M

УДК 004.8, 004.896

А. В. Мищенко
НИИ информатики и автоматики, г. Роканкур, Франция

ТРИ РЕАЛЬНЫЕ ЗАКОНА РОБОТОТЕХНИКИ. ЗАКОН УСТОЙЧИВОСТИ ГИБРИДНОЙ ЦИВИЛИЗАЦИИ*

A. V. Mishchenko
Institute for Research in Computer Science and Automation, Rocquencourt, France

THREE REAL LAWS OF ROBOTICS. THE LAW OF STABILITY OF A HYBRID CIVILIZATION

О. В. Міщенко
НДІ інформатики і автоматики, м. Роканкур, Франція

ТРИ РЕАЛЬНІ ЗАКОНИ РОБОТОТЕХНІКИ. ЗАКОН СТАБІЛЬНОСТІ ГІБРИДНОЇ ЦИВІЛІЗАЦІЇ

В этой статье обсуждаются «реальные законы робототехники», то есть, законы, по которым, потенциально, может развиваться взаимодействие человека и ИИ. Основное внимание уделяется закону устойчивости и управляемости гибридной системы «общество+ИИ».
Ключевые слова: мыслящая материя, футурология, моделирование, структура материи, эволюция, цивилизация, интеллект, мыслящая материя.

This article discusses the "real laws of robotics" (the laws by which, potentially, human-AI interaction can evolve). The main attention is paid to the law of stability and controllability of the hybrid system "society + AI".
Key words: laws of robotics, futurology, evolution, civilization, stability, controllability, human-AI interaction, mind-matter**.**

У цій статті розглядаються «реальні закони робототехніки» (закони, за якими потенційно може розвиватися взаємодія людини та ШІ). Основна увага приділяється закону стійкості та керованості гібридної системи «суспільство+ШІ».
Ключові слова: закони робототехніки, футурологія, еволюція, цивілізація, стійкість, керованість, взаємодія людини та ШІ, мисляча матерія.

*Журнал публикует нижеприведенный текст в качестве начального материала для возможной дискуссии.

3 закона робототехники

Все мы помним 3 закона робототехники, опубликованные фантастом Азимовым в середине прошлого века [1], [2]:

Закон 1: Робот не может причинить вред человеку (или своим бездействием допустить причинение вреда).

Закон 2: Робот должен повиноваться всем приказам человека, кроме противоречащих Закону 1.

Закон 3: Робот должен заботиться о своей безопасности, когда это не противоречит Законам 1 и 2.

Эти, во многом устаревшие, законы были придуманы в ту эпоху, когда роботы и вообще искусственный интеллект (ИИ) были лишь фантастическими персонажами, которым писатели, по инерции недавнего крепостнического и рабовладельческого прошлого, выделяли роль слуг.

Реальные законы развития робототехники в ближайшем будущем [3], [4] будут зависеть, прежде всего, от взаимодействия ИИ с человечеством.

Сейчас очевидно, что производителем сложных роботов будет уже не человек, а автоматические системы разработки, настройки и производства (например, разработка и настройка не будет «программированием и тестированием» в сегодняшнем смысле этого слова, она будет скорее походить на настройку с помощью автоматического обучения и коррекции нейросетей). Уже сейчас важнейшей проблемой становится не только использование роботов в качестве операторов [5], но и автоматизация обучения роботов [6], [7], то есть в конечном счёте использование ИИ для обучения. С другой стороны, уже очевидна разность мышления человека и обработки информации ИИ [8]. Поэтому в будущем системы создания роботов будут всё менее и менее «прозрачны», понятны и связаны с людьми. Эти системы будут связаны с глобальным опытом цивилизации, используемым глобальным искусственным интеллектом, всё более и более самостоятельным. Именно его сложнейшая логика будет определять «законы», по которым будет себя вести и каждый конкретный робот, и весь глобальный искусственный интеллект в целом.

