

**ИИ:**

НЕОБЪЯСНИМЫЙ, НЕПРЕДСКАЗУЕМЫЙ,  
НЕКОНТРОЛИРУЕМЫЙ

ROMAN V. YAMPOLSKIY,PHD

**AI**

UNEXPLAINABLE, UNPREDICTABLE,  
UNCONTROLLABLE

CRC Press  
2024

РОМАН ЯМПОЛЬСКИЙ

**ИИ:**

НЕОБЪЯСНИМЫЙ, НЕПРЕДСКАЗУЕМЫЙ,  
НЕКОНТРОЛИРУЕМЫЙ

Перевод с английского

**NOVA**  
CREATIVE GROUP

Москва  
2025

УДК 004.8  
ББК 16.6  
Я57

Перевод с английского *Камилла Ахметова*

**Ямпольский Р.**

Я57 ИИ : Необъяснимый, непредсказуемый, неконтролируемый / Роман Ямпольский. — М.: Nova Creative Group, 2025. — 358 с.

ISBN 978-5-908058-00-1

Книга посвящена фундаментальным ограничениям, с которыми сталкивается человечество в попытках понять, контролировать и безопасно использовать искусственный интеллект (ИИ), особенно его продвинутую форму — сильный ИИ. Автор подробно исследует три ключевых свойства ИИ — необъяснимость, непредсказуемость и неконтролируемость, — показывая, что это не просто его характеристики, а потенциально непреодолимые барьеры для общества.

Автор книги, доктор Роман Ямпольский, профессор Университета Луисвилля и один из ведущих мировых специалистов по безопасности ИИ, предлагает читателю уникальный взгляд на проблему, объединяя технические, философские и этические подходы исследования в цельное, мощное высказывание.

Книга написана доступным языком и рассчитана как на специалистов в области ИИ, так и на широкий круг читателей, интересующихся будущим технологий и их воздействием на общество.

УДК 004.8  
ББК 16.6

Все права защищены.  
Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельца авторских прав.

© All Rights Reserved

Authorised translation from the English language edition published by CRC Press, a member of the Taylor & Francis Group LLC

© Издание на русском языке, дизайн.  
Nova Creative Group, 2025

ISBN 978-5-908058-00-1 (рус.)  
ISBN 978-1032576268 (англ.)

# Содержание

|                                                                                      |           |
|--------------------------------------------------------------------------------------|-----------|
| Предисловие .....                                                                    | 11        |
| Благодарности .....                                                                  | 13        |
| Об авторе.....                                                                       | 14        |
| <b>Глава 1 Введение.....</b>                                                         | <b>16</b> |
| 1.1. Введение.....                                                                   | 16        |
| 1.2. Задача контроля ИИ.....                                                         | 17        |
| 1.3. Что мешает контролировать ИИ.....                                               | 19        |
| 1.4. Определение безопасного ИИ.....                                                 | 21        |
| 1.5. Управляемость ИИ.....                                                           | 22        |
| 1.6. Выводы.....                                                                     | 23        |
| 1.7. Об этой книге.....                                                              | 24        |
| Использованные материалы .....                                                       | 26        |
| <b>Глава 2 Непредсказуемость.....</b>                                                | <b>30</b> |
| 2.1. Введение в проблему непредсказуемости.....                                      | 30        |
| 2.2. Предсказуемость: что мы можем предвидеть.<br>Обзор литературных источников..... | 33        |
| 2.3. Когнитивная неместимость.....                                                   | 36        |
| 2.4. Выводы.....                                                                     | 37        |
| Использованные материалы .....                                                       | 39        |
| <b>Глава 3 Необъяснимость и непостижимость.....</b>                                  | <b>44</b> |
| 3.1. Введение.....                                                                   | 45        |
| 3.2. Обзор литературных источников.....                                              | 46        |
| 3.3. Необъяснимость.....                                                             | 49        |
| 3.4. Непостижимость.....                                                             | 52        |
| 3.5. Выводы.....                                                                     | 55        |
| Примечания.....                                                                      | 56        |
| Использованные материалы .....                                                       | 57        |

