**Э. Юдковский**

# Искусственный интеллект как позитивный и негативный фактор глобального риска

**Eliezer Yudkowsky**

# Artificial Intelligence as a Positive and Negative Factor in Global Risk

Выходит в 2007 году в сборнике *Риски глобальной катастрофы* под редакцией Ника Бострома и Милана Цирковича, Оксфорд.

Оригинал: http://www.singinst.org/reading/corereading/

Forthcoming in *Global Catastrophic Risks,* eds. Nick Bostrom and Milan Cirkovic
Draft of August 31, 2006
Eliezer Yudkowsky
(yudkowsky@singinst.org)
Singularity Institute for Artificial Intelligence
Palo Alto, CA

## Введение [(1)]

До сих пор основной опасностью искусственного интеллекта (ИИ) было то, что люди слишком рано делали вывод, что они его понимают. Разумеется, эта проблема не ограничена ИИ. Jacques Monod пишет: «забавным аспектом теории эволюции является то, что каждый думает, что понимает её». (Monod, 1974.) Мой отец, физик, жаловался на людей, придумывавших свои собственные физические теории: «интересно знать, почему люди не придумывают свои собственные теории химии?» (Но они делают.) Тем не менее, проблема является особенно актуальной в области ИИ. Область ИИ имеет репутацию того, что она даёт огромные обещания и не исполняет их. Большинство наблюдателей считают, что ИИ

[1] Сноска (1): Я благодарю Michael Roy Ames, Eric Baum, Nick Bostrom, Milan Cirkovic, John K Clark, Emil Gilliam, Ben Goertzel, Robin Hanson, Keith Henson, Bill Hibbard, Olie Lamb, Peter McCluskey и Michael Wilson за их комментарии, предложения и критику. Нет необходимости говорить, что все оставшиеся ошибки в этой статье – мои.

труден, и это на самом деле так. Но запутанность не происходит из трудности. Трудно сделать звезду из водорода, но звездная астрофизика не имеет ужасающей репутации обещать сделать звезду и затем не смочь. Критическим выводом является не то, что ИИ труден, а то, что, по неким причинам, людям очень легко думать, что они знают об искусственном интеллекте гораздо больше, чем на самом деле.

В моей другой статье о *рисках глобальной катастрофы* «**Систематические ошибки мышления, потенциально влияющие на суждения о глобальных рисках**», я начинаю с замечания, что немногие люди предпочли бы нарочно уничтожить мир; сценарий же уничтожения Земли по ошибке кажется мне очень беспокоящим. Немногие люди нажмут кнопку, которая, как они точно знают, вызовет глобальную катастрофу. Но если люди склонны быть абсолютно уверены, что кнопка делает нечто, совершенно отличное от её реального действия, это действительно причина для тревоги.

Гораздо труднее писать о глобальных рисках искусственного интеллекта, чем о систематических ошибках мышления. Ошибки мышления – это твёрдо установленное знание; достаточно процитировать литературу. ИИ – это не твёрдо установленное знание; ИИ относится к передовым исследованиям, а не к учебникам. И, по причинам, объясняющимся в следующей главе, проблема *глобальных рисков* в связи с искусственным интеллектом фактически не обсуждается в существующей технической литературе.

Я вынужден анализировать тему со своей точки зрения, делать мои собственные выводы и делать всё, от меня зависящее, чтобы доказать их в ограниченном пространстве этой статьи. Дело не в том, что я пренебрегаю необходимостью цитировать существующие источники на эту тему, но в том, что таких источников, несмотря на все мои попытки их найти, обнаружить не удалось (на январь 2006 года).

Соблазнительно игнорировать ИИ в этой книге, потому что это наиболее трудная тема для обсуждения. Мы не можем обратиться к статистическим данным, чтобы вычислить маленькую годовую вероятность катастрофы, как в случае астероидных ударов. Мы не можем использовать вычисления на основании точных, точно подтверждённых моделей, чтобы исключить некие события или установить бесконечно малые верхние границы их вероятности, как в случае возможных физических катастроф. Но это делает катастрофы с ИИ ещё более беспокоящими, а не менее.

Эффекты систематических ошибок мышления, как оказалось, имеют тенденцию *увеличиваться* при недостатке времени, занятости ума или недостатке информации. Это говорит, что *чем труднее аналитическая задача*, тем важнее избежать или ослабить систематическую ошибку. Поэтому я *усиленно рекомендую* прочесть статью «Систематические ошибки мышления, потенциально влияющие на оценку глобальных рисков» (http://www.proza.ru/texts/2007/03/08-62.html) до прочтения этой статьи.

**1. Систематическая ошибка, связанная с антропоморфизмом.**

Когда нечто очень широко распространено в нашей повседневной жизни, мы принимаем это как само собой разумеющееся вплоть до того, что забываем о существовании этого. Представьте себе сложную биологическую адаптацию, состоящую из 10 необходимых частей.

Если каждый из 10 генов независим и имеет 50% частоту в наборе генов - то есть каждый ген имеется только у половины особей вида – тогда в среднем только одна особь из 1024 будет обладать полнофункциональной адаптацией. Меховая шуба не является значительным эволюционным приобретением, пока окружающая среда не начнёт подвергать организмы отбору холодом. Точно так же, если ген Б зависит от гена А, тогда ген Б не имеет значительного преимущества, пока ген А не станет надёжной частью *генетического* окружения. *Сложное*, *взаимозависимое* устройство должно быть у *всех* сексуально воспроизводящихся видов; оно не может развиться в противном случае. (Tooby и Cosmides, 1992.) Одна малиновка может иметь более гладкие перья, чем другая, но у обеих должны быть крылья. Естественный отбор, двигаемый разнообразием, сужает это разнообразие. (Sober, 1984.) В каждой известной культуре люди испытывают грусть, отвращение, ярость, страх и удивление (Brown, 1991), и передают эти эмоции одними и теми же выражениями лица. У нас у всех под капотом один и тот же мотор, хотя мы можем и быть раскрашены разными красками; этот принцип эволюционные психологи называют *психическим единством человечества*. (Tooby and Cosmides, 1992). Это описание и объясняется, и требуется законами эволюционной биологии.

Антрополог не будет восторженно писать о новооткрытом племени: «Они едят еду! Они дышат воздухом! Они используют инструменты! Они рассказывают друг другу истории!» Мы, люди, забываем, как мы подобны друг другу, живя в мире, который напоминает нам только о наших различиях.

Люди научились моделировать других людей, - чтобы соревноваться и кооперироваться со своими сородичами. Это было надёжным инструментом в мире наших предков, где любой сильный ум, который вам попадался, был тоже человеком. Мы развили способность понимать наших ближних путём *эмпатии*, помещая себя на их место; для этого то, что моделируется, должно быть похоже на моделирующего. Не удивительно, что люди часто очеловечивают, – то есть ожидают человекоподобных качеств от того, что не является человеком. В фильме «Матрица» (братья Вачовские, 1999) представитель искусственного интеллекта Агент Смит вначале кажется совершенно холодным и собранным, его лицо неподвижно и неэмоционально. Но позже, допрашивая человека Морфеуса, Агент Смит даёт выход свому отвращению к человечеству – и его лицо выражает общечеловеческое выражение отвращения. Опрашивание своего собственного ума работает хорошо, в качестве инстинкта адаптации, когда вам нужно предсказывать других людей.

Но если вы исследуете некий другой процесс оптимизации, - если вы, например, теолог 18 века William Paley – то тогда антропоморфизм – это липучка для мух для неосторожных учёных, столь липкая западня, что нужен Дарвин, чтобы из неё выбраться.

Эксперименты по исследованию антропоморфизма показали, что испытуемые часто антропоморфизируют неосознанно, вопреки своим базовым установкам. Barrett и Keil (1996) провели эксперименты на субъектах, исповедовавших веру в неантропоморфные качества Бога – что Бог может быть более чем в одном месте одновременно, или одновременно наблюдать множество вещей. Barrett and Keil предложили этим испытуемым истории, в которых Бог спасает людей от утопления. Испытуемые отвечали на вопросы об историях или пересказывали их своими словами, таким образом, что это предполагало, что Бог был только в одном месте в одно время и выполнял задачи последовательно, а не параллельно. К счастью для целей нашего исследования, Barrett и Keil в другой группе использовали в прочих отношениях аналогичные истории о суперкомпьютере по имени "Uncomp". Например, чтобы изобразить свойство всезнания, говорилось, что сенсоры Uncomp’а покрывают каждый квадратный сантиметр земли, и никакая информация не теряется. Испытуемые в этих условиях всё равно демонстрировали сильный антропоморфизм, хотя и значительно меньший, чем в «группе Бога». С нашей точки зрения, главным результатом является то, что хотя люди сознательно полагали ИИ не подобным человеку, они по-прежнему представляли себе такие сценарии, как если бы ИИ был человекоподобным (хотя и не настолько человекоподобным, как Бог).

Ошибка антропоморфизма подкрадывается незаметно: она происходит без нарочного намерения, не осознанно и вопреки очевидному знанию.

В эпоху бульварной научной фантастики обложки журналов часто изображали монструозного инопланетянина – собирательно известного как жукоглазый монстр (ЖГМ) – тащащего привлекательную полуобнажённую женщину. Может показаться, что художник верил, что негуманоидный инопланетянин, с полностью другой эволюционной историей, может сексуально желать женщину-человека. Такие ошибки происходят не из-за того, что люди явным образом рассуждают подобно следующему: «Все умы, скорее всего, возбуждаются похожим образом, и поэтому, вероятно, ЖГМ находит женщину-человека сексуально привлекательной». Скорее, художник просто не *задался* вопросом о том, действительно ли гигантский жук *воспринимает* женщин-людей привлекательными. Наоборот, полуобнажённая женщина *является сексуальной* – изначально, потому что это неотъемлемо присущее ей свойство. Те, кто делают эту ошибку, не думают об уме насекомообразного существа; они концентрируются на задранных одеждах женщины. Если бы одежды не были задраны, женщина была бы менее сексуальна, но ЖГМ этого не понимает. (Это частный случай глубокой, запутывающей и чрезвычайно распространённой ошибки, которую E. T. Jaynes назвал ошибочностью, связанной с умственной проекцией *(*mind projection fallacy). (Jaynes and Bretthorst, 2003.) Jaynes, специалист по байесовской теории достоверности, определил «ошибочностью, связанную с умственной проекцией» как ошибку, связанную с тем, что состояния знания перепутаны со свойствами объектов. Например, фраза «*мистический феномен*» подразумевает, что мистичность – это свойство самого феномена. Если я неосведомлен относительно некого феномена, то это факт о моём состоянии сознания, а не о самом феномене.)

Людям нет нужды понимать, что они антропоморфизируют (или хотя бы понимать, что они вовлечены в сомнительный акт предсказания состояния чужого ума) для того, чтобы антропоморфизм повлиял на мышление. Когда мы пытаемся рассуждать о чужом сознании, каждый шаг рассуждений может быть соединён с предположениями, настолько очевидными для человеческого опыта, что мы обращаем на них внимания не больше, чем на воздух или гравитацию. Вы возражаете журнальному иллюстратору: «Не является ли более правдоподобным, что огромный жук-самец будет сексуально желать огромных жуков-самок?» Иллюстратор немного подумает и скажет: «Но даже если бы инопланетные инсектоиды начинали с любви к твёрдым экзоскелетам, после того, как инсектоид повстречает женщину-человека, он вскоре поймёт, что у неё гораздо более мягкая и нежная

кожа. Если у инопланетян имеется достаточно продвинутая технология, они могут генетически изменить себя, чтобы любить мягкую кожу, а не твёрдые экзоскелеты».

Это - ошибочность-один-шаг-назад (fallacy-at-one-remove). После того, как указано на антропоморфичность мышления инопланетянина, журнальный иллюстратор делает шаг назад и пытается представить умозаключения инопланетянина как нейтральный продукт его мышления. Возможно, продвинутые инопланетяне могут перестроить себя (генетически или как-то иначе), чтобы любить мягкую кожу, но захотят ли они? Инопланетянин-инсектоид, любящий жёсткие скелеты, не будет хотеть переделать себя, чтобы любить мягкую кожу вместо этого, – кроме как в случае, если естественный отбор каким-то образом породит в нём определённо человеческое чувство метасексуальности. При использовании длинных сложных цепочек рассуждений в поддержку антропоморфических выводов, каждый шаг таких рассуждений является ещё одной возможностью, чтобы прокралась ошибка.

И ещё одной серьёзной ошибкой является начинать с вывода и искать кажущуюся нейтральной линию рассуждений, ведущую к нему; это называется рационализацией. Если первое, что приходит на ум, при вопросе на эту тему, это образ инсектоида, преследующего женщину-человека, то тогда антропоморфизм является первопричиной этого восприятия, и никакое количество рационализации не изменит этого.

Любой, кто бы хотел уменьшить систематическую ошибку антропоморфизма в себе, должен был бы изучить эволюционную биологию для практики, желательно, эволюционную биологию с математическими выкладками. Ранние биологи часто очеловечивали естественный отбор – они полагали, что эволюция будет делать тоже, что и они сами; они пытались предсказать эффекты эволюции, ставя себя на её место. В результате получался по большей части нонсенс, который начали *изгонять* из биологии только в поздние 1960-е годы, например, это делал Williams (1966). Эволюционная биология предлагает обучение на основе как математики, так и конкретных примеров, помогающие выбить из себя ошибку очеловечивания.

**1.1: Широта пространства возможных устройств ума. (The width of mind design space).**

Эволюция жёстко сохраняет некоторые структуры. В той мере, как развитие других генов опирается на ранее существовавший ген, этот ранний ген полностью цементируется: он не может мутировать, не нарушая множество форм адаптации. Гомеотические (Homeotic) гены – гены, контролирующие развитее структуры тела эмбриона – говорят множеству других

генов, когда активироваться. Мутация гомеотического гена может привести к тому, что эмбрион плодовой мушки разовьётся нормально, за исключением того, что у него не будет головы. В результате гомеотические гены столь точно сохраняются, что многие из них одни и те же у человека и плодовой мушки – они не изменились со времён последнего общего предка человека и насекомых. Молекулярные механизмы синтеза АТФ по существу одни и те же в митохондриях животных, хлоропластах растений и у бактерий; синтез АТФ не претерпел значительных изменений с развития эукариотов 2 миллиарда лет назад.

Любые два устройства ИИ могут быть менее похожи друг на друга, чем вы и садовый цветок петуния.

Термин ИИ относится к гораздо большему *пространству возможностей*, чем термин "Homo sapiens". Когда мы говорим о разных ИИ, мы говорим об умах вообще, или о процессах оптимизации вообще. Представьте себе карту возможных устройств ума. В одном углу маленький кружочек означает всех людей. И вся эта карта находится внутри ещё большего пространства, *пространства процессов оптимизации*. Естественный отбор создаёт сложные функционирующие механизмы не привлекая процесса думания; эволюция находится внутри пространства процессов оптимизации, но за пределами пространства умов.

Этот *гигантский* круг возможностей исключает антропоморфизм как законный способ мышления.

**2: Предсказание и устройство. (Prediction and design).**

Мы не можем спрашивать наш собственный мозг о нечеловеческих процессах оптимизации – ни о насекомоглазых монстрах, ни о естественном отборе, ни об искусственном интеллекте. И как же мы будем продолжать? Как мы можем предсказать, что ИИ будет делать? Я нарочно задаю этот вопрос в форме, которая делает его труднообрабатываемым. При такой постановке проблемы невозможно предсказать, будет ли произвольная вычислительная система выполнять хоть какие-нибудь функции ввода-вывода, включая, например, простое умножение (Rice, 1953.) Так как же возможно, что компьютерные инженеры могут создавать микросхемы, которые надёжно выполняют вычисления? Потому что люди-инженеры нарочно используют те проекты, которые они могут понять.

Антропоморфизм заставляет людей верить, что они могут делать предсказания, не имея никакой другой информации, кроме как о самом факте «интеллектуальности» (intelligence) чего-то – антропоморфизм продолжает генерировать предсказания, не взирая ни на что, в то время как ваш мозг автоматически ставит себя на место этой самой «интеллектуальности». Это может быть одним из факторов вызывающей замешательство истории ИИ, которая происходит не из трудности ИИ как такового, но из загадочной лёгкости обретения ошибочной веры в то, что некий данный дизайн ИИ сработает.

Для того, чтобы сделать утверждение о том, что мост выдержит вес автомобилей в 30 тонн, гражданские инженеры имеют два оружия: выбор изначальных условий и запас прочности для безопасности. Им нет необходимости предсказывать, может ли выдержать вес в 30 тонн произвольная конструкция, но только проект данного конкретного моста, относительно которого они делают это заявление. И хотя это показывает с лучшей стороны инженера, который может вычислить точный вес, который мост может выдержать, также приемлемо вычислить, что мост выдержит автомобили не менее, чем в 30 тонн – хотя для того, чтобы доказать это расплывчатое утверждение строго, может потребоваться большая часть того теоретического понимания, которое входит в точное вычисление.