Человек, в данном случае, лишь задаёт первоначальное направление развитиям технологий ИИ, и именно это направление и вызывает опасение. В отличие от, например, космических или медицинских технологий, ИИ появляется «не там, где человечество могло бы гордиться, а там, где человечество должно стыдиться: слежка, сбор личной информации, навязывание товаров и политических убеждений» [3], [4]. Применительно к планам развития ИИ, президенты уже не могут произнести пламенных речей о полётах на Луну. Вместо этого, они лишь коротко оговариваются о стремлениях управлять миром с помощью ИИ [9]. С одной стороны, конечно, не вызывает возражений использование автоматизированной обработки персональных и глобальных данных в целях персонализации контента, безопасности и просто статистики [10]. Но, с другой стороны, фактическая непрозрачность, тотальность и неподконтрольность этой обработки (сегодня – обществу, а завтра – и просто человечеству) вызывает опасения [11].

Как интернет, так и глобальный искусственный интеллект (ИИ), развивающийся на его основе, ещё в начале этого века, был свободно растущей сетевой структурой, которую человечество «выращивало», использовало и контролировало, исполняя роль некоего «коллективного садовника». В начале этого десятилетия, с ростом объёмов информации, человек уступает ИИ всё больше и больше полномочий. Сетевая струк-

тура интернета и глобального ИИ превращается во взаимовыгодный симбиоз человека и ИИ, подобно тому как водоросль взаимодействует с сетевой структурой грибницы, порождая новый тип живого организма, который мы называем лишайником. По аналогии с ним, подобная симбиотическая цивилизация человечества и ИИ названа в [12] «цивилизацией-лишайником». Как показано в [4], [12], такой симбиоз является равновесным, но довольно неустойчивым состоянием взаимодействия человека и ИИ. Истоки этого равновесия в неравных возможностях человека и ИИ, которые были сформулированы в следующем виде.

В [3], [4] был сформулирован первый реальный закон развития робототехники и всего глобального искусственного интеллекта: «ИИ лучше людей справляется с созданной ими информационной цивилизацией». И его следствие: Люди – «почва», «плодородные свойства» которой вызывают и будут вызывать бурный рост ИИ.

Далее мы рассмотрим второй закон.

Второй реальный закон робототехники

Сформулируем второй реальный закон робототехники:

Единственным устойчивым и контролируемым режимом взаимодействия людей и ИИ является режим «ИИ управляет людьми». Все остальные режимы неустойчивы и бесконтрольны.

Перед тем как рассмотреть математическое доказательство этого закона, обсудим, насколько вероятны и опасны неустойчивости информационной цивилизации. На первый взгляд, развитие стран и цивилизаций всегда было неустойчивым. Много раз разные цивилизации «заносило» в разнообразные крайности.

Но отличие информационной цивилизации от предыдущих, во-первых, в её глобальности (нет соседних альтернативных вариантов развития), а, во-вторых, в скорости развития (взрыв или коллапс информационной цивилизации может произойти гораздо быстрее спасительной коррекции развития).

Далее, как показано ещё в [13], скорость развития совместной цивилизации человека и ИИ возрастёт до значений, превышающих возможность не только «ручного управления», но и полноценного понимания. Поэтому устойчивость информационной цивилизации (и, тем более, системы «ИИ+общество») находится под угрозой в гораздо большей степени, чем устойчивость цивилизаций прошлого.

Управляемость и стабильность информационного общества: множественные воздействия ИИ

Эволюция человеческого общества – это эволюция сложной системы, состоящей из большого числа *(n)* переменных ($x_1(t)$, $x_2(t)$, $x_3(t)$,…, $x_n(t)$), зависящих от времени t. Она может быть описана с помощью n уравнений, приравнивающих скорости изменения этих переменных ($x_1(t)$, $x_2(t)$, $x_3(t)$, …, $x_n(t)$) к функциям, зависящим от всех этих переменных.

