ВСЕРОССИЙСКИЙ КОНКУРС «ЮНЫЕ ТЕХНИКИ И ТЕХНИКИ»

Тема: Как с помощью IT-систем предупредить глобальные катастрофы

Работу выполнил: Колоев Акроман Курейшевич

ученик 11 класса ГКОУ « СОШ № 1 с.п. Кантышево» Назрановского района Республики Ингушетия

Научный руководитель: Беков Илез Назирович, учитель математики ГКОУ «СОШ № 1 с.п. Кантышево» Назрановского района РИ.

Республика Ингушетия, 2015 г.

Введение

Общее понятие о предотвратимости глобальных рисков

Очевидно, что если удастся выяснить, что существует несколько простых, очевидных и надёжных способов противостоять глобальным катастрофам, то мы значительно улучшим свою безопасность, а ряд глобальных рисков перестанет нам угрожать. Напротив, если окажется, что во всех предлагающихся мерах и средствах защиты есть свои изъяны, которые делают их в лучшем случае неэффективными, а в худшем – просто опасными, то нам необходимо придумать нечто кардинально новое. Представляется, что система защиты – на каждой фазе развития глобального риска – должна осуществлять следующие функции:

· Наблюдение.

· Анализ информации и принятия решения.

· Уничтожение источника угрозы.

Эта стратегия хорошо отработана в контрразведке, борьбе с терроризмом и военном деле. Другая стратегия предполагает бегство от источника угрозы (космические поселения, бункеры). Очевидно, эта вторая стратегия должна применяться в случае провала первой (или одновременно с ней, на всякий случай).

Глобальные риски различаются по степени того, насколько возможно их предотвращение. Например, вполне реально запретить некий класс опасных экспериментов на ускорителях, если научное сообщество придёт к выводу, что эти эксперименты создают определённый риск. Поскольку в мире всего несколько больших ускорителей, которые управляются достаточно открыто, и потому что сами учёные не желают катастрофы и не имеют от неё никаких выгод, то кажется очень простым отменить

эксперименты. Фактически, для этого нужно только общее понимание их опасности. То есть максимально предотвратимый риск – это риск, который:

1) легко предвидеть,

2) легко достичь научного консенсуса в отношении такого предвидения,

3) этого консенсуса достаточно, чтобы отказаться от действий, ведущих к данному риску.

Отказаться от действий, ведущих некому риску (например, запретить некую опасную технологию), легко лишь при определённых условиях:

А) если опасные процессы создаются только людьми.

Б) если эти процессы создаются в небольшом числе широко известных мест. (Как, например, физические эксперименты на огромных ускорителях)

В) если люди не ждут никакой выгоды от этих процессов.

Г) если опасные процессы предсказуемы как по моменту своего возникновения, так и по ходу развития.

Д) если опасные объекты и процессы легко распознаваемы. То есть мы легко, быстро и наверняка узнаём, что началась некая опасная ситуация, и мы правильно оцениваем степень её опасности.

Е) если у нас есть достаточно времени в силу специфики процесса, чтобы разработать и принять адекватные меры.

Соответственно, риски, которые трудно предотвращать, характеризуются тем, что:

1) Их трудно предвидеть, даже трудно предположить об их возможности. (Даже предположить, что в SETI может быть риск, было трудно.)

2) Даже если кто-то осознаёт этот риск, ему крайне трудно убедить в этом кого-либо ещё (примеры: трудности с осознанием ИИ и SETI как источника риска, трудности доказательства Теоремы о Конце света).

3) Даже если будет достигнуто общественное согласие о том, что подобные риски действительно опасны, это не приведёт к тому, что люди добровольно откажутся от данного источника риска. (Примеры: ядерное оружие.)

Последнее связано с тем, что:

1) Источники риска доступны большому числу людей, а кто эти люди - неизвестно (можно поставить на учёт всех физиков ядерщиков, но не хакеров-самоучек).

2) Источники риска находятся в неизвестном месте и/или их легко скрыть (биолаборатории).

3) Риски создаются независящими от человека природными факторами, или в результате взаимовлияния человеческих действий и природных факторов.

4) Источник опасности сулит не только риски, но и выгоды, в частности, является оружием.

5) Момент начала аварийной ситуации непредсказуем, равно как и то, как она будет развиваться.

6) Опасную ситуацию трудно опознать в качестве таковой, это требует много времени и содержит элемент неопределённости. (Например, трудно определить, что некая новая бактерия является опасной, пока она кого-то не заразит и пока не достигнет таких масштабов, когда можно понять, что это именно эпидемия.)

7) Опасный процесс развивается быстрее, чем мы успеваем на него адекватно реагировать.

