Этика конструирования систем искусственного интеллекта: мировые тенденции и российский подход к проблеме

Карпов Валерий Эдуардович
Вице-президент Российской ассоциации искусственного интеллекта

В последние годы активно если не развиваются, то, по крайней мере, широко обсуждаются темы этики в искусственном интеллекте (ИИ), угроз, происходящих от ИИ, различного рода гуманитарных аспектов создания систем ИИ. Будучи крайне старой темой, появившейся задолго до того, как было сформулировано само понятие ИИ и в большинстве сводящейся к пресловутому "бунту машин", эти вопросы сегодня получает новое звучание. При этом, несмотря на множество спекуляций и безграмотность, в этом звучании появляются вполне здравые рассуждения, связанные с некоторыми аспектами проектирования и применения технических систем, поведение которых является значимым для человека. В первую очередь эта тенденция связана, разумеется, с ростом числа критически важных, потенциально опасных систем, работающих автономно.

В докладе будет рассмотрены вопросы соотношения этики и искусственного интеллекта с критических позиций. Здесь важно подчеркнуть, что наблюдается некоторая путаница в самой постановке вопроса. Чаще всего, как будет видно ниже, речь идет об этичности применения систем искусственного интеллекта в тех или иных областях. Мы же далее будем говорить об этических аспектах функционирования таких систем, о том, насколько их поведение может быть обусловлено этическими парадигмами, нормами, представлениями. При этом мы постараемся не только отыскать некоторое рациональное зерно в рассуждениях об этике ИИ, но и сформулировать некоторые технические задачи, решение которых представляется значимым для данного вопроса.

Важным аспектом является рассмотрение интеллектуальной системы не только как когнитивной, но и активной сущности. В этом плане определяющим свойством такой системы является возможность осуществления воздействия на окружающий мир и прежде всего – социум. Иными словами, вопросы этики и ИИ сводятся к тому, что мы имеем дело с искусственной системой, реализующей процессы планирования, целеполагания и выбора того или иного поведения (ИИ-система или ИИС). При этом выбор, осуществляемый системой, должен определяться некоторыми этическими императивами и нормами в самом широком смысле. Например, этические нормы могут трактоваться как некоторые эвристики, которыми руководствуется ИИС при совершении выбора того или иного действия, формирования системы оценок, целевых функций и проч.

С прагматической точки зрения исследования в области этики ИИС приводят в конечном итоге к созданию различного рода стандартов и последующей сертификации ИИС. И здесь возникают три важнейшие проблемы.

Первая касается конструктивной формализации этических норм в форме, пригодной для описания функционирования конкретных программно-аппаратных комплексов. Вторая

проблема – это способность объективного (инструментального, прямого или косвенного, основанного на анализе поведения и т.п.) контроля соответствия компонент ИИС этическим нормам. Третья – это, то какое влияние окажут в будущем эти стандарты и не будут ли они нести жестко ограничивающую роль, которая будет только тормозить развитие ИИС.

В докладе будет проведен небольшой экскурс в историю вопроса этических проблем искусственного интеллекта, а также обсуждена инициатива IEEE по этически обусловленному проектированию систем ИИ. Кроме того, мы постараемся обсудить некоторые конструктивные аспекты такого этически обусловленного проектирования, особое внимание уделив вопросам имеющегося математического аппарата и, главное, формальной постановке задачи.

Опасность «думающих машин»

1) А. Тьюринг (Turing, 1950).

Обсуждает постулат «машины не могут делать ошибок» и отмечает, что в сложных машинах, ошибки могут быть детерминированы неадекватностью исходных данных, при этом машины максимально точно выполнит все математические операции по их обработки. Пример, когда интеллектуальная машина получает неадекватные исходные данные и далее совершает какие-либо неэтичные или "ужасающие действия" со временем стал одним из наиболее активно демонстрируемых не только в профессиональных сообществах, но и в популярной культуре и кинематографе.

2) Н. Винер (Wiener, 1960), (Wiener, 1965).

Основная мысль: машины могут быть опасны для человека и непредсказуемы. Даже понимая в деталях как работает машина, оператор может не успеть понять, что ее рассуждения идут к негативному сценарию или даже не успеть понять, что машина уже «приняла решение» и работает над осуществлением этого сценария.

30 Негативные сценария развития ИИ и угрозы (Bostrom & Yudkowsky, 2011), (Havens, 2016).

Инициатива IEEE

IEEE (Institute of Electrical and Electronics Engineers). Глобальная инициатива для исследований в области этики ИИ. Результатом таких исследований должны стать технические нормативные документы, регламентирующие разработку и внедрение систем ИИ с требованиями к их этическому поведению. Первым таким документом стал проект общих рекомендаций для разработчиков ИИ, посвященный тому, как разработчикам начать ориентироваться на этические проблемы в процессе разработки своих продуктов (IEEE, 2016) – "Ethically Alligned Design" (в вольном переводе – "Этически обусловленное проектирование"). В нем собраны основные ближнесрочные угрозы, связанные с внедрением автономных систем на базе ИИ, которые на сегодня отмечены в научной литературе. Помимо перечисления угроз IEEE обращает внимание на необходимость изменений в подготовке специалистов – разработчиков программных продуктов, использующих ИИ.

В целом документ (IEEE, 2016) является первым шагом к переносу рассуждений об этики ИИ из области научных исследований в практическое русло. Очевидно, что этот документ сам по себе пока обозначает круг проблем и дает только первичные идеи по их решению, однако это уже весомая основа от которой будут строится последующие, в том числе – нормативные, документы, разрабатываемые IEEE.

Карпов Валерий Эдуардович

Вице-президент Российской ассоциации искусственного интеллекта,
к.т.н., руководитель Отделения нейрокогнитивных наук и интеллектуальных систем
начальник Лаборатории робототехники НИЦ «Курчатовский институт»

Россия, 123182, Москва, пл. Академика Курчатова, 1
+7 (499) 196-71-00 (доб. 3370)

Источники

- Bostrom, N., & Yudkowsky, E. (2011). The Ethics of Artificial Intelligence. Cambridge Handbook of Artificial Intelligence, 1–20. http://doi.org/10.1016/j.mpmed.2010.10.008
- Havens, J. (2016). Heartificial Intelligence: Embracing Our Humanity to Maximize Machines. New York.
- IEEE. (2016). Ethically Alligned Design. IEEE. Retrieved from http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf
- Turing, A. M. (1950). Computing machinery and Intellicence. Mind, 54(236), 433–460.
- Wiener, N. (1960). Some Moral and Technical Consequences of Automation. Science, 131(3410), 1355–1358. http://doi.org/10.1126/science.132.3429.741
- Wiener, N. (1965). Cybernetics: or, Control and communication in the animal and the machine. M.I.T. Press.