План предотвращения глобальных рисков, могущих привести к полному вымиранию человечества

Временая шкала и шаги

Exact dates depend on the pace of technological progress and of our activity to prevent risks

Ближайшее десятилетие (2015 – 2025) → Вторая четверть 21 века (2025-2050) → Вторая половина 21 века (2050-2100)

Шаг 1
Планирование

Шаг 2
Подготовка

Шаг 3
Первый уровень защиты на низком технологическом уровне

Шаг 4
Защита на высоком технологическом уровне

Шаг 5
Достижение неуничтожимости цивилизации с почти нулевым уровнем глобальных рисков

Plan A
Prevent the catastrophe

План A1.1
Международная система контроля

Исследования

Сбор информации
- Создание полного списка рисков
- Создание карты рисков
- Оценка вероятности
- Анализ способов защиты и создание плана действий
- Создание, проверка и продвижение долгосрочной модели будущего (эта карта основана на модели будущего, определяемого экспоненциальном развитием технологий)
- Инвестиции в долгосрочные исследования по прогнозированию
- Определить, какие риски наиболее вероятны, и какие проще всего предотвратить.
- Создание вики и форума по тематике глобальных рисков, которые позволят привлечь к себе лучшие умы, но при этом они должны оставаться широко доступными публике и хорошо модерироваться

Помощь в процессе
- Образование «спасателей мира»: выбор лучших студентов, предоставление им необходимых курсов, денег и задач
- Решение проблемы того, что ученые игнорируют исследования друг друга (проблема "высокомерия" спасителей мира)
- Интеграция различных подходов к проблематике глобальных рисков
- Неограниченное финансирование исследований глобальных рисков
- Помощь лучшим ученым в области (Бострон) в создании высококачественных исследований по глобальным рискам

Дополнительные области для изучения
- Изучение существующей системы принятия решений в ООН
- Общая теория безопасности и предотвращения риска
- Создать список когнитивных искажений в области глобальных рисков и устранить их
- Перевод лучших книг и статей на тему глобальных рисков и на основные языки

Общественное движение

Наука
- Группа авторитетных ученых, разделяющих правильное понимание проблемы и с низким уровнем внутренних конфликтов
- Научная структура: журнал, вики, форум, панель
- Продуктивное сотрудничество между учеными и обществом на основе доверия

Общество
- Убеждение общества (статьи, книги, форумы, разные медиа)
- Общественная поддержка, уличные протесты (антиядерные протесты в 80-е)

Политика
- Политические деятели и партии, лоббирование
- Написание рекомендаций, участие в законотворчестве

Международная кооперация

ООН
- Государства делегируют часть своих полномочий в ООН для предотвращения глобальных рисков
- Непосредственное сотрудничество государств без участия ООН
- Ответственность по предотвращению х-рисков переходит на группу суперсил (США, ЕС, БРИКС).
- Подписание договора о рисках.
- Международное соглашение о контроле глобальных рисках, которое позволит
- наказывать людей, ответственных за рост вероятности глобальных рисков,
- за недооценку рисков,
- за заговоры о глобальных рисках,
- за пренебрежение глобальными рисками.
- награждать людей за снижение глобальных рисков, обнаружение новых рисков и за работу над их предотвращением.
- Международные агентства, посвященные определенным рискам (ВОЗ, МАГАТЭ) и создание новых для новых рисков

Стимулы к интеграции
- Некая малая катастрофа помогает объединиться человечеству для противостояния ей (пандемия, астероид, локальная ядерная война), либо повышает осознание проблемы глобальных рисков

Контроль над рисками

Запреты
- Международные запреты опасных технологий
- Заморозка опасных проектов на 30 лет (SETI, ITER, LHC) на основе принципа предосторожности
Сосредоточить все био, нано и ИИ исследования в контролируемых центрах