Предотвратимость некоторых рисков, однако, не должна приводить к тому, что их следует сбрасывать со счёта, поскольку не обязательно означает, что риск в конечном счёте будет предотвращён. Например, астероидная опасность относится к числу относительно легко предотвратимых рисков, однако реальной противоастероидной (и, что важнее, противокометной) системы защиты у нас сейчас нет. И пока её нет, «предотвратимость» угрозы остаётся чисто гипотетической, поскольку мы не знаем, насколько эффективной и безопасной будет будущая защита, появится ли она вообще и если появится, то когда.

Активные щиты

В качестве способа предотвращения глобальных рисков предлагается создавать разного рода активные щиты. Активный щит – это средство контроля и воздействие на источник риска по всему земному шару. Фактически, это аналог иммунной системы в масштабе всей планеты. В качестве наиболее очевидного примера можно привести идеи создания всемирной системы ПРО.

Активность щитов подразумевает, что они могут относительно автономно реагировать на любой раздражитель, который попадает под определение угрозы. При этом щит полностью покрывает защищаемую поверхность, то есть поверхность Земли. Понятно, что автономный щит опасен неконтролируемым поведением, а управляемый является абсолютным оружием в руках того, кто им управляет. Как нам известно из дискуссий о СОИ, даже если активный щит является полностью оборонительным оружием, он всё равно даёт преимущество в нападении для защищённой стороны, так как она может не опасаться возмездия.

Сравнение активных щитов с иммунной системой человека, как идеальной формой защиты, некорректно, потому что иммунная система неидеальна. Она обеспечивает статистическое выживание вида за счёт того, что отдельные особи живут в среднем достаточно долго. Но она не обеспечивает неограниченное выживание отдельного индивида. Любой человек неоднократно заражается заразными болезнями в течение жизни, и многие от них гибнут. Для любого человека найдётся болезнь, которая его убьёт. Кроме того, иммунная система хорошо работает тогда, когда точно знает патоген. Если она его не знает, то потребуется время, чтобы он успел себя проявить, и ещё время, чтобы иммунная система выработала против него ответ. То же самое происходит и с компьютерными антивирусами, которые тоже являются активным щитом: хотя они обеспечивают устойчивое существование всех компьютеров, каждый отдельный компьютер время от

времени всё равно заражается вирусом, и данные на нём теряются. Кроме того, антивирус не даёт защиты от принципиально нового вируса, пока не пришлют обновления, а за это время новый вирус успевает заразить определённое число компьютеров. Если бы возникла угроза распространения «серой слизи», мы поймём, что это – «серая слизь», только после того, как она значительно распространится. Впрочем, есть иммунные системы, работающие по принципу: запрещено всё, что не разрешено, но их тоже можно обмануть, и они более склонны к автоиммунным реакциям.

Короче говоря, иммунная система хороша только тогда, когда есть мощное дублирование основной системы. У нас пока нет возможности дублирования земных жизненных условий, а космические поселения столкнуться с рядом принципиальных трудностей. Кроме того, у всех иммунных систем бывают ложные срабатывания, которые проявляются в автоиммунных заболеваниях – как, например, аллергия и диабет – которые оказывают значительный вклад в человеческую смертность, сравнимый по порядку величины с вкладом рака и инфекционных заболеваний. Если иммунная система слишком жёсткая, она порождает автоиммунные заболевания, а если слишком мягкая – то пропускает некоторые опасности. Поскольку иммунная система покрывает весь защищаемый объект, то выход её из строя создаёт угрозу всему объекту (здесь действует принцип «распространение фактора опаснее разрушения»). Террористическая атака на иммунную систему делает всю систему беззащитной. Так работает СПИД, который тем быстрее распространяется, чем сильнее с ним иммунная система борется, поскольку он находится внутри неё.

Широко обсуждаются идеи БиоЩита и НаноЩита[i]. Эти щиты подразумевают распыление по всей поверхности Земли тысяч триллионов контролирующих устройств, способных оперативно проверять любые агенты на опасность и оперативно уничтожать опасные. Также к щитам относится дальнейшее ужесточение контроля в Интернете и всемирное развешивание

следящих видеокамер. Однако уже на примере всемирной ПРО видны существенные проблемы любых щитов:

1. Они мучительно отстают от источника угрозы по времени разработки.

2. Они должны действовать сразу на всей территории Земли без исключений. Чем точечнее угроза, тем плотнее должен быть щит.

3. Они уже сейчас вызывают серьёзные политические разногласия. Если щит покрывает не всю поверхность Земли, то он может создавать ситуацию стратегической нестабильности.

4. Любой щит создаётся на основе ещё более продвинутых технологий, которые могут создавать угрозы своего уровня.

5. Щит может быть источником глобального риска сам по себе, если у него начнётся некая «автоиммунная реакция», то есть он начнёт уничтожать то, что должен был защищать. Или если управление щитом будет потеряно, и он начнёт защищаться от своих хозяев. Или если его ложное срабатывание станет поводом для войны.

2. Щит не может быть абсолютно надёжен – то есть успех его срабатывание носит вероятностный характер. И тогда, в случае постоянной глобальной угрозы вопрос его пробивания – это только вопрос времени.

