



TOT SAMYI
МАГАЗИН

[Искусственный интеллект. Этапы. Угрозы. Стратегии](#)

Выбрать

Рациональные и убедительные доводы Ника Бострома об опасности создания искусственного интеллекта заставят задуматься даже заядлых скептиков.

Евгений Касперский,
генеральный директор «Лаборатории Касперского»

Ник Бостром

Искусственный интеллект

ЭТАПЫ. УГРОЗЫ. СТРАТЕГИИ

Оглавление

Предисловие партнера	11
Неоконченная история о воробьях	14
Введение	16
Глава первая. Прошлые достижения и сегодняшние возможности.	19
Модели роста и история человечества	19
Завышенные ожидания	22
Путь надежды и отчаяния	25
Последние достижения	34
Будущее искусственного интеллекта — мнение специалистов	45
Глава вторая. Путь к сверхразуму	50
Искусственный интеллект	51
Полная эмуляция головного мозга человека	61
Усовершенствование когнитивных способностей человека	71
Нейрокомпьютерный интерфейс	85
Сети и организации	91
Резюме	94
Глава третья. Типы сверхразума	96
Скоростной сверхразум	97
Коллективный сверхразум	99
Качественный сверхразум	103
Прямая и опосредованная досягаемость	105
Источники преимущества цифрового интеллекта	106
Глава четвертая. Динамика взрывного развития интеллекта	111
Время и скорость взлета	111
Сопrotивляемость	117
Пути, не подразумевающие создания машинного интеллекта	117
Пути создания имитационной модели мозга и искусственного интеллекта	119
Сила оптимизации и взрывное развитие интеллекта	128
Глава пятая. Решающее стратегическое преимущество	133
Получит ли лидирующий проект абсолютное преимущество?	134
Насколько крупным будет самый перспективный проект?	138
Система контроля	140
Международное сотрудничество	143
От решающего преимущества — к синглтону	144

Глава шестая. Разумная сила, не имеющая себе равной	149
Функциональные возможности и непреодолимая мощь	150
Сценарий захвата власти сверхразумом	156
1. Фаза приближения к критическому моменту	157
2. Рекурсивная фаза самосовершенствования.	157
3. Фаза скрытой подготовки.	157
4. Фаза открытой реализации	158
Власть над природой и другими действующими силами	162
Глава седьмая. Намерения сверхразума	169
Связь между интеллектом и мотивацией	169
Инструментальная конвергенция	175
Самосохранение	175
Непрерывная последовательность целей	176
Усиление когнитивных способностей	178
Технологическое совершенство.	180
Получение ресурсов.	181
Глава восьмая. Катастрофа неизбежна?	184
Экзистенциальная катастрофа как неизбежное следствие взрывного развития искусственного интеллекта?	184
Вероломный ход	186
Пагубные отказы	191
Порочная реализация	192
Инфраструктурная избыточность	195
Преступная безнравственность	201
Глава девятая. Проблемы контроля	203
Две агентские проблемы.	203
Методы контроля над возможностями	206
Изоляционные методы	207
Стимулирующие методы	209
Методы задержки развития	216
Методы «растяжек»	218
Методы выбора мотивации	220
Метод точной спецификации.	221
Метод приручения	225
Метод косвенной нормативности.	226
Метод приумножения	227
Резюме.	228
Глава десятая. Оракулы, джинны, монархи и инструменты	230
Оракулы	230
Джинны и монархи.	234