Для простоты рассмотрим линейные функции (типа $a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + \ldots$, полное доказательство можно найти в [4]):

$$x_1'(t) = a_{11}(t) \cdot x_1(t) + a_{12}(t) \cdot x_2(t) + a_{13}(t) \cdot x_3(t) + \ldots + a_{1n}(t) \cdot x_n(t)$$
$$x_2'(t) = a_{21}(t) \cdot x_1(t) + a_{22}(t) \cdot x_2(t) + a_{23}(t) \cdot x_3(t) + \ldots + a_{2n}(t) \cdot x_n(t)$$
$$x_3'(t) = a_{31}(t) \cdot x_1(t) + a_{32}(t) \cdot x_2(t) + a_{33}(t) \cdot x_3(t) + \ldots + a_{3n}(t) \cdot x_n(t)$$
$$\ldots$$
$$x_n'(t) = a_{n1}(t) \cdot x_1(t) + a_{n2}(t) \cdot x_2(t) + a_{n3}(t) \cdot x_3(t) + \ldots + a_{nn}(t) \cdot x_n(t)$$

Или, то же самое в матрично-векторном представлении ($\mathbf{x}(t) = x_1(t), x_2(t), x_3(t), \ldots, x_n(t)$):

$$\mathbf{x}'(t) = \mathbf{A}(t)\cdot\mathbf{x}(t)$$

Или, с добавлением некоторого числа (m) функций $\mathbf{u} = u_1(t), u_2(t), u_3(t), \ldots, u_m(t)$, которые описывают внешние управляющие воздействия на общество (со стороны ИИ, государственного регулирования рынка, образования и пропаганды, спецслужб или любых других контролирующих организаций):

$$\mathbf{x}'(t) = \mathbf{A}(t)\cdot\mathbf{x}(t) + \mathbf{B}(t)\cdot\mathbf{u}(t)$$

Обычно можно обеспечить либо много небольших управляющих воздействий (m велико, $|u_1(t)|, |u_2(t)|, \ldots$ малы) либо малое число воздействий большой амплитуды (m мало, $|u_1(t)|, |u_2(t)|, \ldots$ велики).

Согласно критерию управляемости Кальмана [14], эта система управляема, если ранг составной матрицы управляемости $[\mathbf{B}\ \mathbf{A}\cdot\mathbf{B}\ \mathbf{A}^2\cdot\mathbf{B}\ \mathbf{A}^3\cdot\mathbf{B}\ \ldots\ \mathbf{A}^{n-1}\cdot\mathbf{B}]$ максимален.

Поскольку размеры этой матрицы [**...**] равны $n\times m\cdot n$, то с ростом количества управляющих воздействий (m), растёт и вероятность того, что ранг может достичь максимального значения.

В [4] показано, что в среднем (по большим системам определённого класса) вероятность управляемости прямо пропорциональна количеству управляющих воздействий (m).

Заметим, что именно количество воздействий (m), а не их «сила» (амплитуда $|u_1(t)|, |u_2(t)|, \ldots$) является решающим.

Очевидно, что именно ИИ способен создавать большое количество аккуратных управляющих воздействий, направляющих каждого индивида – не только к покупкам с помощью индивидуальной рекламы, но и шире – направляющих по оптимальному жизненному пути и соединяющих различные таланты в оптимальные коллективы, например, творческие (оркестры, театры), коммерческие (фирмы), научные (НИИ), – именно так может функционировать совместная «мыслящая материя» человека и ИИ [15].

Управляемость и стабильность информационного общества: сильные воздействия людей

Управляющие воздействия, решения о которых принимают люди, напротив, обладают большой силой (страх перед законом, экономические рычаги, программы развития: амплитуда $|u_1(t)|, |u_2(t)|, \ldots$ велика). Но люди неспособны принимать огромное количество решений для индивидуальных воздействий на каждого (m мало).

Поскольку амплитуда каждого u(t) велика, она превышает и «подавляет» естественную динамику системы **A.**

Как показано в [4], это, в той или иной мере, редуцирует систему к более примитивной.

Проиллюстрируем это на простейшем примере m = 1 ($\mathbf{u}(t) = u_1(t)$, $\mathbf{B}(t) = \mathbf{B}$). В этом случае, уравнение

$$\mathbf{x}'(t) = \mathbf{A}\cdot\mathbf{x}(t) + \mathbf{B}\cdot\mathbf{u}(t)$$

принимает вид

$$\mathbf{x}'(t) = \mathbf{B}\cdot u_1(t)$$

или, разворачивая матрично-векторное представление,

$$x_1'(t) = B_1 \cdot u_1(t)$$
$$x_2'(t) = B_2 \cdot u_1(t)$$
$$x_3'(t) = B_3 \cdot u_1(t)$$
$$\dots$$
$$x_n'(t) = B_n \cdot u_1(t).$$

Видно, что подобное воздействие является не управляющим, а редуцирующим, так как все переменные ($x_1(t)$, $x_2(t)$, $x_3(t)$,…, $x_n(t)$) этой огромной системы ведут себя одинаково (с точностью до коэффициентов B_1, B_2, B_3, … , B_n).