Гражданские инженеры придерживаются высоких стандартов в предсказании того, что мосты выдержат нагрузку. Алхимики прошлого придерживались гораздо более низких стандартов в предсказании того, что последовательность химических реагентов трансформирует свинец в золото. Какое количество свинца в какое количество золота? Каков причинный механизм этого процесса? Вполне понятно, почему исследователь-алхимик хотел золото больше, чем свинец, но почему данная последовательность реагентов превращает свинец в золото, а не золото в свинец или свинец в воду?

Ранние исследователи ИИ полагали, что искусственная нейронная сеть из слоёв пороговых устройств, обученная посредством обратного распространения, будет «интеллектуальной» (intelligent). Использованное при этом мышление, обусловленное результатом (wishful thinking), ближе к алхимии, чем к гражданском строительству. Магия входит в список человеческих универсалий Дональда Брауна (Brown, 1991); наука – нет. Мы инстинктивно не понимаем, что алхимия не работает. Мы инстинктивно не различаем строгие рассуждения и хорошее рассказывание историй. Мы инстинктивно не замечаем ожидание положительных результатов, висящее в воздухе. Человеческий вид возник посредством естественного отбора, функционирующего посредством неслучайного сохранения случайных мутаций.

Один из путей к глобальной катастрофе – когда кто-то нажимает кнопку, имея ошибочное представление о том, что эта кнопка делает – когда ИИ возникнет посредством подобного сращения работающих алгоритмов, с исследователем, не имеющим глубокого понимания, как вся система работает. Нет сомнения, они верят, что ИИ будет дружественным, без ясного представления о точном процессе, вовлечённом в создание дружественного поведения, или какого-либо детального понимания того, что они имеют в виду под дружественностью. Несмотря на то, что ранние исследователи ИИ имели сильно ошибочные, расплывчатые ожидания об интеллектуальности своих программ, мы можем представить, что этим исследователям ИИ удалось сконструировать интеллектуальную программу, но они имели сильно ошибочные расплывчатые ожидания относительно дружественности своих программ.

Не знание того, как сделать дружественный ИИ, не смертельно само по себе, в том случае, если вы знаете, что вы не знаете. Именно *ошибочная* вера в то, что ИИ будет дружественным, означает очевидный путь к глобальной катастрофе.

**3: Недооценка силы интеллекта. (Underestimating the power of intelligence).**

Мы склонны видеть индивидуальные различия вместо общечеловеческих качеств. Поэтому, когда кто-то говорит слово «интеллект» (intelligence), мы думаем скорее об Эйнштейне, чем о людях. Индивидуальные различия в человеческом интеллекте имеют стандартное обозначение, известные как G-фактор Шпеермана (Spearman's G-factor), это - спорная интерпретация твёрдых экспериментальных фактов о том, что различные тесты интеллекта высоко коррелируют друг с другом, а также с результатами в реальном мире, такими, как суммарный доход за жизнь. (Jensen, 1999.) G-фактор Шпеермана является статистической абстракцией индивидуальных различий в интеллекте между людьми, которые, как вид, гораздо более интеллектуальны, чем ящерицы. G-фактор Шпеермана выводится из миллиметровых различий в высоте среди представителей вида гигантов.

Мы не должны путать G-фактор Шпеермана с *общечеловеческой интеллектуальностью*, то есть нашей способностью обрабатывать широкий круг мыслительных задач, непостижимых для других видов. Общая интеллектуальность – это межвидовое различие, комплексная адаптация и общечеловеческое качество, обнаруживаемое во всех известных культурах. Возможно, ещё нет академического согласия об интеллектуальности, но нет сомнения в

существовании, или силе, такой вещи, которая должна быть объяснена. Есть что-то такое в людях, что позволяет нам оставлять следы ботинок на Луне.

Но слово «интеллектуальность» обычно вызывает образы голодающего профессора с IQ в 160 единиц и миллиардера-главу компании с IQ едва ли в 120. В действительности, существуют различия в индивидуальных способностях помимо качеств из «книжек про карьеру», которые влияют на относительный успех в человеческом мире: энтузиазм, социальные навыки, музыкальные таланты, рациональность. Отметьте, что каждый из названных мною факторов является когнитивным. Социальные навыки присущи мозгу, а не печени. И – шутки в сторону – вы не обнаружите много глав компаний, ни даже профессоров академии, которые были бы шимпанзе. Вы не обнаружите много ни прославленных мыслителей, ни художников, ни поэтов, ни лидеров, ни опытных социальных работников, ни мастеров боевых искусств, ни композиторов, которые были бы мышами. Интеллектуальность – это основание человеческой силы, мощь, которая наполняет другие наши искусства.

Опасность перепутать общую интеллектуальность с g-фактором состоит в том, что это ведёт к колоссальной недооценки потенциального воздействия ИИ. (Это относится как к недооценке потенциально хороших воздействий, равно как и плохих воздействий.) Даже фраза «трансгуманистический ИИ» или «искусственный суперинтеллект» по-прежнему может создавать впечатление о «ящике с книгами как сделать карьеру»: ИИ, который реально хорош в когнитивных задачах, обычно ассоциируется с «интеллектуальностью», подобной шахматам или абстрактной математике. Но не со сверхчеловеческой убедительностью, или со способностью гораздо лучше, чем люди, предсказывать и управлять человеческими институтами, или нечеловечески умом в формулировании длительных стратегий. Так что, может, нам следует подумать не об Эйнштейне, а о политическом и дипломатическом гении 19 века Отто фон Бисмарке? Но это только малая часть ошибки. Весь спектр от деревенского идиота до Эйнштейна, или от деревенского идиота до Бисмарка, уменьшается в маленькую точку на отрезке между амёбой и человеком.

Если слово «интеллектуальность» ассоциируется с Эйнштейном, а не с людьми, то может показаться осмысленным заявление, что интеллектуальность не имеет отношения к ружьям, как если бы ружья росли на деревьях. Может показаться осмысленным заявление о том, что интеллектуальность не имеет ничего общего с деньгами, как если бы мыши использовали деньги. Человеческие существа начинали, не обладая большими активами зубов, когтей, вооружений, или каких-либо других преимуществ, которые были ежедневной валютой для других видов. Если вы взгляните на людей с точки зрения остальной экосферы, не было

никакого намёка на то, что мягкие розовые твари в конце концов закроют себя в бронированные танки. Мы создали поле битвы, на котором мы победили львов и волков. Мы не сражались с ними посредством когтей и зубов; у нас было собственное представление о том, что действительно важно. Такова сила творчества.

Винж (Vinge, 1993) уместно замечает, что будущее, в котором существуют умы, превосходящие человеческие, отличается *качественно*. ИИ – это не удивительный блестящий дорогой гаджет, рекламируемый в свежайших выпусках технических журналов. ИИ не принадлежит к тому же графику, который показывает прогресс в медицине, производстве и энергетике. ИИ – это не то, что вы можете небрежно добавить в *люмпен-футуристический* сценарий будущего с небоскрёбами и летающими машинами и нанотехнологическими красными кровяными клетками, которые позволяют вам задержать дыхание на 8 часов. Достаточно высокие небоскрёбы не могут начать проектировать сами себя. Люди достигли господства на Земле не из-за того, что задерживали дыхание дольше, чем другие виды.

Катастрофический сценарий, произрастающий из недооценки силы интеллекта, заключается в том, что некто создаст кнопку, не достаточно заботясь о том, что эта кнопка делает, потому что он не думает, что эта кнопка достаточно сильна, чтобы повредить ему. Или, поскольку недооценка силы интеллекта ведёт к пропорциональной недооценке силы Искусственного Интеллекта, то (в настоящая время микроскопическая) группа озабоченных исследователей и поставщиков грантов и отдельных филантропов, занимающихся рисками существованию, не будет уделять достаточно внимания ИИ.
Или широкое поле исследований ИИ не будет уделять достаточно внимания рискам сильного ИИ, и в силу этого хорошие инструменты и твёрдые установления для Дружественности окажутся недоступными, когда возникнет возможность создавать мощные интеллекты.

И также следует заметить – поскольку это тоже влияет на глобальные риски – что ИИ может быть мощным решением для других глобальных рисков, и по ошибке мы можем игнорировать нашу лучшую надежду на выживание. Утверждение о недооценке потенциального воздействия ИИ симметрично относительно потенциально хороших и потенциально плохих воздействий. Именно поэтому название этой статьи – «Искусственный интеллект как позитивный и негативный фактор глобального риска», а не «Глобальные риски Искусственного интеллекта». Перспектива ИИ влияет на глобальные риски более сложным образом; если бы ИИ был чистой помехой, ситуация была бы проще.

**4: Способности и мотивы. (Capability and motive).**

Есть один вид ошибочности, часто встречающийся в дискуссиях об ИИ, особенно об ИИ сверхчеловеческих способностей. Кто-нибудь говорит: «Когда технологии продвинутся достаточно далеко, мы будем способны создавать интеллекты, далеко превосходящие человеческие. Очевидно, что размер ватрушки, который вы можете испечь, зависит от вашего интеллекта. Суперинтеллект может создавать гигантские ватрушки – ватрушки, размером с города – боже мой, будущее будет полно гигантских ватрушек!» Вопрос в том, захочет ли суперинтеллект создавать огромные ватрушки. Видение образа ведёт прямо от *возможности* к *реализации*, без осознавания необходимого промежуточного элемента – мотива. Следующие цепочки рассуждений, рассматриваемые в изоляции без подтверждающего доказательства, все являются примером Ошибочности Гигантской Ватрушки:

- Достаточно сильный ИИ может преодолеть любое человеческое сопротивление и истребить человечество. (И ИИ решит сделать это.) Поэтому мы не должны строить ИИ.
- Достаточно сильный ИИ может создать новые медицинские технологии, способные спасти миллионы человеческих жизней. (И он решит сделать это.) Поэтому мы должны создать ИИ.
- Когда компьютеры станут достаточно дёшевы, огромное большинство работ будет выполняться ИИ более легко, чем людьми. Достаточно сильный ИИ даже будет лучше нас в математике, конструировании, музыке, искусстве и во всех других работах, которые нам кажутся важными (И ИИ решит выполнять эти работы.) Таким образом, после изобретения ИИ, людям будет больше нечего делать, и мы будем голодать или смотреть телевизор.

**4.1: Процессы оптимизации. (Optimization processes)**

Вышеприведенный разбор ошибочности Гигантской Ватрушки имеет органически присущий ему антропоморфизм – а именно, идею о том, что мотивы разделимы; подразумеваемое предположение о том, что, говоря о «способностях» и «мотивах», мы разрываем связность реальности. Это удобный срез, но антропоморфический.

Для того, чтобы рассмотреть проблему с более общей точки зрения, я ввёл концепцию *процесса оптимизации*: системы, которая поражает маленькие цели в большом пространстве поиска, чтобы порождать согласованные эффекты в реальном мире.

Процесс оптимизации направляет будущее в определённые регионы возможного. Когда я посещаю удалённый город, мой друг из местных вызывается отвезти меня в аэропорт. Я не знаю окрестностей. Когда мой друг выезжает на перекрёсток, я не могу предсказать его повороты, ни в последовательности, ни по отдельности. Но я могу предсказать результат непредсказуемых действий моего друга: мы прибудем в аэропорт. Даже если дом моего друга находится в другом месте города, так что моему другу придётся совершить совершенно другую последовательность поворотов, я могу с той же степенью уверенности предсказать, куда мы конце концов прибудем. Не странная ли эта ситуация, научно говоря? Я могу предсказать *результат* процесса, будучи неспособным предсказать ни один из его *промежуточных* этапов. Я буду называть область, в которую процесс оптимизации направляет будущее, *целью оптимизации*.

Рассмотрим автомобиль, например, Тойоту Кароллу. Из всех возможных комбинаций атомов, которые её составляют, только бесконечно малая часть будет работающим автомобилем. Если вы будете собирать атомы в случайном порядке, много
*много* возрастов вселенной пройдёт, пока вам удастся собрать автомобиль. Малая доля пространства проектов описывает автомобили, которые мы могли бы признать как более быстрые, более эффективные и более безопасные, чем Королла. Таким образом, Королла не является оптимальной с точки зрения целей своего конструктора. Но Королла является, однако, *оптимизированной*, поскольку конструктор должен был попасть в сравнительно бесконечно малую область в пространстве возможных конструкций, только чтобы создать работающий автомобиль, не говоря уже о машине качества Короллы. Вы не можете даже построить эффективную тележку, распиливая доски случайно и сколачивая их по результатам броска монеты. Чтобы достичь такой малой цели в пространстве конфигураций, необходим мощный оптимизационный процесс.

Понятие о «процессе оптимизации» является *предсказательно полезным*, поскольку легче понять цель процесса оптимизации, чем его пошаговую динамику. Обсуждение Короллы выше неявно *предполагает*, что конструктор Короллы пытался создать «автомобиль», средство транспорта. Это предположение следует сделать явным, но оно не ошибочно и оно очень полезно для понимания Короллы.

### 4.2: Наведение на цель. (Aiming at the target.)

Есть соблазн спросить, что ИИ будет хотеть, забывая о том, что пространство умов-вообще гораздо больше, чем малая человеческая точка. Следует сопротивляться соблазну

распространить количественные ограничения на все возможные умы. Рассказчики историй накручивают сказки об отдалённой и экзотичной земле, называемой Будущее, говоря, каким будущее *должно быть*. Они делают *предсказания*. Они говорят: «ИИ нападёт на людей с помощью армий марширующих роботов» или «ИИ изобретёт лекарство от рака». Они не предлагают сложных отношений между изначальными условиями и результатами – так они могли бы потерять аудиторию. Но мы нуждаемся в понимании соотношений, чтобы управлять будущим, направляя его в область, приятную человечеству. Если не рулить, мы рискуем попасть туда, куда нас занесёт.

Главный вызов состоит не в том, чтобы предсказать, что ИИ атакует людей с помощью армий роботов, или, наоборот, введёт лекарство от рака. Задача состоит даже не в том, чтобы сделать это предсказание для произвольного устройства ИИ. Скорее, задача состоит в том, чтобы выбрать и создать такой процесс оптимизации, чьи позитивные эффекты могут быть твёрдо доказаны.

Я *усиленно* призываю своих читателей не начинать придумывать причины, почему универсальный процесс оптимизации должен быть дружественным. Естественный отбор не является дружественным, ни ненавидит вас, ни оставляет вас в одного. Эволюция не может быть так антропоморфизирована, она не работает, как вы.

Многие биологи до 1960-х годов ожидали, что естественный отбор создаст полный набор всех хороших вещей, и выдумывали всевозможные усложнённый причины, почему он должен сделать это. Они были разочарованы, поскольку естественный отбор сам по себе не начинает со знания, что от него хотят приятного человеку результата, и затем не придумывает сложные пути, чтобы создать приятные результаты, используя давление отбора. Таким образом, события в природе были результатами совершенно других по своим причинам процессов, чем те, что приходили в голову биологам до 1960-х годов, и поэтому предсказания и реальность расходились.

Мышление, привязанное к цели (wishful thinking), добавляет детали, ограничивает предсказания и таким образом отягощает невозможностью. Как насчёт инженера гражданских сооружений, который надеется, что мост не упадёт? Следует ли инженеру доказывать это тем, что мосты обычно не падают? Но природа сама по себе не предлагает разумных причин, почему мосты не должны падать. Скорее, это инженер преодоляет тяжесть недостоверности (burden of improbability) посредством специфического выбора, направляемого специфическим пониманием. Инженер начинает с намерения создать мост. Затем он использует строгую теорию, чтобы выбрать конструкцию моста, которая бы выдерживала автомобили. Затем строит реальный мост, чья структура отражает рассчитанный проект. И в результате реальная структура выдерживает автомобили. Таким

образом достигается гармония предсказанных позитивных результатов и реальных позитивных результатов.

**5: Дружественный ИИ. (Friendly AI).**

Было бы очень здорово, если бы человечество знало, как создать мощный оптимизационный процесс с неким частным результатом. Или, говоря более общими словами, было бы здорово, если бы мы знали, как создать хороший ИИ (nice AI).

Для того, чтобы описать *область знания*, необходимого, чтобы взяться за этот вызов, я предложил термин «Дружественный ИИ». Этот термин я отношу не только к самой методике, но также и к её продукту – то есть к ИИ, созданному со специфической мотивацией. Когда я использую термин Дружественный в любом из этих двух смыслов, я пишу его с большой буквы, чтобы избегать путаницы с обычным смыслом слова «дружественный».

Типичная реакция на это людей, которую я часто встречал, заключалась в немедленном заявлении, что Дружественный ИИ невозможен, потому что любой достаточно сильный ИИ сможет модифицировать свой собственный исходный код так, чтобы разорвать любые наложенные на него ограничения.