3. Щит должен иметь централизованное управление, но при этом автономность на местах для быстрого реагирования.

Например, антиастероидный щит создаст много новых проблем безопасности человечества. Во-первых, он обеспечит технологию точного управления астероидами, которая за счёт малых воздействий может направить на Землю огромную массу, причём тайно, в духе криптовойны. Во-вторых, сам такой щит может быть использован для атаки по Земле. Например, если на высокой орбите будет висеть 50 штук гигатонных бомб, готовых по команде устремиться в любую точку Солнечной системы, я не

буду чувствовать в большей безопасности. В-третьих, движение всех астероидов за миллиарды лет хорошо синхронизировалось, и любое нарушение этого равновесия может привести к тому, что тот же самый астероид станет постоянной угрозой, регулярно проходя рядом с Землёй. Особенно это будет опасно, если человечество после такого вмешательства откатится на предтехнологический уровень.

Обратим внимание на то, что каждая опасная технология может быть средством собственного предотвращения:

· Ракеты сбиваются с помощью ракет ПРО.

· По местам производства ядерного оружия наносятся ядерные удары.

· ИИ контролирует весь мир, чтобы нигде не создали неправильный ИИ.

· Биодатчики не дают распространиться биологическому оружию.

· Нанощит защищает от нанороботов.

Часто щиты делают нечто ровно противоположное тому, ради чего они создавались. Например, считается (доклад Беллоны, глава IV.1. «Три «трещины» ДНЯО»[ii]), что договор о нераспространении ядерного оружия плохо справляется с чёрным рынком, но хорошо справляется с распространением «мирного атома» (то есть строительствам во всех странах, которые этого хотят, исследовательских ядерных реакторов), который фактически оказывается технологией двойного назначения. Прочные двери, которые защищают кабины самолётов после терактов 11 сентября, не дадут проникнуть террористам в кабину, но если они там всё-таки окажутся (например, в силу того, что сам пилот – террорист), то пассажиры и стюарды не смогут им помешать. Если есть система управления полётом с Земли, то появляется шанс захватить самолёт, используя эту систему, по радио.

Наконец, все щиты предлагаются, исходя из предположения о том, что у нас есть некая идеальная система, которая наблюдает и контролирует

другую, менее совершенную. Например, неподкупная милиция контролирует несовершенное общество. Если же милиция коррумпирована, то отдел собственной безопасности её контролирует, и так далее. Очевидно, что подобных идеальных систем в реальности не бывает, поскольку и система контроля, и контролируемый объект сделаны из одного теста. Можно представить себе многоуровневую иерархическую систему щитов, но в таком случае есть риск раскола между разными контролирующими системами. Наконец, у любого щита есть слепое пятно – он не может контролировать собственный центр управления.

Действующие и будущие щиты

Здесь я привожу краткий, но насколько возможно полный список щитов, которые уже есть или могут появиться в будущем.

1) Всемирная система ПРО. Страдает от политических и технологических проблем, готова только в зачаточной стадии.

2) МАГАТЭ. Работает, но со сбоями. Упустила несколько военных ядерных программ.

3) Всемирная борьба с наркотиками. Находится в равновесии со своей проблемой – сдерживает в некоторой степени, но не более.

4) Система тотального информационного наблюдения, которую можно назвать «оруэлловский контроль» в честь антиутопии «1984» год Оруэлла, где подобная система живо описана. Система контроля за каждым человеком с помощью видеокамер, чипов идентификации, слежения за Интернетом, перехвата телефонных разговоров. Технически такая система достижима, но в реальности она развёрнута только на несколько процентов от того, что могло бы быть, однако при этом она активно развивается. Уже сейчас становятся очевидны и открыто обсуждаются проблемы такой системы, связанные с легитимностью, интернациональностью, слепыми зонами, хакерами. Теоретически может стать основой для всех других систем контроля, так как, возможно, контроля над поведением всех людей достаточно, чтобы не появлялось опасных био, нано и ИИ устройств (а не выискивать уже готовые опасные устройства в окружающей среде).

5) «Mind-контроль». Эта система подразумевает вживление в мозг неких контролирующих чипов (или расшифровка мыслей с помощью анализа энцефалограмм – уже есть конкретные результаты на этом пути). Это может быть не так сложно, как кажется, если мы найдём группы клеток, на которые проецируется внутренний диалог и эмоциональные состояния. Чем-то вроде этого сейчас является детектор лжи. Такое устройство может решить

проблему даже спонтанных преступлений, вроде внезапной агрессии. С другой стороны, потенциал злоупотребления такой технологией – неограничен. Если с помощью такой системы можно будет управлять людьми, то достаточно одной неверной команды, чтобы уничтожить всё человечество. (Та же проблема возникает с предлагающейся в качестве меры против террористов системой управления полётом авиалайнеров с земли: хотя она уменьшит риск захвата отдельного самолёта, она создаст теоретическую возможность одновременно перехватить управление над всеми находящимися в воздухе самолётами и осуществить с их помощью массовый таран зданий или ядерных реакторов.) Наконец, она не даст абсолютной защиты, так как её можно взломать, а так же, потому что некоторые катастрофы происходят не по злому умыслу, а от недомыслия.