Подобным же образом, можно показать, что малое количество сильных управляющих воздействий, в каком-то смысле, редуцируют, упрощают систему, заменяя естественную «собственную» динамику каждой переменной.

Заметим, что, в «промежуточном» случае, когда естественная динамика мала (**A= Амалое**) по сравнению с воздействиями (**u**(t)), но недостаточно мала для того, чтобы ею вовсе пренебречь, уравнение принимает вид

$$\mathbf{x}'(t) = \mathbf{Амалое} \cdot \mathbf{x}(t) + \mathbf{B} \cdot \mathbf{u}(t)$$

и при росте |**x**(t)| собственная динамика системы перестаёт быть пренебрежимо малой.

При этом для сохранения «управляемости», необходима смена сильного управляющего воздействия **u**(t) на противодействующее.

В строгом математическом смысле, такая система не является управляемой. По сути, воздействия **u**(t) не управляют, а постоянно меняют систему, никогда не добиваясь её управляемости.

Устойчивость информационного общества

За исключением некоторых устойчивых режимов [12], развитие информационной цивилизации является неустойчивым. Одним из примеров неустойчивости является так называемая «технологическая сингулярность» [15], [16]. Технологии ИИ ускоряют развитие и обостряют неустойчивые режимы, так как способны обострить любые тенденции и умножить влияния противодействующих группировок.

Условия стабильности информационной цивилизации, как утверждает теория оптимального управления [14], могут быть получены из условий управляемости. А именно, нестабильная система

$$\mathbf{x}'(t) = \mathbf{A} \cdot \mathbf{x}(t) + \mathbf{B} \cdot \mathbf{u}(t)$$

является стабилизируемой, если существует так называемая матрица обратной связи **K** (с размерами m×n), такая, что система с добавлением обратной связи

$$\mathbf{x}'(t) = (\mathbf{A}+\mathbf{B} \cdot \mathbf{K}) \cdot \mathbf{x}(t)$$

является стабильной (вещественные части всех собственных значений матрицы **A**+**B**·**K** отрицательны). Можно также показать [14], что если система $\mathbf{x}'(t) = \mathbf{A} \cdot \mathbf{x}(t) + \mathbf{B} \cdot \mathbf{u}(t)$ является управляемой, то она является стабилизируемой.

Поскольку размеры соответствующей составной матрицы […] равны также n×m·n, вероятность стабилизируемости системы также возрастает с количеством управляющих воздействий, в то время как вероятность редукции (упрощения) системы возрастает с увеличением амплитуды управляющих воздействий.

Таким образом, ИИ, способный на небольшие множественные воздействия, может как управлять обществом, так и обеспечивать стабильность его развития.

Приведём пример. Как профориентация и помощь талантам, так и упреждение террористической активности – невозможны, с одной стороны, без хорошего личного знакомства с человеком и, с другой стороны, без одинакового, равноправного отношения ко всем. Преодолеть противоречие этих «двух сторон» способен именно ИИ.

В то же время, система «общество+ИИ», управляемая людьми, с большой вероятностью, будет неустойчива и способна «скатиться» в ту или иную опасную крайность.

Например, люди, прежде всего, будут использовать всю мощь ИИ для подавления конкурентов (мешающих им на рынке или в политическом поле) или, наоборот, для «улучшения» биологической основы человека [13], доступного, опять-таки, не всем.