Первую логическую несообразность, которую вы тут можете отметить – это ошибочность Гиганстской Ватрушки. Любой ИИ, имеющий свободный доступ к своему исходному коду, в принципе, будет обладать способностью изменить свой код таким образом, что изменится его цель оптимизации. Но это не означает, что ИИ имеет побуждение изменить свои собственные побуждения. Я не стану сознательно глотать пилюлю, которая побудит меня наслаждаться убийствами, потому что я в настоящем предпочитаю, чтобы мои собратья - люди не умирали.

Но что если я попытаюсь изменить себя и сделаю ошибку? Когда компьютерные инженеры доказывают пригодность чипа – что есть хорошая идея, если в чипе 155 миллионов транзисторов, и вы не можете выпустить патч потом – инженеры используют руководимую человеком и проверяемую машинами формальную проверку. Замечательным свойством формального математического доказательства является то, что доказательство из 10 миллиардов шагов в той же мере надёжно, что и доказательство из 10 шагов. Но человеческие существа недостойны доверия в том, чтобы следить за проверкой из 10 миллиардов шагов; у нас слишком высокие шансы пропустить ошибку. Современные техники доказывания теорем не достаточно умны, чтобы спроектировать и проверить целый компьютерный чип сами по себе – современные алгоритмы испытывают экспоненциальный рост по мере увеличения пространства поиска. Люди-математики могут доказывать теоремы гораздо более сложные, чем те, что могут осилить современные программы-доказыватели,

без того, чтобы быть поверженными экспоненциальным взрывом. Но люди-математики неформальны и ненадёжны; время от времени кто-то находит ошибку в принятом ранее неформальном доказательстве. Выход состоит в том, что люди-инженеры направляют программы-доказыватели на *промежуточные* шаги доказательства. Человек выбирает следующую лемму, и сложный доказыватель теорем генерирует формальное доказательство, и простой проверяльщик сверяет шаги. Таким образом современные инженеры создают надёжные механизмы со 155 миллионами независимых частей.

Проверка корректности работы компьютерного чипа требует синергии человеческого интеллекта и компьютерных алгоритмов, поскольку *сейчас* ни того, ни другого недостаточно. Возможно, подлинный ИИ будет использовать подобную *комбинацию способностей*, когда будет модифицировать свой собственный код – будет обладать как способностью вводить объёмные проекты без того, чтобы потерпеть поражение от экспоненциального роста, так и способностью проверить свои шаги с высокой надёжностью. Это один из путей, которым подлинный ИИ может оставаться познаваемо (knowably) стабильным в своих целях даже после выполнения большого количества самоисправлений.

Эта статья не будет разъяснять приведённую выше идею в деталях. (Также см. Schmidhuber 2003 на связанную с данной тему.) Но следует подумать об этом вызове, и изучить его с привлечением наилучших доступных технических данных, до того, как объявлять его невозможным – особенно, если большие ставки зависят от ответа. Неуважительно по отношению к человеческой изобретательности объявлять проблему неразрешимой без внимательного и творческого рассмотрения. Это очень сильное заявление: сказать, что вы не можете сделать нечто – что вы *не можете* построить летающую машину тяжелее воздуха, что вы *не можете* извлечь полезную энергию из ядерных реакций, что вы *не можете* летать на Луну. Такие заявления являются универсальными обобщениями, относящимися к любому возможному подходу к решению этой проблемы, который кто-либо придумал или придумает. Требуется всего один противоположный пример, чтобы опровергнуть универсальное обощение. Утверждение о том, что Дружественный (или дружественный) ИИ *теоретически невозможен*, осмеливается относиться к *любым возможным* устройствам ума и *любым возможным* процессам оптимизации – включая человеческие существа, которые тоже имеют ум, и многие из которых хорошие (nice) и хотят быть ещё лучше. На настоящий момент имеется неограниченное количество расплывчато убедительных аргументов, почему Дружественный ИИ может быть не под силу человеку, и всё же гораздо вероятнее, что проблема разрешима, но никто не соберётся решить её вовремя. Но не следует слишком быстро списывать проблему, особенно учитывая масштаб ставок.

## 6: Техническая неудача и философская неудача. (Technical failure and philosophical failure.)

Бостром (Bostrom, 2001) определяет глобальную катастрофу (existential catastrophe) как такую, которая истребляет возникшую на Земле разумную жизнь или необратимо *повреждает часть её потенциала*. Мы можем разделить потенциальные ошибки в попытках создания Дружественного ИИ на две неформальные категории, *техническую ошибку* и *философскую ошибку*. Техническая состоит в том, что вы пытаетесь создать ИИ, и он не работает так, как должен – вы не смогли понять, как работает на самом деле ваш собственный код. Философская неудача заключается в попытке построить неправильную вещь, так что даже если вы достигните успеха, вы всё равно не сможете никому помочь или облагодетельствовать человечество. Нет необходимости говорить о том, что одна ошибка не исключает другую.

Граница между двумя случаями тонка, поскольку большинство философских ошибок гораздо легче объяснить при наличии технического знания. В теории вы должны сначала заявить, что вы *хотите*, а затем обрисовать, *как* вы это достигните. На практике часто требуется глубокое техническое понимание, чтобы очертить то, что вы хотите.

### 6.1: Пример философской ошибки. (An example of philosophical failure.)

В конце 19 века многие честные и интеллигентные люди выступали за коммунизм, исходя только из лучших побуждений. Люди, которые первыми ввели, распространили и усвоили коммунистическую идею (meme) были, по строгому историческому счёту, идеалистами. У первых коммунистов не было предупреждающего примера Советской России. *В то время, без преимущества знания задним числом, это должно было звучать как весьма хорошая идея.* После революции, когда коммунисты пришли к власти и были отравлены ею, в игру могли вступить другие мотивы; но это не было предсказано первыми идеалистами, несмотря на то, насколько это могло быть предсказуемо. Важно понимать, что автор огромной катастрофы не должен быть злым или особо тупым. Если мы отнесём любую трагедию насчёт зла или особенной глупости, мы посмотрим на себя, правильно обнаружим, что мы не злы и не особенно тупы и скажем: «Но ведь это никогда не случится с *нами*».

Первые коммунисты думали, что эмпирическим последствием их революции будет то, что жизнь людей должна улучшиться: рабочие больше не будут работать долгие часы на изнурительной работе и получать за это мало денег. Это оказалось не совсем так, мягко говоря. Но то, что, по мнению первых коммунистов, должно было получится, не сильно

отличалось от того, что, по мнению сторонников других политических систем, должно было быть эмпирическим последствием их любимой политической системы. Они думали, что люди будут счастливы. Они заблуждались.

Теперь представим, что кто-то запрограммирует «Дружественный» ИИ на построение коммунизма, или либертарианства, или анархо-феодализма, или любой другой *любимой-политической-системы*, веря, что это осуществит утопию. Любимые политические системы людей порождают сияющие солнца позитивных эмоций, так что предложение будет казаться действительно хорошей идеей для предлагающего.

Мы можем наблюдать здесь программистскую ошибку на моральном или этическом уровне – скажем, в результате того, что кто-то доверяет себе столь высоко, что неспособен принять в расчет собственную подверженность ошибкам, отказываясь рассмотреть возможность того, что, например, коммунизм может быть ошибочным в конечном счёте. Но на языке байсовской теории решений, существует дополнительный технический взгляд на проблему. С точки зрения теории решений выбор в пользу коммунизма происходит из комбинации эмпирической веры и ценностного суждения. *Эмпирическая* вера состоит в том, что введение коммунизма приведёт к определённому результату или классу результатов: люди станут счастливее, работать меньше часов и обладать большим материальным богатством. Это, в конечном счёт, эмпирическое предсказание: даже его часть о счастье относится к реальным состояниям мозга, хотя её трудно измерить. Если вы введёте коммунизм, это результат будет или достигнут, или нет. Ценностное суждение состоит в том, что этот результат удовлетворяет или предпочтителен в текущих обстоятельствах. При другой *эмпирической* вере о *действительных последствиях* коммунистической системы в реальном мире, решение может претерпеть соответствующие изменения.

Мы можем ожидать, что подлинный ИИ, Искусственный Универсальный Интеллект, будет способен изменять свои эмпирические верования. (Или свою вероятностную модель мира и т.д.) Если бы каким-то образом Чарльз Баббадж (Charles Babbage) жил до Николая Коперника, и если бы каким-то образом компьютеры были бы изобретены до телескопов, и каким-то образом программисты той эпохи сконструировали бы Искусственный Универсальный Интеллект, из этого не следует, что ИИ верил бы всегда, что Солнце вращается вокруг Земли. ИИ может преодолеть фактическую ошибку своих программистов, в случае, если программисты понимают теорию умозаключений лучше, чем астрономию. Чтобы создать ИИ, который *откроет* орбиты планет, программистам не нужно знать математику Ньютоновской механики, а только математику Байсовой теории вероятности.

Недомыслие программирования ИИ для введения коммунизма, или любой другой политической системы, состоит в том, что вы программируете средства, а не цель. Вы программируете определённые решения без возможности их переработать после обретения улучшенного эмпирического знания о результатах коммунизма. Вы даёте ИИ готовое решение без того, чтобы обучить его, как создать заново (re-evaluate), - на более высоком уровне понимания, - исходно ошибочный процесс, который создал это решение.

Если я играю в шахматы против более сильного игрока, я не могу предсказать *точно*, где мой оппонент сделает ход против меня – если бы я мог предсказать, я бы, по необходимости, был бы так же силён в шахматах сам. Но я могу предсказать конечный результат, а именно выигрыш другого игрока. Я знаю область возможных будущ, куда мой оппонент направляется, что позволяет мне предсказать конец пути, даже если я не могу видеть дороги. Когда я нахожусь в наиболее творческом состоянии, это тогда, когда труднее всего предсказать мои действия и *легче* всего предсказать *последствия* моих действий. (Предполагая, что вы знаете и понимаете мои цели.) Если я хочу сделать игрока в шахматы, превосходящего человека, я должен запрограммировать поиск выигрышных ходов. Мне не следует программировать конкретные шаги, потому что в этом случае шахматный игрок не будет чем-либо лучше меня. Когда я начинаю поиск, я по необходимости жертвую своей способностью предсказать точный ответ заранее. Чтобы получить по настоящему хороший ответ, вы должны пожертвовать своей способностью предсказать ответ, но не своей способностью сказать, каков вопрос.

Такая путаница, как непосредственное программирование коммунизма, вероятно, не соблазнит программиста универсального ИИ, который говорит на языке теории решений. Я бы назвал это философской ошибкой, но обвинил бы в этом недостаток технического знания.

**6.2: Пример технической неудачи. (An example of technical failure.)**

«Вместо законов, ограничивающих поведение интеллектуальных машин, мы должны дать им эмоции, которые будут руководить их обучением поведению. Они должны хотеть, чтобы мы были счастливы и процветали, - что есть эмоция, которую мы называем любовью. Мы можем спроектировать интеллектуальные машины так, что их основная, врождённая эмоция будет безусловная любовь ко всем людям. В начале мы можем сделать относительно простые машины, которые научатся распознавать выражения счастья и несчастья на человеческом лице, человеческие голоса и человеческий язык жестов. Затем мы можем жёстко привязать

результат этого обучения в качестве изначально присущих эмоциональных ценностей более сложным интеллектуальным машинам, позитивно подкрепляемым, когда мы счастливы, и негативно – когда несчастливы. Машины могут обучиться алгоритмам приблизительного предсказания будущего, как, например, инвесторы используют сейчас обучающиеся машины, чтобы предсказать будущие цены облигаций. Таким способом мы можем запрограммировать интеллектуальные машины обучиться алгоритмам предсказания будущего человеческого счастья, и использовать эти предсказания, как эмоциональные ценности».

**Bill Hibbard (2001), Сверх-интеллектуальные машины (Super-intelligent machines.)**

Однажды американская армия захотела использовать нейронную сеть для автоматического обнаружения закамуфлированных танков. Исследователи натренировали нейронную сеть на 50 фотографиях закамуфлированных танков среди деревьев, и на 50 фото деревьев без танков. Используя стандартные методики контролируемого обучения, исследователи обучили нейронную сеть взвешиванию, которое правильно опознавало тренировочный набор – ответ «да» - для 50 фотография закамуфлированных танков, и ответ «нет» для 50 фотографий леса. Это не гарантировало, ни даже означало, что новые образцы будут классифицированы правильно. Нейронная сеть могла обучиться ста отдельным случаям, которые могли не обобщаться ни на одну новую задачу. Предусмотрительные исследователи сделали в начале 200 фото, 100 фото танков и 100 деревьев. Они использовали только 50 из каждой группы для тренировочного набора. Исследователи запустили в нейронную сеть оставшиеся 100 фото, и без дальнейшей тренировки нейронная сеть распознала все оставшиеся фотографии правильно. Успех подтвердился! Исследователи направили законченную работу в Пентагон, откуда её вскоре вернули, жалуясь, что в их собственной серии тестов нейронная сеть была не лучше, чем случай, в отборе фотографий.

Оказалось, что в наборе данных исследователей фотографии закамуфлированных танков были сделаны в облачные дни, тогда как фотографии чистого леса были сделаны в солнечные дни. Нейронная сеть обучилась различать облачные и солнечные дни вместо того, чтобы научиться различать закамуфлированные танки от пустого леса[2].

[2] Эта история, хотя и известная, и часто цитируемая, может быть апокрифической. Я не нашёл сообщения из первых рук. Для отчёта без ссылок см. Crochat и Franklin (2000) или

Технический провал имеет место, когда код не делает то, что, вы думаете, он делает, хотя он четно выполняет то, на что вы его запрограммировали. Одни и те же данные могут соответствовать разным моделям. Допустим, что мы обучаем нейронную сеть различать улыбающиеся человеческие лица и отличать их от хмурящихся лиц. Будет ли эта сеть распознавать маленькую картинку смеющегося лица как такой же аттрактор, как и смеющееся человеческое лицо? Если ИИ, жёстко фиксированный на таком коде, обретёт власть – и Hibbard (2001) говорит о сверхинтеллекте – не закончит ли галактика тем, что будет покрыта малюсенькими молекулярными картинками улыбающихся лиц? [3]

Эта форма провала особенно опасна, потому что система *выглядит* работающей в одном контексте, и проваливается при смене контекста. Создатели «определителя танков» обучали свою нейронную сеть до тех пор, пока она не начинала правильно распознавать данные, затем проверили сеть на дополнительных данных (без дальнейшего обучения). К несчастью, данные и для обучения, и для проверки содержали предположение, которое относилось ко всей информации, использованной в разработке, но не к ситуациям реального мира, где нейронная сеть была призвана работать. В истории с определителем танков это предположение состояло в том, что танки фотографируются в облачные дни.

Предположим, мы стремимся создать усиливающийся ИИ. Этот ИИ будет иметь фазу развития, когда люди-программисты будут сильнее его – не только в смысле физического контроля над электропитанием ИИ, но в смысле, что люди-программисты умнее, хитрее и более творческие, чем этот ИИ. Мы предполагаем, что в течение фазы развития программисты будут обладать способностью изменять исходный код ИИ без его согласия. После этого момента мы должны полагаться на установленную до того систему целей, потому что, если ИИ заработает достаточно непредсказуемым образом, то он сможет

---

http://neil.fraser.name/writing/tank/. Ошибки такого рода являются предметом больших реалистических рассмотрений при создании и тестировании нейронных сетей.

[3] Bill Hibbard, после просмотра черновика этой статьи, написал ответ, доказывающий, что аналогии с проблемой «классификатора танков» не применима к подкрепляющему обучению в целом. Его критика может быть найдена здесь: http://www.ssec.wisc.edu/~billh/g/AIRisk_Reply.html. Мой ответ: http://yudkowsky.net/AIRisk_Hibbard.html. Hibbard также отмечает, что предложение Hibbard (2001) заменено предложением Hibbard (2004). Последнее предлагает двухуровневую систему, в которой выражения согласия со стороны людей подкрепляют распознавание счастья, и распознанное счастье подкрепляет стратегии поведения.

активно сопротивляться нашим попыткам корректировать его – и если ИИ умнее человека, то, скорее всего, он победит.

Попытки контролировать растущий ИИ посредством тренировки нейронной сети, чтобы *создать его систему целей* сталкиваются с проблемой большой *смены контекста* при переходе от стадии развития ИИ к стадии после его развития (postdevelopmental stage). На стадии развития, ИИ может быть только способен создавать реакции, попадающие в категорию «улыбающихся человеческих лиц», решая предоставленные людьми задачи, как задумали его создатели. Вскоре, когда ИИ станет сверхчеловечески интеллектуален и создаст свою собственную нанотехнологическую инфраструктуру, он станет способен создавать столь же притягательные для него стимулы, покрывая всю галактику маленькими улыбающимися лицами.