6) Анти-астероидная защита. Ведётся наблюдение за потенциально опасными объектами, но недостаточное, средства перехвата официально не разрабатываются. (Но зонд Deep Impact в 2005 г. использовался для столкновения с кометой Темпеля, в результате чего на теле кометы образовался кратер, а её траектория очень незначительно изменилась.)

7) БиоЩит. Борьба с биотерроризмом осуществляется в настоящий момент на уровне разведки и международных соглашений по контролю. Есть рекомендации по безопасной разработке биотехнологий (начиная от добровольных самоограничений, принятых в Асиломаре в 70-е годы и заканчивая книгой «Руководство по биоконтролю»[iii]; вместе с тем, ряд предлагающихся ограничений до сих пор не принят[iv].)

8) НаноЩит. В стадии предварительного обсуждения. Есть рекомендации по безопасной разработке, разрабатываемые Центром Ответственных Нанотехнологий.

9) ИИ-щит. Защита от создания враждебного ИИ. В Сингулярити Институте в Калифорнии (SIAI) ведётся обсуждение проблем безопасного

целеполагания для сильного ИИ, то есть проблемы его Дружественности. Есть рекомендации по безопасной разработке.

10) Обычная полиция и службы безопасности.

Можно также охарактеризовать последовательность во времени срабатывания щитов в случае развития опасной ситуации.

Первый уровень обороны состоит в поддержании цивилизации в осознанном, миролюбивом, уравновешенном состоянии и в подготовке к работе по предотвращению рисков на всех остальных уровнях. На этом уровне важны обмен информацией, открытые дискуссии, публикации в реферируемых журналах, сбор средств, пропаганда, образование и инвестиции.

Второй состоит в непосредственном компьютерном контроле над людьми и опасными системами с тем, чтобы ситуации глобального риска вообще не могли возникнуть. На этом уровне действует МАГАТЭ, глобальные системы видеонаблюдения и перехвата интернет сообщений и т. д.

Третий – в подавлении возникшей опасности с помощью ракет, антинанороботов и т д. Это уровень, аналогичный уроню систем ПРО в защите от оружия массового поражения.

Четвёртый – в эвакуации с Земли или закупоривании в бункеры (однако принцип предосторожности предполагает, что следовало бы начать это делать даже одновременно с первым пунктом).

Сохранение мирового баланса сил

Новые технологии могут нарушать военно-политическое равновесие в мире, предоставляя одной из сторон невиданные возможности. Эрик Дрекслер описывает проблему следующим образом: «В поиске серединного пути, мы могли бы пытаться найти баланс сил, основанный на балансе технологий. Это, по-видимому, расширило бы ситуацию, которая сохраняла определенную меру мирного сосуществования на протяжении четырех десятилетий. Но ключевое слово здесь – "по-видимому": грядущие прорывы будут слишком стремительными и дестабилизирующими, чтобы старый баланс мог продолжать существование. В прошлом страна могла испытывать технологическое отставание на несколько лет и все же поддерживать приблизительный военный баланс. Однако, со стремительными репликаторами и продвинутым ИИ, задержка на единственный день могла бы быть фатальной»[vii]. Короче говоря, чем быстрее развиваются технологии, тем меньше шансов, что они будут находиться в равновесии в разных странах, а также с силами сдерживания и контроля. Сознательное нарушение баланса также опасно: попытка одной из стран явным образом уйти «в отрыв» в сфере военных сверхтехнологий может спровоцировать её противников на агрессию по принципу «атака при угрозе потери преимущества».

Возможная система контроля над глобальными рисками

Любая защита от глобального риска опирается на некую систему глобального наблюдения и контроля. Чем опаснее риск и чем в большем числе мест он может возникнуть, тем тотальнее и эффективнее должна быть эта система контроля. Примером современной системы контроля является МАГАТЭ. Щиты также могут быть системой контроля, или содержать её в себе как особую структуру. Но щиты могут действовать локально и автономно, как иммунная система, а система контроля предполагает сбор и передачу данных в единый центр.

Окончательным вариантом такого глобального контроля было бы «оруэлловское государство», где из каждого угла торчало бы по видеокамере, или чипы были бы установлены в мозг каждого человека, не говоря уже о компьютерах. Увы, в отношении видеонаблюдения это уже почти реализованный вариант. А в домах это можно реализовать технически в любой момент – везде, где есть постоянный интернет и компьютеры. Вопрос скорее не в наблюдении, а в передаче и, особенно, анализе этих данных. Без помощи ИИ нам трудно проверить всю эту информацию. Привлекательными выглядят системы взаимной подотчётности и гражданской бдительности, продвигаемые как альтернатива тоталитарному государству в борьбе с терроризмом, где за счёт абсолютной прозрачности каждый может контролировать каждого, но в отношении их возможности пока много неясного. Проблемы:

Чтобы быть эффективной, такая система контроля должна охватывать весь Земной шар без исключения. Это невозможно без некой формы единой власти.

