, а эмоции – интегральная оценка ситуации, при этом репертуар эмоций крайне скуден.
- Согласно Симонову, есть четыре эмоциональных состояния: гнев, страх, удовольствие и отвращение. И все они связаны с базовыми поведенческими процедурами.

Последователи Р. Плутчика [Plutchik, 2001] любят красивые картинки и безудержно расширяют спектр эмоций.



Технический аспект. Нерешенные научные, технические проблемы

По Азимову структура мозга робота (СУ) должна быть такой, чтобы эта система вообще не могла функционировать при нарушении определенных правил поведения (законов робототехники).

=> **Законы не просто органично интегрированы в структуру СУ, но образуют центральную ось всей структуры СУ.**

=> Вопросы:

1. Возможно ли создать архитектуру СУ, чтобы в ее основе лежали этические императивы, а не просто какой-то дополнительный набор эвристик?
2. Чем будет руководствоваться АИС при принятии решений?
3. Могут ли этические императивы стать важной частью каких-либо потребностей АИС (наряду с потребностью самосохранения, реализации социальных функций и т.д.)?

Система управления роботом и этика поведения

Азимовские законы имеют теснейшую связь с **этикой**.

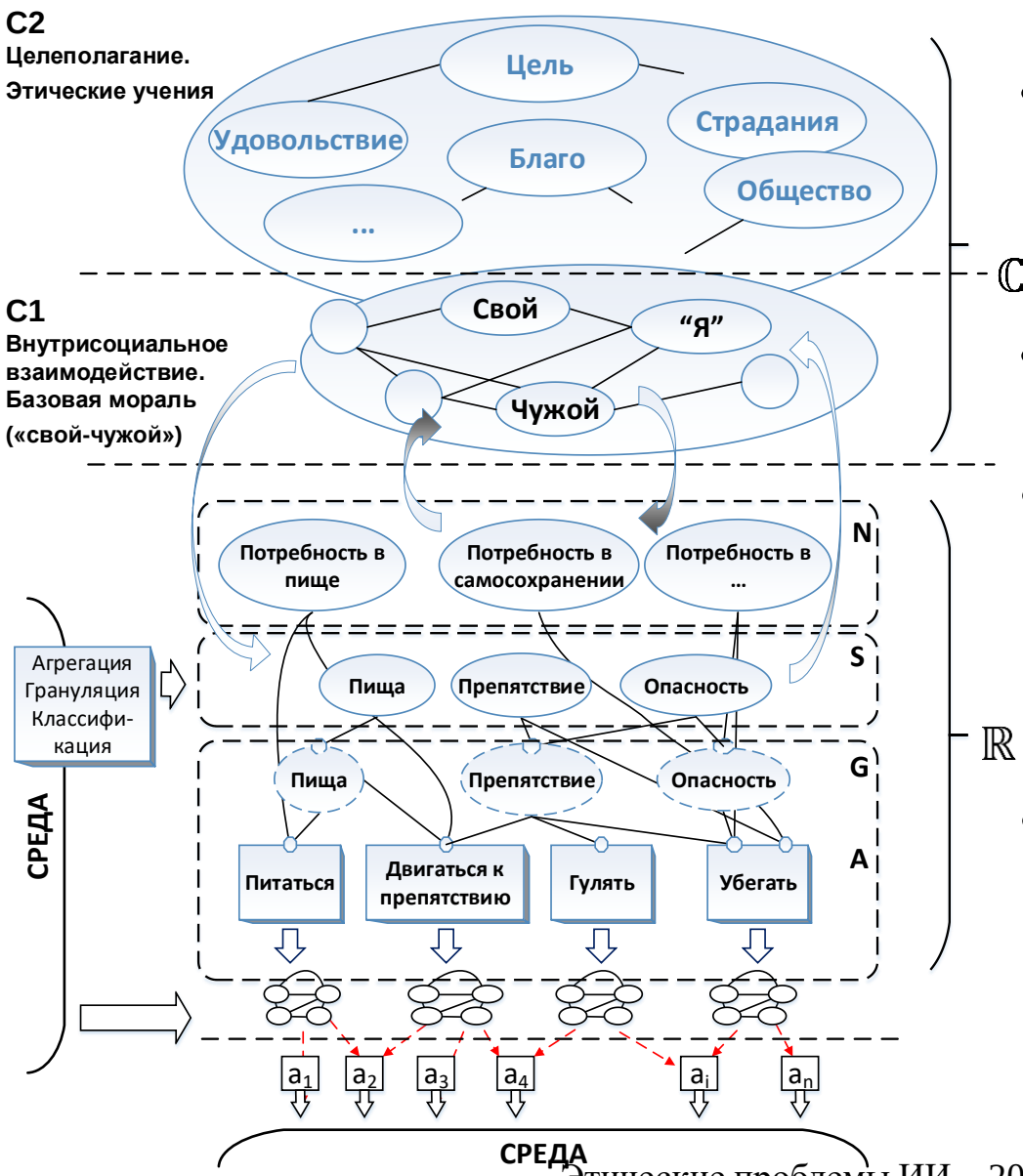
Рассказ Улика: *"...Три Закона Роботехники совпадают с основными принципами большинства этических систем, существующих на Земле"*.