Контроль
- Контроль над распространением знаний массового уничтожения (ДНК вирусов) или белого шума в области ; интернет-цензура
- Контроль над суицидальными и террористическими настроениями людей, занятых в опасных областях
- Инвестиции в системы безопасности: образование, сооружения, контроль, проверки , экзамены
- Международные системы контроля (МАГАТЭ)
- Сканирование интернета, наблюдение из космоса
- Специализированный ИИ или экспертная система по глобальным рискам как советник

Стимулы
- Дифференциальное развитие технологий (первыми – повышающие безопасность, Бостром)
- Законы и экономические стимулы (Richard Posner, квоты на углекислый газ)

Создание всемирного центра по предотвращению глобальных рисков
- Компьютерный контроль (всемирная вездесущая система видеонаблюдения и контроля)
- Всемирный центр быстрого реагирования на риски, полиция по х-рискам
- Способность организовать в случае необходимости сильный глобальный скоординированный ответ на ожидаемый экзистенциальные риски (Бостром)
- Роботы для ликвидации ЧС связанных с био- и ядерными опасностями

Активные щиты
- Геоинженерия против глобального потепления
- Всемирная ПРО – система защиты от ракет и астероидов
- Наношит – распределенная система контроля над опасными репликаторами
- Биощит – всемирный аналог иммунной системы
- Майнд-щит – контроль опасных мемов, идей, суицидальных и террористических настроений
- Слияние государства, интернета и всемирного ИИ в единую систему контроля, безопасности и управления
- Удаление источников рисков на большое расстояние от Земли (ставить опасные эксперименты на Плутоне)

Синглетон
- Единый центр принятия глобальных решений
- Всемирная система управления, основанная на искусственном интеллекте
- СверхИИ, решающий задачи предотвращения всех мыслимых рисков и обеспечивающий бессмертие и развитие всех людей
- Освоение Солнечной системы, создание сферы Дайсона, полеты к звездам
- Освоение Галактики
- Слияние со Вселенной, победа над "тепловой смертью" Вселенной

Проактивный и реактивный подходы

Реактивный подход: реакция на наиболее актуальные и самые заметные риски.
Плюсы: приемлемые сроки, видимые результаты, правильное распределение ресурсов, инвестиции в противостояние реальным рискам
Хорош для медленно развивающихся рисков
Минусы: не позволяет реагировать на новые, быстроразвивающиеся и внезапные опасности.
Риски в этом подходе ранжированы по срочности

Проактивный подход: Предвидение будущих рисков и создание многоуровневой защиты
Плюсы: Позволяет предвидеть принципиально новые риски и построить защиту от рисков заранее
Минусы: Возможно, что часть инвестиций пойдет на борьбу с воображаемыми рисками, нет видимого результата, необходимость отсеивать множество бредовых идей
Риски в этом подходе ранжированы по вероятности.
Эта карта подразумевает проактивный подход, а сейчас в целом работает реактивный

Идеологическая нагрузка новых технологий

Частный космос – идея о Марсе как о запасном доме для человечества
Электрические атомобили – устойчивое развитие
Роботы-автомобили – идея о рисках опасного ИИ
Facebook – идея эмпатии, дружбы всех со всеми и взаимного контроля

Суть в том, чтобы новая монопольная технология несла с собой некую идеологическую нагрузку

Практические шаги для предотвращения отдельных рисков

- Технологии улавливания радиоактивной пыли

Биотех:
- Разработка лучших рекомендаций по безопасной биотехнологии, с разъяснениями почему именно они необходимы.
- Синтезаторы ДНК постоянно подключены к интернету по зашифрованному соединению и проверяются на синтез опасных ДНК-последовательностей
- Финансирование открытых лабораторий с высоким уровнем биозащиты, например, чистых комнат с отрицательным давлением
- Лучшие законы, обеспечивающие карантин в случае угрозы пандемии
- "Создать и накопить запасы новых лекарств или вакцин, мониторить биологические образцы и возникающие болезни, и усилить способность медицины на местах реагировать на пандемии" (Матени)