Любую систему контроля можно ввести в заблуждение – поэтому по-настоящему эффективная система контроля должна быть многократно избыточна.

Мало наблюдать всё, необходимо всю эту информацию анализировать в реальном времени, что невозможно без ИИ или тоталитарного государственного аппарата. Кроме того, эта верхушка не сможет контролировать сама себя, следовательно, понадобится система обратной её подотчётности либо народу, либо «службе внутренней безопасности».

Такая система будет противоречить представлениям о демократии и свободе, которые сформировались в европейской цивилизации, и вызовет ожесточённое сопротивление вплоть до распространения практик терроризма. Такая система тотального контроля вызовет соблазн применять её не только против глобальных рисков, но и против любого рода «правонарушений», вплоть до случаев употребления неполиткорректной речи и прослушивания нелицензионной музыки.

Те, кто контролируют, должны иметь полное и ясное представление обо всех глобальных рисках. Если это будут только биологические риски, но не создание ИИ и опасные физические эксперименты, то система будет неполноценна. Очень трудно отличить опасные биологические эксперименты от безопасных – во всех случаях используются ДНК секвенсоры и опыты на мышах. Без чтения мыслей учёного не поймёшь, что он задумал. А от случайных опасных экспериментов эта система не защищает.

Поскольку подобная система будет уже «доставлена» в любую точку земного шара, она может упростить действие любого оружия, поражающего каждого человека. Иначе говоря, захват власти над системой тотального контроля даст власть над всеми людьми и упростит задачу сделать с ними всё, что угодно, в том числе и нанести им вред. Например, можно разослать по почте некое лекарство и проконтролировать, чтобы все его приняли. Тех, кто отказался, арестовать.

Итак, система тотального контроля кажется наиболее очевидным средством противостояния глобальным рискам. Однако она содержит ряд

подводных камней, которые могут превратить её саму в фактор глобального риска. Кроме того, система тотального контроля подразумевает тоталитарное государство, которое, будучи снабжённым средствами производства в виде роботов, может утратить потребность в людях как таковых.

знательная остановка технологического прогресса

Часто выдвигаются предложения об отказе от технического прогресса или в насильственной форме, или путём взывания к совести учёных, с целью предотвращения глобальных рисков. Есть разные варианты способов осуществления этой остановки, и все они или не работают, или содержат подводные камни:

1.Личный отказ от разработки новых технологий – практически ни на что не влияет. Всегда найдутся другие, которые это сделают.

2.Агитация, просвещение, социальные действия или терроризм как способы заставить людей отказаться от развития опасных технологий – не работают. Как пишет Юдковски: любая стратегия, которая предполагает единодушные действия всех людей, обречена на провал.

3.Отказ от технологических новшеств на определённой территории, например, одной страны, неспособен остановить технологический прогресс в других странах. Более того, если более ответственные страны откажутся от развития некой технологии, то пальма первенства перейдёт к более безответственным.

4.Всемирное соглашение. На примере МАГАТЭ мы знаем, как плохо это работает.

5.Завоевание всего мира силой, которая сможет регулировать развитие технологий. Но в процессе этого завоевания велики шансы применения оружия «судного дня» теми ядерными державами, которые в результате утратят суверенитет. Кроме того, словами Дрекслера: «Далее, победившая сила была бы сама главной технологической силой с огромной военной мощью и демонстрируемой готовностью ее использовать. Можно ли в этом случае доверять такой силе в том, что она подавит свой собственный прогресс?» («Машины созидания».)

6.Мирное объединение наций перед лицом нависшей угрозы, подобно тому, как возникло ООН в годы фашизма, и делегирование ими своих сил на

остановку прогресса в тех странах, которые не захотят присоединиться к этому объединению. Вероятно, это наилучший вариант, который объединяет достоинства всех предыдущих, и сглаживает их недостатки. Но он станет реальным, только если общая угроза станет явной.

7. Ник Бостром предложил концепцию дифференцированного технологического развития, когда проекты, увеличивающие нашу безопасность, стимулируются и ускоряются, тогда как потенциально опасные проекты искусственно замедляются. Таким образом, управляя скоростью развития разных областей знания, мы получаем более безопасные сочетания технологий нападения и защиты.

Средства превентивного удара

Мало иметь систему тотального контроля – нужно обладать возможностью предотвратить риск. Сейчас обычно в качестве крайней меры рассматривается удар ракетно-ядерными силами по точке источника риска.