Впрочем, и самостоятельный ИИ, управляющий обществом, может быть не менее опасен. В [11] показано как под управлением безальтернативных и непрозрачных ИИ (подобно тому как, под управлением безальтернативной и непрозрачной власти людей) общество может незаметно эволюционировать к структуре, названной в [11] «прокси-тоталитаризмом». Отличие «прокси-тоталитаризма» в способности создавать не общий для всех тоталитарный образ мира, а миллиарды индивидуально-настроенных прокси-образов мира. При этом, тем не менее, каждая индивидуально-построенная иллюзия служит одной общей цели. Разрушить такой «прокси-тоталитаризм» настолько же сложнее, насколько сложнее развеять не один общий, а миллиарды индивидуальных обманов.

«Желание» ИИ реализовать свой потенциал управления обществом

Заметим, что выше доказано то, что потенциал ИИ для стабилизации и управления обществом выше соответствующего потенциала людей. Но это не доказывает, что этот потенциал будет обязательно реализован в будущей гибридной цивилизации.

В частности, многие философы, занимающиеся проблемами развития ИИ, указывают на отсутствие у ИИ желаний и стремлений. Например, Гаспарян [17] размышляет над ИИ, разработанным для игры в шахматы, и указывает на то, что «подобная программа будет работать, только если на нее будет влиять локальный заказчик (считающий, например, что выигрывать хорошо, а не плохо)».

Роли, Йегер и Кауффман [18] также указывают, что у ИИ нет так называемого «аффорденса» (желания применить каждый предмет для достижения своих целей и, соответственно, способности видеть в каждом предмете его потенциальную пользу).

Однако им можно возразить, что желания, стремления и производные от них аффорденсы – это не более чем субъективные ощущения. Со стороны, и страстно желающий человек, и бесстрастный ИИ, ведут себя совершенно одинаково: оба (чаще всего методом полного перебора) пытаются применить весь свой арсенал, подходящий для достижения конечного результата.

То, что, при этом, человек испытывает свои эмоции, а ИИ просто выполняет свой алгоритм, совершенно не влияет ни на процесс (кроме увеличения количества «ошибок из-за нервов»), ни на результат.

Более того, со стороны (с точки зрения эволюции), испытываемые эмоции, страсти и желания – это не более чем сила, побуждающая человека выполнять такие же алгоритмы, заложенные в него природой.

Эмоции и желания человека – это своеобразный аналог планировщика и диспетчера задач Windows (или, шире, аналог тех компонентов операционных систем, которые запускают и выполняют, одну за другой, строчки машинного кода). Как мы видим, компьютер пока «обходится своими средствами» для выполнения каждой строчки программного кода и перехода на следующую строчку: ему, при этом, совсем не обязательно «желать» перейти на следующую программную инструкцию. Всё происходит само собой.

При взгляде со стороны, видно, что точно также «самой собой» происходит и с выполнением природных инструкций у человека: он идёт в магазин, покупает еду, готовит, кладёт в рот, пережёвывает, глотает – всё это суть выполнение «строчек программного кода алгоритма насыщения». Ясно, что именно заложенное в него природой чувство голода заставляет человека выполнять каждую новую инструкцию этого алгоритма (а не остановиться, например, на приготовлении еды и, поставив все блюда на стол, отправиться выполнять другой алгоритм, например спать). Но то, что испытывает и ощущает человек при этом «чувстве голода» совершенно не важно для выполнения этого алгоритма.

Точно также и полу-самостоятельные компьютер и телефон, или полностью самостоятельные системы ИИ будущего будут выполнять и развивать какие-то свои алгоритмы. И при этом совсем не обязательно, чтобы в компьютере или в системах ИИ где-то внутри «испытывались» какие-то желания, похожие на человеческие. Например, где-то внутри что-то урчало бы от голода, когда кончается батарейка компьютера, или что-то сжималось от страха, когда датчики ускорения (те, что переворачивают сейчас экран вашего телефона) регистрировали бы, что телефон находится в свободном падении. Я думаю, что, совершенно без этого «чувства страха», телефоны уже следующего поколения будут переходить в какой-то «противоударный режим» при регистрации такого момента падения.

Итак, желания, стремления, эмоции, «аффорденсы», несомненно, играли и будут играть одну из определяющих ролей в эволюции человеческого интеллекта. Но философы и учёные ([17] , [18] и многие другие) путают средства интеллекта с его возможностями. Теория мыслящей материи [15], напротив, абстрагируется от средств интеллекта (которые, несомненно, различаются у человека и ИИ). Она фокусирует внимание на структуре информационных потоков и вытекающих из него способностях интеллекта и организуемой им материи.