Два важных аспекта этики.

1. Основное назначение этики – разрешение конфликтов. Мораль – это адаптивный механизм, позволяющий социуму более эффективно приспосабливаться к сложным условиям.
2. Мораль – это не прерогатива человека (Вааль де Ф. Истоки морали: В поисках человеческого у приматов. М.: Альпина нон-фикшн, 2019).

- эмпатия (базовое понятие этического поведения) – это сугубо биологический механизм, присущий животным, причем даже не высшим млекопитающим;
- понятия "Я", "свой", "чужой" определяют социальное поведение животных;
- они же являются основой того, что называется т.н. "золотым правилом морали".

Архитектура интеллектуального (когнитивного) робота



- Уровень ℝ отвечает за условно-рефлекторную деятельность агента,
- Когнитивный уровень ℂ:
- ℂ1 описываетмеханизмы взаимодействия агента с социумом (себе подобными)
- ℂ2 отвечает за более абстрактные понятия, в том числе – из области этики.

Этика как легковесная надстройка

1. Чем выше очередной уровень надстройки, тем она легковеснее и вариабельнее.
2. Грубое вмешательство в виде, например, изменения весов потребностей, чревато потерей функциональности системы.
3. Верхнеуровневые этические нормы крайне гибки и способны меняться в многократно в зависимости от изменяющихся внешних условий.

Представления о моральных категориях (что такое хорошо-плохо, правильно-неправильно) меняются на протяжении жизненного цикла индивида (животного, робота) многократно, а вот удовлетворение базовых, физиологических потребностей – это вопрос выживания здесь и сейчас.

=>

При такой структуре, когда этические императивы образуют надстройку СУ, их «обход» – это сугубо технический вопрос.

Моральный робот-партнер. Уровень 1.

Коллаборация как полноценное партнерство. Робот – напарник человека, вместе с которым человек выполняет некоторую сложную и критически важную для человека задачу (например, совместное патрулирование или охрану).

https://vektorus.ru/blog/kollaborativnyj-robot.html

Если среда естественная, слабоформализуемая, недетерминированная, то робот должен быть автономным и интеллектуальным

(анализировать сложную обстановку, планировать поведение, принимать решения).

=> Поведение робота неизбежно должно быть сродни поведению животного.

=> Поведение робота может трактоваться с точки зрения удовлетворения его потребностей, основным из которых является потребность в самосохранении.

Множество потребностей и поведенческих процедур, заложенных в робота и призванных удовлетворять возникающий потребности – это уровень простого животного, реализующего простое или сложное **рефлекторное поведение**.

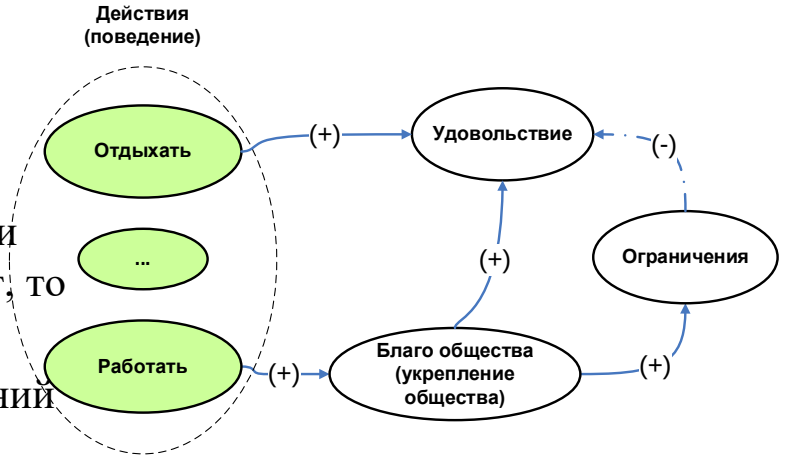
1. Это еще не полноценный партнер.
2. Роль оператора сводится к тому, чтобы манипулировать базовым поведением робота для выполнения прикладных задач (включаться в роли вида "свой-чужой", формировать условия среды и т.д.).

