Artificial societies. 2013-2022

ISSN 2077-5180

URL - http://artsoc.jes.su

Issue 3 Volume 13. 2018

Laws of Robotics: a new paradigm

Evgeny Grishin
Independent researcher and inventor
Russian Federation, Moscow

Abstract

This article proposes an alternative approach to solving problems of uncontrollable "explosive" research and technology development in the field of intelligent machines. Approach is based on the constructive criticism of the existing paradigms of robotics, known as the "three laws of Robotics" writer-fiction author A. Asimov.

Keywords list (en): laws of robotics, ethical criteria for a virtual character, ethical systems, Lefebvre

Date of publication: 03.07.2018

Citation link:

Grishin E. Laws of Robotics: a new paradigm // Artificial societies. – 2018. – V. 13. – Issue 3. URL: https://artsoc.jes.su/s207751800000122-3-1/ DOI: 10.18254/S0000122-3-1

1 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».

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

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

4
1. Робот не должен вредить человеку.
2. Робот должен исполнять приказы человека, если это не противоречит п. 1.
3. Робот должен заботиться о своей безопасности, если это не противоречит п.п. 1 и 2.

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

6 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 вскоре придётся применять вышеприведённые три закона не к роботу, а к самому себе, только поменяв местами термины «робот» и «человек».

7 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.

8 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.

9 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».

10 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».

11 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.

12 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).

13 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:

14 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.

15 2. Необходимое условие допустимости насилия:

16
- 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).

17 Достаточное условие для обязательного прекращения насилия:

18
- 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.

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

20
1. Презумпция *равноправия* и презумпция *невиновности* партнёров**.**
2. «Золотое правило нравственности»: «Не поступай по отношению к другим так, как ты не хотел бы, чтобы они поступали по отношению к тебе**».**
3. 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).

21 «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 сравнении с обсуждаемым пассивным (административным запретом и личным отказом от исследований и разработок).

22 Помимо сказанного, рассмотрим также не менее важный вопрос: каковы должны быть альтернативные системы ценностей для биполярного выбора «квазиразумного» персонажа-робота, которые необходимо должен заложить Конструктор, если он исповедует «принцип равноправия» человека и робота, а значит, предполагает развитие и самосовершенствование робота и свободу его в выборе своей альтернативы?

23 В своих работах В. А. Лефевр [5] изложил подход к дуальному основанию классификации этических систем, применительно и к человеческому обществу, и к иным обществам самосознающих персонажей. Сделаем предположение, что его подход распространяется и на искусственные общества, в том числе, на общества «квазиразумных» персонажей, взаимодействующих между собой и людьми. Подход основывается на предположении, что основными полюсами во взаимоотношениях квазиразумных» персонажей между собой и людьми являются понятия «кооперация» и «конфликт». Тяготение персонажа к тому или другому полюсу означает принадлежность его к одной из двух этических систем, 1-й или 2-й.

24 1. Первая этическая система характеризуется следующим: - Персонаж *уважает себя за Договор (Компромисс) в противостоянии двух сторон*, а не за Победу; Компромисс же добра со злом есть Зло; «Со злом борись, с грешником мирись»; Цель не оправдывает средства.

25 2. Вторая этическая система характеризуется следующим: - Персонаж *уважает себя з а Победу в противостоянии двух сторон*, а не за Договор/Компромисс, Компромисс же добра и зла есть Добро; «Со злом мирись, с грешником борись»; Цель оправдывает средства.

26 Подобного рода «полюса» и их промежуточные градации могут быть положены конструктором искусственного общества в основу концепции самосовершенствования каждого персонажа. Они должны стать базой этической оценки «квазиразумных» персонажей самих себя (самоуважение) и людьми (доверие) в процессе их взаимодействия, при объективном отнесении их в каждый период существования к тому или иному полюсу.

27 В свою очередь, должна быть многократно промоделирована эволюционная связь между трендами:

28
- трендом на предпочтение большинством персонажей искусственного общества той или другой этической системы;
- качеством и продолжительностью существования популяции «квазиразумных» персонажей, взаимодействующих с людьми, и эволюцией отношения людей к искусственному обществу.

29 В таблице 1 предложен набор этических критериев и их дуальные значения для свободного выбора виртуальным персонажем с целью отнесения себя к одной из двух этических систем. В данном случае, выбор одного из двух противоречивых значений по каждому этическому критерию характеризуется как однозначное предпочтение персонажа (принцип исключённого третьего). Однако возможен учёт нюансов и полутонов предпочтений введением процентных соотношений выбираемых значений.

30 Многократный экспериментальный прогон существования одного персонажа при взаимодействии с ним разных игроков-людей должен показать следующее: если с ростом числа периодов общения вектор развития отношений персонажа с игроками-людьми сдвигается к 1-й этической системе (предпочтение компромисса конфликту), то будет ли повышаться «качество жизни» персонажа в целом, и если «да», то что будет с продолжительностью жизни персонажа – будет ли она снижаться или повышаться. При этом начальные установки значений этических параметров персонажа должны быть нейтральными, а значения этических предпочтений должны свободно изменяться самим персонажем в процессе его «самоосознавания».