Окружающая среда и ресурсы:
- Захват метана и CO2 из атмосферы, может быть с помощью бактерий.
- Инвестировать в биоразнообразие производства продуктов питания, чтобы недопустит распространение вредителей: 1) лучшие законы о карантине 2) переносное оборудование для быстрой идентификации чужеродных включений в грузах продуктов 3) дальнейшее развитие способов нехимической стериализации
- Мягкий контроль рождаемости (женское образование)
- Солнечная и ветро энергетика

- Система отклонения астероидов без помощи ядерных взрывов.

"Закрытие" некоторых рисков
- Универсальные вакцины
- Обнаружение опасных астероидов или доказательство их отсутствия (проект WISE)
- Переход на энергосберегающие, экологичные и возобновляемые технологии, снижение выбросов
- Остановка LHC
- Интеграция стран-изгоев

Война за объединение планеты
- Прямая оккупация всех стран мира, разгром соперников (вторая мировая, холодная война)
- Использование одной из сверхтехнологий для мгновенного захвата власти над всей планетой и установления вечного мира (нанотех)
- Шантаж оружием судного дня
- Уничтожение одной страной всех остальных (Сев. Корея? биологическое оружие)

Фатальная ошибка в работе всемирной системы контроля
- Ошибочная команда заставляет защитные системы атаковать людей
- Геоинженерия идет насмарку
- Всемирный тоталитаризм ведет мир к катастрофе

Глобальная катастрофа

Глобальная катастрофа

План A1.2
Децентрализованный мониторинг рисков

Трансформация ценностей

Ценность безопасности
- Распространение ценности продления жизни и ценности глобальной безопасности
- "Можно доказать, что с этической точки зрения снижение экзистенциальных рисков гораздо более важно, чем любое другое общественное благо" (Бостром)
- Ценность нерушимости цивилизации становится первым приоритетом на всех уровнях: в образовании, на личном уровне и в качестве цели каждого народа

Распространение этой ценности
- Ослабление радикально-религиозных (ИСИС) и националистических ценностей
- Распространение трансгуманизма
- Изменение картины будущего
- Фильмы, романы и другие произведения искусства, которые честно изображают глобальные риски и мотивируют к их предупреждению
- Дни памяти и осознанности: День Земли, Петров День, день астероидной опасности
- Преподование в школах теории х-рисков, рациональности и теории безопасности

Повышение интеллекта и моральности людей

Интеллект
- Ноотропики, стимуляция мозга и генная терапия для повышения IQ
- Образование, рациональность, снижение когнитивных искажений
- Много здравомыслящих людей, способных самоорганизоваться и противостоять рискам

Эмпатия и позитивные эмоции
- Высокий уровень моральности и эмпатии, чтобы гении не стали супертеррористами
- Устранение деструктивных верований, опасного поведения и эгоизма
- Устранение депрессии и негативных эмоций, которые могут вести к суицидальному или небезопасному поведению
- Инженерное просветление: использовать данные науки о мозге, чтобы объединить людей и сделать их менее агрессивными; открыть мир духовных переживаний для каждого (психоактивные вещества, стимуляция мозга)

Мораль
- Предотвращение худших форм капитализма: предпочтение краткосрочной выгоды, нарушение правил и пренебрежение безопасностью ради выгоды
- Усиление лучших моральных качеств: честности, заботы, верности слову, не убий (но без религии)

Предотвращение Холодной войны, Третьей мировой войны и локальных ядерных войн
- Международная институция, в которой конфликты могут разрешаться мирным способом и путем диалога,
- Секретная служба, способная проникать в правительства разных стран и не давать им вступать в военный конфликт
- Большой проект, который может объединить человечесто, например, предотвращение пандемии
- Интеграция стран-изгоев на основе диалога
- Горячие линии между ядерными странами
- Антиядерное и антивоенное движение
- Примат международного права как инструмента решения конфликтов
- Мирная интеграция национальных государств
- "Не жать на красное": кооперативная теория принятия решений в мировой политике
- Не балансировать на грани войны для достижения своих целей
- Предотвратить распространение ядерного оружия

