ЗАКОНЫ РОБОТОТЕХНИКИ: НОВАЯ ПАРАДИГМА

Гришин Е.А.

Независимый консультант

bg987@mail.ru

В статье предлагается альтернативный подход к разрешению проблемы неуправляемости «взрывного» развития исследований и технологий в области «умных» машин. Подход основывается на конструктивной критике существующей парадигмы робототехники, известной как «три закона робототехники» писателя-фантаста А.Азимова.

По общепринятому мнению, основная проблема искусственного интеллекта, требующая первоочередного решения, заключается в том, как сделать машину «разумной».

На наш взгляд, проблема искусственного интеллекта, требующая первоочередного решения, заключается в том, как сделать машину «нравственной».

По крайней мере, если решать именно проблему «нравственности» автомата, это очевидно предполагает и решение проблемы «разумности» автомата.

Но обратное - неочевидно.

Писатель-фантаст А.Азимов сформулировал известные три закона робототехники [1], выглядящие в вольном пересказе так:

Робот не должен вредить человеку.

Робот должен исполнять приказы человека, если это не противоречит п. 1.

Робот должен заботиться о своей безопасности, если это не противоречит п.п. 1 и 2.

Эти законы сформулированы в предположении, что робот – это существо более низкого порядка по отношению к человеку. Доказательство: поставьте на место термина «робот» термин «собака»…

В то же время существуют предположения (Билл Джой, руководитель научного отдела компании Sun Microsystem, [2]), что создание самообучающихся, саморазвивающихся и самовоспроизводящихся автоматов приведёт к необратимым последствиям: автоматы станут умнее и многочисленнее людей, человеческий контроль их поведения и размножения окажется невозможным по определению, зато станет возможным контроль ими поведения человека.

Добавим, что, во всяком случае, автоматы не останутся существами «более низкого порядка».

Тогда, при сохранении традиционного подхода (три закона робототехники А.Азимова), не исключена возможность, что человеку вскорости придётся применять вышеприведённые три закона не к роботу, а к самому себе, только поменяв местами термины «робот» и «человек».

В описанных условиях назревающего технологического «апокалипсиса» в работе [2] предлагаются варианты его предотвращения. Не вдаваясь подробно в их рассмотрение, скажем лишь, что они сводятся к двум вариантам: добровольному отказу учёных и разработчиков от участия в проектах генно-/нано-/робото-технологий (ГНР-технологий) и созданию мирового правительства, реализующего тотальный запрет на исследования, разработки и применение ГНР-технологий.

При всей важности упомянутых подходов, на наш взгляд, целесообразно также рассмотреть решение поставленных проблем, в частности, в робототехнике, исходя из реальных профессиональных возможностей исследователей и разработчиков. При этом не надеясь, что иные варианты решения проблем увенчаются успехом.

Нам представляется необходимой конструктивная замена парадигмы взаимоотношений «человек-робот», выраженной в трёх законах робототехники А.Азимова.

Главный принцип, который следовало бы положить в основу новой парадигмы, должен быть принцип *равноправия* разумных существ, человека и робота (если угодно, пока – существа «квазиразумного»).

Правильнее было бы даже сказать – презумпция *равноправия* и презумпция *невиновности*.

Если согласиться с предложенным главным принципом, то вторая позиция новой парадигмы могла бы повторить «Золотое правило нравственности» (из Нагорной проповеди Иисуса [4]): «*Не поступай по отношению к другим так, как ты не хотел бы, чтобы они поступали по отношению к тебе*».

Изложенный подход применим лишь к ситуации компромисса во взаимоотношениях пары «человек-робот», в которой адекватными формами будут просьба, предложение, совет и т.д.

Для ситуации конфликта более характерны такие формы вербальных взаимоотношений, как требование, приказ, угроза, а также форма физических взаимоотношений – насилие.

Ситуации конфликта не столь уж редки среди априори равноправных партнёров – людей. Исходя из этого, разумно предположить, что они могут иметь место и во взаимоотношениях «человек-робот».

Попытаемся разобрать понятие «насилие» применительно к рассматриваемому случаю взаимоотношений «человек-робот».

Известна заочная полемика русского философа И.А. Ильина с графом Л.Н. Толстым по поводу его концепции «непротивления злу насилием» [3].

В ней И.А. Ильин формулирует своё понимание насилия как средства, неочевидно всегда служащего злой цели, и доказывает моральную и социальную необходимость насилия в совершенно конкретных ситуациях (в частности, для пресечения преступления).

