Plan of Action to Prevent Human Extinction Risks

Steps and timeline

Exact dates depend on the pace of technological progress and on our risk-prevention activity

First quarter of 21st century (2015—2025) → Second quarter of 21st century (2025—2050) → Second half of the 21st century (2050—2100)

Step 1
Planning

Step 2
Preparation

Step 3
First level of defence on low-tech level

Step 4
Second level of defence on high-tech level

Step 5
Reaching indestructibility of civilization with negligible annual probability of global catastrophe (1×10^{-10} chance per year)

More probable

Plan A – Prevent catastrophe

Plan A – Control risks

Plan A1.1 – International Control System

Research

Information gathering

- Create and promote **long-term future model**
- **Comprehensive list of risks**
- **Probability assessment**
- **Prevention roadmap**
- **Determine most probable risks** and most easy to prevent ones
- **Create x-risks wiki and x-risks internet forum** which would attract the best minds, but is also open to everyone and well moderated

Assistance

- Solve the problem of different x-risk-aware scientists ignoring each other ("world savior's arrogance" problem)
- **Integrate different lines of thinking about x-risks**
- Lower the barriers to entry
- **Unconstrained funding of x-risk research** for many different approaches
- Help best thinkers in the field (Bostrom) produce high quality x-risk research
- **Education of "world saviours"**: choose best students, provide them courses, money and tasks

Additional study areas

- **Study existing system of decision-making in UN**, hire a lawyer
- Create a general theory of safety and risk prevention
- **Create full list of x-risk-related cognitive biases** and work to prevent them (Yudkowsky)
- **Translate best x-risk articles** and books into common languages

Social support

Science

- Rising support **inside academic community** by high level signaling
- **Cooperative** scientific community with shared knowledge and productive rivalry
- Productive cooperation between scientists and society **based on trust**
- Scientific attributes:
  - **peer-reviewed journal,**
  - **conferences,**
  - **inter-governmental panel,**
  - **international institute**

Popularization

- **Articles, books, forums, media**: show public that x-risks are real and diverse, but we could and should act

Politics

- Public support: **street action (anti-nuclear protests in 80s)**
- Political support: **lobbyism**
- **Political parties** for x-risk prevention
- Write policy recommendations

Reactive and Proactive approach

Reactive: **React on most urgent and most visible risks**.
Pros: Good timing, visible results, proper resource allocation, invest only in real risks.
Good for slow risks, like pandemic.
Cons: can't react to fast emergencies (AI, asteroid, collider failure).
Risk are ranged by urgency

Proactive: **Envision future risks and build multilevel defence**.
Pros: Good for coping with new risks, enough time to build defense.
Cons: Misinvestment in fighting with non-real risks, no clear reward, challenge in identifying new risks and discounting mad ideas.
Risks are ranged by probability

International cooperation

Super-UN

- **All states contribute to the UN to fight certain global risks**
- States **cooperate directly** without UN
- **Superpowers take responsibility** for x-risk prevention and sign a treaty
- **International law about x-risks** which will punish people for rising risk: underestimating it, plotting it, risk neglect, as well as reward people for lowering x-risks, finding new risks, and for efforts in their prevention.
- **International agencies dedicated to certain risks** (old ones like WHO and new ones for new risks)

Stimulus

- **A smaller catastrophe could help unite humanity** (pandemic, small asteroid, local nuclear war)
- Some movement or event that will cause a **paradigmatic change** so that humanity becomes more existential risk aware

Practical steps to cope with certain risks

Biosecurity:

- Developing **better guidance on safe bio-technology**
- DNA synthesizers are constantly connected to the Internet and **all newly created DNA is checked** for dangerous fragments
- **Funding things like clean-rooms** with negative pressure
- **Better quarantine laws** for travelling during pandemic
- "**Develop and stockpile** new drugs and vaccines, monitor biological agents and **emerging diseases**, and strengthen the capacities of **local health systems** to respond to pandemics" (Matheny)

Environment

- **Methane and CO2 capture**, may be by bacteria
- Invest in biodivercity of food supply chain, prevent pest spread: 1) better food quarantine law, 2) **portable equipment for instant identification** of alien inclusions in medium bulks of foodstuffs, and 3) further development of nonchemical ways of sterilization.
- Mild birth control (female education, contraceptives)
- **Solar and wind energy**

Nukes

- **Asteroid deflection technology without nukes**
- **Improve nuclear diplomacy**
- Instruments in tech to **capture radioactive dust**