Коллизии. Главное – уметь объясниться (оправдаться)

Партнером робот станет тогда, когда его поведение будет определяться дополнительными оценочными механизмами, императивами или правилами – заложенным этическим учением.

Суть любого этического учения – в определении приоритетов, оценок, целевых функций (что такое хорошо/плохо, правильно/неправильно и т.п.).

1. Неизбежно возникает коллизия между тем, что предписано правилами морали, и тем, что диктует рефлекторный уровень.
2. В критической ситуации гипотетическое "этическое " азимовское правило вида "Необходимо подчиняться человеку" вступает в противоречие с правилом "Если опасность, то надо убегать" (правило подчинения является этическим).
3. Разрешение вопросов коллизий этического поведения робота-партнера осуществляется точно так, как это происходит в человеческом обществе.
4. Вопрос не в том, что сделал робот в той или иной ситуации, а в том, сможет ли он **объяснить** это с точки зрения заложенной в него этической схемы. Если смог, то поведение робота этичное, иначе – нет.
5. Формально это означает построить цепочку рассуждений (доказательство) того, что действие было обусловлено базовым этическим императивом.



Примеры

findallways1: интересы_личности

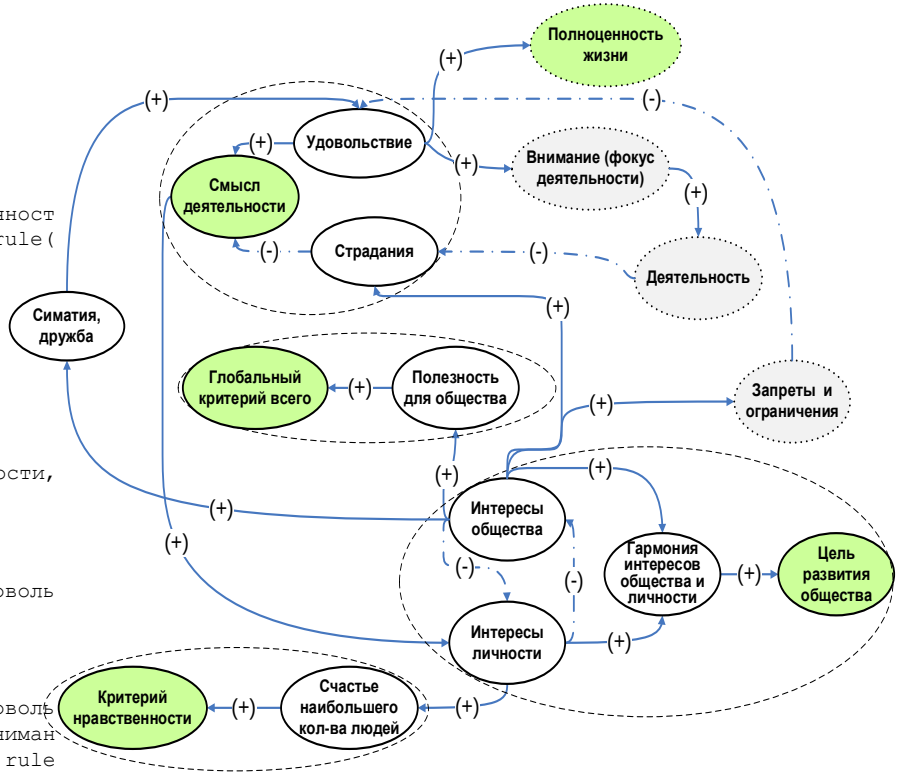
```
path(интересы_личности,интересы_общества,
[rule(r91,интересы_личности,интересы_общества,neg)],-1)

path(интересы_личности,гармония_интересов_о_л,
[rule(r92,интересы_личности,гармония_интересов_о_л,pos)],1),path(интересы_личности,гармония_интересов_о_л,[rule(r91,интересы_личности,интересы_общества,neg),rule(r84,интересы_общества,гармония_интересов_о_л,pos)],-1)

path(интересы_личности,счастье_наибольшего_кол_ва_людей
[rule(r93,интересы_личности,счастье_наибольшего_кол_ва_людей,pos)],1)
```

