

# Этические проблемы искусственного интеллекта

*исследовательский семинар*

*Российской ассоциации искусственного интеллекта,  
кафедры этики философского факультета*

*МГУ имени М. В. Ломоносова и  
Института философии Российской академии наук*

**Заседание в Институте философии РАН  
16 ноября 2022 года**

# СВЯЗЬ РОБОЭТИКИ С ТЕОРИЕЙ ИНТЕЛЛЕКТА

Алексеева Ирина Юрьевна, доктор философских наук,  
ведущий научный сотрудник сектора философских проблем  
социальных и гуманитарных наук

Института философии РАН

**«...роботизация, автоматизация и развитие всех форм искусственного интеллекта... должны стать центральной проблемой философского дискурса современности»**

**Анатолий Ракитов, 2019**

(Ракитов А. И. Философия, роботы, автоматы и зримое будущее // Философия и общество, № 3. С. 47)

# ***Gianmarco Veruggio***

**The Birth of Roboethics**

*Scuola di Robotica*

*C.P. 4124, P.za Monastero, 4 - 16149 Genova Italy*

*First International Symposium on*

***Roboethics***

*The ethics, social, humanitarian and ecological aspects of  
Robotics*

*30<sup>th</sup> - 31<sup>rd</sup> January 2004, Villa Nobel, Sanremo, Italy*

# *Roboethics*



# Робоэтика=этика робототехники

1. Вопросы профессионального поведения людей, участвующих в создании робототехники и ее эксплуатации (поведение и профессиональное самосознание ученого, инженера, менеджера...)
2. Вопросы интеграции роботов (прежде всего, автономных) в человеческое общество, а также взаимодействия роботов между собой (этический статус автономного робота с искусственным интеллектом, «моральная машина» и «машинная мораль», сообщество автономных роботов...)

Возможное название для группы 2

«Этиконика»? «Этоника»?

В 60-е годы исследования Поспелова и Пушкина по искусственному интеллекту называли «психоникой» (по аналогии с бионикой).

Семинар по психонике работал в Московском энергетическом Институте

Осадчий Петр Семенович. К вопросу о принципах профессиональной этики инженера. СПб, тип. А. Бенке, 1911.

«Кодекс профессиональной этики электротехников». V Всероссийский электротехнический съезд. Москва, 1909. VI (электротехническое) отделение Императорского русского технического общества. Журнал «Электричество» за 1909 г. (№ 5 и № 4).

# Карел Чапек. РУР

РУР – «Россумские универсальные роботы»

Пьеса написана в 1920 году, премьера состоялась в 2021 году.

Одно из действующих лиц - **ДОКТОР ГАЛЛЕМАЙЕР** -- руководитель Института психологии и воспитания роботов.

- «зачем делают роботов! Для работы».
- «живые, наделенные интеллектом рабочие машины»
- «Роботы – не люди... Механически они совершеннее нас, они обладают невероятно сильным интеллектом, но у них нет души».
- «Лига гуманности... Ее цель... защищать роботов и обеспечить им... хорошее отношение»
- «Неплохая цель. С машинами надо обращаться хорошо. Ей-богу, это мне нравится. Я не люблю испорченных вещей»

# РУР

«...когда рабочие восстали против роботов и начали разбивать их, и когда люди дали роботам оружие против восставших, и роботы истребили столько людей... И потом, когда правительства превратили роботов в солдат и было столько войн»

## Айзек Азимов. Робби

1940. «Странная нянька» (журнал *Super Science Stories* ,  
рассказ «*Strange Playfellow*»).

Первый закон робототехники: робот не может  
причинить вреда человеку

# Законы робототехники Айзека Азимова

1. Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинён вред.
2. Робот должен повиноваться всем приказам, которые даёт человек, кроме тех случаев, когда эти приказы противоречат Первому Закону.
3. Робот должен заботиться о своей безопасности в той мере, в которой это не противоречит Первому или Второму Законам.

# Проекция «законов робототехники» на область профессиональной этики инженера

1. Создание техники, безопасной для человека.
2. Создание техники, служащей целям человека при условии обеспечения его безопасности.
3. Создание техники долговечной, соответствующей современным потребностям и безопасной.

# Влада Леушина

*Этика искусственного интеллекта в стандартах и рекомендациях.*

Доклад на семинаре 9 февраля 2022 года

«В докладе рассматриваются разрабатываемые и опубликованные нормативные документы международных и региональных организаций, в которых приводится схожее абстрактное описание этических принципов, которым должен соответствовать ИИ».

# «Кодекс этики искусственного интеллекта», 26.10.2021

[https://a-ai.ru/wp-content/uploads/2021/10/Кодекс\\_этики\\_в\\_сфере\\_ИИ\\_финальный.pdf](https://a-ai.ru/wp-content/uploads/2021/10/Кодекс_этики_в_сфере_ИИ_финальный.pdf)

«Акторы ИИ должны принимать необходимые меры, направленные на сохранение автономии и свободы воли человека в принятии им решений, права выбора и в целом сохранения *интеллектуальных способностей человека как самостоятельной ценности* и системообразующего фактора современной цивилизации».

