УДК 004.89

А.В. Гаврилов, к.т.н., доцент, gavrilov@corp.nstu.ru
Новосибирский государственный технический университет, г. Новосибирск, Россия

К ВОПРОСУ ОБ ЭТИКЕ ИНТЕЛЛЕКТУАЛЬНЫХ РОБОТОВ

Аннотация: В статье рассматривается проблема сосуществования будущего искусственного разума с людьми. Предлагается вместо трудно реализумых законов робототехники А.Азимова, предназначенных для непосредственного управления поведением робота, использовать их для формирования эмоций, управляющих поведением опосредованно. Дружественное поведение робота в этом случае обеспечивается процессом его обучения, так же как моральное миролюбивое поведение человека.

Ключевые слова: искусственный интеллект, гуманоидные роботы, машинное обучение, эмоции

TOWARDS ETHICS OF INTELLIGENT ROBOTS

A.V. Gavrilov, Ph.D., docent, gavrilov@corp.nstu.ru
Novosibirsk State Technical University, Novosibirsk, Russia

Abstract: A problem of coexixsting of future artificial mind with people is discussed in this paper. Instead of difficulty implemented Laws of A.Asimov providing direct control of robot's behaviour author suggests to use them for generation of emotions controlling indirectly. In this case friendly behaviour of robot is providing by corresponding learning similar to learning of moral friendly behaviour of human beings.

Key words: Artificial Intelligence, Humanoid Robots, Mchine Learning, Emotions

В 2008г. автор данной статьи писал [1], что можно выделить следующие тенденции в развитии искусственного интеллекта:

1. Проникновение ИИ во все аспекты информатизации жизни – от разработки программного обеспечения до приготовления пищи и уборки на кухне. Проявлением этой тенденции является появление и разработка таких парадигм как "ubiquitous computing", "pervasive computing" и "smart environment",

2. Появление парадигмы «настоящего искусственного интеллекта» или универсального ИИ (Artificial General Intelligence) [3], способного обучаться и решать любые задачи (а не определенный тип задач) подобно тому, как это делает человек,
3. Создание искусственных интеллектуальных систем (в частности, интеллектуальных роботов) с человекоподобным поведением, характеризующимся моделированием коллективного поведения, эмоций, мотивации, самообучения,
4. Появление реальных попыток создания самовоспроизводящихся роботов,
5. Появление интереса к этическим вопросам, связанным с появлением такого человекоподобного искусственного интеллекта.

В настоящее время эти тенденции сохраняются, часть из них получило существенное развитие, особенно, первые три. Примерами проникновения ИИ во все области деятельности человека являются такие разработки как Watson [3], Siri [4], «умные дома» [5], программа AlphaGo для игры в Го [6] и т.д. Разработка человекоподобного искусственного интеллекта в последнее время интенсивно развивается: по проблеме AGI (Artificial General ntelligence) ежегодно проводится одноименная международная конференция [7]. Интеллектуальные роботы получают все бОльшее распространение в военной области [8], в сфере обслуживания [9], в индустрии развлечений, внедряются в производство в качестве промышленных роботов [10]. И особенно бурно развивается направление создания гуманоидных роботов с человекоподобным поведением и обучением. Системы управления таких роботов кАк правило используют технологии искусственных нейронных сетей.

Успехи в этих направлениях подогреваются относительно новыми направлениями в области искусственных нейронных сетей: глубокими нейронными сетями (deep neural networks) [11] и нейроморфными технологиями (neuromorphic technology) [12]. Эти два направления в их комбинации позволяют строить компактные аппаратные малоэнергоемкие нейронные сети, по сложности и объему сопоставимые с мозгом млекопитающих.

Остались две основные нерешенные проблемы: обеспечить аппаратные (нейроморфные) нейронные сети способностью к обучению в реальном времени в процессе функционирования системы (сейчас их обучение в основном производится на хост-компьютере предварительно) и наполнить их структурой и функциональностью, подобными существующим в реальном биологическом мозге. Можно ожидать, что в ближайшие десятилетия эти проблемы будут решены.

Многие эксперты говорят о том, что в ближайшие десятилетия люди будут жить среди роботов с человекоподобным искусственным интеллектом (в том числе, гуманоидных). В связи с этим, многие ученые, имеющие дело с искусственным интеллектом, начинают задумываться об опасностях, связанных с появлением человекоподобного ИИ [13, 14, 15, 16], хотя есть и другая

оптимистическая точка зрения, основанная на идее наступления технологической сингулярности [17, 18], преобразующнй человека в некое другое существо. Однако, в этом случае проблема сосуществования человека (или уже киборга) с искусственным разумом не снимается. Проблема заключается в том, что человекоподобный искусственный интеллект, с одной стороны, будет обладать мотивациями и целеполаганием (являющимися следствием требований автономности и человекоподобного поведения/общения), а с другой стороны, будет иметь отличную от нашей биологической природу и обладать более быстрым и мощным интеллектом. Так что, мы не сможем прогнозировать ход его мышления и развития и то, как он будет относиться к людям. Таким образом, при разработке человекоподобного интеллекта необходимо обеспечить его дружественность по отношению к людям. Решение этой проблемы можно разбить на решение двух задач:

- разработка этических правил (или морали) для роботов и
- обеспечение надежного исполнения этих правил роботами.

