Этические проблемы искусственного интеллекта

Исследовательский семинар

1 октября 2019, ИФ РАН
18:30–20:30

Р.Г. Апресян (Институт философии РАН)
В.Э. Карпов (НИЦ «Курчатовский институт», МФТИ)

Теоретические и методологические предпосылки обсуждения проблем этики искусственного интеллекта

Тезисы

Уже предварительные обсуждения проблематики, связанной с «этикой искусственного интеллекта» обнаруживают расхождения в понимании «этики», «морали», «искусственного интеллекта», «робота». Вступая в междисциплинарные обсуждения, никто не может надеяться на инерции и интуиции понимания. Понимание надо вырабатывать для данного междисциплинарного обсуждения. С целью междисциплинарной ясности. Вряд ли кто-то сможет в наших обсуждениях не исходить из своего исследовательского и теоретического опыта, но каждому придется разъяснять другим и себе, что имеется в виду под «этикой», «моралью», «искусственным интеллектом», «роботом».

Понятие морали вырабатывается в западно-европейской культуре для интегративной репрезентации комплекса нормативных (например, справедливость, забота и др.) и дескриптивных (например, императивность, автономия, универсальность и др.) идей. Годятся ли они для приложения к функционированию автономных интеллектуальных систем (АИС)?

О каком аспекте функционирования АИС мы говорим? Об их программном «начале» (детерминистском или когнитивном), об их взаимодействии с человеком (с живыми существами вообще), об их взаимодействии друг с другом? Если о программном начале АИС детерминистского типа, то вопрос, кажется, сводится к знакомому – к этическим принципам инженерно-конструкторской деятельности вообще. Уместно ли предположить, что в случае программирования когнитивных АИС этические принципы претерпевают изменение? Когда же мы переходим к вопросам взаимодействия между АИС и человеком/людьми и тем более взаимодействия АИС между собой, видоизменяются ли как-либо составляющие понятия морали и морального кодекса? Видоизменяются ли они в зависимости от характера тех практических задач, которые решает АИС?

Судя по специальной литературе и медиа, в центре обсуждений этических проблем АИС – (а) алгоритмы решений моральных дилемм, которые предстоят АИС, (б) негативные социальные последствия от широкого внедрения АИС в жизнь человека и общества. Не упускаются ли в этих обсуждениях какие-то важные другие этические проблемы, связанные с АИС?

Обычная человеческая мораль – это средство согласования частных интересов ради блага людей и сообщества с помощью ценностно-императивных инструментов. Что это за инструменты? – имеющие императивную природу ценности, например, *непричинение вреда, равенство, соблюдение оправданных интересов других, исполнение обещаний, поддержание договоренностей, соучастие (солидарность, помощь), забота*. Они направлены на согласование частных интересов, и это значит – на предотвращение конфликтов или, если конфликты произошли, их конструктивное разрешение.

А в какие конфликты рискуют попасть АИС? В условиях какой деятельности, какого взаимодействия? – Не об этом ли должен быть этический кодекс для АИС?

***

Сегодня эти вопросы переходят во вполне осязаемые, прагматические сферы. Так, с 2016 года идет работа по созданию этических стандартов для интеллектуальных и автономных систем. 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.

Кроме того, вполне устоявшимися направлениями стали исследования социогуманитарных аспектов систем искусственного интеллекта и робототехники. Причем зачастую речь идет не только об этической стороне вопросов взаимодействия АИС с человеком, но и о том, насколько применимы понятия морали для описания механизмов взаимодействия внутри сообществ искусственных агентов – роботов.

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 АИС.