Радикальные социальные изменения
- нето-кратия,
- уход дикого капитализма,
- научный подход к политике на основании теории систем,
- всемирная демократия на основе интернет-голосования,
- рост горизонтальных связей между людьми

Децентрализованный мониторинг рисков
- **Прозрачное общество**: каждый может следить за каждым (Дэвид Брин)
- в открытом пространстве группы добровольцев-виджилянтов ищут признаки опасной активности
- группы хакеров вроде "Anonymous" превращают закрытое информационное пространство в открытое и ищут там опасную активность

- **Локальный контроль**:
- местная полиция может отслеживать терроризм;
- местные органы здравоохранения могут обнаруживать начало возникновения новой болезни
- взаимный контроль в профессиональных сообществах: ученые могут обнаруживать чужие ошибки
- мониторинг интернет запросов для выявления подозрительных патернов
- активные граждане, бьющие тревогу (whistleblower)

- **Мониторинг косвенных признаков:**
- искать дым, а не огонь
- использование в этом систем машинного обучения

- **Сетевой контроль**:
- сеть организаций, занимающихся предотвращением х-рисков без единого центра
- создание личных протоколов безопасности для каждого сотрудника с коротким перечнем пунктов

- **Экономические стимулы**
- торговля углеродными квотами
- призы за каждый обнаруженный и предотвращенный риск

План A2
Дружественный ИИ

Продвижение
- Исследование теории ИИ (MIRI)
- Пропаганда важности создания дружественного ИИ
- Фандрайзинг
- Распространение принципов рационального мышления (Less Wrong)
- Замедление опасных ИИ-проектов (скупка умов, рецессии)
- Образование в области дружественного ИИ

Теория Дружественного ИИ
- Теория ценностей и теория принятия решений
- Доказанная безопасность ИИ
- Способ сохранения системы целей при самоапгрейте
- Теория проста для понимания и легко реализуема на практике

Практическое распространение теории
- Донесение теории Дружественного ИИ до основных команд, создающих ИИ, и убеждение их в ее применении
- Тестирование ее на игрушечных моделях
- Создание узкого ИИ, эмулящй мозга, эксперименты в области загрузки систем целей в ИИ.

Зародыш AI
- Создание и запуск небольшого ИИ-проекта, способного к самооптимизации и основанного на теории Дружественного ИИ

Сверхинтеллект
- ИИ становится сверх-ИИ и быстро берет под контроль всю планету, а затем устраняет смерть, страдания и глобальные риски

- Возможно, создание AI Nanny – помогает предотвратить глобальные риски, но не более того

Недружественный ИИ
- Уничтожает всех людей и реализует ценности, мало имеющие отношения к человеческим (paperclip maximazer)
- Однако некая форма разумной жизни сохраняется и несет определенную информацию о человечестве и наследует некоторые из его целей и свойств

План A3
Повышение неуничтожимости

Повышение устойчивости цивилизации
- Системы с мощной защитой от дурака
- Повышение разнообразия и устойчивости людей и их жилищ
- Универсальные превентивные средства (мощная медицина, вакцины)
- Накопление резервов (запасы продовольствия, минералов, энергии, механизмов, знаний)
- Децентрализованная система гражданской обороны, включающая в себя временные убежища, системы очистки воды и воздуха, датчики радиации, аптечки, обучение людей и т. д. (Герман Кан)

Полезные идеи для уменьшения масштабов катастрофы

Ограничение последствий катастрофы посредством замедления ее развития и ограничения областей, которые ей поражены:
- создание механизмов реализации карантина
- постройка инфраструктуры для быстрого производства вакцин в ответ на возникающие угрозы,
- накопление запасов важных медицинских препаратов и контрмер.