В то же время из работы [3] остаётся невыясненным, как же может быть сформулирован критерий моральной оправданности насилия для совершающего его человека.

Сделаем свою попытку сформулировать критерий моральной оправданности насилия для субъекта, совершающего насилие. При этом будем помнить о заданной проблематике - отношения человека и робота (если посчитать робота равноправным субъектом во взаимоотношениях с человеком).

Итак, некий Субъект, который расценивает насилие как последнее средство в ряду приемлемых, совершает его в отношении другого Субъекта, именуемого в данном случае Объектом.

Моральным оправданием насилия Субъекта для себя самого в этом случае может служить следующее рассуждение:

1. Цель насилия – необходимость срочного пресечения действий Объекта, которые, по убеждению Субъекта, несут угрозу негативных последствий для общества и для самого Объекта, или которые несут угрозу положительным тенденциям для общества и для самого Объекта.
2. Необходимое условие допустимости насилия:
   Искреннее желание Субъекта не делать зла (сделать добро) Объекту и согласие уважать его права как личности **(**добро как бескорыстный отказ от своей пользы в пользу другого с целью передачи ему стремления также делать добро**)**.
   Обязательное доведение до сознания Объекта: целей пресечения, нежелания сделать ему зло (желания сделать ему добро) и подтверждения уважения его прав (п.п.1 и 2, дефис 1).
   Обязательные попытки выяснить мотивацию Объекта, а также то, присутствует ли в мотивации действий Объекта аналогичное моральное оправдание его собственных действий (п.п. 1 и 2, дефис 1).

3. Достаточное условие для обязательного прекращения насилия:
   Убеждение Субъекта в отсутствии угрозы в действиях Объекта
   Отсутствие убеждённости (неуверенность) Субъекта в наличии угрозы в действиях Объекта, особенно при выяснении, что у Субъекта имеются свои моральные оправдания его собственных (насильственных) действий.

Вышеприведённые рассуждения позволяют сформулировать три позиции новой парадигмы взаимоотношений человека и робота:

Презумпция *равноправия* и презумпция *невиновности* партнёров**.**

«Золотое правило нравственности»: «*Не поступай по отношению к другим так, как ты не хотел бы, чтобы они поступали по отношению к тебе*»**.**

Условие *нравственной допустимости насилия:* насилие одного субъекта над другим допустимо для него лишь при наличии *морального оправдания* перед собой в соответствии с определениями (п.п. 1, 2 и 3).

«Ортодоксально» уничижительное отношение человека к роботу, заложенное классиками жанра, возможно, было оправдано для соответствующего периода развития роботостроения. Но современный сознательный отход от него поможет сконцентрировать конструкторскую мысль создателей «разумных» машин в более адекватном направлении исследований.

Таковым направлением нам сегодня представляется разрешение возможных будущих проблем закладкой принципов паритетности отношений «человек-робот» на нравственных началах. Проблемы связаны с опасностью неуправляемого развития робототехники. Предлагаемое направление представляется активным и более действенным способом решения, в отношении с обсуждаемым пассивным (административным запретом и личным отказом от исследований и разработок).

Заключение

Всё вышеизложенное следует воспринимать как ещё одну попытку найти решение проблемы ожидаемой сингулярности в исследованиях и разработках «умных» машин.

Попытка представляет собой альтернативу по отношению к вариантам решений, ориентированным на личные и административные запреты на исследования и разработки. Она заключается в замене конструкторской парадигмы робототехники.

Сутью и смыслом замены парадигмы является принципиальный акцент на изначальное конструирование нравственных основ взаимодействия равноправных партнёров – человека и робота. Это представляется необходимым и достаточным условием, обязывающим «разумного» робота вырабатывать собственные внутренние этические ограничения.

Литература

1. Азимов А. // Я, Робот, 1950.
2. Джой Билл // Почему будущему мы не нужны. «Wired», Апрель 2000.
   http://www.wired.com/wired/archive/8.04/joy.htm
   http://chuma21.narod.ru/doc23/statyi/futdntneedus.html
3. Ильин И.А. // Путь к очевидности. Москва, Республика, 1993, 431 с.
4. Википедия // Золотое правило нравственности.
   http://ru.wikipedia.org/wiki/%D0%97%D0%BE%D0%BB%D0%BE%D1%82%D0%BE%D0%B5_%D0%BF%D1%80%D0%B0%D0%B2%D0%B8%D0%BB%D0%BE_%D0%BD%D1%80%D0%B0%D0%B2%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D1%81%D1%82%D0%B8