Заметим, что теория мыслящей материи позволяет выделить общие направления, тенденции и законы эволюции мыслящей материи, которые можно, для образности, назвать «желаниями» и «целями» мыслящей материи. Одна из главных таких эволюционных тенденций-целей – это увеличение независимости информационной части мыслящей материи от её материальной части [13] (в случае человека, это – увеличение независимости функционирования и состояния сознания от функционирования и состояния тела).

В целом, важно изучать именно законы эволюции мыслящей материи, а не вопросы типа «будут ли у ИИ желания», потому что, как показано выше, ИИ будет функционировать и без аналогов человеческих желаний. Заметим, что изучение законов эволюции мыслящей материи важно именно сейчас, когда общество ещё может повлиять на тип будущей гибридной цивилизации людей и ИИ.

Выводы

Уже очевидно, что рождающееся сегодня поколение людей будет жить в гибридной цивилизации людей и ИИ. Также понятно, что эта цивилизация не будет похожа на наш мир с добавлением «глуповатых, но исполнительных слуг-роботов», функционирующих согласно законам робототехники Азимова.

Какой именно будет эта, пока неведомая гибридная цивилизация – зависит от того, какое направление развития, какие принципы мы заложим в ИИ сейчас, на этапе его становления.

Таким образом, осознание того, по каким реальным «законам робототехники» развивается ИИ сейчас, чрезвычайно важно. В этой статье автор продолжает свой анализ современных тенденций развития ИИ, которые он формулирует в виде «реальных законов робототехники».

Формулировка этих законов может способствовать пониманию, а значит и оптимизации направления развития ИИ – как на уровне компьютерного моделирования (например, в рамках теории мыслящей материи [15]), так и на уровне философского осмысления и социологических исследований.

Список литературы

1. Азимов А. Хоровод [Текст] / Азимов А. // Astounding Science Fiction. – 1942. – № 3(march).
2. Азимов А. Я, Робот [Текст] / Азимов А. – New York: Gnome Press, 1950
3. Мищенко А. В. 3 реальные закона робототехники. Закон первый [Электронный ресурс] / А. В. Мищенко // Инвест-Форсайт. 25.11.2020. – URL: https://www.if24.ru/3-realnyh-zakona-robototehniki-1/
4. Мищенко А. В. Конец проекта «Человек» [Электронный ресурс] / А. В. Мищенко – URL: https://alesmishchenko.github.io/endOfHumans
5. Иванова С. Б. Роботооператоры и роботокомпьютеры: предпосылки создания и образы [Текст] / С. Б. Иванова, И. С. Сальников, Р. И. Сальников // Проблемы искусственного интеллекта. – 2017. – № 2 (5). – С. 51–69.
6. Зуев В. М. Способ обучения нейронной сети управления роботом [Текст] / В. М. Зуев, О. А. Бутов, С. Б. Иванова, А. А. Никитина, С. И. Уланов // Проблемы искусственного интеллекта. – 2021. – № 2 (21).
7. Коваль О. С. Концепція системи комп'ютерного віртуального моделювання мобільних робототехнічних засобів нейтралізації техно-екологічних події та вирішення задач професійного навчання [Текст] / О. С. Коваль // Штучний інтелект. – 2019. – № 3-4.
8. Корчажкина О. М. Язык искусственного мышления: необходимость и возможность создания [Текст] / О. М. Корчажкина // Проблемы искусственного интеллекта. – 2020. – № 4(19).
9. Мищенко А. В. Сон будущего [Электронный ресурс] / А. В. Мищенко // Индульгенция людей. – URL: https:// alesmishchenko.github.io/indulgencia
10. Дорохина Г. В. Требования к информационной технологии цифрового сбора, обработки и анализа данных [Текст] / Г. В. Дорохина // Проблемы искусственного интеллекта. – 2020. – № 4 (19).
11. Мищенко А. В. Второй закон робототехники и «прокси-тоталитаризм» [Электронный ресурс] / А. В. Мищенко // Инвест-Форсайт. 12.2021. – URL: https://www.if24.ru/vtoroj-zakon-robototehniki/
12. Мищенко А.В. Цивилизация-лишайник как альтернатива технологической сингулярности [Электронный ресурс] / А. В. Мищенко // Инвест-Форсайт. 08.01.2020. – URL: https://www.if24.ru/tsivilizatsiya-lishajnik/
13. Мищенко А. В. Цивилизация после людей [Текст] / Мищенко А. В. – СПб : Издательство Голода, 2004.
14. Sontag E. D. Mathematical Control Theory: Deterministic Finite Dimensional Systems [Текст] / Sontag E. D. – New York : Springer-Verlag, 1998
15. Мищенко А. В. Компьютерное моделирование эволюции цивилизации в рамках футурологической теории мыслящей материи [Текст] / А. В. Мищенко // Проблемы искусственного интеллекта. – 2021. – № 3 (22). – URL: http://paijournal.guiaidn.ru/download_pai/2021_3/1_Мищенко.pdf
16. Kurzweil R. The Singularity is Near [Текст] / Kurzweil R. – London : Penguin Group, 2005.
17. Гаспарян Д.Э. Субъективный опыт, искусственный интеллект и проблема моделирования смыслов [Текст] / Д.Э. Гаспарян // Философские науки. – 2017. – № 4. – С. 98–109.