ИИ-инструменты	238
Сравнительная характеристика	245
Глава одиннадцатая. Сценарии многополярного мира	249
О лошадях и людях	250
Заработная плата и безработица	250
Капитал и социальное обеспечение	252
Мальтузианские условия в исторической перспективе	254
Рост населения и инвестиции	256
Жизнь в цифровом мире	259
Добровольное рабство, случайная смерть	260
В высшей степени тяжелый труд как высшая степень счастья.	264
Аутсорсеры, лишенные сознания?	267
Эволюция — путь наверх или не обязательно?	270
А потом появится синглтон?	275
Второй переход	275
Суперорганизмы и эффект масштаба	277
Объединение на договорных началах	280
Глава двенадцатая. Выработка ценностей	287
Проблема загрузки системы ценностей	287
Естественный отбор	290
Обучение с подкреплением	292
Ассоциативная модель ценностного приращения	293
Строительные леса для мотивационной системы	296
Обучение ценностям	298
Вариации имитационной модели	308
Институциональное конструирование	310
Резюме	317
Глава тринадцатая. Выбор критериев выбора	319
Необходимость в косвенной нормативности	319
Когерентное экстраполированное волеизъявление	322
Некоторые комментарии	323
Целесообразность КЭВ.	325
Дополнительные замечания	329
Модели, основанные на этических принципах	331
Делай то, что я имею в виду	335
Перечень компонентов	337
Описание цели	338
Принятие решений	339
Эпистемология, или Познание мира	341
Ратификация, или Подтверждение	343
Выбор правильного пути	345

Глава четырнадцатая. Стратегический ландшафт	347
Стратегия научно-технологического развития	348
Различные темпы технологического развития	348
Предпочтительный порядок появления	350
Скорость изменений и когнитивное совершенствование	353
Технологические связи	358
Аргументация от противного	361
Пути и возможности	363
Последствия прогресса в области аппаратного обеспечения	363
Следует ли стимулировать исследования в области полной эмуляции головного мозга?	366
С субъективной точки зрения — лучше быстрее	371
Сотрудничество	372
Гонка и связанные с ней опасности	372
О пользе сотрудничества	376
Совместная работа	381
Глава пятнадцатая. Цейтнот	384
Крайний срок философии	384
Что нужно делать?	386
В поисках стратегии	387
В поисках возможностей	388
Конкретные показатели	389
Все лучшее в человеческой природе — шаг вперед!	390
Примечания	392
Библиография	459
Список сокращений	488
Благодарности	489
Об авторе	491

Предисловие партнера

...У меня есть один знакомый, — сказал Эдик. — Он утверждает, будто человек — промежуточное звено, необходимое природе для создания венца творения: рюмки коньяка с ломтиком лимона.

Аркадий и Борис Стругацкие. Понедельник начинается в субботу

Компьютеры, а точнее алгоритмы, опирающиеся на непрерывно растущие вычислительные мощности, лучше людей играют в шахматы, шашки и нарды. Они очень неплохо водят самолеты. Они смогли пройти тест Тьюринга, убедив судей в своей «человечности». Однажды таксист в Дублине — городе, где расположены европейские штаб-квартиры многих глобальных IT-компаний, — сказал мне, что приветствует бурное развитие технологического сектора своей страны, но потом с сожалением добавил: «Одна беда — из-за этих умных ребят довольно скоро таксисты будут не нужны». Автомобили без водителей, управляемые компьютерами, уже проходят испытания на обычных дорогах в нескольких странах. По мнению философа Ника Бострома, чью книгу вы держите в руках, — все это звенья одной цепи и довольно скоро из-за развития компьютерных технологий нам всем, человеческому роду, может прийти конец.

Автор считает, что смертельная угроза связана с возможностью создания искусственного интеллекта, превосходящего человеческий разум. Катастрофа может разразиться как в конце XXI века, так и в ближайшие десятилетия. Вся история человечества показывает: когда происходит столкновение представителя нашего вида, человека разумного, и любого другого, населяющего нашу планету, побеждает тот, кто умнее. До сих пор умнейшими были мы, но у нас нет гарантий, что так будет длиться вечно.

Ник Бостром пишет, что если умные компьютерные алгоритмы научатся самостоятельно делать еще более умные алгоритмы, а те, в свою очередь, еще более умные, случится взрывной рост искусственного интеллекта, по сравнению с которым люди будут выглядеть приблизительно как сейчас муравьи рядом с людьми, в интеллектуальном смысле, конечно. В мире появится новый, хотя и искусственный, но сверхразумный вид. Неважно, что ему «придет в голову», попытка сделать всех людей счастливыми или решение остановить антропогенное загрязнение мирового океана наиболее эффективным путем, то есть уничтожив человечество, — все равно сопротивляться этому у людей возможности не будет. Никаких шансов на противостояние в духе кинофильма про Терминатора, никаких перестрелок с железными киборгами. Нас ждет шах и мат — как в поединке шахматного компьютера «Дип Блю» с первоклассником.