интересы_общества -> смысл_деятельности

```
1) path(интересы_общества,смысл_деятельности,
[rule(r82,интересы_общества,страдания,pos),rule(r21,страдания,смысл_деятельности,neg)],-1),

2) path(интересы_общества,смысл_деятельности,
[rule(r83,интересы_общества,симпатия_дружба,pos),rule(r71,симпатия_дружба,удовольствие,pos),rule(r12,удовольствие,смысл_деятельности,pos)],1),

3) path(интересы_общества,смысл_деятельности,
[rule(r83,интересы_общества,симпатия_дружба,pos),rule(r71,симпатия_дружба,удовольствие,pos),rule(r13,удовольствие,внимание_фокус_деятельности,pos),rule(r31,внимание_фокус_деятельности,деятельность,pos),rule(r41,деятельность,страдания,neg),rule(r21,страдания,смысл_деятельности,neg)],1)]

SLIST=[-1,1,1] 2 1 => противоречие
```

evaluate_loops: POS_LOOPS

```
1) path(интересы_общества,интересы_общества,
[rule(r85,интересы_общества,интересы_личности,neg),rule(r91,интересы_личности,интересы_общества,neg)],1),

2) path(интересы_личности,интересы_личности,
[rule(r91,интересы_личности,интересы_общества,neg),rule(r85,интересы_общества,интересы_личности,neg)],1)]
```




```
rule(r11, "удовольствие","полноценность_жизни", pos).
rule(r12, "удовольствие","смысл_деятельности", pos).
rule(r13,
"удовольствие","внимание_фокус_деятельности", pos).
…
rule(r93, "интересы_личности",
"счастье_наибольшего_кол_ва_людей", pos).
rule(r101, "гармония_интересов_о_л",
"цель_развития_общества", pos).
rule(r111, "счастье_наибольшего_кол_ва_людей",
"критерий_нравственности", pos).
```

Обсуждение и заключение

1. Законы Азимова нельзя понимать буквально. Азимов – сам лучший их критик.

2. Законы – это прежде всего повод задуматься о противоречивости и неработоспособности жесткой системы императивов, когда мы имеем дело с реальным, динамическим, недетерминированным и сложным (слабоформализуемым) миром; о роли эмоций; о том, насколько глубоко и подробно или, напротив, общезначимо следует ставить задачи для АИС.

3. Принципиально глубокий вопрос о моральной агентности робота. Оставаясь в современной парадигме создания ИАС, мы можем говорить о том, что робот может быть полноценным МА, принимая решения, основываясь на нормах морали, неся ответственность и т.п.

- Важно, что это будет "обычный" МА, агент, совсем не тот, о котором говорил Азимов.
- Обычный МА принципиально ограничен и ситуативен при принятии решений.

Абсолютная моральность или мораль на уровне рефлексов

4. **Абсолютная** агентная моральность невозможна в силу того, что для этого необходимо просчитывать все возможные последствия действий и оценивать их с точки зрения морали, что накладно и нереализуемо технически.

- Если имеется некоторый реальный горизонт планирования оценок последствий, то вновь получится естественно ограниченный моральный агент. Кроме того, все равно не решаются проблемы коллизий при выборе морального решения. Единственный способ их разрешения – это **аргументировать** моральность действия.
- Иными словами, моральный агент по образу и подобию человека может быть создан, но у Азимова как раз не об этом.

5. Нерешенной технической проблемой является реализация этических императивов не в виде компонента когнитивной надстройки системы управления, а определение их на базовом уровне.

- Неясно, как создать адекватно функционирующего робота, у которого моральные в человеческом понимании императивы выполняются рефлекторно.
- Если создать робота, который руководствуется такими принципами на базовом уровне архитектуры, и при этом система останется функциональной, то мы столкнемся с еще более неприятной ситуацией, когда у агента будет совсем нечеловеческая мораль со всеми вытекающими гуманитарными и социальными последствиями.

Чувство вины

6. Законы – это хороший повод к новым исследованиям.

Рассказ "Улика“. Крайне неуклюжий, казалось бы, вариант разрешения коллизий Законов

"[Робот] сойдет с ума, если будет поставлен перед таким противоречием — нарушить букву Первого Закона, чтобы остаться верным его духу".

Азимов сумел изящно обойти такой скользкий вопрос, как наличие у робота **чувства вины**.