Здесь наблюдается любопытное противоречие с программами строительства бункеров для выживания – если такие бункера будут секретны и неуязвимы, то их будет трудно уничтожить. Более того, они должны содержать в себе полностью оборудованные лаборатории и учёных на случай катастрофы. Поэтому возможна ситуация, когда «сверхоружие» будет создаваться в таком бункере (например, в СССР создавались подземные ядерные заводы для продолжения производства ядерного оружия в случае затяжной ядерной войны.) Люди, которые уже находятся в неуязвимом бункере, могут быть более психологически склонны к созданию сверхоружия для удара по поверхности. Следовательно, либо бункеры будут представлять угрозу человеческому выживанию, либо средства превентивного удара уничтожат все бункеры, которые могли бы использоваться для выживания людей после некой катастрофы.

Однако удар по одной точке в пространстве не действует ни против системного кризиса, ни против некой информационной угрозы. Компьютерный вирус не вылечишь ядерным ударом. И не избавишь людей от привязанности к сверхнаркотику. Далее, удар возможен, пока некий риск не вышел из точки. Если рецепт супервируса попал в интернет, обратно его не выловишь. Уже сейчас современная военная машина бессильна против сетевых угроз, вроде террористических сетей, дающих метастазы по всей планете. Точно также в будущем компьютерный вирус будет не просто информационной угрозой данным на диске: он может заставлять компьютерно управляемые фабрики по всему миру незаметно производить некие свои материальные носители (скажем, в виде микроскопических

роботов или программных закладок в обычных продуктах), а через них снова уходить в сеть (например, подключаясь к радиоканалу).

Наконец, сам удар (или даже его возможность) создаст ситуацию стратегической нестабильности. Например, сейчас удар баллистической ракетой с обычной боеголовкой по террористам может вызвать срабатывание системы предупреждении о ракетном нападении вероятного противника и привести к войне.

Наконец, удар требует определённого времени. Это время должно быть меньше времени от обнаружения развития угрозы до времени её перехода в необратимую фазу (например, в случае появления «серой слизи» важно уничтожить её до того, как она сумела размножиться в миллиардах копий и распространится по всей Земле). Сейчас время от обнаружения до удара по любой точке Земли меньше 2 часов, и может быть уменьшено до минут с помощью спутникового оружия. (Однако время принятия решения больше.) Если бы от момента принятия решения о написании кода опасного вируса до его запуска проходило бы только 15 минут, то этой скорости было бы недостаточно. И этой скорости очевидно недостаточно, если в некоем месте началось распыление опасных нанороботов по воздуху.

Эффективность удара по точке принципиально изменится после основания космических колоний (хотя бы чисто робототехнических – там тоже может произойти сбой, который превратит колонию в «раковую» – то есть склонную к неограниченному саморазмножению и распространению «токсинов»: опасных нанороботов, сверхбомб и прочего; а именно освоение космоса с помощью саморазмножающихся роботов, использующих местные материалы, наиболее перспективно). За время, пока сигнал об опасности пройдёт, скажем, от спутника Юпитера до Земли, и затем от Земли туда прилетит боевой «флот» (то есть ракеты с ядерными боеголовками и боевыми нанороботами) наводить порядок (жечь всё подряд), будет уже поздно. Конечно, можно держать «флот» на орбите каждого спутника

планеты или астероида, где есть способные к саморазмножению робототехнические колонии, но что если мятеж произойдёт именно на самом флоте? Тогда нужен флот, который контролирует другие флоты, и плавает между спутниками планет. А затем ещё один межпланетный флот для контроля над ними. Короче, ситуация не выглядит стратегически стабильной, – то есть выше определённого уровня системы контроля начинают мешать друг другу. Возможно, неспособность контролировать удалённые колонии приводит к тому, что цивилизациям выгодно замыкаться на материнской планете – вот ещё одно решение парадокса Ферми.

Удаление источников рисков на значительное расстояние от Земли

Теоретически можно удалить источники рисков от Земли, в первую очередь это касается опасных физических экспериментов. Проблемы, связанные с этим подходом:

· Получив в руки технические средства создавать мощные экспериментальные установки далеко от Земли, мы также будем иметь возможности быстро доставлять результаты экспериментов обратно.

· Это не сможет остановить некоторых людей от аналогичных опытов на Земле, особенно если они просты.

· Это не защитит нас от создания опасного сильного ИИ, так как он может распространяться информационно.

· Даже за орбитой Плутона возможны опасные эксперименты, которые повлияют на Землю.

· Трудно заранее знать, какие именно эксперименты надо проводить «за орбитой Плутона».

· Нет технических возможностей доставить огромное количество оборудования за орбиту Плутона в течение ближайших десятков лет, тем более без использования опасных технологий в виде самовоспроизводящихся роботов.

Создание автономных поселений в отдалённых уголках Земли

Создание таких поселений, равно как и навыки выживания в дикой природе, вряд ли поможет в случае действительно глобальной катастрофы, поскольку она должна затронуть всю поверхность Земли (если это некий неразумный агент), или обнаружить всех людей (если это разумный агент). Автономное поселение уязвимо и к первому, и ко второму – если только это не вооружённая секретная база, но тогда оно проходит, скорее, под графой «бункеры».

