Семинар "Этические проблемы ИИ", 17.12.2019, ИФ РАН

Применимо ли понятие морали к отношениям между искусственным агентами?

В.Э. Карпов

Karpov.ve@gmail.com

НИЦ "Курчатовский институт", МФТИ

Одним из направлений групповой робототехники являются т.н. модели социального поведения. Ключевой задачей этого направления является создание конструктивных моделей, позволяющих реализовывать феномены социальной организации в группах искусственных агентов. Основная идея парадигмы моделей социального поведения (МСП) заключается в том, чтобы рассматривать принципы организации сообществ роботов с точки зрения некоторого универсального адаптационного механизма.

В докладе ставится вопрос о правомерности применимости понятия морали к описанию поведения искусственных агентов – роботов или аниматов. С одной стороны, мораль может рассматривается как один из механизмов групповой и индивидуальной адаптации посредством социальной организации поведения. С другой стороны, правила, определяющие моральность поведения, являются крайне удобным способом управления поведением социума. Не затрагивая такие фундаментальные механизмы системы управления (СУ) индивида, как параметры и структура, моральные установки, будучи весьма гибкими и вариабельными, позволяют эффективно определять поведение индивида и социума в целом.

В докладе описывается потребностно-эмоциональная архитектура СУ анимата, а также рассматриваются некоторые базовые механизмы управления его поведением, основанные на принципах паразитического манипулирования. Приводится описание таких механизмов социального поведения живых и искусственных агентов, как когезия, контагиозное поведение и др. Описывается такой элемент СУ анимата, как Я (Я-познающее, субъективное Я, С.Я.). Введение С.Я. позволяет, в том числе, реализовать такие феномены, как подражательное поведение и социальное обучение. Все эти механизмы вкупе с сигнальной коммуникацией и потребностно-эмоциональной СУ, являются основой для реализации такого феномена, как эмпатия. В основе разработанных моделей лежит понятие степени близости наблюдаемого контрагента (конспецифика) к субъекту. Делается предположение, что подражательное поведение, социальное обучение и эмпатия можно рассматривать как некий базис для поведения, оцениваемого с точки зрения морали.

В докладе обсуждаются три основных вопроса, определяющих содержание морального поведения аниматов: (1) зачем такое поведение нужно, (2) какова целевая функция (или основной регулятив), определяющая поведение, (3) каковы механизмы, лежащие в его основе. Постулируется, что: (1) в перечень механизмов, определяющих нравственное поведение, наряду с прочими, входят социальное обучение и эмпатия и (2) основной мотивацией морального поведения (то, на что направлено золотое правило морали) является максимизация эмоционального уровня контрагента, на которого направлено воздействие индивида. Поскольку знак и величина эмоции непосредственно определяется существующими потребностями агента, то можно сделать весьма "механистический" вывод: действие основного морального регулятива направлено на удовлетворения потребностей индивида. Такая неизбежная вульгаризация – это результат попытки переноса сугубо технических механизмов (вплоть до адаптационных моделей) на принципиально плохо формализуемую область – моральную философию.

В любом случае, если продолжать исследование аспектов моральности поведения аниматов "снизу", с "технической" стороны, то остается открытым целый ряд вопросов, таких, например, как:

1. Насколько нравственные правила определяются подражательным поведением, социальным обучением и способностью к эмпатии?
2. Насколько целесообразно интерпретировать формы поведения анимата именно с точки зрения основного нравственного правила?
3. Что подлежит регуляции в "моральном поведении"? Только ли характеристика близости "свой-чужой"?
4. Какова значимость фактора необходимости совместной деятельности для решения сложных задач в зависимости от свойств среды обитания и индивидуальных потребностей?

Эти вопросы требуют своего разрешения, причем вне зависимости от положений моральной философии, а оставаясь в рамках парадигмы моделей социального поведения роботов.