Таким образом, ИИ кажется работающим правильно на стадии разработки, но создаёт катастрофические результаты, когда он становится умнее программистов(!)

Есть соблазн подумать: «Но наверняка ИИ будет знать, что это не то, что мы имеем в виду?» Но код не *дан* ИИ, чтобы он его просмотрел и вернул, если выяснится, что он работает неправильно. Код и есть ИИ. Возможно, приложив достаточно усилий и понимания, мы можем написать код, который следит, чтобы мы не написали неправильный код – легендарная DWIM-инструкция, которая среди программистов означает делай-то-что-я-имею-в-виду. (Do-What-I-Mean. (Raymond, 2003.)) Но требуются усилия, чтобы описать механику работы DWIM, и нигде в предложении Хиббарда нет упоминаний о создании ИИ, который делает то, что мы имеем в виду, а не то, что мы говорим. Современные чипы не выполняют DWIM над своим кодом; это не автоматическое свойство. И если у вас проблемы с самим DWIM, вы пострадаете от последствий. Предположим, например, что DWIM был определён так, чтобы максимизировать удовлетворение программиста от своего кода; когда этот код запустится как сверхинтеллект, он может переписать мозги программиста, чтобы он был максимально удовлетворён этим кодом. Я не говорю, что это неизбежно; я только говорю, что Делай-то-что-я-имею-в-виду – это большая и не тривиальная техническая проблема на пути к Дружественному ИИ.

**7: Темпы усиления интеллекта. (Rates of intelligence increase.)**

С точки зрения глобальных рисков, одно из наиболее критических обстоятельств в связи с ИИ, это то, что ИИ может усилить свой интеллект *чрезвычайно быстро*. Очевидная причина

подозревать такую возможность – это рекурсивное само-улучшение (Good, 1965) ИИ становится умнее, в том числе умнее в отношении написания внутренней когнитивной функции ИИ, так что ИИ может переписать свою существующую когнитивную функцию, чтобы она работала лучше. Это сделает ИИ ещё умнее, в том числе умнее в отношении задачи переделывания себя, так что он сделает ещё больше улучшений.

Люди *по большому счёту* не могут улучшать себя рекурсивно. В *ограниченном* объёме мы себя улучшаем: мы учимся, мы тренируемся, мы затачиваем свои навыки и знания. В небольшом отношении эти само-улучшения улучшают нашу способность улучшаться. Новые открытия могут увеличить нашу способность делать дальнейшие открытия – в этом смысле знание само себя питает. Но есть более нижний уровень, которого мы даже не коснулись. Мы не переписываем человеческий мозг. Мозг является, в конечном счёте, источником открытий (the source of discovery), и наши мозги сейчас почти такие же, как они были 10 тысяч лет назад.

Похожим образом, естественный отбор улучшает организмы, но процесс естественного отбора не улучшает сам себя – по большому счёту. Одна адаптация может открыть дорогу к дополнительным адаптациям. В этом смысле адаптация питает сама себя. Но даже когда генетический океан (pool) кипит, там всё равно присутствует нижестоящий нагреватель, а именно процессы рекомбинации, мутации и селекции, которые сами себя не перепроектируют. Несколько редких нововведений увеличили скорость эволюции самой по себе, например, появление половой рекомбинации. Но даже пол не изменил сущностной природы эволюции: её отсутствие абстрактного интеллекта, её зависимость от случайных мутаций, её слепоту и постепенность, её сосредоточенность на частоте аллелей. Точно также появление науки не изменило сущностного характера человеческого мозга: его лимбическое ядро, церебральный кортекс, его префронтальные собственные модели (prefrontal self-models), его характеристическую скорость в 200 ГЦ.

ИИ может переписать свой код с самого начала – он может изменить лежащую в основе динамику процесса оптимизации. Такой процесс оптимизации будет закручиваться гораздо сильнее, чем эволюционные накапливающие адаптации, равно как и человеческие накапливающиеся знания. Главным последствием с точки зрения наших целей является то, что ИИ может совершить огромный прыжок в интеллектуальности после достижения некого порога критичности.

Часто встречающееся скептическое мнение об этом сценарии, – который Good (1965) назвал «интеллектуальным взрывом» - происходит из того, что прогресс в области ИИ имеет репутацию очень медленного.
Здесь полезно рассмотреть свободную историческую аналогию об одном неожиданном открытии. (Дальнейшее взято главным образом из Rhodes, 1986.)

В 1933 году лорд Эрнст Резерфорд заявил, что никто не должен ожидать, что когда-нибудь удастся извлечь энергию из распада атома: «Любой, кто искал источник энергии в трансформации атомов, говорил вздор». В те времена требовались дни и недели работы, чтобы расщепить небольшое количество ядер.

Вскоре, в 1942 году, на теннисном корте под Стаг Филдом около университета Чикаго физики строят агрегат в форме гигантской шарообразной дверной ручки из чередующихся слоёв графита и урана, намереваясь запустить первую само-поддерживающуюся ядерную реакцию. За проект отвечает Энрико Ферми.

Ключевым числом для реактора является K,  эффективный фактор умножения нейтронов: то есть среднее значение числа нейтронов из реакции деления, которое вызывает другую реакцию деления. Пока K меньше единицы, реактор является субкритическим. При K >=1 реактор должен поддерживать критическую реакцию. Ферми рассчитал, что реактор достигнет K=1 при числе слоёв между 56 и 57.

Рабочая группа, руководимая Гербертом Андерсоном, закончила 57 слой в ночь 1 декабря 1942 года. Контрольные стержни - бруски дерева, покрытые поглощающей нейтроны кадмиевой фольгой, - предохраняли реактор от достижения критичности. Андерсон убрал все стержни, кроме одного и замерил радиацию реактора, подтвердив, что реактор готов к цепной реакции на следующий день. Андерсон вставил все стержни, запер их на висячие замки, запер теннисный корт и пошёл домой.

На следующий день, 2 декабря 1942 года, ветреным и морозным Чикагским утром, Ферми начал окончательный эксперимент. Все, кроме одного, стержни были подняты. В 10:37 Ферми приказал поднять последний контролирующий стержень на половину высоты. Счётчики Гейгера застучали чаще, и самописец дёрнулся вверх. «Это не то, - сказал Ферми, - график дойдёт до вот этой точки и выровняется», - указывая на точку на графике. Через несколько минут самописец дошёл до указанной точки, и не пошёл выше. Через несколько

минут Ферми приказал поднять стержень ещё на один фут. Опять радиация усилилась, но затем выровнялась. Стержень подняли ещё на 6 дюймов, затем ещё и ещё.

В 11:30 медленный подъём самописца прервался колоссальным ПАДЕНИЕМ - защитный контролирующий стержень, запущенный ионизационным датчиком, активировался и опустился в реактор, который был всё ещё некритичен. Ферми тихо приказал команде сделать перерыв на обед.

В два часа пополудни команда собралась снова, вынула и заперла защитный стержень, и вывела контролирующий стержень на его последнюю позицию. Ферми сделал несколько измерений и вычислений, и затем опять начал процесс подъёма стержня небольшими шагами. В 15:25 Ферми приказал поднять стержень ещё на 12 дюймов. «Это должно дать результат», - сказал Ферми. «Сейчас она станет самоподдерживающейся. График будет расти и расти, не выравниваясь».

**Герберт Андерсон рассказывает (Rhodes, 1986):**

«В начале вы могли слышать звук нейтронного счётчика, щёлк-щёлк. Затем щёлчки стали появляться всё чаще и через некоторое время они слились в рёв; счётчик за ними больше не успевал. Теперь надо было переключаться на графический регистратор. Но когда это было сделано, все уставились во внезапной тишине на возрастающее отклонение пера самописца. Это была значительная тишина. Каждый понимал значительность этого переключения; мы были на режиме высшей интенсивности и счётчики больше не могли справляться с этой ситуацией. Снова и снова шкала самописца должна была сменяться, чтобы подстраиваться под интенсивность нейтронов, которая возрастал всё более и более быстро. Внезапно Ферми поднял свою руку. «Реактор достиг критичности», - объявил он. Никто из присутствующих не имел на этот счёт никаких сомнений».

Ферми дал проработать реактору 28 минут, при скорости удвоения интенсивности нейтронов в две минуты. Первая критическая реакция имела К в 1,0006. Но даже при К=1.0006 реактор был контролируем только потому, что некоторые из нейтронов из деления урана задерживаются – они получаются при распаде короткоживущих продуктов деления. На каждые 100 распадов $U_{235}$ 242 нейтрона испускаются почти мгновенно (0,0001 сек) и 1,58 нейтронов испускаются в среднем через десять секунд. Поскольку среднее время жизни нейтрона ~0.1 секунды, что означает 1200 поколений за 2 минуты, и время удвоения в 2

минуты, потому что умножение 1.0006 на 1200 примерно даёт 2. Ядерная реакция, являющаяся мгновенно критичной (prompt critical), достигает критичности без вклада отложенных нейтронов. Если бы реактор Ферми был бы мгновенно критичным с k=1.0006, интенсивность нейтронов удваивалась бы каждую десятую долю секунды.

Первая мораль этой истории состоит в том, что смешение скорости *исследований* ИИ со скоростью реального ИИ подобно смешению скорости физических исследований со скоростью ядерных реакций. Происходит смешение карты и территории. Потребовались годы, чтобы построить этот первый реактор, усилиями небольшой группы физиков, которые не публиковали много пресс-релизов. Но когда реактор был построен, интересные события произошли на временной шкале ядерных взаимодействий, а не на временной шкале человеческого общения. В ядерной области элементарные взаимодействия происходят гораздо быстрее, чем срабатывают человеческие нейроны. Тоже может быть сказано о транзисторах.

Другая мораль в том, что есть колоссальная разница между ситуацией, когда одно самоулучшение запускает в среднем 0.9994 дальнейших самоулучшений, и когда одно самоулучшение запускает 1.0006 дальнейших самоулучшений. Ядерный реактор перешёл порог критичности не потому, что физики внезапно заложили в него много дополнительного вещества. Физики вводили вещество медленно и равномерно. Даже если имеется гладкая кривая интеллектуальности мозга как функции оптимизационного давления, оказанного до того на этот мозг, то кривая рекурсивного самоулучшения может содержать огромный скачок.

Есть и другие причины, по которым ИИ может совершить внезапный огромный скачок в интеллектуальности. Вид *Homo sapiens* совершил большой прыжок в эффективности интеллекта, как результат естественного отбора, оказывавшего более-менее равномерное давление на гоминидов в течение миллионов лет, постепенно расширяя мозг и лобовую кору, настраивая программную архитектуру. Несколько десятков тысяч лет назад интеллект гоминидов пересёк некий ключевой порог и сделал огромный прыжок в эффективности в реальном мире; мы перешли от пещер к небоскрёбам за мгновение ока эволюции. Это произошло при неизменном давлении отбора – не было большого прыжка в оптимизирующей силе эволюции, когда появились люди. Наша соответствующая мозговая архитектура тоже развивалась плавно – объём нашего черепа не увеличился вдруг на два порядка величины. Так что может так случиться, что даже если ИИ будет развивать снаружи силами людей-инженеров, кривая его интеллектуальной эффективности совершит резкий скачок.

Или, возможно, некто построит прототип ИИ, который покажет некие многообещающие результаты, и эта демо-версия привлечёт дополнительные 100 миллионов долларов венчурного капитала, и на эти деньги будет закуплено в тысячу раз больше суперкомпьютеров. Я сомневаюсь, что усиление оборудования в 1000 раз приведёт к чему-либо подобному усилению интеллектуального потенциала в 1000 раз – но само это сомнение не надёжно при отсутствии какой-либо возможности произвести какие-либо аналитические вычисления.

В сравнении с шимпанзе, человек имеет трёхкратное преимущество в мозге и шестикратное – в лобовой коре, что означает (а) программы важнее оборудования и (б) малые увеличения оборудования могут поддержать большие улучшения программного обеспечения. И есть ещё один момент, который надо рассмотреть.

В конечном счёте, ИИ может сделать *кажущийся* резким скачок в интеллектуальности только по причине антропоморфизма, то есть человеческой склонности думать о «деревенском идиоте» и Эйнштейне как о крайних границах интеллектуальной шкалы, а не как о почти неразличимых точках на шкале умов-вообще.
Любой объект, более немой, чем немой человек, может показаться нам просто немым. Можно представить «стрелу ИИ», медленно ползущую по шкале интеллекта, проходящую уровни мыши и шимпанзе, и при этом ИИ остаётся всё ещё немым, потому что ИИ не может свободно говорить или писать научные статьи, и затем стрела ИИ пересекает тонкую грань между ультра-идиотом и Эйнштейном в течение месяца или такого же малого периода. Я не думаю, что этот сценарий убедителен, в основном, потому что я не ожидаю, что кривая рекурсивного самоулучшения будет ползти линейно. Но я не буду первым, кто укажет, что ИИ – это двигающаяся цель. Как только веха достигнута, она перестаёт быть ИИ. Это может только вдохновлять промедление.

Давайте допустим, для продолжения дискуссии, что, исходя из всего, что мы знаем (и это кажется мне реально возможным), ИИ обладает способностью совершить внезапный, резкий, огромный скачок в интеллектуальности. Что из этого следует? Первое и главное: из этого следует, что реакция, которую я часто слышал: «Нам не следует заботится о Дружественном ИИ, потому что у нас ещё нет самого ИИ» - неверна или просто самоубийственна. Мы не можем полагаться на то, что у нас будут заранее предупреждающе сигналы до того, как ИИ будет создан; прошлые технологические революции обычно не телеграфировали о себе людям, жившим в том время, что бы потом ни говорилось.

Математика и техника Дружественного ИИ не появится из ниоткуда, когда она будет нужна; требуются годы, чтобы установить твёрдые основания. И мы должны разрешить проблему Дружественного ИИ до того, как универсальный ИИ появится, а не после; мне даже не следует говорить об этом. Будут трудности с Дружественным ИИ, потому что поле исследований ИИ само по себе имеет мало согласия и высокую энтропию. Но это не значит, что мы не должны беспокоиться о Дружественном ИИ. Это означает, что будут трудности. Эти два утверждения, к сожалению, даже отдалённо не эквивалентны.

Возможность резкого скачка в интеллектуальности также требует высоких стандартов для техники Дружественного ИИ. Техника не может полагаться на способность программиста наблюдать ИИ *против его воли*, переписывать ИИ *против его воли*, угрожать превосходящей военной силой, ни на то, что программисты смогут контролировать «кнопку вознаграждения», которую умный ИИ отберёт у программистов, и так далее. В действительности, никто не должен исходить из этих предположений. Необходимой защитой является ИИ, который не хочет вам повредить. Без этого ни одна дополнительная защита не является безопасной. Ни одна система не является безопасной, если она ищет способы разрушить свою безопасность. Если ИИ повредит человечеству в любом смысле, вы должны были сделать что-то неправильно на очень глубоком уровне, искривив свои основные посылки. Вы делаете дробовик, направляете его на свою ступню и спускаете крючок. Вы осознанно приводите в движение некую когнитивную динамику, которая, при некоторых обстоятельствах, будет стремиться вам повредить. Это – неправильное поведение для данной динамики; напишите вместо этого код, который делает что-то другое.

Примерно по тем же причинам, программисты Дружественного ИИ должны предполагать, что ИИ имеет полный доступ к своему исходному коду. Если ИИ хочет модифицировать себя, чтобы не быть больше Дружественным, Дружественность уже потерпела неудачу в тот момент, когда ИИ создал такое намерение. Любое решение, которое полагается на то, что ИИ не будет способен модифицировать сам себя, будет разрушено тем или иным способом, и будет разрушено даже в том случае, если ИИ решит никогда себя не модифицировать. Я не говорю, что это должна быть единственная предосторожность, но главной и незаменимой предосторожностью будет то, что вы создадите ИИ, который не захочет вредить человечеству.

Чтобы избежать ошибочности Гигансткой Ватрушки, мы должны сказать, что способность улучшать себя не означает выбора делать это. Успешное *воплощение* техники

Дружественного ИИ может создать ИИ, который обладает потенциалом расти более быстро, но выбирающего вместо этого расти медленнее и по более управляемой кривой.

Даже в этом случае, после того, как ИИ пройдёт критический порог рекурсивного самоулучшения, вы окажетесь действующими в гораздо более опасном режиме. Если дружественность потерпит неудачу, ИИ может решить направиться с полной скоростью в сторону самоулучшения – метафорически говоря, он станет мгновенно критичным.