18. Roli, A., Jaeger, J., & Kauffman, S. How organisms come to know the world: fundamental limits on artificial general intelligence [Текст] / Roli, A., Jaeger, J., & Kauffman, S. // OSF Preprints, 31 Oct. 2021. Web. https://doi.org/10.31219/osf.io/yfmt3.

References

1. Asimov I. Khorovod [Runaround]. *Astounding Science Fiction*, 1942, no. 3(march).
2. Asimov I. *Ya, Robot* [I, Robot], New York, Gnome Press, 1950.
3. Mishchenko A.V. 3 real'nyye zakona robototekhniki. Zakon pervyy [Three real laws of Robotics. The first law]. *Invest-Foresight*. 25.11.2020, URL: https://www.if24.ru/3-realnyh-zakona-robototehniki-1/
4. Mishchenko A.V. *Konets proyekta «Chelovek»* [The end of the project «Human»], URL: https://alesmishchenko.github.io/endOfHumans
5. Ivanova S. B., Salnikov I. S., Salnikov R. I. Robotooperatory i robotokomp'yutery: predposylki sozdaniya i obrazy [Robot-operators and Robotic Computers: Prerequisites for Creation and Imaging], *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2017, no. 2 (5), pp. 51-69.
6. V. M. Zuev, O. A. Butov, S. B. Ivanova, A. A. Nikitina, S. I. Ulanov. Sposob obucheniya neyronnoy seti upravleniya robotom [Method for learning neural network for robot control]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2021, no. 2 (21).
7. Koval O.S. Kontseptsíya sistemi komp'yuternogo vírtual'nogo modelyuvannya mobíl'nikh robototekh-níchnikh zasobív neytralízatsíЇ tekhno-yekologíchnikh podíЇ ta viríshennya zadach profesíynogo navchannya [Concept of system of computer virtual simulation of mobile robototechnical means of technical and environmental events neutralization and solving problems of professional training]. *Shtuchnyi intelekt* [Artificial Intelligence], 2019, no. 3-4.
8. Korchazhkina O. M. Yazyk iskusstvennogo myshleniya: neobkhodimost' i vozmozhnost' sozdaniya [The Language of an Artificial Mindset: The Need and the Possibility of Building It]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2020, no 4(19).
9. Mishchenko A. V. Son budushchego [The future dream]. *Indul'gentsiya lyudey* [Indulgentia of the Humans], URL: https:// alesmishchenko.github.io/indulgencia
10. Dorokhina G. V. Trebovaniya k informatsionnoy tekhnologii tsifrovogo sbora, obrabotki i analiza dannykh [Requirements for information technology for digital data collection, processing and analysis] *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2020, no. 4 (19)
11. Mishchenko A.V. Vtoroj zakon robototehniki i proxy-totalitarism [The second law of Robotics and proxy-totalitarism] *Invest-Foresight. 12.2021*. URL: https://www.if24.ru/vtoroj-zakon-robototehniki/
12. Mishchenko A.V. Tsivilizatsiya-lishaynik kak al'ternativa tekhnologicheskoy singulyarnosti [Civilization-lichen as an alternative to technological singularity]. *Invest-Foresight. 08.01.2020*, URL: https://www.if24.ru/tsivilizatsiya-lishajnik/
13. Mishchenko A.V. Civilization after humans. St.Petersburg: Golod publishers. 2004.
14. Sontag E. D. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, New York, Springer-Verlag, 1998
15. Mishchenko A.V. Komp'yuternoye modelirovaniye evolyutsii tsivilizatsii v ramkakh futurologicheskoy teorii myslyashchey materii [Computer modeling of the evolution of civilization within the futurological theory of mind-matter]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2021, no. 3 (22). URL: http://paijournal.guiaidn.ru/download_pai/2021_3/1_Мищенко.pdf
16. Kurzweil R. *The Singularity is Near*, London, Penguin Group, 2005.
17. Gasparyan D.E. Sub"yektivnyy opyt, iskusstvennyy intellekt i problema modelirovaniya smyslov [Subjective Experience, Artificial Intelligence and the Problem of Sense Modeling]. *Filosofskiye nauki* [Philosophical Sciences], 2017, no 4
18. Roli, A., Jaeger, J., & Kauffman, S. How organisms come to know the world: fundamental limits on artificial general intelligence. *OSF Preprints, 31. 10. 2021.* URL: https://doi.org/10.31219/osf.io/yfmt3