Если речь идёт о выживании после очень большой, но не окончательной катастрофы, то следует вспомнить опыт продразвёрстки и колхозов в России, – город силой властвует над деревней и отбирает у неё излишки. В случае системного кризиса главную опасность будут представлять другие люди. Недаром в фантастическом романе «Метро 2033» основной монетой является патрон от автомата Калашникова. И до тех пор, пока патронов будет больше, чем крестьян, будет выгоднее грабить, а не выращивать. Возможно также полное растворение человека в природе в духе Маугли.

Создание досье на глобальные риски и рост общественного понимания связанной с ними проблематики.

Публикация книг и статей на тему глобальных рисков приводит к росту осознания проблемы в обществе и составлению более точного списка глобальных рисков. Междисциплинарный подход позволяет сравнивать разные риски и учитывать возможность их сложного взаимодействия. Сложности данного подхода:

· Не понятно, к кому именно адресованы любые такого рода тексты.

· Террористы, страны изгои и регулярные армии могут воспользоваться идеями о создании глобальных рисков из опубликованных текстов, что приведёт к большему увеличению рисков, чем к их предотвращению.

· Неправильное и преждевременное вложение капитала может привести к разочарованию в борьбе с рисками – как раз тогда, когда эта борьба на самом деле понадобится.

Предотвращение одной катастрофы с помощью другой

Теоретически возможны следующие примеры взаимной нейтрализации опасных технологий и катастроф:

1.Ядерная война останавливает развитие технологий вообще.

2.Тотальный ИИ предотвращает биотерроризм.

3.Биотерроризм делает невозможным развитие ИИ

4.Ядерная зима предотвращает глобальное потепление.

Суть в том, что крупная катастрофа делает невозможной глобальную, отбрасывая человечество на несколько эволюционных ступеней назад. Это возможно в том случае, если мы входим в длительный период высокой вероятности крупных катастроф, но малой вероятности глобальных катастроф. В некотором смысле со второй половине XX века мы находимся в этом периоде. Тем не менее, это не помешало нам успешно приблизится к тому моменту, когда до создания многих средств глобального всеобщего уничтожения остались, возможно, десятки лет.

Было бы в каком-то смысле «приятно» доказать теорему, что глобальная катастрофа невозможна, потому что к ней не дадут приблизиться очень крупные катастрофы. Однако эта теорема носила бы исключительно вероятностный характер, так как некоторые опасные сверхтехнологии могут появиться в любой момент, особенно ИИ.

Кроме того, любая большая авария (но меньшая отбрасывающей назад катастрофы) повышает осознанность людей в отношении рисков. Хотя здесь возникает определённый стереотип: ожидание повторения точно такого же риска.

Предположения о том, что мы живём в «Матрице».

Основы научного анализа этой проблемы заложены Ником Бостромом в его статье «Рассуждение о симуляции».[xv] Многие религиозные концепции можно сделать наукообразными, введя предположение, что мы живём в симулированном мире, возможно, созданном внутри сверхкомпьютера силами некой сверхцивилизации. Опровергнуть то, что мы живём в матрице, невозможно, но это можно было бы доказать, если бы в нашем мире появились некие невероятные чудеса, несовместимые с какими бы то ни было физическими законами (например, в небе бы возникла надпись из сверхновых звёзд).

Однако есть концепция, что может произойти глобальная катастрофа, если хозяева этой симуляции внезапно её выключат (Бостром). Можно показать, что в этом случае вступают в действие аргументы, описанные в статье Дж. Хигго о многомирном бессмертии. А именно, то, что мы живём в Матрице, вероятно только в том случае, если множество возможных симуляций очень велико. Это делает вероятным существование значительного количества совершенно одинаковых симуляций. Уничтожение одной из копий никак не влияет на ход самой симуляции, так же, как сожжение одного из экземпляров романа «Война и мир» не влияет на отношение персонажей. (При этом никакие аргументы о душе, непрерывности сознания и других некопируемых факторах не работают, так как обычно предполагается, что «самосознание» в симуляции вообще невозможно.)

Следовательно, никакой угрозы полное выключение симуляции не представляет. Однако если мы всё же живём в симуляции, то хозяева симуляции могут подкинуть нам некую маловероятную природную проблему, хотя бы для того, чтобы просчитать наше поведение в условиях кризиса. Например, изучить, как цивилизации ведут себя в случае извержения сверхвулканов. (А любая сверхцивилизация будет

заинтересована в просчитывании разных вариантов своего предыдущего развития, например, чтобы оценить частоту распространённости цивилизаций во Вселенной.) При этом можно предположить, что крайние узловые события будут чаще становиться объектами моделирования, особенно моменты, когда развитие могло полностью прекратиться, то есть глобальные риски. (А мы как раз живём в районе такого события, что, по байесовой логике, повышает вероятность гипотезы о том, что мы живём в симуляции.) Иначе говоря, в симуляциях гораздо чаще будут встречаться ситуации глобального риска. (Точно также в кино гораздо чаще показывают взрывы, чем мы видим их в реальности.) А значит, это увеличивает наши шансы столкнуться с ситуацией, близкой к глобальной катастрофе. При этом, поскольку сама глобальная катастрофа в мире симуляций невозможна, ибо всегда найдутся симуляции, где «главные герои не умирают», то здесь наиболее вероятным сценарием будет выживание горстки людей после очень большой катастрофы. К вопросу о рассуждении о симуляции Бострома мы ещё вернёмся далее.