Я склонен предполагать *потенциально* произвольно большие прыжки в интеллектуальности, потому что это (а) консервативное предположение; (б) это отвергает предложения построить ИИ без реального понимания его; и (с) большие скачки потенциала (large potential jumps) кажутся мне наиболее вероятными в реальном мире. Если я обнаружу некую область знаний, в которой консервативной точкой зрения по поводу перспектив управления рисками предполагается медленное улучшение ИИ, тогда я потребую, чтобы этот план не стал катастрофическим, если ИИ замедлится на около-человеческой стадии на годы или дольше. Это не та область, относительно которой бы мне хотелось предлагать узкие интервалы уверенности.

**8: Оборудование. (Hardware.)**

Люди склонны думать о больших компьютерах как о ключевом факторе ИИ. Это, мягко говоря, очень сомнительное утверждение. Не-футурологи, обсуждая ИИ, говорят обычно о прогрессе компьютерного оборудования, потому что его легко измерить – в отличие от понимания интеллекта. Не потому что здесь нет прогресса, а потому что этот прогресс не может быть выражен в аккуратных графиках компьютерных презентаций. Трудно сообщать об улучшениях в понимании, и поэтому об этом меньше сообщают. Вместо того, чтобы думать о «минимальном» уровне оборудования, которое «необходимо» для ИИ, задумаемся лучше о минимальном уровне понимания исследователя, который уменьшается по мере улучшения оборудования. Чем лучше компьютерное оборудование, тем меньше понимания вам нужно, чтобы построить ИИ. Крайним случаем является естественный отбор, который использовал удивительные количества грубой компьютерной силы, чтобы создать человеческий интеллект, не используя никакого понимания, только неслучайное сохранение случайных мутаций.

Увеличивающаяся компьютерная мощность делает изготовление ИИ проще, но нет очевидных причин, по которым увеличивающая компьютерная мощь поможет сделать ИИ

Дружественным. Возрастающая сила компьютеров делает более простым применение грубой силы, а также совмещение плохопонятных, но работающих техник. Закон Мура устойчиво снижает барьер, который предохраняет нас от построения ИИ без глубокого понимания мышления.

Приемлемо провалиться в попытках создания как ИИ, так и Дружественного ИИ. Приемлемо достичь успеха и в ИИ, и в Дружественном ИИ. Что неприемлемо – это создать ИИ и провалиться в создании Дружественного ИИ. Закон Мура делает именно последнее гораздо проще. «Проще», но, слава богу, не просто. Я сомневаюсь, что ИИ будет прост, когда его, наконец, построят – просто потому что есть группы людей, которые приложат огромные усилия, чтобы построить ИИ, и одна из них достигнет успеха, когда ИИ, наконец, станет возможным достичь посредством колоссальных усилий.

Закон Мура является посредником (interaction) между Дружественным ИИ и другими технологиями, что добавляет часто пропускаемый глобальный риск к другим технологиям. Мы можем представить себе, что молекулярная нанотехнология развивается силами мягкого многонационального правительственного консорциума и им удалось успешно избежать опасностей физического уровня нанотехнологий. Они непосредственно не допустили случайное распространение репликатора, и с гораздо большими трудностями разместили глобальную защиту на местах против враждебных репликаторов; они ограничили доступ к базовому уровню нанотехнологии, в то же время распространяя настраиваемые наноблоки и так далее. (См. Phoenix и Treder, в этом же сборнике.) Но, тем не менее, нанокомпьютеры становятся широко распространены, потому что предпринятые ограничения обходятся, или потому что никаких ограничений не введено. И затем кто-то добивается грубой силой ИИ, который не Дружественен, и дело закончено. Этот сценарий является особенно беспокоящим, потому что невероятно мощные нанокомпьютеры будут среди первых, простейших и кажущихся безопаснейшими применений нанотехнологии.

Как насчёт регуляторного контроля над суперкомпьютерами? Я бы определённо не стал на него полагаться, чтобы предотвратить создание ИИ; вчерашние суперкомпьютеры - это завтрашние лэптопы. Стандартный ответ на предложение о регулировании состоит в том, что когда нанокомпьютеры будут вне закона, только стоящие вне закона люди будут ими обладать.

Трудно доказать, что предполагаемые преимущества от ограничения распространения перевешивают неизбежные риски от неточного распространения. Я сам точно не буду

выступать в пользу регулятивных ограничений на использование суперкомпьютеров для исследований ИИ; это предложение сомнительной полезности будет встречено в штыки всем ИИ-сообществом. Но в том маловероятном случае, если это предложение будет принято – что весьма далеко от текущего политического процесса – я не буду прикладывать значительных усилий, чтобы бороться с ним, поскольку я не думаю, что хорошим ребятам нужен доступ к современным им суперкомпьютерам. Дружественный ИИ – это не про грубую силу.

Я могу представить регулирующие органы, эффективно контролирующие небольшой набор сверхдорогих компьютерных ресурсов, которые *нынче называются* суперкомпьютеры. Но компьютеры везде. Это не похоже на ядерное нераспространение, где основное направление – это контроль плутония и обогащённого урана. Исходные материалы для ИИ уже есть везде. Эта кошка так далеко выскочила из мешка, что она уже в ваших наручных часах, сотовом телефоне и посудомоечной машине. Это тоже является особенным и необычным фактором ИИ как глобального риска. Мы отделены от рискованного процесса не большими видимыми установками, такими как изотопные центрифуги или ускорители частиц, но только недостаточным знанием. Если использовать слишком драматичную метафору, это подобно тому, как если бы субкритические массы обогащённого урана приводили бы в движение машины и корабли по всему миру *до того*, как Лео Сцилард впервые подумал бы о цепной реакции.

**9: Угрозы и перспективы. (Threats and promises.)**

Это рискованное интеллектуальное предприятие, - пытаться предсказать конкретно, как именно благожелательный ИИ поможет человечеству, или недружественный ИИ повредит. Здесь есть риск систематической ошибки наложения: каждая добавленная деталь обязательно уменьшает общую вероятность всей истории, но испытуемые склонны приписывать большую вероятность историям, которые включают чёткие добавленные детали. (См. Элизер Юдковский. Систематические ошибки в рассуждениях, потенциально влияющие на оценку глобальных рисков.) Есть риск – почти наверняка – потерпеть неудачу, пытаясь вообразить сценарий будущего; и есть риск ошибочности Гигантской Ватрушки, который превращается из возможности в мотивирующую силу (that leaps from capability to motive).

Тем не менее, я попробую очертить угрозы и перспективы. Будущее имеет репутацию совершать подвиги, которые прошлое считало невозможными. Будущие цивилизации даже нарушали то, что прошлые цивилизации считали (неверно, разумеется) законами физики.

Если пророки 1900 года – и даже не думайте о 1000 годе – пытались ограничить силу человеческой цивилизации через миллиард лет, то некоторые из названных ими невозможностей были бы преодолены до конца столетия; превращение свинца в золото, например. Мы помним, что будущие цивилизации удивляли прошлые цивилизации, и поэтому стало клише, что мы не можем накладывать ограничений на своих праправнуков. И всё же все в 20 веке, в 19 веке и в 11 веке мы были людьми.

Мы можем различить три семейства ненадёжных метафор для представления возможностей превосходящего человека ИИ:

- метафора G-фактора: вдохновлена различиями индивидуального уровня интеллекта между людьми. ИИ будет патентовать новые технологии, публиковать прорывные статьи, делать деньги на фондовом рынке или возглавлять политические блоки.
- историческая метафора: вдохновлена знанием различий между прошлыми и будущими человеческими цивилизациями. ИИ быстро введёт набор возможностей, который обычно связывается с человеческой цивилизацией через сто или тысячу лет: молекулярную нанотехнологию; межзвёздные путешествия; компьютеры, выполняющие $10^{25}$ операций в секунду.

- Видовая метафора: вдохновлена различиями в архитектуре мозга между видами. ИИ овладеет магией.

Метафора G-фактора наиболее популярна в современной футурологии: когда люди думают об интеллектуальности, они думают о человеческом гении, а не о людях вообще. В историях о враждебном ИИ G-метафоры ответственны за «хорошую историю» в духе Бострома: а именно, за оппонента, достаточно могущественного, чтобы создать драматическое напряжение, но не достаточно могущественного, чтобы мгновенно истребить героев, как мух, и, в конечном счёте, достаточно слабого, чтобы проиграть в последних главах книги. Голиаф против Давида – пример хорошей истории, но Голиаф против плодовой мушки – нет.

Если мы предполагаем метафору G-фактора, то риски глобальной катастрофы в этом сценарии относительно умеренные: враждебный ИИ – не большая угроза, чем враждебный человеческий гений.

Если мы предполагаем множественность ИИ, то тогда мы имеем метафору конфликта между племенем ИИ и человеческим племенем. Если племя ИИ выиграет в военном конфликте и истребит людей, то это глобальная катастрофа по типу Взрыва (Bostrom, 2001). Если племя

ИИ будет доминировать над миром экономически и обретёт эффективный контроль над судьбой возникшей на Земле разумной жизни, но цели ИИ не будут для нас интересными или стоящими, то это будет катастрофа в духе Визг, Хныкание или Хруст. Но насколько вероятно, что ИИ преодолеет весь огромный разрыв от амёбы до деревенского идиота, и затем остановится на уровне человеческого гения? Быстрейший из наблюдавшихся нейронов срабатывает 1000 раз в секунду; быстрейший аксон передаёт сигналы со скоростью 150 метров в секунду, в пол-миллионную долю от скорости света; каждая операция синапса рассеивает примерно 15 000 аттоджоулей, что в миллион раз больше термодинамического минимума для необратимых вычислений при комнатной температуре ($kT_{300} \ln(2) = 0.003$ аттоджоулей на бит). Физически возможно построить мозг, вычисляющий в миллион раз быстрее человеческого, без уменьшения размера, работы при низких температурах, применения обратимых вычислений и квантового компьютера. Если человеческий ум будет таким образом ускорен, субъективный год размышлений завершится за 31 физическую секунду во внешнем мире, и тысячелетие пролетит за восемь с половиной часов. Винж (Vinge, 1993) назвал такие ускоренные умы «слабым сверхинтеллектом»: ум, думающий как человек, но гораздо быстрее.

Мы предполагаем, что возникнет чрезвычайно быстрый ум, установленный в сердцевине человеческой технологической цивилизации, которая будет существовать в это время. Провалом воображения было бы сказать: «Не важно, как быстро он думает, он может влиять на мир только со скоростью своих манипуляций; он не может управлять машинами быстрее, чем он приказывает человеческим рукам работать; поэтому быстрый ум – это не великая опасность». Нет такого закона природы, по которому физические операции должны тянуться секундами. Характерное время для молекулярных реакций измеряется в фемтосекундах, иногда в пикосекундах.

Drexler (1992) проанализировал контролируемые молекулярные манипуляторы, которые будут выполнять $>10^6$ молекулярных операций в секунду – отметьте это в связи с основной темой о «миллионократном ускорении». (Наименьшим физически значимым приращением времени обычно считается интервал Планка, $5 \cdot 10^{-44}$ секунды, и на этой шкале даже танцующие кварки кажутся статуями.)

Представьте себе, что человечество было бы заперто в ящике и могло бы воздействовать на окружающий мир только посредством заморожено медленных движений щупалец пришельца, или механических рук, которые бы двигались со скоростью несколько микрон в

секунду. Тогда мы бы сконцентрировали всю нашу творческую силу на поисках наикратчайшего пути построить быстрые манипуляторы во внешнем мире. Размышляя о быстрых манипуляторах, немедленно вспоминаешь о молекулярной нанотехнологии – хотя могут быть и другие пути. Какой кратчайший путь может привести вас к нанотехнологии в медленном внешнем мире, если у вас есть эоны времени, чтобы обдумывать каждый ход? Ответ в том, что я не знаю, потому что у меня нет эонов времени на размышления. Вот один из воображаемых быстрых путей:

- разрешить проблему свёртки белков, до той степени, чтобы быть способным создавать последовательности ДНК, чьи свёрнутые пептидные последовательности будут соответствовать специфическим функциональным задачам в сложных химических реакциях.

- Отправить по E-mail набор последовательностей ДНК в одну или несколько лабораторий, которые предлагают синтез ДНК, секвенсирование пептидов и доставку по FedEx. (Много лабораторий предлагают сейчас такие услуги, и некоторые рекламируют время в 72 часа на полный цикл.)

- Найти по крайней мере одного человека, соединённого с Интернетом, которого можно оплатить, запугать шантажом или задурить соответствующей историей, чтобы он получил отправления по FedEx и смешал их в специальной среде.

- Синтезированные белки сформируют очень примитивную «влажную» наносистему, вроде рибосомы, способную принимать внешние инструкции; возможно, модулированные акустические волны, направляемые динамиком на мензурку.

- Использовать эту невероятно примитивную наносистему, чтобы построить более сложную систему, развивающуюся затем в молекулярную нанотехнологию – или дальше.

Полное время на всю процедуру имело бы, вероятно, порядок недель с момента, когда быстрый интеллект смог бы решить проблему сворачивания белков. Разумеется, этот сценарий целиком выдумал я. Возможно, за 19 500 лет субъективного времени (одна неделя физического времени при ускорении в миллион раз) я бы нашёл более простой путь. Возможно, вы можете заплатить за быструю курьерскую доставку вместе FedEx. Возможно, существуют технологии, или небольшие модификации существующих технологий, которые синергетически соединяются с простыми белковыми механизмами. Возможно, если вы

достаточно умны, вы можете использовать волновые электрические поля, чтобы изменять пути реакций в существующих биохимических процессах.  Я не знаю. Я не настолько умён.

Задача состоит в том, чтобы связать вместе ваши способности – аналогом чему в реальном мире является комбинирование слабых уязвимостей в компьютерной системе для получения корневого доступа. Если один путь перекрыт, вы выбираете другой, всегда ища способы увеличить свои возможности и использовать их взаимоусиливающим образом (in synergy). Подразумеваемая цель – построить быструю инфраструктуру, то есть средства манипулировать внешним миром в большом масштабе за малое время. Молекулярная нанотехнология удовлетворяет этим критериям, во-первых, потому что её элементарные операции происходят быстро, и, во вторых, потому что имеется готовый набор совершенных частей – атомов – которые могут быть использованы для самореплицирования и экспоненциального роста нанотехнологической инфраструктуры. Путь, обсуждённый выше, подразумевает ИИ, получающий скоростную инфраструктуру в течение недели – что звучит быстро для человека с 200 Гц нейронами, но является гораздо бОльшим временем для ИИ.

Как только ИИ обретает быструю инфраструктуру, дальнейшие события происходят по шкале времени ИИ, а не по человеческой временной шкале. (Кроме того случая, когда ИИ предпочтёт действовать в человеческой временной шкале.) С молекулярной нанотехнологией, ИИ может (потенциально) переписать всю Солнечную систему без какого-либо сопротивления.

Недружественный ИИ с молекулярной инфраструктурой (или другой быстрой инфраструктурой) не должен беспокоиться об армиях марширующих роботов, или шантаже или тонких экономических вмешательств. Недружественный ИИ обладает способностью переделать всё вещество Солнечной системы согласно своей цели оптимизации. Для нас будет фатальным, если этот ИИ не будет учитывать при своём выборе то, как эта трансформация повлияет на существующие сейчас системы, такие как биология и люди. Этот ИИ не ненавидит вас, ни любит, но вы сделаны из атомов, которые он может использовать как-то по-другому. ИИ работает на другой временной шкале, чем вы; к тому моменту, когда ваши нейроны закончат думать слова «я должен сделать нечто», вы уже проиграли. Дружественный ИИ плюс молекулярная нанотехнология предположительно достаточно сильны, чтобы разрешить любую проблему, которая может быть разрешена путём перемещения атомов или творческого мышления. Следует соблюдать предосторожность в отношении возможных ошибок воображения: лечение рака – это популярная современная цель для филантропии, но из этого не следует, что Дружественный

ИИ с молекулярной нанотехнологией скажет сам себе: «Теперь я буду лечить рак». Возможно, лучшее описание проблемы состоит в том, что человеческие клетки непрограммируемы. Если решить эту проблему, то это излечит рак как частный случай, а заодно диабет и ожирение. Быстрый, позитивный интеллект, владеющий молекулярной нанотехнологией, обладает силой избавиться от болезней, а не от рака.

Последнее семейство метафор связано с видами, и основывается на межвидовых различиях интеллекта. Такой ИИ обладает магией – не в смысле заклинаний или снадобий, но в том смысле, как волк не может понять, как работает ружьё, или какого рода усилия требуются, чтобы изготовить ружья, или природу человеческой силы, которая позволяет нам придумывать ружья.

Винж (Vinge, 1993) пишет: «Сильное сверхчеловечество (strong superhumanity) будет не просто разогнанным до большой скорости эквивалентом человеческого ума. Трудно сказать, чем именно сверхчеловечество будет, но разница, вероятно, будет глубокой. Представьте себе ум собаки, работающий на огромной скорости. Дадут ли тысячелетия собачей жизни хотя бы один человеческий инсайт?»