За последнюю сотню-другую лет достижения науки у одних пробуждали надежду на решение всех проблем человечества, у других вызывали и вызывают безудержный страх. При этом, надо сказать, обе точки зрения выглядят вполне оправданными. Благодаря науке побеждены страшные болезни, человечество способно сегодня прокормить невиданное прежде количество людей, а из одной точки земного шара можно попасть в противоположную меньше чем за сутки. Однако по милости той же науки люди, используя новейшие военные технологии, уничтожают друг друга с чудовищной скоростью и эффективностью.

Подобную тенденцию — когда быстрое развитие технологий не только приводит к образованию новых возможностей, но и формирует небывалые угрозы, — мы наблюдаем и в области информационной безопасности. Вся наша отрасль возникла и существует исключительно потому, что создание и массовое распространение таких замечательных вещей, как компьютеры и интернет, породило проблемы, которые было бы невозможно вообразить в докомпьютерную эру. В результате появления информационных технологий произошла революция в человеческих коммуникациях. В том числе ею воспользовались разного рода киберпреступники. И только сейчас человечество начинает постепенно осознавать новые риски: все больше объектов физического мира управляются с помощью компьютеров и программного обеспечения, часто несовершенного, дырявого и уязвимого; все большее число таких объектов имеют связь с интернетом, и угрозы

кибермира быстро становятся проблемами физической безопасности, а потенциально — жизни и смерти.

Именно поэтому книга Ника Бострома кажется такой интересной. Первый шаг для предотвращения кошмарных сценариев (для отдельной компьютерной сети или всего человечества) — понять, в чем они могут состоять. Бостром делает очень много оговорок, что создание искусственного интеллекта, сравнимого с человеческим разумом или превосходящего его, — искусственного интеллекта, способного уничтожить человечество, — это лишь вероятный сценарий, который может и не реализоваться. Конечно, вариантов много, и развитие компьютерных технологий, возможно, не уничтожит человечество, а даст нам ответ на «главный вопрос жизни, Вселенной и всего такого» (возможно, это и впрямь окажется число 42, как в романе «Автостопом по Галактике»). Надежда есть, но опасность очень серьезная — предупреждает нас Бостром. На мой взгляд, если вероятность такой экзистенциальной угрозы человечеству существует, то отнестись к ней надо соответственно и, чтобы предотвратить ее и защититься от нее, следует предпринять совместные усилия в общемировом масштабе.

Завершить свое вступление хочется цитатой из книги Михаила Веллера «Человек в системе»:

Когда фантастика, то бишь оформленная в образы и сюжеты мысль человеческая, долго и детально что-то повторяет — ну так дыма без огня не бывает. Банальные голливудские боевики о войнах людей с цивилизацией роботов несут в себе под шелухой коммерческого смотра горькое зернышко истины.

Когда в роботы будет встроена передаваемая программа инстинктов, и удовлетворение этих инстинктов будет встроено как безусловная и базовая потребность, и это пойдет на уровень самовоспроизводства — вот тогда, ребята, кончай бороться с курением и алкоголем, потому что будет самое время выпить и закурить перед ханой всем нам.

*Евгений Касперский,
генеральный директор «Лаборатории Касперского»*

Обратите внимание!



[Перегляд позитивного мислення](#)



[Психологія. Люди, концепції, експерименти](#)



[Парадокс страсти. Она его любит, а он ее нет](#)



[Санта действительно существует? Философское расследование](#)



[Голая статистика. Самая интересная книга о самой скучной науке](#)



[Чарунки долі](#)



[В активном поиске](#)



[Женщина. Руководство для мужчин](#)



[Книга радости. Как быть счастливым в меняющемся мире](#)