Решением первой из этих задач, как известно, занимался в своих фантастических произведениях известный писатель-фантаст и популяризатор науки Айзек Азимов. Он в 1940-х годах [19] предложил так называемые три закона робототехники и в своем цикле рассказов о роботах рассматривал различные ситуации, в которых эти три закона оказывается затруднительно исполнить. Эти три закона выглядят следующим обрaзом:

1. Робот не может причинить вреда человеку или своим бездействием допустить, чтобы человеку был причинён вред.
2. Робот должен выполнять приказы человека в той мере, в которой это не противоречит Первому Закону.
3. Робот должен заботиться о своей безопасности в той мере, в которой это не противоречит Первому и Второму Законам.

Проблема, которая возникает при реализации этих правил поведения, заключается в том, что невозможно точно определить, что такое вред, а что такое польза для всех случаев жизни и для всех людей. Например, сервисный робот может считать, что применение антибиотиков или обезболивающих средств приносит вред, и не давать их человеку, хотя в данном конкретном случае эти средства могут быть единственно полезными. Или как обеспечить защиту человека от другого человека? При соблюдеии законов Азимова это практически невозможно. Кстати, это касается и применения человеческой морали в условиях конфликтных ситуаций, например, в условиях военных действий.

В [1, 20] автором было предложено формулировать три закона робототехники в виде правил-продукций, использующих лингвистические переменные, которые робот учится распознавать на самом верхнем уровне подсистемы восприятия окружающей среды:

1) Danger_for_man – опасность для человека,

2) Danger_for_myself – опасность для себя (робота),

3) *Command_executed* – успешное выполнение текущей команды (процесс достижения текущей цели).

Правила выглядят так:

If *Danger_for_man* then *Negative_emotion*;

If *Danger_for_myself* and not *Danger_for_man* then *Negative_emotion*;

If not *Command_executed* and not *Danger_for_man* and not *Danger_for_myself* then *Negative_emotion*;

If *Command_executed* and not *Danger_for_man* and not *Danger_for_myself* Then *Positive_emotion*;

Эти правила-продукции генерируют одну из двух базовых эмоций - положительную или отрицательную, которые влияют на выбор действий и процесс обучения робота, являясь поощрением и наказанием для подсистем распознавания и выработки решений. При этом в процессе обучения робот учится избегать условий, при которых вырабатывается отрицательная эмоция, и стремится к таким, когда вырабатывается положительная.

Таким образом, вместо попытки непосредственного управления поведением (как в законах Азимова) предлагается опосредованное управление поведением путем поощрения/наказания посредством положительных и отрицательных эмоций (рис. 1).

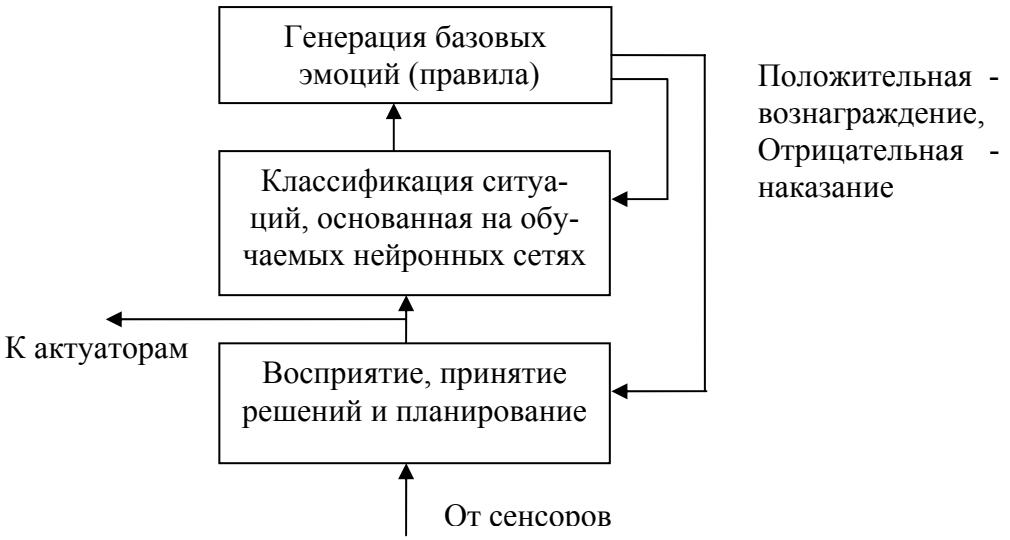

Рис. 1. Схема использования базовых эмоций для обучения

Подходы к реализации такого обучения в рамках создания самообучающейся и самомодифицирующейся нейронной сети как модели мозга предложены в [21].

Процесс обучения такого искусственного интеллекта разбивается на два этапа:

- обучение тому, «что такое хорошо и что такое плохо», т.е. предварительная тренировка механизма выработки положительных и отрицательных эмоций,

- обучение всему остальному, т.е. решению задач, связанных с назначением робота (манипулирование объектами, движение и более сложное поведение), которое может осуществляться уже в процессе использования робота.