Видовая метафора является ближайшей аналогией а приори, но она не очень пригодна для создания детальных историй. Главный совет, которая даёт нам эта метафора, состоит в том, что нам лучше всего всё-таки сделать Дружественный ИИ, что есть хороший совет в любом случае. Единственную защиту, которую она предлагает от враждебного ИИ – это вообще его не строить, что тоже очень ценный совет. Абсолютная власть является консервативным инженерным предположением в отношении Дружественного ИИ, который был неправильно спроектирован. Если ИИ повредит вам с помощью магии, его Дружественность в любом случае ошибочна.

**10: Локальные стратегии и стратегии большинства (Local and majoritarian strategies)**

Можно классифицировать предлагающиеся стратегии снижения риска следующим образом:

- стратегии, требующие единодушной кооперации – стратегии, которые могут быть повержены отдельными вредителями или небольшими группами.

- стратегии, которые требуют совместного действия большинства (majoritarian strategy): большинства законодателей в одной стране, или большинства голосующих людей, или

большинства стран в ООН: стратегии, требующие большинства, но не всех людей из некой большой группы, чтобы действовать определённым образом.

- Стратегии, которые требуют локальных действий – концентрации воли, таланта и финансирования, которая достигает порогового значения для некоторой конкретной задачи.

Единодушные стратегии не работоспособны, что не мешает людям продолжать предлагать их.

Мажоритарные стратегии (стратегии большинства) иногда работают, если у вас есть десятилетия на то, чтобы сделать свою работу. Следует создать движение, и пройдут годы до его признания в качестве силы в публичной политике и до его победы над оппозиционными фракциями. Мажоритарные стратегии занимают значительное время и требуют огромных усилий. Люди уже старались это сделать, и история помнит несколько успехов. Но будьте настороже: исторические книги имеют тенденцию селективно концентрироваться на тех движениях, которые имели влияние, в отличие от большинства, которое никогда ни на что не влияло. Здесь есть элемент удачи и изначальной готовности публики слушать. Критические моменты этой стратегии включают элементы, лежащие за пределами нашего контроля. Если вы не хотите посвятить всю свою жизнь продвижению некой мажоритарной стратегии, не беспокойтесь; и даже целиком посвящённой жизни недостаточно.

Обычно, локальные стратегии наиболее убедительны. Не легко получить 100 миллионов долларов обеспечения, и всеобщей политической перемены тоже нелегко достичь, но всё же гораздо легче получить 100 миллионов, чем продвинуть глобальную политическую перемену. Два предположения, выдвигаемые в пользу мажоритарной стратегии в отношении ИИ:

- Большинство из Дружественных ИИ может эффективно защитить человеческий вид от неДружественного ИИ.

- Первый построенный ИИ не может сам по себе нанести катастрофический ущерб.

Это повторяет по существу ситуацию в человеческой цивилизации до создания ядерного и биологического оружия: большинство людей сотрудничают во всемирной социальной структуре, а вредители могут причинить определённый, но не катастрофический ущерб.

Большинство исследователей ИИ не хотят построить неДружественный ИИ. Если кто-то знает, как сделать стабильный Дружественный ИИ – если проблема не находится полностью за пределами современных знаний и техники – исследователи будут учиться успешным результатам друг у друга и повторять их. Законодательство может (например) потребовать от исследователей публиковать свои стратегии Дружественности или наказывать тех исследователей, чьи ИИ причинили ущерб; и хотя эти законы не предотвратят всех ошибок, они могут гарантировать, что большинство ИИ будут построены Дружественными.

Мы можем также представить сценарий, который предполагает простую локальную стратегию:

- первый ИИ не может сам по себе причинить катастрофический ущерб.

- Если даже хотя бы один Дружественный ИИ появится, этот ИИ вместе с человеческими учреждениями может отогнать любое количество неДружественных ИИ.

Этот лёгкий сценарий выдержит, если человеческие институты смогут надёжно отличать Дружественный ИИ от неДружественного и дадут могущую быть отменённой власть в руки Дружественного ИИ. Тогда мы сможем собрать и выбрать наших союзников. Единственное требование состоит в том, чтобы проблема Дружественного ИИ была разрешима (В противовес тому, что бы быть полностью за пределами человеческих возможностей.)

Оба из вышеприведённых сценариев предполагают, что первый ИИ (первый мощный, универсальный ИИ) не может сам по себе причинить глобально катастрофический ущерб. Более конкретные представления, которые это предполагают, используют G-метафору: ИИ как аналог особо одарённым людям. В главе 7 о скоростях усиления интеллекта, я указал несколько моментов, почему следует подозревать огромный, быстрый скачок в интеллектуальности.

- расстояние от идиота до Эйнштейна, которое выглядит большим для нас, является маленькой точкой на шкале умов вообще.

- Гоминиды сделали резкий скачок в эффективности во внешнем мире, несмотря на то, что естественный отбор оказывал примерно равномерное давление на их геном.

- ИИ может впитать колоссальное количество дополнительного оборудования после достижения определённого уровня компетентности (то есть, съесть интернет).

- Существует критический порог рекурсивного самоулучшения. Одно самоулучшение, дающее приращение в 1,0006 раз, качественно отличается от самоулучшения, дающего приращение в 0,9994 раза.

Как описано в главе 9, достаточно сильному ИИ может потребоваться очень короткое время (с человеческой точки зрения), чтобы достичь молекулярной нанотехнологии, или другой формы быстрой инфраструктуры. Теперь мы можем представить себе всё значение того, кто начнёт первым (the first-mover effect) в суперинтеллекте. Эффект начавшего первым состоит в том, что исход возникшей на Земле разумной жизни зависит в первую очередь от особенностей (makeup) того ума, который первым достигнет определённого ключевого порога интеллектуальности – такого, как критичности (criticality) самоулучшения. Два необходимых предположения таковы:

- Первый ИИ, который достиг некого критического порога (то есть критичности самоулучшений), будучи неДружественным, может истребить человеческий вид.

- Если первый ИИ, который достигнет этого уровня, будет Дружественным, то он сможет не допустить возникновения враждебных ИИ или причинения ими вреда человеческому виду; или найдёт другие оригинальные пути, чтобы обеспечить выживание и процветание возникшей на Земле разумной жизни.

Более, чем один сценарий соответствует эффекту начавшего первым. Каждый из следующих примеров отражает другой ключевой порог:

- Пост-критический, самоулучшающийся ИИ достигает сверхинтеллекта в течение недель или меньше. Проекты ИИ достаточно редки, так что ни один другой ИИ не достигает критичности до того, как начавший первым ИИ становится достаточно сильным, чтобы преодолеть любое сопротивление. Ключевым порогом является критический уровень самоулучшения.

- ИИ-1 разрешает проблему свёртывания белков на три дня раньше ИИ-2. ИИ-1 достигает нанотехнологии на 6 часов раньше, чем ИИ-2. С помощью быстрых манипуляторов веществом ИИ-1 может (потенциально) отключить исследования и разработку ИИ-2 до её

созревания. Бегуны близки, но тот, кто первым пересекает финишную черту – побеждает. Ключевым порогом здесь является быстрая инфраструктура.

- тот ИИ, который первым поглощает интернет, может (потенциально) не допустить в него другие ИИ. Затем, посредством экономического доминирования, скрытых действий или шантажа или превосходящих способностей к социальной манипуляции, первый ИИ останавливает или замедляет другие ИИ проекты, так что никакого другого ИИ не возникает. Ключевой порог – поглощение уникального ресурса.

Человеческий вид, Homo sapiens, является начавшим первым. С точки зрения эволюции, наши кузены – шимпанзе – отстают от нас только на толщину волоса. Homo sapiens заполучили все технологические чудеса, потому что мы попали сюда немного раньше. Эволюционные биологи всё ещё пытаются выяснить порядок ключевых порогов, потому что начавшие первыми виды должны были первыми пересечь столь много порогов: речь, технология, абстрактное мышление. Мы всё ещё пытаемся понять, что первым вызвало эффект домино. Результат состоит в том, что Homo Sapiens движется первым без нависшего сзади соперника. Эффект движущегося первым предполагает теоретически локальную стратегию (задачу, реализуемую, в принципе, исключительно местными усилиями), нот при этом вызывает к жизни технический вызов чрезвычайной трудности. Нам нужно правильно создать Дружественный ИИ только в одном месте и один раз, а не каждый раз везде. Но создать его нужно правильно с первой попытки, до того, как кто-то построит ИИ с более низкими стандартами.

Я не могу произвести точных вычислений на основании точно подтверждённой теории, но моё мнение сейчас состоит в том, что резкие прыжки в интеллектуальности возможны, вероятны и являют собой доминирующую возможность. Это не та область, в которой я хотел бы давать узкие интервалы уверенности, и поэтому стратегия не должна потерпеть катастрофу – то есть не оставить нас в ситуации худшей, чем раньше, - если резкий прыжок в интеллектуальности не произойдёт. Но гораздо более серьёзной проблемой являются стратегии, представляемые для медленно растущего ИИ, которые терпят катастрофу, если здесь есть эффект движущегося первым. Это более серьёзная проблема, потому что:

- Более быстро растущий ИИ является более сложной технической задачей.

- Подобно автомобилю, едущему по мосту для грузовиков, ИИ, спроектированный, чтобы оставаться Дружественным в экстремально сложных условиях (предположительно) остаётся Дружественным в менее сложных условиях. Обратное неверно.

- Быстрые скачки в интеллектуальности контр-интуитивны с точки зрения обычной социальной жизни. Метафора G-фактора для ИИ является интуитивной, притягательной, заверяющей и, по общему согласию, требующей меньше конструктивных ограничений.

- Моя нынешняя догадка состоит в том, что кривая интеллектуальности содержит огромные, резкие (потенциально) скачки.

Моя теперешняя стратегическая точка зрения имеет тенденцию фокусироваться на трудном локальном сценарии: первый ИИ должен быть Дружественным. С этой мерой предосторожности, если никаких быстрых прыжков в ИИ не произойдёт, можно переключиться на стратегию, которая сделает большинство ИИ Дружественными. В любом случае, технические усилия, которые ушли на подготовку к экстремальному случаю появления первого ИИ, не сделают нам хуже.

Сценарий, который требует невозможной – требующей единодушия – стратегии:

- Единственный ИИ может быть достаточно силён, чтобы уничтожить человечество, даже несмотря на защитные меры Дружественных ИИ.

- Ни один ИИ недостаточно могуществен, чтобы остановить людей-исследователей от создания одного ИИ за другим (или найти другой творческий путь решения проблемы.).

Хорошо, что этот баланс возможностей кажется невероятным a приори, потому что при таком сценарии мы обречены. Если вы выкладываете на стол колоду карт одна за другой, вы рано или поздно выложите туза треф.

Та же проблема относится и к стратегии намеренного конструирования ИИ, которые выбирают не увеличивать свои способности выше определённого уровня. Если ограниченные ИИ недостаточно сильны, чтобы победить неограниченных, или предотвратить их возникновение, то тогда ограниченные ИИ вычёркиваются из уравнения. Мы участвуем в игре, до тех пор, пока мы не вытащим сверхинтеллект, независимо оттого, что это – туз червей или туз треф. Мажоританые стратегии работают, только если

невозможно для одиночного вредителя причинить катастрофический ущерб. Для ИИ эта возможность является свойством самого пространства возможных проектов (design space) – эта возможность не зависит от человеческого решения, равно как скорость света или гравитационная константа.

**11: ИИ и усиление человеческого интеллекта. (AI versus human intelligence enhancement)**

Я не нахожу достоверным, что Homo sapiens будут продолжать существовать в неограниченном будущем, тысячи или миллионы или миллиарды лет, без того, чтобы возник хотя бы один ум, который бы прорвал верхний предел интеллектуальности. И если так, то придёт время, когда люди впервые встретятся с вызовом более умного, чем человек, интеллекта. И если мы выиграем первый уровень схватки, то человечество сможет взывать к более умному, чем человек, интеллекту в следующих раундах схватки.

Возможно, мы скорее выберем другой путь, чем ИИ, более умный, чем человек, - например, будем улучшать людей вместо этого. Чтобы рассмотреть крайний случай, допустим, что кто-то говорит: «Перспектива ИИ меня беспокоит. Я бы предпочёл, чтобы, до того, как какой-либо ИИ был сконструирован, отдельный люди были бы отсканированы в компьютеры, нейрон за нейроном, и затем усовершенствованы, медленно, но наверняка, пока они не станут сверх-умными; и это та основа, на которой человечество должно сразиться с вызовом суперинтеллекта».

Здесь мы сталкиваемся с двумя вопросами: Возможен ли этот сценарий? И если да, то желателен ли он? (Разумно задавать вопросы именно в такой последовательности, по причинам рациональности: мы должны избегать эмоциональной привязки к привлекательным возможностям, которые не являются реальными возможностями.)

Представим, что некий человек сканирован в компьютер, нейрон за нейроном, как предлагает Moravec (1988). Отсюда однозначно следует, что использованная компьютерная мощность значительно превосходит вычислительную мощность человеческого мозга. Согласно гипотезе, компьютер выполняет детальную симуляцию биологического человеческого мозга, исполняемую с достаточной точностью, чтобы избежать каких-либо обнаружимых высокоуровневых эффектов от системных низкоуровневых ошибок.

Каждый биологический аспект, который любым образом влияет на переработку информации, мы должны тщательно симулировать с достаточной точностью, чтобы общий ход процесса

был изоморфен оригиналу. Чтобы симулировать беспорядочный (messy) биологический компьютер, каким является человеческий мозг, мы должны иметь гораздо больше полезной компьютерной силы, чем воплощено в самом беспорядочном человеческом мозге.

Наиболее вероятный способ, который будет создан, чтобы сканировать мозг нейрон за нейроном – с достаточным разрешением, чтобы захватить любой когнитивно важный аспект нейронной структуры – это развитая молекулярная нанотехнология. (4)

(сноска 4) Albeit Merkle (1989) предполагает, что нереволюционное развитие технологий считывания, таких как электронная микроскопия или оптические срезы (optical sectioning) может быть достаточно для загрузки целого мозга.

Молекулярная нанотехнология, возможно, позволит создать настольный компьютер с общей вычислительной мощностью, превосходящей суммарную мозговую мощь всей человеческой популяции. (Bostrom 1998; Moravec 1999; Merkle и Drexler 1996; Sandberg 1999.) Более того, если технология позволит нам сканировать мозг с достаточной точностью, чтобы выполнять этот скан в качестве кода, это означает, что за несколько лет до того эта технология была способна создать невероятно точные картины процессов в нейронных сетях, и, предположительно, исследователи сделали всё от них зависящее, чтобы понять их. Более того, чтобы проапгрейдить загруженное – трансформировать скан мозга, чтобы усилить интеллект ума внутри него – мы обязательно должны понимать во всех деталях высокоуровневые функции мозга, и какой полезный вклад они делают в интеллект.

Более того, люди не созданы для того, чтобы их улучшали, ни внешние нейробиологи, ни посредством рекурсивного самоулучшения изнутри. Естественный отбор не создал человеческий мозг удобным для людей-хакеров. Все сложные механизмы в мозгу адаптированы для работы в узких параметрах конструкции мозга. Допустим, вы можете сделать человека умнее, не говоря уже о сверхинтеллекте; останется ли он вменяемым (sane)? Человеческий мозг очень легко разбалансировать; достаточно изменить баланс нейротрансмиттеров, чтобы запустить шизофрению или другие расстройства. В Deacon (1997) представлено отличное описание эволюции человеческого мозга того, как деликатно элементы мозга сбалансированы, и как это отражается в дисфункциях современного мозга. Человеческий мозг немодифицируем конечным пользователем.

Всё это делает весьма невероятным, что первое человеческое существо будет сканировано в компьютер и вменяемо усовершенствовано до того, как кто-нибудь где-нибудь первым

построит ИИ. В тот момент, когда технология впервые станет способна осуществить загрузку, это потребует невообразимо больше компьютерной мощности и гораздо лучшей науки о мышлении, чем требуется, чтобы построить ИИ. Построить Боинг 747 с нуля непросто. Но проще ли:

- начать с существующего дизайна биологической птицы
- и путём пошаговых добавлений модифицировать этот дизайн через серию успешных стадий
- где каждая стадия независимо жизнеспособна
- так что в конечном итоге мы имеем птицу, растянутую до размеров 747ого
- которая на самом деле летает
- также быстро, как 747
- и затем провести серию трансформаций реальной живой птицы
- не убивая её и не причиняя ей невыносимых страданий.