Risk control

Technology bans

- **International ban on dangerous technologies** or **voluntary relinquishment** such as not creating new strains of flu
- **Freezing potentially dangerous projects** for 30 years
- Lowering international confrontation
- Lock down risky domains beneath piles of bureaucracy, paperwork and safety requirements

Control

- **Concentrate all bio, nano and AI research in several controlled centers**
- **Control over dissemination of knowledge of mass destruction** (viral DNA) or **create white noise in the field**
- Control over suicidal or terrorist ideation in people working on risky subjects
- Investments in general safety: education, facilities, control, checks, exams

Surveillance

- **International control systems (like IAEA)**
- **Internet scanning, monitoring from space**

Elimination of certain risks

- **Universal vaccines,** UV cleaners
- Asteroid detection (WISE) **prove that no dangerous asteroids exist**
- **Transition to renewable energy**, cut emissions, carbon capture
- **Stop LHC,** SETI and METI **until AI** is created
- **Rogue countries integrated** or prevented from having dangerous weapons programs and advanced science

Worldwide risk prevention authority

- **Center for quick response** to any emerging risk: x-risk police
- **Worldwide video-surveillance** and control
- "The ability **when necessary to mobilise a strong global coordinated response** to anticipated existential risks" (Bostrom)
- **Peaceful unification of the planet** based on a system of international treaties
- **Robots for emergency** liquidation of bio- and nuclear hazards
- **Narrow AI** based expert system on x-risks, Oracle AI

Planetary unification war

- **War for world domination**
- One country uses **bio-weapons to kill all world's population except its own** which is immunized
- Use of a **super-technology** (like nano-tech) to quickly **gain global military dominance**
- **Doomsday Machine blackmail** for world domination

Global catastrophe

Active shields

- **Geoengineering** against global warming
- Worldwide missile defense and **anti-asteroid shield**
- **Nanoshield** – distributed system of control of hazardous replicators
- **Bioshield** – worldwide immune system
- Dangerous memes control (existential terrorism prevention)
- Knowledge of mass destruction control
- Merger of the state, Internet and worldwide AI in **worldwide monitoring, security and control system**
- **Isolation of risk sources at a great distance from Earth;**
- **Do scientific experiments in the ways that are close to natural events**

Fatal mistakes in world control system

- **Wrong command** makes shield system attack people
- **Geo-engineering** goes awry
- **World totalitarianism** leads to catastrophe

Global catastrophe

Singleton

- "A world order in which there is a single decision- making agency at the highest level" (Bostrom)
- Worldwide security system based on AI
- Super AI which prevents all possible risks and provides immortality and happiness to humanity
- Colonization of the solar system, interstellar travel and Dyson spheres
- Colonization of the Galaxy
- Exploring the Universe

Ideological payload of new technologies

Space tech - Mars as backup
Electric car - sustainability
Self-driving cars - risks of AI and value of human life
Facebook - empathy and mutual control
Open source - transparency
Computer games and brain stimulation - virtual world

Design new monopoly tech with special ideological payload

Plan A1.2 – Decentralized monitoring of risks

Values transformation

The value of existential risk reduction

- "A moral case can be made that existential risk reduction is strictly more important than any other global public good" (Bostrom)
- **The value of indestructibility of civilization should be first priority on all levels:** in education, on a personal level, and as a goal of every nation
- Improve public desire for life extension and global security

Dissemination of this value

- Reduction of **radical religious** (ISIS) or nationalistic values
- Popularity of **transhumanism**
- Movies, novels and other **works of art** that honestly depict x-risks and motivate their prevention
- **Memorial and awareness days:** Earth day, Petrov day, Asteroid day
- **Education** in schools on x-risks, safety, and rationality topics; raising the sanity waterline

Improving human intelligence and morality

Intelligence

- Nootropics, brain stimulation, and gene therapy for **higher IQ**
- **New rationality**: Bayesian probability theory, interest in long term future, LessWrong
- Fighting **cognitive biases**
- Many **rational**, positive, and cooperative **people** are needed to reduce x-risks (effective altruists)

Empathy

- **High empathy for new geniuses** is needed to prevent them from becoming superterrorists
- **Lower proportion of destructive beliefs**, risky behaviour, and selfishness
- **Engineered enlightenment**: **use brain science** to make people more united, less aggressive; open realm of spiritual world to everybody

Morality

- **Prevent worst forms of capitalism:** the desire for short term monetary reward
- **Promote best moral qualities:** honesty, care, non-violence