|                                                                                |     |
|--------------------------------------------------------------------------------|-----|
| <b>Глава 4. Непроверяемость</b>                                                | 64  |
| 4.1. О наблюдателях и проверяющих                                              | 64  |
| 4.2. Историческая перспектива                                                  | 65  |
| 4.3. Классификация инструментов проверки                                       | 68  |
| 4.4. Непроверяемость                                                           | 72  |
| 4.5. Непроверяемость программ                                                  | 74  |
| 4.6. Выводы и направления для дальнейших исследований                          | 77  |
| Примечания                                                                     | 78  |
| Использованные материалы                                                       | 79  |
| <b>Глава 5. Невозможность владения</b>                                         | 85  |
| 5.1. Введение                                                                  | 85  |
| 5.2. Препятствия владению                                                      | 87  |
| 5.3. Выводы                                                                    | 90  |
| Использованные материалы                                                       | 91  |
| <b>Глава 6. Неконтролируемость</b>                                             | 94  |
| 6.1. Введение                                                                  | 96  |
| 6.2. Проблема контроля ИИ                                                      | 98  |
| 6.3. Предшествующие работы                                                     | 112 |
| 6.4. Доказывание неконтролируемости                                            | 125 |
| 6.5. Мультидисциплинарные данные, свидетельствующие<br>о неконтролируемости ИИ | 130 |
| 6.6. Доказательства неконтролируемости ИИ в исследованиях<br>безопасности ИИ   | 145 |
| 6.7. Обсуждение                                                                | 168 |
| 6.8. Выводы                                                                    | 172 |
| Примечания                                                                     | 176 |
| Использованные материалы                                                       | 176 |
| <b>Глава 7. Пути к опасному ИИ</b>                                             | 198 |
| 7.1. Классификация путей к опасному ИИ                                         | 198 |
| 7.2. Выводы                                                                    | 207 |
| Использованные материалы                                                       | 209 |
| <b>Глава 8. Сбои</b>                                                           | 213 |
| 8.1. Введение                                                                  | 213 |
| 8.2. Сбои ИИ                                                                   | 215 |
| 8.3. Безопасность ИИ                                                           | 226 |
| 8.4 Кибербезопасность и безопасность ИИ                                        | 228 |
| 8.5. Выводы                                                                    | 230 |

|                                                                                |            |
|--------------------------------------------------------------------------------|------------|
| Примечания.....                                                                | 231        |
| Использованные материалы .....                                                 | 235        |
| <b>Глава 9. Правосубъектность.....</b>                                         | <b>240</b> |
| 9.1. Введение в правосубъектность ИИ.....                                      | 240        |
| 9.2. Эгоистичные мемы.....                                                     | 241        |
| 9.3. Принижение человеческого достоинства .....                                | 243        |
| 9.4. Взлом правовой системы.....                                               | 244        |
| 9.5. Безопасность человека.....                                                | 245        |
| 9.6. Выводы .....                                                              | 250        |
| Примечания.....                                                                | 251        |
| Использованные материалы .....                                                 | 252        |
| <b>Глава 10. Сознание .....</b>                                                | <b>256</b> |
| 10.1. Введение в проблему сознания.....                                        | 256        |
| 10.2. Тест на наличие квалиа.....                                              | 260        |
| 10.3. У компьютеров могут быть иллюзии, поэтому они обладают<br>сознанием..... | 263        |
| 10.4. Цель сознания.....                                                       | 266        |
| 10.5. Сознание и искусственный интеллект .....                                 | 269        |
| 10.6. Выводы и предположения.....                                              | 272        |
| Примечание.....                                                                | 275        |
| Использованные материалы .....                                                 | 275        |
| <b>Глава 11. Персонализированные вселенные .....</b>                           | <b>289</b> |
| 11.1. Введение в проблему ценностного согласования<br>нескольких агентов.....  | 289        |
| 11.2. Индивидуальные моделируемые вселенные.....                               | 291        |
| 11.3. Преимущества и недостатки персонализированных<br>вселенных .....         | 294        |
| 11.4. Выводы.....                                                              | 296        |
| Примечание.....                                                                | 297        |
| Использованные материалы .....                                                 | 297        |
| <b>Глава 12. Человек ≠ сильный ИИ .....</b>                                    | <b>304</b> |
| 12.1 Введение.....                                                             | 304        |
| 12.2. Предшествующие работы.....                                               | 306        |
| 12.3. Человек и сильный ИИ — не эквивалентные понятия.....                     | 308        |
| 12.4. Выводы.....                                                              | 312        |
| Примечание.....                                                                | 313        |
| Использованные материалы .....                                                 | 314        |

|                                                                      |     |
|----------------------------------------------------------------------|-----|
| <b>Глава 13. Скептицизм</b> .....                                    | 319 |
| 13.1. Скептицизм в отношении рисков ИИ:<br>введение в проблему ..... | 319 |
| 13.2. Виды скептиков в отношении рисков ИИ .....                     | 323 |
| 13.3. Аргументы в пользу скептицизма в отношении рисков ИИ .....     | 327 |
| 13.4. Меры противодействия скептицизму в отношении рисков ИИ .....   | 341 |
| 13.5. Выводы .....                                                   | 345 |
| Примечания .....                                                     | 347 |
| Использованные материалы .....                                       | 348 |