Я не хочу сказать, что это никогда не может быть сделано. Я хочу сказать, что проще сделать 747, и, имея уже 747-ой, метафорически говоря, апгрейдить птицу. «Давайте просто увеличим птицу до размеров 747-ого» не выглядит в качестве разумной стратегии, избегающей контакта с устрашающе сложной теоретической мистерией аэродинамики. Может быть, в начале, всё, что вы знаете о полёте – это то, что птица обладает загадочной сущностью полёта, и что материалы, из которых вы должны построить 747ой просто лежат здесь на земле. Но вы не можете слепить загадочную сущность полёта, даже если она уже имеется в птице, до тех пор, пока она не перестанет быть загадочной сущностью для вас. Вышеприведённый довод предложен как нарочито экстремальный случай. Основная идея в том, что у нас нет абсолютной свободы выбирать путь, который выглядит позитивным и утешительным, или который будет хорошей историей для научно-фантастического романа. Мы ограничены тем, какие технологии будут, скорее всего, предшествовать другим. Я не против сканирования человеческих существ в компьютеры и делания их умнее, но кажется чрезвычайно маловероятным, что это будет полем, на котором люди впервые столкнутся с вызовом превосходящего человеческий интеллекта. Из различных ограниченных наборов технологий и знаний, требуемых, чтобы загружать и усовершенствовать людей, можно выбрать:

- апгрейдить биологические мозги на месте (например, добавляя новые нейроны, которые полезным образом встраиваются в работу);
- или продуктивно связать компьютеры с биологическими человеческими мозгами.
- или продуктивно связать мозги людей друг с другом
- или сконструировать ИИ.

Далее, это одно дело усилить среднего человека, сохраняя его здравомыслие, до IQ 140, и другое – развить Нобелевского лауреат до чего-то за пределами человеческого. (Отложим в сторону каламбуры по поводу IQ или Нобелевских призов как меры совершенного интеллекта; простите меня за мои метафоры.) Приём пирацетама (или питьё кофеина) может сделать, а может и не сделать, по крайней мере, некоторых людей умнее; но это не сделает вас существенно умнее Эйнштейна. Это не даёт нам никаких существенных новых способностей; мы не переходим на следующие уровни проблемы; мы не пересекаем верхние границы интеллекта, доступного нам, чтобы взаимодействовать с глобальными рисками. С точки зрения управления глобальными рисками, любая технология улучшения интеллекта, которая не создаёт (позитивного и вменяемого) сознания, буквально более умного, чем человек, ставит вопрос о том, стоило ли, возможно, те же время и усилия более продуктивно потратить на то, чтобы найти чрезвычайно умных современных людей и натравить их на ту же самую проблему. Более того, чем дальше вы уходите от «естественных» границ конструкции человеческого мозга – от наследственного состояния мозга, к которому отдельные компоненты мозга адаптированы – тем больше опасность личного безумия. Если улучшенные люди существенно умнее обычных, это тоже глобальный риск. Сколько ущерба усовершенствованные в сторону зла люди может причинить? Насколько они творческие? Первый вопрос, который мне приходит в голову: «Достаточно творческие, чтобы создать свой собственный рекурсивно улучшающийся ИИ?» Радикальные техники улучшения человеческого интеллекта поднимают свои вопросы безопасности. Опять, я не говорю, что эти проблемы технически не разрешимы; только указываю на то, что эти проблемы существуют. ИИ имеет спорные вопросы, связанные с безопасностью; тоже касается и усовершенствования человеческого интеллекта. Не всё, что лязгает – это ваш враг, и не всё, что хлюпает – друг. С одной стороны, позитивный человек начинает со всей огромной моральной, этической и структурной сложности, которая описывает то, что мы называем «дружественным» решением. С другой стороны, ИИ может быть спроектирован для стабильного рекурсивного самоулучшения и нацелен на безопасность: естественный отбор не создал человеческий мозг с множеством кругов мер предосторожности, осторожного процесса принятия решений и целыми порядками величины полей безопасности.

Улучшение человеческого интеллекта это самостоятельный вопрос, а не подраздел ИИ; и в этой статье нет места, чтобы обсуждать его в деталях. Стоит отметить, что я рассматривал как улучшение человеческого интеллекта, так и ИИ в начале своей карьеры, и решил сосредоточить свои усилия на ИИ. В первую очередь, потому что я не ожидал, что полезные, превосходящие человеческий уровень техники улучшения человеческого интеллекта

появятся достаточно вовремя, чтобы существенно повлиять на развитие рекурсивно самоулучшающегося ИИ. Я буду рад, если мне докажут, что я не прав в отношении этого. Но я не думаю, что это жизнеспособная стратегия – нарочно выбрать не работать над Дружественным ИИ, пока другие работают над усовершенствованием человеческого интеллекта, в надежде, что усовершенствованные люди решат проблему Дружественного ИИ лучше. Я не хочу вовлекаться в стратегию, которая потерпит катастрофическое поражение, если усовершенствование человеческого интеллекта потребует больше времени, чем создание ИИ. (Или наоборот.) Я боюсь, что работа с биологией займёт слишком много времени – здесь будет слишком много инерции, слишком много борьбы с плохими конструкторскими решениями, уже сделанными естественным отбором. Я боюсь, что регуляторные органы не одобрять экспериментов с людьми. И даже человеческие гении тратят годы на обучение своему искусству; и чем быстрее улучшенный человек должен учиться, тем труднее улучшить кого-либо до этого уровня.

Я буду приятно удивлён, если появятся улучшенные люди (augmented humans) и построят Дружественный ИИ раньше всех. Но тот, кто хотел бы видеть этот результат, должен, вероятно, тяжело трудиться над ускорением технологий улучшения интеллекта; будет трудно убедить меня замедлиться. Если ИИ по своей природе гораздо более сложен, чем усиление интеллекта, то никакого вреда не будет; если же построение 747-ого естественным путём проще, чем увеличивание птицы до его размеров, то промедление будет фатальным. Имеется только небольшая область возможностей, внутри которой намеренный отказ от работы над Дружественным ИИ может быть полезен, и большая область, где это будет неважно или опасно. Даже если усиление человеческого интеллекта возможно, здесь есть реальные, сложные вопросы безопасности; мне следовало бы серьёзно задаться вопросом, хотим ли мы, чтобы Дружественный ИИ предшествовал усилению интеллекта, или наоборот.

Я не приписываю высокой достоверности утверждению, что Дружественный ИИ проще, чем усовершенствование человека, или что он безопаснее. Есть много приходящих на ум путей улучшить человека. Может быть, существует техника, которая прощу и безопаснее, чем ИИ, достаточно мощная, чтобы оказать влияние на глобальные риски. Если так, я могу переключить направление своей работы. Но я желал указать на некоторые соображения, которые указывают против принимаемого без вопросов предположения, что улучшение человеческого интеллекта проще, безопаснее и достаточно мощно, чтобы играть заметную роль.

**12: Взаимодействие ИИ и других технологий. (Interactions of AI with other technologies)**

Ускорение желательной технологии – это локальная стратегия, тогда как замедление опасной технологии – это трудная мажоритарная стратегия. Остановка или отказ от нежелательной технологии имеет тенденцию требовать невозможную единодушную стратегию. Я предлагаю думать не в терминах развития или неразвития некоторых технологий, но в терминах прагматичных доступных возможностей ускорять или замедлять технологии; и задаваться вопросом, в границах этих возможностей, какие технологии мы бы предпочли бы видеть развитыми до или после одна другой.

В нанотехнологиях, обычно предлагаемая цель состоит в развитии защитных щитов до появления наступательных технологий. Я очень обеспокоен этим, поскольку заданный уровень наступательной технологии обычно требует гораздо меньших усилий, чем технология, которая может защитить от него. Наступление превосходило оборону в течение большей части человеческой истории. Ружья были созданы за сотни лет до пуленепробиваемых жилетов. Оспа была использована как орудие войны до изобретения вакцины от оспы. Сейчас нет защиты от ядерного взрыва; нации защищены не благодаря обороне, превосходящей наступательные силы, а благодаря балансу угроз наступления. Нанотехнологии оказались по самой своей природе сложной проблемой. Так что, должны ли мы предпочесть, чтобы нанотехнологии предшествовали развитию ИИ, или ИИ предшествовал развитию нанотехнологий? Заданный в такой форме, это несколько мошеннический вопрос. Ответ на него не имеет ничего общего с присущей нанотехнологиям проблемностью в качестве глобального риска, или с собственной сложностью ИИ. В той мере, в какой мы беспокоимся о порядке возникновения, вопрос должен звучать: «Поможет ли ИИ нам справиться с нанотехнологиями? Помогут ли нанотехнологии нам справится с ИИ?»

Мне кажется, что успешное создание ИИ существенно поможет нам во взаимодействии с нанотехнологиями. Я не вижу, как нанотехнологии сделают более простым развитие Дружественного ИИ. Если мощные нанокомпьютеры сделают проще создание ИИ, без упрощения решения самостоятельной проблемы Дружественности, то это – негативное взаимодействие технологий. Поэтому, при прочих равных, я бы очень предпочёл, чтобы Дружественный ИИ предшествовал нанотехнологиям в порядке технологических открытий. Если мы справимся с вызовом ИИ, мы сможем рассчитывать на помощь Дружественного ИИ в отношении нанотехнологий. Если мы создадим нанотехнологии и выживем, нам всё ещё будет предстоять принять вызов взаимодействия с ИИ после этого.

Говоря в общем, успех в Дружественном ИИ должен помочь в решении почти любой другой проблемы. Поэтому, если некая технология делает ИИ не проще и не труднее, но несёт собой определённый глобальный риск, нам следует предпочесть, при прочих равных, в первую очередь встретиться с вызовом ИИ. Любая технология, увеличивающая доступную мощность компьютеров, уменьшает минимальную теоретическую сложность, необходимую для создания ИИ, но нисколько не помогает в Дружественности, и я считаю её в сумме негативной. Закон Мура для Безумной Науки: каждые 18 месяцев минимальный IQ, необходимый, чтобы уничтожить мир, падает на один пункт. Успех в усилении человеческого интеллекта сделает Дружественный ИИ проще, а также поможет в других технологиях. Но улучшение людей не обязательно безопаснее, или проще, чем Дружественный ИИ; оно также не находится в реалистически оцененных пределах наших возможностей изменить естественный порядок возникновения улучшения людей и Дружественного ИИ, если одна из технологий по своей природе гораздо проще другой.

**13: Ход прогресса в области Дружественного ИИ. (Making progress on Friendly AI)**

«Мы предлагаем, чтобы в течение 2 месяцев, десять человек изучали искусственный интеллект летом 1956 года в Дармутском колледже, Ганновер, Нью Гемпшир. Исследование будет выполнено на основе предположения, что любой аспект обучения или любое другое качество интеллекта может быть в принципе столь точно описано, что может быть сделана машина, чтобы симулировать его. Будет предпринята попытка узнать, как сделать так, чтобы машины использовали язык, формировали абстракции и концепции, разрешали бы те проблемы, которые сейчас доступны только людям, и улучшали себя. Мы полагаем, что возможно существенное продвижение в одной или нескольких из этих работ, если тщательно подобранная группа учёных проработает над этим вместе в течение лета».
---- McCarthy, Minsky, Rochester, и Shannon (1955).
Предложение Дартмутского Летнего Исследовательского Проекта по Искусственному Интеллекту являет собой первое зафиксированное употребление фразы «Искусственный Интеллект». У них не было предыдущего опыта, который мог бы их предупредить, что проблема трудна. Я бы назвал искренней ошибкой то, что они сказали, что «значительное продвижение может быть сделано», а не есть «есть небольшой шанс на значительное продвижение». Это специфическое утверждение относительно трудности проблемы и времени решения, которое усиливает степень невозможности. Но если бы они сказали «есть небольшой шанс», у меня бы не было возражений. Как они могли знать?

Дартмутское предложение включало в себя, среди прочего, следующие темы: лингвистические коммуникации, лингвистические рассуждения, нейронные сети, абстрагирование, случайность и творчество, взаимодействие с окружением, моделирование мозга, оригинальность, предсказание, изобретение, открытие и самоулучшение.

(Сноска 5) Это обычно правда, но не есть универсальная истина. В последней главе широко использован учебник «Искусственный интеллект: современный подход». (Russell and Norvig 2003) (Включающий раздел «Этика и риски Искусственного интеллекта», упоминающий взрыв интеллекта по I.J.Good и Сингулярность, и призывающий к дальнейшим исследованиям.) Но и к 2006 году это отношение является скорее исключением, чем правилом.

Теперь мне кажется, что ИИ, способный к языкам, абстрактному мышлению, творчеству, взаимодействию с окружением, оригинальности, предсказаниям, изобретению, открытиям, и, прежде всего, к самоулучшению, находится далеко за пределами того уровня, на котором он должен быть так же и Дружественным. В Дармутском предложении ничего не говорится о построении позитивного / доброго / благоволящего ИИ. Вопросы безопасности не обсуждены даже с целью отбросить их. И это в то искреннее лето, когда ИИ человеческого уровня казался прямо за углом. Дармутское предложение было написано в 1955 году, до Асиломарской (Asilomar) конференции по биотехнологии, детей, отравленных тамиламидом во время беременности, Чернобыля и 11 Сентября. Если бы сейчас идея искусственного интеллекта был бы предложена в первый раз, кто-то доложен был бы постараться выяснить, что конкретно делается для управления рисками. Я не могу сказать, хорошая это перемена в нашей культуре или плохая. Я не говорю, создаёт ли это хорошую или плохую науку. Но сутью остаётся то, что если бы Дартмутское предложение было бы написано 50 лет спустя, одной из его тем должна была бы стать безопасность.

В момент написания этой статьи в 2006 году, сообщество исследователей ИИ по-прежнему не считает Дружественный ИИ частью проблемы. Я бы хотел цитировать ссылки на этот эффект, но я не могу цитировать отсутствие литературы. Дружественный ИИ отсутствует в пространстве концепций, а не просто не популярен или не финансируем. Вы не можете даже назвать Дружественный ИИ пустым местом на карте, поскольку нет понимания, что что-то пропущено. (5) Если вы читали научно-популярные/полутехнические книги, предлагающие, как построить ИИ, такие как «Гёдель, Эшер, Бах». (Hofstadter, 1979) или «Сообщество сознаний» (Minsky, 1986), вы можете вспомнить, что вы не видели обсуждения Дружественного ИИ в качестве части проблемы. Точно так же я не видел обсуждения

Дружественного ИИ как технической проблемы в технической литературе. Предпринятые мною литературные изыскания обнаружили в основном краткие нетехнические статьи, не связанные одна с другой, без общих ссылок за исключением «Трёх законов Робототехники» Айзека Азимова. (Asimov, 1942.) Имея в виду, что сейчас уже 2006 год, почему не много исследователей ИИ, которые говорят о безопасности? У меня нет привилегированного доступа к чужой психологии, но я кратко обсужу этот вопрос, основываясь на личном опыте общения.

Поле исследований ИИ адаптировалось к тому жизненному опыту, через который оно прошло за последние 50 лет, в частности, к модели больших обещаний, особенно способностей на уровне человека, и следующих за ними приводящих в замешательство публичных провалов. Относить это замешательство к самому ИИ несправедливо; более мудрые исследователи, которые не делали больших обещаний, не видели триумфа своего консерватизма в газетах. И сейчас невыполненные обещания тут же приходят на ум, как внутри, так и за пределами поля исследований ИИ, когда ИИ упоминается. Культура исследований ИИ адаптировалась к следующему условию: имеется табу на разговоры о способностях человеческого уровня. Есть ещё более сильное табу против тех, кто заявляет и предсказывает некие способности, которые они ещё не продемонстрировали на работающем коде.

У меня сложилось впечатление, что каждый, кто заявляет о том, что исследует Дружественный ИИ, косвенным образом заявляет, что его проект ИИ достаточно мощен, чтобы быть Дружественным.

Должно быть очевидно, что это не верно ни логически, ни философски. Если мы представим себе кого-то, кто создал реальный зрелый ИИ, который достаточно мощен для того, чтобы быть Дружественным, и, более того, если, в соответствии с нашим желаемым результатом, этот ИИ действительно является Дружественным, то тогда кто-то должен был работать над Дружественным ИИ годы и годы. Дружественный ИИ – это не модуль, который вы можете мгновенно изобрести, в точный момент, когда он понадобится, и затем вставить в существующий проект, отполированный дизайн которого в остальных отношениях никак не изменится.

Поле исследований ИИ имеет ряд техник, таких как нейронные сети и эволюционное программирование, которые росли маленькими шажками в течение десятилетий. Но нейронные сети непрозрачны – пользователь не имеет никакого представления о том, как

нейронные сети принимают свои решения – и не могут быть легко приведены в состояние прозрачности; люди, которые изобрели и отшлифовали нейронные сети, не думали о долгосрочных проблемах Дружественного ИИ. Эволюционное программирование (ЭП) является стохастическим, и не сохраняет точно цель оптимизации в сгенерированном коде; ЭП даёт вам код, который делает то, что вы запрашиваете – большую часть времени в определённых условиях, но этот код может делать что-то на стороне. ЭП – это мощная, всё более зрелая техника, которая по своей природе не подходит для целей Дружественного ИИ. Дружественный ИИ, как я его представляю, требует рекурсивных циклов самоулучшения, которые абсолютно точно сохраняют цель оптимизации.