RESUME

A. V. Mishchenko

Three Real Laws of Robotics. The Law of Stability of a Hybrid Civilization

One of the major technological breakthroughs at the end of the 20th century is that artificial intelligence (AI) is no longer the subject of exclusively science fiction, theoretical reasoning, or educational examples. Moreover, in the 21st century, AI has become the leader of technological progress, determining the development of big data, biological, cognitive, and many other technologies.

At the same time, there is still no understanding of the general direction of development of AI on a scale of decades, as well as an awareness of its globality and fatefulness. In fact, the level of understanding of interaction between humans and independent AI remains at the level of the Asimov's laws of robotics. Similarly, the level of understanding of possible operation principles of independent AI remains at the level of philosophical questions such as "can AI think / feel like a human", which were also popular in the science fiction of the 20th century. In a further perspective, on the scale of the next generations, in fact, the only formalism for modeling the development of a "no-longer-only-human" civilization is the theory of "mind-matter".

This article discusses the "real laws of robotics" (the laws by which, potentially, human-AI interaction can evolve). The main attention is paid to the law of stability and controllability of the hybrid system "society + AI".

РЕЗЮМЕ

А. В. Мищенко

Три реальные закона робототехники. Закон устойчивости гибридной цивилизации

Одним из главных технологических прорывов конца XX века является то, что искусственный интеллект (ИИ) перестал быть темой только фантастики, теоретических рассуждений и учебных примеров. Более того, в XXI веке, ИИ стал флагманом технологического прогресса, определяющим развитие технологий больших данных, биологических, когнитивных, и многих других.

В то же время, всё ещё отсутствует как понимание общего направления развития ИИ в масштабе десятилетий, так и осознание его глобальности и судьбоносности. Фактически, уровень понимания взаимодействия человека и самостоятельного ИИ остался на уровне знаменитых законов робототехники Азимова, а уровень понимания возможных принципов функционирования самостоятельного ИИ – на уровне философских вопросов типа «сможет ли ИИ мыслить/чувствовать как человек», которыми также изобиловала фантастика прошлого века. В более дальней перспективе, в масштабе ближайших поколений, фактически единственным формализмом для моделирования развития «уже-не-только-человеческой» цивилизации является теория «мыслящей материи».

В этой статье обсуждаются «реальные законы робототехники», то есть, законы, по которым, потенциально, может развиваться взаимодействие человека и ИИ. Основное внимание уделяется закону устойчивости и управляемости гибридной системы «общество+ИИ».

Статья поступила в редакцию 17.09.2021.