# Предисловие

«Сотворили его люди, но управлять им они не могут», — писал Джон Стейнбек в своем романе «Гроздь гнева» почти сто лет назад. И хотя речь шла о банке, сегодня эти слова можно отнести и к проблематике искусственного интеллекта (AI).

Новостной поток, который сопровождает всемирную гонку в создании искусственного сверхинтеллекта, постоянно приносит нам сообщения о превосходстве машин над человеком в решении задач. Эти успехи не только создают завышенные ожидания к разработчикам AI, но и вызывают опасения за будущее человечества среди ученых и инженеров.

Обеспокоенность эта серьезна — пионеры отрасли готовы к радикальным изменениям карьеры, чтобы привлечь внимание общества к проблеме контроля над AI. И хотя технология создается нами, людьми, сможем ли мы ее контролировать? Один из героев книги Ямпольского, американский специалист по AI Элиезер Юлковский, предупреждает, что «наибольшая опасность искусственного интеллекта заключается в том, что люди слишком рано решают, что понимают его».

Более того, если посмотреть в историческом плане, то первая, по-настоящему значимая технология, оказавшаяся в руках человека, — расщепление атома. Она разрабатывалась по инициативе государств, и до сих пор использование атомной энергии тщательно контролируется во всех странах. Вторая — преобразование информации. Однако есть одно важное отличие — впервые радикальная, трансформационная технология развилась и сосредоточивалась в частных компаниях.

В публичном поле сегодня представлено множество мнений о контролируемости AI. Но ситуация больше напоминает библиотеку после

ремонта — книги на полках хорошие, но не на своих местах. На этом фоне работа доктора Ямпольского появилась в правильный момент: ученый предлагает несколько методов и классификаций для ответа на вопрос, можем ли мы контролировать AI. И кажется, это именно то, что нужно современному читателю, — тщательно и аккуратно разобратся в теме, используя научный подход. Обрисовывая долгосрочную перспективу, автор дает однозначный ответ — на существующей технологической парадигме мы не можем полностью и однозначно контролировать AI и последствия его применения.

Однако отнести Ямпольского к техноскептикам нельзя. Со свойственной ученым основательностью он отделяет зерна от плевел: надо использовать слабый AI там, где это полезно, например в медицине, и тщательно изучать все последствия, где есть риски. В случае искусственного суперинтеллекта любая ошибка может стать последней для человечества. И чтобы избежать катастрофических сценариев, необходимы исследования новых подходов к разработке моделей. Ямпольский считает, что любое определение AI критически ошибочно, если в нем не используется слово «безопасность».

Именно поэтому книга достойна внимания руководителей технологических компаний, разрабатывающих и внедряющих AI. Кроме того, ее будет полезно прочитать всем исследователям, вовлеченным в создание AI, — осознание небезопасности этой технологии должно стать необходимым шагом для всех дальнейших работ. Единственное, что мы не можем себе позволить, — игнорировать риски необъяснимого, непредсказуемого и неконтролируемого AI.

Альберт Ефимов, к. филос. н.,  
Вице-президент — директор  
управления исследований  
и инноваций  
ПАО Сбербанк  
Заведующий кафедрой  
инженерной кибернетики  
Университет науки  
и технологий МИСИС

*Моему другу Яну Таллинну, человеку,  
сделавшему для мира больше, чем вы  
когда-либо узнаете.*