Наиболее сильные современные техники ИИ, так, как они были развиты, отполированы и улучшены с течением времени, имеют основополагающую несовместимость с требованиями Дружественного ИИ, как я их сейчас понимаю. Проблема Y2K, исправить которую оказалось очень дорого, хотя это и не было глобальной катастрофой, - точно так же произошла из неспособности предвидеть завтрашние проектные требования. Кошмарным сценарием является то, что мы можем обнаружить, что нам всучили каталог зрелых, мощных, публично доступных техник ИИ, которые соединяются, чтобы породить неДружественный ИИ, но которые нельзя использовать для построения Дружественного ИИ без переделывания всей работы за три десятилетия с нуля. В поле исследований ИИ довольно вызывающе открыто обсуждать ИИ человеческого уровня, в связи с прошлым опытом этих дискуссий. Есть соблазн поздравить себя за подобную смелость, и затем остановиться. После проявления такой смелости обсуждать трансчеловеческий ИИ кажется смешным и ненужным. (Хотя нет выделенных причин, по которым ИИ должен был бы медленно взбираться по шкале интеллектуальности, и затем навсегда остановиться на человеческой точке.) Осмеливаться говорить о Дружественном ИИ, в качестве меры предосторожности по отношению к глобальному риску, будет на два уровня смелее, чем тот уровень смелости, на котором выглядишь нарушающим границы и храбрым.

Имеется также резонное возражение, которое согласно с тем, что Дружественный ИИ является важной проблемой, но беспокоится, что, с учётом нашего теперешнего понимания, мы просто не на том уровне, чтобы обращаться с Дружественным ИИ: если мы попытаемся разрешить проблему прямо сейчас, мы только потерпим поражение, или создадим анти-науку вместо науки. И об этом возражении стоит обеспокоится. Как мне кажется, необходимые знания уже существуют – что возможно изучить достаточно большой объём существующих знаний и затем обращаться с Дружественным ИИ без того, чтобы вляпаться лицом в кирпичную стену – но эти знания разбросаны среди множества дисциплин: Теории

решений и эволюционной психологии и теории вероятностей и эволюционной биологии и когнитивной психологии и теории информации и в области знаний, традиционно известной как «Искусственный интеллект»… Не существует также учебной программы, которая бы подготовила большой круг учёных для работ в области Дружественного ИИ.

«Правило десяти лет» для гениев, подтверждённое в разных областях – от математике до тенниса – гласит, что никто не достигает выдающихся результатов без по крайней мере десяти лет подготовки. (Hayes, 1981.) Моцарт начал писать симфонии в четыре года, но это не были моцартовские симфонии – потребовалось ещё 13 лет, чтобы Моцарт начал писать выдающиеся симфонии. (Weisberg, 1986.) Мой собственный опыт с кривой обучения подкрепляет эту тревогу. Если нам нужны люди, которые могут сделать прогресс в Дружественном ИИ, то они должны начать тренировать сами себя, всё время, за годы до того, как они внезапно понадобятся.

Если завтра Фонд Билла и Мелинды Гейтс выделит сто миллионов долларов на изучение Дружественного ИИ, тогда тысячи учёных начнут переписывать свои предложения по грантам, чтобы они выглядели релевантными по отношению к Дружественному ИИ. Но они не будут искренне заинтересованы в проблеме – свидетельство чему то, что они не проявляли любопытства к проблеме до того, как кто-то предложил им заплатить. Пока Универсальный ИИ немоден и Дружественный ИИ полностью за пределами поля зрения, мы можем предположить, что каждый говорящий об этой проблеме искренне заинтересован в ней. Если вы вбросите слишком много денег в проблему, область которой ещё не готова к решению, излишние деньги создадут скорее анти-науку, чем науку – беспорядочную кучу фальшивых решений.

Я не могу считать этот вывод хорошей новостью. Мы были бы в гораздо большей безопасности, если бы проблема Дружественного ИИ могла бы быть разрешена путём нагромождения человеческих тел и серебра. Но на момент 2006 года я сильно сомневаюсь, что это годится – область Дружественного ИИ, и сама область ИИ, находится в слишком большом хаосе. Если кто-то заявляет, что мы не можем достичь прогресса в области Дружественного ИИ, что мы знаем слишком мало, нам следует спросить, как долго этот человек учился, чтобы придти к этому заключению. Кто может сказать, что именно наука не знает? Слишком много науки существует в природе, чтобы один человек мог её выучить. Кто может сказать, что мы не готовы к научной революции, опережая неожиданное? И если мы не можем продвинуться в Дружественном ИИ, потому что мы не готовы, это не означает, что нам не нужен Дружественный ИИ. Эти два утверждения вовсе не эквивалентны!

И если мы обнаружим, что не можем продвинуться в Дружественном ИИ, мы должны определить, как выйти из этой ситуации как можно скорее! Нет никаких гарантий в любом случае, что раз мы не можем управлять риском, то он должен уйти.

И если скрытые таланты юных учёных будут заинтересованы в Дружественном ИИ по своему собственному выбору, тогда, я думаю, будет очень полезно с точки зрения человеческого вида, если они смогут подать на многолетний грант, чтобы изучать проблему с полной занятостью. Определённое финансирование Дружественного ИИ необходимо, чтобы это сработало – значительно большее финансирование, чем это имеется сейчас. Но я думаю, что на этих начальных стадиях Манхетенский проект только бы увеличил долю шума в системе.

**Заключение**

Однажды мне стало ясно, что современная цивилизация находится в нестабильном состоянии. I.J. Good предположил, что взрыв интеллекта описывает динамическую нестабильную систему, вроде ручки, точно сбалансированной, чтобы стоять на своём кончике. Если ручка стоит совершенно вертикально, она может оставаться в прямом положении, но если ручка отклоняется даже немного от вертикали, гравитация потянет её дальше в этом направлении, и процесс ускорится. Точно так же и более умные системы будут требовать меньше времени, чтобы сделать себя ещё умнее.

Мёртвая планета, безжизненно вращающаяся вокруг своей звезды, тоже стабильна. В отличие от интеллектуального взрыва, истребление не является динамическим аттрактором – есть большая разница между «почти исчезнуть» и «исчезнуть». Даже в этом случае, тотальное истребление стабильно.

Не должна ли наша цивилизация, в конце концов, придти в один из этих двух режимов? Логически, вышеприведённое рассуждение содержит проколы. Например, Ошибочность Гигантской Ватрушки: умы не бродят слепо между аттракторами, у них есть мотивы. Но даже если и так, то, я думаю, наша альтернатива состоит в том, что или стать умнее, или вымереть.

Природа не жестока, но равнодушна; эта нейтральность часто выглядит неотличимой от настоящей враждебности. Реальность бросает перед вами один выбор за другим, и когда вы

сталкиваетесь с вызовом, с которым не можете справиться, вы испытываете последствия. Часто природа выдвигает грубо несправедливые требования, даже в тех тестах, где наказание за ошибку – смерть. Как мог средневековый крестьянин 10 века изобрести лекарство от туберкулёза? Природа не соизмеряет свои вызовы с вашими умениями, вашими ресурсами, или тем, сколько свободного времени у вас есть, чтобы обдумать проблему. И когда вы сталкиваетесь со смертельным вызовом, слишком сложным для вас, вы умираете. Может быть, неприятно об этом думать, но это было реальностью для людей в течение тысяч и тысяч лет. Тоже самое может случится и со всем человеческим видом, если человеческий вид столкнётся с несправедливым вызовом.

Если бы человеческие существа не старели, и столетние имели бы такой же уровень смерти, как и 15-летиние, люди всё равно не были бы бессмертными. Мы будем продолжать существовать, пока нас поддерживает вероятность. Чтобы жить даже миллион лет в качестве не стареющего человека в мире, столь рискованном, как наш, вы должны каким-то образом свести свою годовую вероятность смерти почти к нулю. Вы не должны водить машину, не должны летать, вы не должны пересекать улицу, даже посмотрев в обе стороны, поскольку это всё ещё слишком большой риск. Даже если вы отбросите все мысли о развлечениях и бросите жить ради сохранения своей жизни, вы не сможете проложить миллионолетний курс без препятствий. Это будет не физически, а умственно (cognitively) невозможно.

Человеческий вид Homo sapiens не стареет, но не бессмертен. Гоминиды прожили так долго только потому, что не было арсенала водородных бомб, не было космических кораблей, чтобы направлять астероиды к Земле, не было воённых биолабораторий, чтобы создавать супервирусов, не было повторяющихся ежегодных перспектив атомной войны или нанотехнологической войны или отбившегося от рук ИИ. Чтобы прожить какое-либо заметное время, мы должны свести каждый из этих рисков к нулю. «Довольно хорошо» - это недостаточно хорошо, для того, чтобы прожить ещё миллион лет.

Это выглядит как несправедливый вызов. Этот вопрос обычно не был в компетенции исторических человеческих организаций, не зависимо от того, насколько они старались. В течение десятилетий США и СССР избегали ядерной войны, но не были в этом совершенны; были очень близкие к войне моменты, например, Кубинский ракетный кризис в 1962 году. Если мы предположим, что будущие умы будут являть ту же смесь глупости и мудрости, ту же смесь героизма и эгоизма, как те, о ком мы читаем в исторических книгах – тогда игра в

глобальный риск практически закончена; она была проиграна с самого начала. Мы можем прожить ещё десятилетие, даже ещё столетие, но не следующие миллион лет.

Но человеческие умы – это не граница возможного. Homo sapiens представляет собой первый универсальный интеллект. Мы были рождены в самом начале вещей, на рассвете ума. В случае успеха, будущие историки будут оглядываться назад и описывать современный мир как труднопреодолимую промежуточную стадию юности, когда человечество стало достаточно смышлёным, чтобы создать себе страшные проблемы, но недостаточно смышлёным, чтобы их решить.

Но до того как мы пройдём эту стадию юности, мы должны, как юноши, встретится с взрослой проблемой: вызовом более умного, чем человек, интеллекта. Это – выход наружу на высоко-моральной стадии жизненного цикла; путь, слишком близкий к окну уязвимости; это, возможно, самый опасный однократный риск, с которым мы сталкиваемся. ИИ – это единственная дорога в этот вызов, и я надеюсь, что мы пройдём эту дорогу, продолжая разговор. Я думаю, что, в конце концов, окажется проще сделать 747ой с нуля, чем растянуть в масштабе существующую птицу или пересадить ей реактивные двигатели.

Я не хочу преуменьшать колоссальную ответственность попыток построить, с точной целью и проектом, нечто, более умное, чем мы сами. Но давайте остановимся и вспомним, что интеллект – это далеко не первая вещь, встретившаяся человеческой науке, которая казалась трудна для понимания. Звёзды когда-то были загадкой, и химия, и биология. Поколения исследователей пытались и не смогли понять эти загадки, и они обрели имидж неразрешимых для простой науки. Когда-то давно, никто не понимал, почему одна материя инертна и безжизненна, тогда как другая пульсирует кровью и витальностью. Никто не знал, как живая материя размножает себя или почему наши руки слушаются наших ментальных приказов. Лорд Кельвин писал:

«Влияние животной или растительной жизни на материю находится бесконечно далеко за пределами любого научного исследования, направленного до настоящего времени на него. Его сила управлять перемещениями движущихся частиц, в ежедневно демонстрируемом чуде человеческой свободной воли и в росте поколений за поколением растений из одного семечка, бесконечно отличается от любого возможного результата случайной согласованности атомов». (Цитировано по MacFie, 1912.)

Любое научное игнорирование освящено древностью. Любое и каждое отсутствие знаний уходит в прошлое, к рассвету человеческой любознательности; и эта дыра длится целые эпохи, выглядя неизменной, до тех пор, пока кто-то не заполняет её. Я думаю, что даже склонные ошибаться человеческие существа способны достичь успеха в создании Дружественного ИИ. Но только если разум перестанет быть сакральной тайной для нас, как жизнь была для Лорда Кельвина. Интеллект должен перестать быть любым видом мистики, сакральным или нет. Мы должны выполнить создание Искусственного Интеллекта как точное приложение точного искусства. И тогда, возможно, мы победим.

**Библиография**


Asimov, I. 1942. Runaround. Astounding Science Fiction, March 1942.

Barrett, J. L. and Keil, F. 1996. Conceptualizing a non-natural entity: Anthropomorphism in God concepts. Cognitive Psychology, 31: 219-247.

Bostrom, N. 1998. How long before superintelligence? Int. Jour. of Future Studies, 2.

Bostrom, N. 2001. Existential Risks: Analyzing Human Extinction Scenarios. Journal of Evolution and Technology, 9.

Brown, D.E. 1991. Human universals. New York: McGraw-Hill.

Crochat, P. and Franklin, D. (2000.) Back-Propagation Neural Network Tutorial. http://ieee.uow.edu.au/~daniel/software/libneural/

Deacon, T. 1997. The symbolic species: The co-evolution of language and the brain. New York: Norton.

Drexler, K. E. 1992. Nanosystems: Molecular Machinery, Manufacturing, and Computation. New York: Wiley-Interscience.

Ekman, P. and Keltner, D. 1997. Universal facial expressions of emotion: an old controversy and new findings. In Nonverbal communication: where nature meets culture, eds. U. Segerstrale and P. Molnar. Mahwah, NJ: Lawrence Erlbaum Associates.

Good, I. J. 1965. Speculations Concerning the First Ultraintelligent Machine. Pp. 31-88 in Advances in Computers, vol 6, eds. F. L. Alt and M. Rubinoff. New York: Academic Press.

Hayes, J. R. 1981. The complete problem solver. Philadelphia: Franklin Institute Press.

Hibbard, B. 2001. Super-intelligent machines. ACM SIGGRAPH Computer Graphics, 35(1).

Hibbard, B. 2004. Reinforcement learning as a Context for Integrating AI Research. Presented at the 2004 AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research.

Hofstadter, D. 1979. Gödel, Escher, Bach: An Eternal Golden Braid. New York: Random House

Jaynes, E.T. and Bretthorst, G. L. 2003. Probability Theory: The Logic of Science. Cambridge: Cambridge University Press.

Jensen, A. R. 1999. The G Factor: the Science of Mental Ability. Psycoloquy, 10(23).

MacFie, R. C. 1912. Heredity, Evolution, and Vitalism: Some of the discoveries of modern research into these matters – their trend and significance. New York: William Wood and Company.

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

Merkle, R. C. 1989. Large scale analysis of neural structure. Xerox PARC Technical Report CSL-89-10. November, 1989.

Merkle, R. C. and Drexler, K. E. 1996. Helical Logic. Nanotechnology, 7: 325-339.

Minsky, M. L. 1986. The Society of Mind. New York: Simon and Schuster.

Monod, J. L. 1974. On the Molecular Theory of Evolution. New York: Oxford.

Moravec, H. 1988. Mind Children: The Future of Robot and Human Intelligence. Cambridge: Harvard University Press.

Moravec, H. 1999. Robot: Mere Machine to Transcendent Mind. New York: Oxford University Press.

Raymond, E. S. ed. 2003. DWIM. The on-line hacker Jargon File, version 4.4.7, 29 Dec 2003.

Rhodes, R. 1986. The Making of the Atomic Bomb. New York: Simon & Schuster.

Rice, H. G. 1953. Classes of Recursively Enumerable Sets and Their Decision Problems. Trans. Amer. Math. Soc., 74: 358-366.

Russell, S. J. and Norvig, P. Artificial Intelligence: A Modern Approach. Pp. 962-964. New Jersey: Prentice Hall.

Sandberg, A. 1999. The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains. Journal of Evolution and Technology, 5.

Schmidhuber, J. 2003. Goedel machines: self-referential universal problem solvers making provably optimal self-improvements. In Artificial General Intelligence, eds. B. Goertzel and C. Pennachin. Forthcoming. New York: Springer-Verlag.

Sober, E. 1984. The nature of selection. Cambridge, MA: MIT Press.

Tooby, J. and Cosmides, L. 1992. The psychological foundations of culture. In The adapted mind: Evolutionary psychology and the generation of culture, eds. J. H. Barkow, L. Cosmides and J. Tooby. New York: Oxford University Press.

Vinge, V. 1993. The Coming Technological Singularity. Presented at the VISION-21 Symposium, sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute. March, 1993.

Wachowski, A. and Wachowski, L. 1999. The Matrix, USA, Warner Bros, 135 min.

Weisburg, R. 1986. Creativity, genius and other myths. New York: W.H Freeman.

Williams, G. C. 1966. Adaptation and Natural Selection: A critique of some current evolutionary thought. Princeton, NJ: Princeton University Press.