Cold War, local nuclear wars and WW3 prevention

- **International conflict management authority like international court** or secret institution
- **Large project which could unite humanity**, like pandemic prevention
- Integrate **rogue countries into global system,** based on dialogue and an appreciation for their values
- **Hotlines** between nuclear states
- **Antiwar and antinuclear movement**
- **International law** as the best instrument of conflict solving
- **Peaceful integration** of national states
- **Cooperative decision theory in international politics** ("do not press on red")
- **Prevent brinkmanship**
- **Prevent nuclear proliferation**

Dramatic social changes

- It could include many interesting but different topics: demise of capitalism, hipster revolution, internet connectivity, global village, dissolving of national states.
- Changing the way that politics works so that the policies implemented actually have empirical backing based on what we know about systems.
- World democracy based on internet voting
- High level of horizontal connectivity between people

Decentralized risk monitoring

- **Transparent society**: everybody could monitor everybody:
  - groups of vigilantes scanning the open web and sensors
  - "Anonymous" style hacker groups: search in encrypted spaces
- **Decentralized control**:
  - **local police** could handle local crime and terrorists;
  - **local health authorities** could find and prevent the spread of disease
  - **mutual control** in professional space
  - **google search control**
  - **whistle-blowers** inform public about risks and dangerous activity
- **Net-based safety** solutions:
  - **ring of x-risk prevention organizations**
  - personal safety instruction for each worker: short and clear
- **Economic stimulus:**
  - carbon emissions trade
  - **prizes for any risk identified and prevented**
- **Monitoring of smoke,** not fire:
  - **search predictors of dangerous activity, using narrow AI**

Plan A2 – Artificial Intlligence

Plan A2.1 – Friendly AI creation

Study and Promotion

- **Study of Friendly AI theory**
- Promotion of Friendly AI (Bostrom and Yudkowsky)
- **Fundraising** (MIRI)
- Slowing other AI projects (recruiting scientists)
- FAI free education, starter packages in FAI

Solid Friendly AI theory

- **Theory of human values and decision theory**
- **Full list of possible ways to create FAI**, and sublist of best ideas
- Proven safe, fail-safe, intrinsically safe AI
- Preservation of the value system during AI self-improvement
- **A clear theory that is practical to implement**

AI practical studies

- **Narrow AI**
- **Human emulations**
- **Value loading**
- **FAI theory promotion to all AI developers**; their agreement to implement it and adapt it to their systems
- **Tests of FAI theory on non self-improving models**

Seed AI

Creation of a small AI capable of recursive self-improvement and based on Friendly AI theory

Specialized AI

Use specialized AI for risk monitoring, risks database, risks prediction and intervention for risk prevention (virus, robots)

Superintelligent AI

- Seed AI **quickly improves itself** and undergoes "hard takeoff"
- It becomes **dominant force** on Earth
- AI **eliminates suffering**, involuntary death, and existential risks
- **AI Nanny** – creating super AI that only acts to prevent other existential risks (Ben Goertzel)

Singleton

Unfriendly AI

- Kills all people and maximizes non-human values (paperclip maximiser)
- People are alive but suffer extensively

Plan A2.2 – AI for risk control

Computer database of risks

Internet forum for integration of knowledge

Specialized AI

Use specialized AI for
risk monitoring,
risks database,
risks prediction and
intervention for risk prevention (virus, robots)

Global non-self-improving AI used for risks control

- Integration of data from different surveillance systems
- Models of future and of the world to predict risks
- Control of active shields

Singleton

Plan A2.3 – Create AI as soon as possible

Technology speedup

- **Differential technological development: develop safety and control technologies first (Bostrom)**
- **Laws and economic stimulus** (Richard Posner, carbon emissions trade)

High-speed Tech Development needed to quickly pass risk window

- Investment in **super-technologies** (nanotech, biotech, Friendly AI)
- **High speed technical progress helps** to overcome slow process of resource depletion
- Invest more in **defensive technologies** than in offensive
- If probability of biorisks and nuclear war is linear in time, then we need to create strong AI quicker.

Plan B – Survive the catastrophe

Limit scope of risk

Plan B1 – Improving Resilience

Improving sustainability of civilization

- **Intrinsically safe critical systems**
- **Growing diversity** of human beings and habitats
- **Universal methods of catastrophe prevention** (resistant structures, strong medicine)
- **Building reserves (food stocks**, seeds, minerals, energy, machinery, knowledge)
- **Widely distributed civil defence**, including:
  - temporary shelters,
  - air and water cleaning systems,
  - radiation meters, gas masks,
  - medical kits
  - mass education