В. Э. КАРПОВ, П. М. ГОТОВЦЕВ, Г. В. РОЙЗЕНЗОН  
К ВОПРОСУ ОБ ЭТИКЕ И СИСТЕМАХ  
ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

*Философия и общество. 2018. № 2. С. 86*

«...мы имеем дело с искусственной системой, реализующей процессы планирования, целеполагания, выбора и осуществления того или иного поведения. ...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 прочего».

# *Карпов Валерий Эдуардович.* **О моральных аспектах адаптивного поведения искусственных агентов**

Заседание семинара «Этические проблемы искусственного интеллекта» 6 марта 2021 г.

Все описанные выше механизмы вводились исключительно и только из соображений технической целесообразности, для решения задач создания эффективных адаптивных механизмов трех уровней:

- (1) Целесообразное поведения технического устройства в сложной, недетерминированной, динамической среде.
- (2) Организация взаимодействия внутри группы агентов. Вплоть до появления форм социальной организации.
- (3) Задача целенаправленного управления поведения социумом.

Решение и действие

«Действие, проникающее в действительность»

Сергей Рубинштейн. «Основы общей психологии»

Теория интеллекта

*«...на фоне рубинштейновской теории, содержащей целостную картину сущностных свойств интеллекта, едва ли не все прочие концепции интеллекта (включая те, что учитывают интеллект искусственный) выглядят как ограничившие внимание лишь отдельными сторонами и моментами интеллектуальной способности и интеллектуального процесса».*

Алексеев А. П., Алексеева И. Ю. Судьба интеллекта и миссия разума: философия перед вызовами эпохи цифровизации». 2021.

В теории Рубинштейна *интеллект* предстает как *сложная способность*, «ядром» которой является способность субъекта выделять в наличной ситуации свойства, связи и отношения, **существенные для действия, соответствующего сложившимся условиям.**

Формирование интеллекта в филогенезе рассматривается Рубинштейном как обусловленное необходимостью регулировать поведение в ситуации, когда недостаточно ни инстинкта, ни навыка, т. е. в *новой, нестандартной ситуации.*

McCarthy J., Minsky M., Rochester N. Shannon C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

URL: <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>

1955 год, заявка на грант для исследований проблемы искусственного интеллекта

*В основе планируемых исследований лежит предположение, что **все характеристики интеллекта, включая способность к обучению, могут быть описаны настолько точно, что смогут воспроизводиться машиной.***

Один из важнейших проблем ИИ: Точная формулировка идеи **мышления как манипулирования словами в соответствии с определенными правилами** и создание на этой основе примеров машинного обобщения.

# В робоэтике

Нестандартная ситуация в среде и системе

Робот-Человек

или

Робот-Робот

Как регулировать отношения и действия?

Возможности естественного интеллекта человека

и искусственного интеллекта робота

Применение интеллекта - *интеллектуальный* (мыслительный) *процесс*, имеющий следующие характерные черты.

- *Осознанность*, проявляющаяся, в частности, в сопоставлении возникающих мыслей с условиями задачи, в проверке мыслей, контроле мыслительного процесса.
- Выполнение *интеллектуальных операций – анализа, синтеза, обобщения*, позволяющих осуществлять переработку данных ощущений и восприятий, расчленять синкретическое единство восприятия, различать случайные совпадения и необходимые связи, переходить от единичного к общему, образовывать понятия.
- *Направленность на решение задачи.*
- *Связь с практикой*, непосредственная или опосредованная. Изменяя действительность, человек познает ее, и познавая действительность – изменяет.
- *Неразрывная связь с эмоциональной сферой* и волевыми усилиями человека.
- *Возможность осознания ошибки.*

Неразрывная связь интеллектуального процесса с эмоциональной сферой и волевыми усилиями человека.

*Рубинштейн: «Мысль, заостренная чувством» и «Мысль, подчиненная деспотическому господству чувства»*

**Искусственные эмоции робота:**

**Как избежать «деспотического господства» чувства над мыслью?**

(РУР: Наделение робота способностью страдать ведет к бунту машин).

## ***Возможность осознания ошибки***

С. Л. Рубинштейн: «...конечно, не сама по себе ошибка, а возможность осознать ошибку является привилегией мысли как сознательного процесса»

Причины ошибок:

- неадекватное представление задачи
- эмоциональные срывы
- автоматизм навыков
- восприятие опосредованных связей как непосредственных
- частичный учет условий
- и другие

## А. В. Разин: Этика ИИ включает этику технических систем будущего.

«Этика непосредственно начинается тогда, когда появляется **способность реагировать на собственные ошибки**, осуществлять рефлексию поведения, учитывая при этом мнения других людей. Такая же принципиальная возможность ошибки должна быть заложена и в работу искусственного интеллекта, чтобы можно было говорить о его этике в собственном смысле слова».

Разин А. В. Этика искусственного интеллекта // Философия и общество, № 1 2019 57–73. С. 57.