# Благодарности

Я хотел бы поблагодарить многих людей за помощь, советы, замечания к моей работе, содействие в исследованиях — или просто за то, что они так или иначе подталкивали мои размышления. Одних мне повезло встретить лично, другие остались моими виртуальными собеседниками — но для идей это неважно, так что все эти люди в равной степени замечательны. Уверен, что я многих упустил, ведь я не обладаю сверхразумом, да и память у меня скверная. Заранее приношу извинения всем, кого я здесь не упомянул. Вклад многих выдающихся ученых я признаю, цитируя их работы, ведь это главное признание для любого исследователя. Большое спасибо вам: Макс Тегмарк, Иэн Гудфеллоу, Кеннет Риган, Эдуард Френкель, Себастьян Зани, Сорен Эльверлин, Мелисса Хелтон, Анна Хусфельдт, Торе Хусфельдт, Дэвид Келли, Дэвид Джилк, Скотт Ааронсон, Роб Бенсингер, Сет Баум, Тони Барретт и Алексей Турчин. Последними, но не в последнюю очередь я хочу поблагодарить Яана Таллинна и Фонд выживания и процветания (Survival and Flourishing Fund), а также Илона Маска и Институт будущего жизни (Future of Life Institute) за частичное финансирование моей работы, посвященной безопасности ИИ.

## Об авторе

Доктор Роман Владимирович Ямпольский — доцент, штатный сотрудник кафедры компьютерно-технических наук Инженерной школы Дж.Б. Спида при Луисвиллском университете. Основатель и действующий руководитель лаборатории кибербезопасности при Луисвиллском университете, автор множества книг, в т. ч. «Искусственный сверхинтеллект. Футуристический подход» («Artificial Superintelligence: A Futuristic Approach»), редактор сборников «Безопасность и защита ИИ» («AI Safety and Security») и «Технологическая сингулярность» («The Technological Singularity»). За время работы в Луисвиллском университете Роман Ямпольский получил звания почетного профессора-преподавателя, профессора года, самого популярного преподавателя своего факультета, эксперта в техническом образовании и награду за выдающееся начало карьеры в образовании, вошел в четверку лучших преподавателей факультета и десятку лучших профессоров года, а также удостоился многих других наград и почетных званий. Роман Ямпольский — старший член Общества изучения сильного и слабого ИИ института IEEE, член Академии наук штата Кентукки, ранее был научным руководителем Machine Intelligence Research Institute (MIRI) и Global Catastrophic Risk Institute (GCRI).

Роман Ямпольский получил степень доктора философии на кафедре компьютерно-технических наук университета штата Нью-Йорк в Буффало. Четыре года был стипендиатом Программы интегрированного обучения аспирантов и исследовательских стажировок Национального научного фонда США (National Science Foundation, NSF). Перед докторантурой Роман Ямпольский получил комбинированную степень бакалавра/магистра информатики (с отличием) в Рочестерском технологическом институте (шт. Нью-Йорк, США). После защиты докторской

диссертации занял должность члена-корреспондента в Центре расширенного пространственного анализа Университетского колледжа Лондона в Лондонском университете. Ранее проводил исследования в Лаборатории прикладных вычислений в Рочестерском технологическом институте и в Центре комплексной биометрии и датчиков университета штата Нью-Йорк в Буффало. Роман Ямпольский — выпускник Университета сингулярности (летняя программа последиplomного образования, 2012 г.) и приглашенный научный сотрудник Института сингулярности (Институт исследований машинного интеллекта).

Основная сфера интересов Романа Ямпольского — безопасность ИИ. Он автор более 200 публикаций, включая многочисленные статьи в журналах и книгах. Его исследования цитируются тысячами ученых и рассматриваются в популярных журналах как в США, так и за рубежом. Работы Романа Ямпольского упоминаются более 10000 раз в различных статьях в СМИ более чем на 40 языках.

# Введение\*

## 1.1. Введение

Стремительный прогресс искусственного интеллекта (ИИ) за последнее десятилетие сопровождался и несколькими громкими провалами [1]. Стала очевидной необходимость следить за тем, чтобы ИИ работал на благо человечества. Так возникла новая сфера исследований — безопасность ИИ [2], охватывающая широкий спектр областей. Число публикаций по безопасности ИИ в последнее время уверенно растет [3–10].

Однако в основе всех этих исследований лежит предположение, что задача контроля ИИ в принципе разрешима, хотя не существует ни строгих математических доказательств тому, ни предпосылок к их появлению. В компьютерных науках принято сначала определить, относится ли задача к числу решаемых, а затем уже вкладывать ресурсы в ее решение.