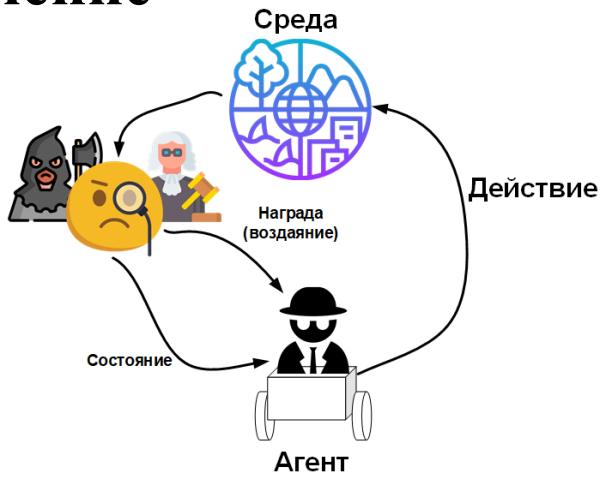
Если робот может быть МА, то неизбежным становится наличие механизма **оценки** моральности его поведения, создание таких механизмов, которые удобно называть **совестью**, чувством вины и т.п.

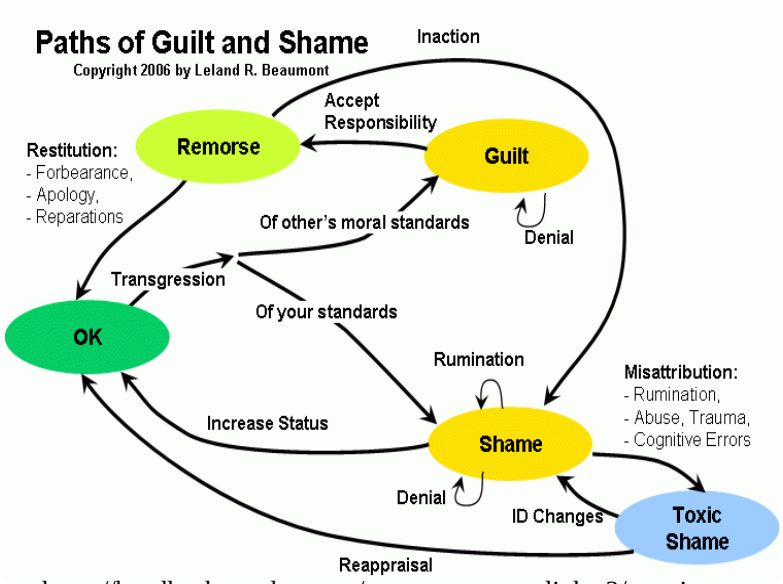
Совесть – способность к моральной оценке.
Вина – результат работы совести.
Стыд – это эмоция, кратковременная вина.
Чувство – длительная и устойчивая форма реакции на мир.
Фрейд З.: мы чувствуем себя виноватыми, когда существует разрыв между нашим моральным сознанием и нашими врожденными желаниями и импульсами.

1. Гришина Е. С. Философия вины: к вопросу классификации // Вологдинские чтения, №60. , 2006. С. 59–67.
2. Демкин А. Чувство вины [Электронный ресурс]. URL: https://onkto.ru/blog/psychology/chuvstvo-viny
3. Фромм Э. Гуманистический психоанализ. СПб.: Питер, 2002. 544 с.
4. Guilt Resources [Электронный ресурс]. URL: https://bandbacktogether.com/master-resource-links-2/emotions-feelings-resources/guilt-resources

Чувство вины и обучение

- Роль чувства вины сводится к реализации процедуры поощрения/наказания, определения штрафных воздействий и прочих оценочных механизмов, требуемых для процедуры обучения АИС.
- Роботы Азимова, будучи МА, обходятся без чувства вины, без обучения в смысле определения моральных последствий своих действий.
- "Обычный" робот действует по рефлекторно-ассоциативной схеме:
  1. В некоторой новой ситуации реализуется некоторая поведенческая процедура, далее оцениваются последствия.
  2. Если возникает чувство вины (оценочный модуль выдает штраф), то в подобных ситуациях робот больше не будет выбирать эту процедуру (сформировалась соответствующая ассоциация).
- Азимовский робот в случае противоречия просто перестает функционировать.





https://bandbacktogether.com/master-resource-links-2/emotions-feelings-resources/guilt-resources/