Рубинштейн: Интеллектуальная способность позволяет выделять в *наличной ситуации* то, что существенно для действия, адекватного *сложившимся условиям*

Конкретизация «ситуации» и «условий»:

- Физические обстоятельства
- Знания и умения
- Потребности, желания, цели, ценности
- Нормы юридические
- Нормы нравственные
- Другое

## «Проблема рамки»

«...как робот будет выбирать релевантную информацию для принятия морального решения? На основе чего он будет отделять релевантную информацию от нерелевантной? Сможет ли робот использовать весь объем необходимой информации, особенно в новых ситуациях, когда речь идет о терабайтах нерелевантных данных?»

Середкина Е.В. Этические аспекты социальной робототехники // Человек. 2020. Т. 31, № 4. С. 121.

Виктор Финн:

Искусственный интеллект (ИИ) как направление исследований занимается аппроксимацией естественного интеллекта (ЕИ), точнее, совокупности способностей, образующих реальный ЕИ.

Аппроксимация (приближение) как метод предполагает замену сложных объектов более простыми.

# Способности ЕИ, по Финну:

- Способность *выделять существенное в наличной ситуации*, упорядочивать наличные данные и знания (необходимый аспект интуиции)
- *Целеполагание*, планирование поведения, порождение последовательности «цель — план — действие».
- Способность к *аргументированному принятию решений* с использованием знаний и рассуждений, соответствующих поставленной цели.
- Способность к *рефлексии, оценке знаний и действий*.
- Способность *выполнять интеллектуальные операции*, включая способность извлекать следствия из имеющихся фактов и знаний, т. е. способности к *рассуждению*, которое может содержать как правдоподобные выводы, используемые для выдвижения гипотез, так и достоверные выводы.
- Способность к *рационализации идей*, стремлению и умению уточнять их как понятия.
- Способность к *синтезу познавательных процедур, фальсификации выдвигаемых гипотез*.
- Способность к *адаптации при изменении жизненных ситуаций*, коррекции взглядов («теории») и поведения, соответствующая способности *осознавать ошибки*.

Рубен Апресян. Теоретические и методологические предпосылки  
обсуждения проблем этики искусственного интеллекта

Резюме выступления на семинаре 1 октября 2019 года

«В дискуссиях по прикладным этическим проблемам, в особенности, когда они оказываются предметом широкого общественного внимания, **нередко встречается нестрогое, а то и просто неправильное употребление этических категорий.** Они используются по инерции обычной и публицистической речи, **без учета традиции из обсуждения в моральной философии и нормативной этике,** сложившихся по их поводу морально-философских и нормативно-этических концепций».

# Проблема сочетаемости традиций теории познания и этики. **Субъектность и агентность**

- Елена Гаврилина. *Агентность интеллектуальных автономных систем в этическом ракурсе: эмпирические и теоретические аспекты. Доклад на семинаре 7 мая 2021*
- Владислав Лекторский. *Субъект, объект, познание*. М., 1980. (Введено понятие предмета-посредника в познании, ЭВМ – один из видов).
- Ирина Алексеева. «Компьютер как квазисубъект знания» – глава в книге «Человеческое знание и его компьютерный образ» (М., 1993).

# Марвин Минский. Общество разума (Society of Mind)

«“Обществом разума” я буду впредь именовать такую схему, в которой каждое сознание представляется состоящим из множества мелких процессов. Указанные процессы мы будем называть *агентами*. Каждый ментальный агент по отдельности выполняет некое простое действие, для чего не требуется ни разум, ни мышление вообще. Тем не менее, когда мы объединяем указанных агентов в сообщества – посредством весьма своеобразных способов, – это ведет к возникновению подлинного интеллекта».

*Минский М.* Сообщество разума / пер. с англ. В. Желнинова. М., 2018. С. 4.

# Марвин Минский. «Машина эмоций» (“The Emotion Machine”)

«В этой книге термин “ресурс” используется для описания того, что в моей предыдущей книге The Society of Mind называлось “агентом”. Я внес это изменение, потому **что слишком многие читатели предполагали, что “агент” – это нечто персонифицированное (как, например, туристический агент), нечто способное действовать независимо, и что ментальные агенты могут сотрудничать почти так же, как это делают люди.**

*Минский М. Машина эмоций / пер. с англ. В. Дегтяревой, А. Курышевой. М., 2020. С. 8–9.*

## *Роботы на Луне*

- Набор роботов (предметы-посредники в деятельности человека)
- Группа роботов (агенты)
- Сообщество роботов (агенты, квазисубъекты, субъекты?)

# Норберт Винер в редакции журнала «Вопросы философии», 1960 год

В “думающих” машинах мы имеем своего рода “механических рабов”. Здесь встает проблема рабства, но без жестокости. Положение оказывается именно проблематичным, ибо мы не можем сочетать две разные вещи, которые требуем от машины: с одной стороны, мы хотим, чтобы “раб” был умным, а с другой стороны – чтобы он был послушным. Эти два требования противоречат друг другу. ... Если мы требуем “ума” от машины, то от самих себя мы должны потребовать еще больше ума»

(Норберт Винер в редакции нашего журнала // Вопросы философии. 1960. № 9. С. 164).