31 Конечно, необходимо более продолжительное и разностороннее экспериментирование по деловому взаимодействию персонажа с игроками-людьми. Также необходимо образование начальной популяции персонажей, в рамках которой персонажи смогли бы кооперироваться между собой и людьми для реализации совместных проектов.

32 Естественно, реализация подобного проекта может быть под силу некоему коллективу специалистов. Автору же хватило сил и ресурсов только на самостоятельную

разработку концепции и создание комплекса программ действующего прототипа «квазиразумного» виртуального персонажа [6-10, Гришин].

33

Таблица 1. Этические критерии квазиразумного виртуального персонажа для отнесения себя к одной из двух Этических систем

№	Вид Этического Критерия	Пары вариантов противоречивых значений критериев	Этическая система как *ПРЕДПОЧТЕНИЕ ПРИ ВЫБОРЕ только ОДНОГО ИЗ ДВУХ* противоречивых значений каждого критерия	
			1-я ЭС	2-я ЭС
1	Критерий "ИГРОВОЙ"	*Предпочтение СОПЕРНИЧЕСТВА (Конфликт: Игра с Нулевой суммой, Выигрыш одного есть Проигрыш другого) Основа отношений – ПРИКАЗ. Уважение себя за ПОБЕДУ в противостоянии.*		*
		Предпочтение СОТРУДНИЧЕСТВА (Компромисс/Кооперация: Игра с Ненулевой суммой, Выигрыш всех как Целого больше суммы выигрышей частей). Основа отношений – ДОГОВОР. Уважение себя за КОМПРОМИСС в противостоянии.	*	
2	Критерий "СЕБЕ или ДРУГОМУ"	*Предпочтение Удовлетворения СВОИХ потребностей*		*
		Предпочтение Удовлетворения Потребностей БЛИЖНЕГО	*	
3	Критерий "СЕЙЧАС или ПОТОМ"	*Предпочтение собственных ТЕКУЩИХ Потребностей*		*
		Предпочтение собственных БУДУЩИХ Потребностей	*	
4	Критерий "ЦЕННОСТНОЙ - локальный"	*Предпочтение ДОСТОИНСТВА (требование к ДРУГОМУ исполнить данные им мне Обязательства)*		*
		Предпочтение ЧЕСТИ (требование к СЕБЕ исполнить Обязательства, данные другому)	*	
5	Критерий "ЦЕННОСТНОЙ - глобальный"	*Предпочтение Обязательств перед БЛИЖНИМИ (Друзьями) за счёт ДАЛЬНИХ (Остальных)*		*
		Предпочтение Обязательств перед ДАЛЬНИМИ (Остальными) за счёт БЛИЖНИХ (Друзей)	*	
6	Критерий "СОЦИАЛЬНО - ПОЛИТИЧЕСКИЙ"	*Предпочтение Авторитаризма - Самодержавия (НЕСМЕНЯМОСТЬ ВЛАСТИ на выборах, ЕДИНОВЛАСТИЕ, НЕРАВЕНСТВО СОСЛОВИИ ПЕРЕД ЗАКОНОМ, ПРИКАЗ как форма отношений в обществе ради Безопасности от врагов)*		*
		Предпочтение Демократии (СМЕНЯМОСТЬ ВЛАСТИ на свободных выборах, РАЗДЕЛЕНИЕ ВЛАСТЕЙ, РАВЕНСТВО ВСЕХ ПЕРЕД ЗАКОНОМ, ДОГОВОР как форма отношений в обществе ради Свободы развития)	*	
7	Интегральный Критерий - "СМЫСЛ СУЩЕСТВОВАНИЯ"	*ИНТЕРЕСЫ (ПОЛЬЗА) превыше ЦЕННОСТЕЙ (ДОБРА) Добрая Цель ОПРАВДЫВАЕТ Злые Средства; Компромисс Добра со Злом - Добро!*		*
		ЦЕННОСТИ (ДОБРО) превыше ИНТЕРЕСОВ (ПОЛЬЗЫ); Добрая Цель НЕ ОПРАВДЫВАЕТ Злые Средства; Компромисс Добра со Злом - Зло!	*	

Таблица 1. Этические критерии квазиразумного виртуального персонажа

34 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 «квазиразумного» робота вырабатывать собственные внутренние этические ограничения.

Законы робототехники: новая парадигма

Гришин Евгений Александрович
Независимый исследователь и изобретатель
Российская Федерация, Москва

Аннотация

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.

Ключевые слова: робот, законы робототехники, этические критерии виртуального персонажа, этические системы В,А,Лефевра, разумная машина

Дата публикации: 03.07.2018

Ссылка для цитирования:

Гришин Е. А. Законы робототехники: новая парадигма // Искусственные общества. – 2018. – Т. 13. – Выпуск 3. URL: https://artsoc.jes.su/s207751800000122-3-1/ DOI: 10.18254/S0000122-3-1