Все признают, что контроль ИИ, возможно, является одной из важнейших проблем, стоящих перед человечеством, — и тем не менее, эта сфера остается малопонятной, недостаточно изученной и плохо исследованной. Задачи компьютерных наук принято делить на решаемые, нерешаемые, неразрешимые или имеющие частичное решение — но мы до сих пор не знаем, к какому виду относится задача контроля ИИ. Мы предполагаем, что в некоторых ситуациях возможны определенные формы контроля, но существует и вероятность того, что

---

\* Фрагменты этой главы ранее были опубликованы в статьях Романа Ямпольского «Об управляемости ИИ» («On Governability of AI») в *AI Governance in 2020 a Year in Review*. June, 2021, и «Определение различий между интеллектом и искусственным интеллектом» (*On Defining Differences Between Intelligence and Artificial Intelligence*) в *Journal of Artificial General Intelligence* 11(2), 68–70. 2020. — *Прим. пер.*

во многих случаях частичного контроля будет недостаточно. Пока мы не понимаем ни природы, ни осуществимости задач контроля ИИ, мы не сможем выработать соответствующую методику [11].

Потенциальные методы контроля сильного ИИ (Artificial General Intelligence, AGI) делят две общие категории: методы, основанные на контроле возможностей, и методы контроля мотивации [12]. Первые призваны минимизировать вред от систем сильного ИИ, помещая их в ограниченную среду, оснащенную механизмами (триггерами) останова. Методы контроля мотивации предполагают разработку систем сильного ИИ, которые будут наделены желанием не причинять вреда даже без ограничения их возможностей. Широко признано, что контроль возможностей — в лучшем случае временная мера и не может считаться долгосрочным решением задачи контроля сильного ИИ [12]. Кроме того, использование средств контроля мотивации может потребоваться еще до запуска системы — на стадии разработки и обучения.

## 1.2. Задача контроля ИИ

Мы определяем проблему контроля ИИ следующим образом:

«Как человечество может безопасным образом осуществлять контроль, используя высшую форму интеллекта с выгодой для себя?»

Это фундаментальная проблема безопасности ИИ, которая сама по себе посвящена обеспечению безопасности интеллектуальных систем для всех заинтересованных сторон.

В настоящее время для обеспечения безопасности ИИ чаще всего рассматривают метод согласования ценностей\*. Тем не менее, несмотря на долгие годы исследований, степень безопасности все еще сложно протестировать или точно измерить даже в программном обеспечении (ПО), которое не используют ИИ [13]. В лучшем случае мы, пожалуй, сможем отличить совершенно безопасную систему от системы, сравнимой по безопасности с человеком, выполняющим

---

\* Тем самым предполагается, что машины можно научить понимать человеческие ценности и цели и действовать в соответствии с ними. — *Прим. пер.*

ту же задачу. Но вряд ли общество смирится с ошибками машин, даже если они возникают не чаще, чем человеческие ошибки. Мы ожидаем, что машины будут работать лучше нас, и не согласимся на частичную безопасность настолько мощных систем. Влияние ИИ (как положительное, так и отрицательное [3]) во многом зависит от его способностей. В том, что касается экзистенциальных угроз, частичная безопасность недопустима.

На первый взгляд кажется, что для решения задачи контроля нужно разработать машину, в точности выполняющую команды человека. Но такая форма контроля неоптимальна — команды могут быть противоречивыми или парадоксальными, человеческие языки неоднозначны [14], а постановка задач бывает некорректной (хотя желательно предусмотреть возможность обратной связи от человека). Предполагается, что для решения этой проблемы ИИ должен стать идеальным советником, который будет избегать проблем с неверным толкованием прямых команд и не будет допускать поступления команд от злоумышленников.

Считается, что отсутствие контроля ИИ приведет к настолько серьезным последствиям, что даже при малейшей вероятности появления враждебного ИИ потенциальный вред приобретет астрономический масштаб — а значит, исследования по безопасности ИИ в любой случае окупятся. Здравый смысл говорит о том, что огромный негативный эффект даже при его малой вероятности означает очень большой вред. И такой вариант развития событий действительно необходимо воспринимать всерьез. К тому же в реальности шансы несогласованности ИИ с целями и нормами его создателей довольно высоки. Это значит, что в отсутствие эффективной программы безопасности ИИ единственно возможный результат — почти гарантированное отрицательное событие с потенциалом экзистенциальной катастрофы. Таким образом, статистика говорит в пользу масштабного исследования безопасности ИИ. Речь о не о сценариях с низкими рисками и высокой отдачей — риски велики, а результатом будет предотвращение худшего. Неудивительно, что многие считают, что это самая серьезная задача, когда-либо стоявшая перед человечеством. Мы выбираем между нашим процветанием и нашим уничтожением. На весах находится судьба мира, поэтому доказательство решаемости или нерешаемости задачи контроля ИИ — действительно важнейшая проблема всех времен.