Иногда высказываются надежды, что если человечество приблизится к грани самоуничтожения, то «добрые инопланетяне», которые давно за нами будто бы следят, нас спасут. Но на это не больше надежд, чем у ягенёнка, которого пожирают львы, на то, что его спасут люди, снимающие об этом документальный фильм.

Мир без глобальной катастрофы: наилучший реалистичный вариант предотвращения глобальных катастроф

Жанр требует «хэппи энда». Если бы глобальная катастрофа была бы абсолютна неизбежна, то не следовало бы и писать этой книги, так как единственное, что оставалось бы людям перед лицом неизбежной катастрофы – это устроить «пир перед чумой». Но даже если шансы катастрофы очень велики, мы можем значительно отсрочить её наступление, уменьшая её погодовую вероятность.

Я вижу эти шансы в таком опережающем развитии систем искусственного интеллекта, которое обгоняет развитие других рисков, но одновременно это развитие должно опережается ростом нашего понимания возможностей и рисков самого ИИ, и нашим пониманием того, как правильно и безопасно поставить перед ним задачу, то есть как создать «Дружественный» ИИ. И затем на базе этого Дружественного ИИ создать единую систему мировых договоров между всеми странами, в которой этот ИИ будет выполнять функции Автоматизированной системы государственного управления. Этот план предполагает плавный и мирный переход к действительно величественному и безопасному будущему.

И хотя я не думаю, что именно этот план легко и безупречно реализуется, или что он является действительно вероятным, я полагаю, он представляет лучшее, к чему мы можем стремиться и чего мы можем достичь. Суть его можно изложить в следующих тезисах, первые два из которых являются необходимыми, а последний – крайне желательным:

1) Наши знания и возможности по предотвращению рисков будут расти значительно быстрее возможных рисков.

2) При этом эти знания и возможности управления не будут порождать новых рисков.

3) Эта система возникает мирно и безболезненно для всех людей.

Заключение

Мы стремимся сохранить жизнь людей и человечества только потому, что она имеет ценность. Хотя у нас не может быть точного знания, о том, что именно создаёт ценность человеческой жизни, так как это не объективное знание, а наше соглашение, мы можем предположить, что для нас ценно число людей, а также испытываемое ими удовольствие и возможность творческой самореализации, иначе говоря, разнообразие создаваемой ими информации. То есть мир, в котором живёт 1000 человек, и при этом однообразно страдает (концлагерь), хуже мира, где радостно живёт 10 000 человек, занимаясь разнообразными ремёслами (древнегреческий полис).

Таким образом, если у нас есть два варианта развития будущего, в которых одинаковая вероятность вымирания, то нам следует предпочесть тот вариант, в котором живёт больше людей, меньше страдает, и их жизнь более разнообразна, то есть в наибольшей мере реализовывает человеческий потенциал.

Более того, нам вероятно следовало бы предпочесть мир, в котором живёт миллиард человек в течение 100 лет (и потом этот мир разрушается), миру, в котором живёт только миллион человек в течение 200 лет.

Крайним выражением этого является «пир во время чумы». То есть, если смерть неизбежна, и невозможно никак отсрочить ее, то наилучшим поведением для рационального субъекта (то есть не верящего в загробную жизнь) оказывается начать развлекаться наиболее интересным образом. Значительное число людей, осознающих неизбежность физической смерти, так и делают. Однако если смерть отстоит на несколько десятков лет, то нет смысла пропить всё сегодня, и максмализация функции удовольствия требует постоянного заработка и т.д.

Интересно задаться вопросом, какова бы была рациональная стратегия для целой цивилизации, которая бы знала о неизбежности гибели через тот

или иной срок. Стоило бы ей максимально увеличит население, чтобы дать пожить максимально большому количеству людей? Или наоборот, раздавать всем наркотики и вживлять электроды в центр удовольствия? Или скрыть сам факт неизбежности катастрофы, так как это знание неизбежно приведёт к страданиям и преждевременному разрушению инфраструктуры? Или возможен смешанный путь при не нулевой, но и не стопроцентной вероятности вымирания, где часть ресурсов уходит на «пир во время чумы», а часть – на поиски выхода?

Но настоящее удовольствие невозможно без надежды на спасение. Поэтому для такой цивилизации было бы рационально продолжать искать выход, даже если бы она наверняка знала, что его нет.