Увеличить время, доступное для подготовки к катастрофе путем улучшения мониторинга и раннего обнаружения:
- поддержка общих исследований в области возможных масштабов рисков биобезопасности и способов их предотвращения
- улучшение и объединение систем наблюдения за возникающими болезнями, чтобы быстрее реагировать на новые угрозы

Провидить всемирные учения по предотвращения глобальной катастрофы

Способность быстро адаптироваться к новым рискам и предвидеть их заранее

Ускоренное развитие, чтобы быстро проскачить зону рисков
- Инвестиции в прорывные технологии (нанотех, био)
- Обогнать медленные процессы деградации и исчерпания с помощью быстрого развития новых методов и технологий
- Больше инвестировать в оборонительные технологии, чем в наступательные

Быстрое достижение индивидуального бессмертия на максимально высоком технологическом уровне
- Нанотехнологическое бессмертное тело
- Разделение человечества на несколько видов, способных жить в космосе: более гибких и разнообразных
- Загрузка сознания в компьютер
- Слияние с ИИ и превращение в него

ИИ, основанный на симуляции мозга его создателя
- Дружествен к его личной системе ценностей
- Естественное развитие системы ценностей при самоапгрейте
- Отказ от неограниченного самоапгрейта

План A4
Расселение в космосе

Временные убежища в космосе
- Косическая станция как временное убежище
- Дешевые и безопасные ракеты и космические аппараты
- Создание космических поселений на Луне и на Марсе (Элон Маск)

Космические колонии на других планетах
- Создание космических колоний на Марсе и на Луне для 100-1000 человек (Elon Musk)

Освоение Солнечной системы
- Самообеспечивающиеся колонии на Марсе и больших астероидах
- Терраформинг планет и астероидов с помощью роботов-репликаторов
- Миллионы независимых колоний внутри астероидов и кометных тел в облаке Оорта

Межзвездные перелеты
- Звездолеты с колонистами типа Орион на ядерном топливе
- Звездолеты на новых физических принципах и с бессмертными людьми
- Зонды-репликаторы, удаляющиеся от Земли на максимальной скорости с ИИ и с человеческими эмбрионами

Распределенное по Галактике человечество
- Множество несвязанных друг с другом космических цивилизаций
- Новые типы рисков, охватывающие большие участки космоса (космические войны, взрывы планет и звезд, ИИ и нанорепликаторы, внеземные цивилизации)

План B
Пережить катастрофу

Подготовка
- Фандрайзинг и убеждение
- Создание инструкций по восстановлению цивилизации (Dartnell "Knowledge")
- Создание кладов со знаниями, семенами и материалами (Doomsday Vault в Норвегии)
- Самообеспечивающиеся группы сюрвайвалистов

Постройка
- Подземные бункеры
- Атомные подводные лодки
- Отдаленные острова

Естественные убежища
- Племена без контакта
- Отдаленные деревни
- Корабли и автономные морские поселения
- Полярные станции в Антарктиде

Готовность к катастрофе
- Команды наготове в бункерах
- Подготовка экипажей и ротация
- Разные типы убежищ
- Замороженные эмбрионы

Миниатюризация для выживания и неуничтожимости
- Колонизация земной коры в микроскопических нанотехнологических телах
- Перенос цивилизации в компьютерную симуляцию внутри маленького самоподдерживающегося компьютера, затерянного в космосе

Восстановление цивилизации после катастрофы
- Восстановление численности населения
- Восстановление науки и техники
- Создание условий невозможности следующей катастрофы

Перезапуск цивилизации
- Может быть несколько циклов перезапуска
- В результате либо окончательная деградация, либо переход в новое устойчивое состояние

План C
Оставить информацию

Капсулы с информацией
- Капсулы на Земле для будущих нечеловеческих цивилизаций с ДНК человека и информацией
- Вечные диски от Long Now Foundation (или "M disks")