### 1.3. Что мешает контролировать ИИ

Для того чтобы контролировать модель сильного ИИ, нам нужны инструменты, которые обеспечат ее объяснимость, предсказуемость и проверяемость [15]. Но возникает вопрос доступности этих инструментов для нас.

- Концепция необъяснимости ИИ предполагает, что невозможно дать на 100% точное и понятное объяснение конкретных решений интеллектуальной системы. Ее дополняет концепция непостижимости ИИ, которая означает, что человек не сможет до конца понять любое совершенно точное объяснение решений интеллектуальных систем [16].
- Непредсказуемость ИИ, которая также мешает достижению безопасности ИИ, — это неспособность человека точно и закономерно предсказывать действия интеллектуальной системы, которые она предпримет для достижения своих целей, даже если известны ее конечные цели [17]. Непредсказуемость связана с необъяснимостью и непостижимостью ИИ, но не совпадает с ними полностью. Она не означает, что невозможен статистический анализ с точностью выше случайной, но указывает на общее ограничение эффективности анализа, особенно ярко выраженное у передовых систем сильного ИИ в новых областях.
- Непроверяемость — фундаментальное ограничение проверки математических доказательств, ПО, поведения интеллектуальных агентов и любых формальных систем [18]. Все более очевидно, что у нас есть лишь вероятная уверенность в правильности математических доказательств и программных решений — и весьма ограниченная возможность проверки интеллектуальных агентов.

Многие исследователи предполагают, что задачу контроля ИИ можно решить, но никаких доказательств или подтверждений тому нет. Поэтому прежде чем приступить к созданию контролируемого ИИ, важно показать, что задача в принципе имеет решение, чтобы не тратить ценные ресурсы понапрасну. Бремя доказательств лежит на тех, кто утверждает, что эта задача вполне решаемая, а отсутствие таких доказательств в настоящее время свидетельствует о том, насколько рискованно разрабатывать сильный ИИ. С большой вероятностью ИИ

не поддается контролю — это можно доказать приведением данной проблемы к задаче контроля со стороны человека.

Есть множество открытых вопросов о контролируемости ИИ: можно ли решить задачу контроля? Возможно ли это в принципе? Возможно ли это на практике? Можно ли при этом добиться достаточного уровня точности? Сколько времени это займет? Успеем ли мы это сделать? Каковы затраты энергии и вычислительных мощностей для данной задачи? На что будет похоже решение? Каким будет минимально осуществимое решение? Как мы поймем, что задача решена? Поддается ли решение масштабированию при дальнейшем развитии системы? Мы считаем, что интеллект без ограничений нельзя контролировать, а интеллект с ограничениями не сможет создавать что-либо новое. Если не контролировать сильный ИИ, то, кто бы ни программировал его, последствия будут катастрофическими для всех и в первую очередь — для его создателей. Неконтролируемый сильный ИИ никому не принесет пользы.

Отсутствуют публикации, подтверждающие, что менее интеллектуальный агент может бесконечно сохранять контроль над более интеллектуальным. Разрабатывая системы, уступающие нам по интеллекту, мы сохраняем контроль над ними, но если системы станут умнее нас, мы потеряем контроль. Если же мы пытаемся сохранить контроль, разрабатывая все более продвинутые интеллектуальные агенты, мы попадаем в «Уловку-22»\*, так как механизм контроля должен быть умнее контролируемого агента или хотя бы иметь тот же уровень интеллекта. Это приведет к тому, что для контроля систем, возможности которых постоянно растут, потребуется целая иерархия интеллектуальных систем.

Более того, задача контроля более мощного интеллекта усложняется и, очевидно, становится недоступной для агентов с постоянным уровнем интеллекта. Контролировать ситуацию будет тот, кто умнее, и он же будет принимать окончательные решения. Насколько нам известно, на данный момент нигде в мире не существует не только действующего механизма контроля ИИ, который можно было бы масштабировать до ИИ человеческого уровня и (когда-нибудь) выше, но даже представления о том, каким мог бы быть прототип такого механизма.

---

\* Логический парадокс, описанный в одноименном романе Джозефа Хеллера, — выполнение условий задачи одновременно обеспечивает невыполнимость задачи. — *Прим. пер.*