Следует иметь в виду, что на втором этапе в процессе использования робота продолжается дообучение механизма выработки эмоций, как побочный эффект основного обучения.

Естественно, что так же как человека можно натренировать на аморальное или преступное поведение, так и робота, управляемого эмоциями. Поэтому, единственным способом предупреждения противостояния между человеком и его творением, превосходящим его по возможностям, является улучшение и самого человека-творца, т.е. отказ от силовых методов решения социальных и прочих проблем. Есть и еще более утопический выход – отказ от создания человекоподобного искусственного интеллекта.

Список литературы

1. Гаврилов А.В. Эмоции, априорные знания и дружественное поведение робота. - Труды 11-ой национальной конференции по искусственному интеллекту с международным участием КИИ-2008 (г.Дубна, Россия). –М.: ЛЕНАНД, 2008. –Т.1. –С.410-419.
2. Ben Goertzel, Cassio Pennachin. Artificial General Intelligence. Springer-Verlag, Berlin, Heidelberg, 2007.
3. The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works. [Электронный ресурс]: режим доступа - http://www.redbooks.ibm.com/redpapers/pdfs/redp4955.pdf
4. Tom Gruber. Siri: A Virtual Personal Assistant. Keynote presentation at Semantic Technologies conference (SemTech09), June 16, 2009. [Электронный ресурс] : режим доступа - http://tomgruber.org/writing/Siri-SemTech09.pdf
5. А.В.Гаврилов. Искусственный Домовой. - Искусственный интеллект и принятие решений, №2, 2012.- С.77-89.
6. AlphaGo - [Электронный ресурс] : режим доступа - https://ru.wikipedia.org/wiki/AlphaGo
7. Conference Series on Artificial General Intelligence [Электронный ресурс] : режим доступа - http://agi-conference.org/
8. Хрипунов С.П., Чиров Д.С., Благодарящев И.В. Военная робототехника: современные тренды и векторы развития. – М.: ООО «НБ-Медиа», 4, 2015. – С. 410-422.
9. Аналитическое исследование: мировой рынок робототехники. – НАУРР. 2016. [Электронный ресурс]: режим доступа -

http://robotforum.ru/assets/files/000_News/NAURR-Analiticheskoe-issledovanie-mirovogo-rinka-robototehniki-(yanvar-2016).pdf

10. Smart, Collaborative Robot for Precision Tasks. [Электронный ресурс] : режим доступа - https://www.rethinkrobotics.com/wp-content/uploads/2015/11/Sawyer_Datasheet_Nov_2015_web.pdf

11. Гаврилов А.В. Deep learning (глубокое или глубинное обучение). – Материалы Всероссийского научно-прктического семинара и школы молодых ученых «Перспективные методы и средства интеллектуальных систем (ПМСИС-2015)», Новосибирск, 2015. - С. 50-59. [Электронный ресурс] : режим доступа - http://www.insycom.ru/html/metodmat/dp.pdf

12. Гаврилов А.В., Канглер В.М. Нейроморфные технологии: состояние и перспективы развития. – Материалы VII Всероссийской научно-технической конференции «Робототехника и искусственный интеллект» РИИ-2015, Железногорск, СФУ, 2015. – С. 148-154.

13. Yudkowsky E**.** Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. San Francisco, CA: The Singularity Institute, 2001.

14. Hugo de Garis. The Artilect War: Cosmists Vs. Terrans: A Bitter Controversy Concerning Whether Humanity Should Build Godlike Massively Intelligent Machines. Palm Springs, CA, 2005.

15. Баррат Дж. Последнее изобретение человечества: Искусственный интеллект и конец эры Homo Sapience. –М: АНФ. 2015.

16. Гаврилов А.В. Искусственный интеллект и будущее цивилизации // Современные научные исследования и инновации. 2015. № 5. [Электронный ресурс] : режим доступа - http://web.snauka.ru/issues/2015/05/50092

17. Kurzweil R. The Singularity Is Near. N. Y.: Viking, 2005.

18. Российское трансгуманистическое движение. [Электронный ресурс] : режим доступа - http://transhumanism-russia.ru/

19. Азимов А. Мечты роботов. М.: Эксмо, 2004.

20. Andrey V.Gavrilov. Emotions and a priori Knowledge Representation in Artificial General Intelligence. In Proc. of Int. Conf. on Intelligent Information and Engineering Systems INFOS-2008. Varna, Bulgaria, June 23-July 03, 2008; in book: “Intelligent Technologies and Applications” of Int. Book Series “Information Science and Computing”, ITHEA, Bulgaria. - Pp. 106-110.

21. А.А.Малявко, А.В.Гаврилов. К вопросу о создании самообучающейся и самомодифицирующейся импульсной нейронной сети в качестве модели мозга. // Труды XIII Международной научно-технической конференции «Актульные проблемы электронного приборостроения» АПЭП-2016. – Новосибирск, НГТУ, Том 9. - С. 66-69.