Теоретические и методологические предпосылки обсуждения проблем этики искусственного интеллекта

В.Э. Карпов

К.т.н., доцент,
начальник лаборатории робототехники НИЦ "Курчатовский институт"

Резюме выступления

Интерес к вопросам этики со стороны искусственного интеллекта (ИИ) и робототехники имеет давнюю историю, причем этот интерес имел двоякий характер. Исторически первым и основным рассматриваемым аспектом были рассуждения об опасности «думающих машин» (Turing, A. Computing machinery and Intelligence (1950) – проблема машинных ошибок, Wiener, N. Some Moral and Technical Consequences of Automation (,1960)– опасность для человека и этические проблемы; Bostrom, N., Yudkowsky, E.. The Ethics of Artificial Intelligence (2011) – негативные сценарии возможного развития ИИ и т.д.). В последние годы такой подход набирает все большую силу, становясь вместе с тем менее осмысленным и конструктивным. Другой подход к проблематике этики в системах ИИ – это попытки формализации понятий морали с целью их имитации в программных и технических системах, см., например, Gensler H.J. Formal Ethics (1996), Лефевр В. Алгебра совести (2003). Достижения в этом направлении выглядят гораздо скромнее.

Вместе с тем, сегодня интерес к этике со стороны ИИ-сообщества начал приобретать несколько иной характер, переходя во вполне осязаемые, прагматические сферы. Так, с 2016 года идет работа по созданию этических стандартов для интеллектуальных и автономных систем (И/АС). Речь идет о глобальной инициативе Института инженеров в области электротехники и электроники (Institute of Electrical and Electronics Engineers, IEEE), в рамках которой разрабатываются стандарты, определяющие этические принципы построения интеллектуальных систем. На данный момент разрабатываемых стандартов уже 18. Спектр их крайне широк: от этических стандартов обработки персональных данных до вопросов эмпатии автономных систем. Разумеется, во многом эти стандарты основаны на анализе возможных опасностей И/АС для человека, однако начинают затрагиваться и иные этические проблемы, имеющие сугубо технический характер использования И/АС.

С другой стороны, в связи с развитием исследований в области коллаборативной (взаимодействующей с человеком) робототехники, исследований группового поведения роботов и т.д., возникают вопросы о целесообразности применения понятий морали для описания механизмов взаимодействия внутри сообществ искусственных агентов – роботов. В этом отношении мораль (этика, нравственность) понимается несколько механистически, как механизм урегулирования конфликтных ситуаций между агентами, а также как способ целеполагания поведения. Так, в рамках такого направления, как моделирование социального поведения в системах групповой робототехники реализованы модели подражательного поведения и социального обучения. При этом в основе этих моделей лежат такие механизмы, сопоставление контрагента с тем, что называется Я, оценка эмоционального состояния и пр. На основании этого делается вывод о возможности моделирования такого механизма, как эмпатия:

Эмпатия = {Эмоции + Отождествление контрагента + Подражательное поведение}

Разумеется, рассмотрение морали с точки зрения регуляции взаимодействия искусственных агентов вызывает множество возражений со стороны этиков. Более того, очевидно, что такой сугубо механистический перенос понятий морали в техническую сферу приводит зачастую к весьма вульгарным «логическим» выводам типа: «Основной мотивацией морального поведения является максимизация эмоционального уровня контрагента, на которого направлено воздействие индивида». Тем не менее, это не снижает остроту и актуальность самой проблемы.

Уже говорилось о том, что вопросы этики применительно к самым разным научно-техническим сферам понимаются в основном с точки зрения опасности применения тех или иных технологий. Однако в области ИИ (И/АС) ситуация совершенно иная. Основное отличие проблематики этики здесь заключается в том, что:

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».

В этом и заключается суть того, что должно быть определено в тех же «этичесих» стандартах по проектированию И/АС. Предмет исследования этически обусловленного проектирования – это И/АС, совершающие выбор того или иного значимого, критически важного для человека или общества действия или решения. При этом нас интересует ситуация, когда совершение этого выбора осуществляется на базе некоторых эвристик, основанных на этических императивах. И это вызывает целый ряд проблем. Прежде всего – понимание того, в чем заключается этичность решения. Итак, принципиально важным являются вопросы конструктивных определений и онтологий. Задача онтологий заключается во взаимном увязывании и согласовании этических и технических понятийных систем.

Примечание. Понятие "онтология" (с маленькой буквы) уже давно применяется в ИИ, причем в отличном от философского понимания смысле. Согласно Т.Груберу (1993) онтология – это эксплицитная спецификация концептуализации. Формально онтология состоит из терминов, организованных в таксономию, их определений и атрибутов, а также связанных с ними аксиом и правил вывода. Этот термин был введен при рассмотрении вопроса взаимодействия интеллектуальных систем (ИС) друг с другом и человеком. Считается, что ИС должны уметь обмениваться между собой имеющимися у них знаниями и представлениями. Для этого описание знаний о мирах, в которых существуют эти ИС, должно быть конкретно, формально и унифицированно. Онтология, как спецификация концептуализации, – это тройка O={C, R, A}, C – это совокупность концептов предметной области, R – совокупность отношений между ними, A – набор аксиом или правил, которые описывают законы и принципы существования концептов. С технической точки зрения онтологии — это базы знаний специального типа, которые могут «читаться» и пониматься, отчуждаться от разработчика и/или физически разделяться их пользователями.

Не менее остра проблема наличия математического аппарата, позволяющего формализовать этические понятия. В некотором смысле можно полагать, что такой аппарат есть. Нечеткие логики, правдоподобные рассуждения, многомерные шкалы и т.п., – все это, возможно, способно в той или иной мере реализовать формализм, необходимый для этически обусловленного проектирования.

Самым же сложным вопросом, видимо, является, проверка «этичности» И/АС. С технической точки зрения основной проблемой является проблема эти-

ческой 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.

Таким образом, можно говорить о взаимном интересе моральной философии и технических наук друг к другу. Вместе с тем, пока мы вряд ли можем говорить о реальных интеграционных процессах. Основная причина этого – проблемы создания единой понятийной базы. Так, в технических исследованиях моральные аспекты зачастую ограничивается их бытовым, интуитивным пониманием. Это связано, прежде всего, с отсутствием формального или хотя бы конструктивного описания основных положений этики. С другой стороны, специалисты в области моральной философии не всегда четко представляют себе суть достижений в области И/АС.