Послания инопланетянам
- Радиопослания, содержащие описания человечества и ДНК людей
- Клады на Луне: замороженный мозг и эмбрионы
- Космические аппараты типа Voyager с информацией о человечестве

Сохранение земной жизни
- Сохранение популяций высокоразвитых животных
- Создание условий для возникновения нового вида разумных существ на Земле
- Направленная панспермия (Марс, Европа, космическая пыль с бактериями)

Роботы-репликаторы в космосе
- Механическая жизнь
- Сохранение информации о человечестве в течение миллиардов лет
- Специализированный безопасный ИИ

Воскрешение иной цивилизацией
- Возникновение цивилизации, разделяющей наибольшую часть человеческих свойств и ценностей
- Точное возвращение к жизни отдельных людей

План D
Невероятные идеи

Спасение нечеловеческим сверхразумом
- Может быть, инопланетяне наблюдают за нами и спасут
- Второе пришествие или любое другое чудо, вера в потусторонний мир и жизнь в нем.
- Радиопослания в космос с просьбой о помощи (если катастрофа неизбежна)

Максимально странная стратегия, чтобы избежать парадокса Ферми
Пустота космоса говорит о том, что все предыдущие цивилизации погибли, следовательно, пути, казавшиеся рациональными или обычными, никуда не вели, и нужно выбрать случайный неожиданный путь, наверняка отличающийся от прошлых

Квантовое бессмертие
- Если верна многомирная интерпретация квантовой механики, то по крайней мере один человек (я-наблюдатель) выживет в любой катастрофе (Моравек, Тегмарк)
- Можно связать мое выживание с выживанием и других людей, ("мы на подлодке")
- В бесконечной вселенной всегда найдутся другие человеческие цивилизации

Предвидение будущего
- Использование квантовых эффектов для предвидения будущего и исключения опасных вариантов
- Использование новых физических эффектов для поиска потенциальных террористов или нест возможных аварий
- Специализированный ИИ предсказывает новые риски и способы их предотвращения

Манипуляция вероятностью с помощью Doomsday argument
- Манипуляция общим числом будущих рождений (Метод UN++, Бостром)
- Уменьшение плотности числа рождений, чтобы выиграть больше времени

Управление симуляцией
- Жить интересной жизнью, чтобы не выключили симуляцию
- Не давать симуляторам догадаться, что мы знаем об их существовании
- Взломать симуляцию и отдавать ей правильные команды
- Вступить в переговоры с симуляторами, просить их

Плохие идеи

Исследование глобальных рисков только увеличивает эти риски
- Не пропагандировать идею глобальной катастрофы
- Внедрение систем контроля только создает новые риски
- Невозможно измерить вероятность глобальной катастрофы и ее изменение

Управляемый регресс
- Использование малой катастрофы для предотвращения большей (Willard Wells)
- Лудизм (Качински)
- остановка развития науки
- Создание экологичной цивилизации без технологий ("World made by hand")
- Снижение общего интеллекта, чтобы никто не мог изобрести ничего плохого

Сокращение населения
- Экономия ресурсов, упрощение контроля
- Пандемии, войны, голод (Мальтус)
- Снижение рождаемости (Гейтс)
- Преднамеренная малая катастрофа (биооружие)

Недружественный ИИ лучше, чем полное вымирание
- Любой ИИ сохранит какую-то информацию о человечестве
- Он будет запускать симуляции человечества, чтобы исследовать вероятность своего возникновения
- Он сохранит долю человеческих ценностей и распространит ее по Вселенной, а также нести отпечаток личности своего создателя

Притягивание позитивного будущего и прочая магия
- Пребывать в максимально позитивном мировоззрении, чтобы "притягивать" безопасное будущее
- Избегание негативных мыслей о конце света и насилии
- Устроить пир во время чумы

Глобальная катастрофа