Useful ideas to limit the scale of a catastrophe

- **Limit the impact of catastrophe** by implementing measures to slow the growth and areas impacted by a catastrophe:
  - technical instruments for implementing **quarantine**,
  - improve the capacity for **rapid production of vaccines** in response to emerging threats
  - **grow stockpiles** of important vaccines
- **Increase preparation time** by improving monitoring and early detection technologies:
  - support general **research on the magnitude of biosecurity risks** and opportunities to reduce them
  - improve and **connect disease surveillance systems** so that novel threats can be detected and responded to more quickly
- **Worldwide x-risk prevention exercises**
- **The ability to quickly adapt** to new risks and envision them in advance

Become AI: Invest in uploading and human like AI

Raise human body resilience by replacing biological parts

- **Nanotech-based immortal body**
- **Diversification of humanity** into several successor species capable of living in space
- **Mind uploading**
- **Integration with AI**

AI based on uploading of its creator

- Friendly to the value system of its creator
- Its values consistently evolve during its self-improvement

Plan B2 – Space Colonization

Temporary asylums in space

- **Space stations** as temporary asylums (ISS)
- Cheap and safe **launch systems**

Space colonies near Earth

- Creation of **space colonies on the Moon** and Mars (Elon Musk) with 100-1000 people

Colonisation of the Solar system

- **Self-sustaining colonies** on Mars and large asteroids
- **Terraforming** of planets and asteroids using self-replicating robots and building space colonies there
- **Millions of independent colonies inside asteroids** and comet bodies in the Oort cloud

Interstellar travel

- "Orion" style, nuclear powered "**generation ships**" with colonists
- Starships which operate **on new physical principles** with immortal people on board
- Von Neumann **self-replicating probes** with human embryos

Interstellar distributed humanity

- Many unconnected human civilizations
- New types of space risks (space wars, planets and stellar explosions, AI and nanoreplicators, ET civilizations)

Plan B3 – Bunkers and asylums

Preparation

- **Fundraising** and promotion
- **Textbook** to rebuild civilization (Dartnell's book "Knowledge")
- **Hoards** with knowledge, seeds and raw materials (Doomsday vault in Norway)
- **Survivalist** communities

Building

- **Underground bunkers,** space colonies
- **Seasteading**
- **Convert existing nuclear submarines** into refuges

Natural refuges

- **Uncontacted tribes**
- **Remote villages**
- **Remote islands**
- **Oceanic ships**
- **Mines**

Readiness

- Crew **training**
- Crews in bunkers
- Crew **rotation**
- **Different types** of asylums
- Frozen embryos

Miniaturization for survival and invincibility

- **Adaptive bunkers** based on nanotech
- **Colonization of Earth's crust** by miniaturized nano-tech bodies
- Moving into **simulation worlds** inside small self-powered computer systems

Rebuilding civilization after catastrophe

- **Rebuilding population**
- **Rebuilding science and technology**
- **Prevention of future catastrophes**

Reboot of civilization

- Several reboots may occur
- Ultimately there will be total collapse or a new level of civilization

Plan C – Leave Backups

Time capsules with information

- Underground **hoards with information** and DNA for future non-human civilizations
- **Eternal disks** from Long Now Foundation (or *M-disks*)

Messages to ET civilizations

- Interstellar **radio messages** with encoded human DNA
- **Hoards on the Moon**, other planets
- **Frozen brains** in cold places
- **Voyager-style spacecrafts** with information about humanity

Preservation of earthly life

- Create conditions for the **re-emergence** of new intelligent life on Earth
- **Directed panspermia** (Mars, Europe, space dust)
- **Preservation of biodiversity** and highly developed animals (apes, habitats)

Robot-replicators in space

- **Mechanical life: nanobots ecosystem, von Neumann probes based on nanobots**
- **Preservation of information** about humanity for billions of years in replicators
- **Safe narrow AI** regulating such robots-replicators

Resurrection by another civilization

- Aliens will create a civilization which has a lot of common values and traits with humans
- Resurrection of concrete people based on information

Plan D – Improbable Ideas

Saved by non-human intelligence

- Maybe extraterrestrials are looking out for us and will save us
- Send radio messages into space asking for help if a catastrophe is inevitable
- Maybe we live in a simulation and simulators will save us

Quantum immortality

- If the many-worlds interpretation of QM is true, an observer may survive any sort of death including any global catastrophe (Moravec, Tegmark)
- It may be possible to make an almost univocal correspondence between observer survival and survival of a group of people (e.g. if all are in submarine)
- Other human civilizations must exist in the infinite Universe

Strange strategy to escape Fermi paradox

Random strategy may help us to escape some dangers that killed all previous civilizations in space

Technological precognition

- Prediction of the future based on advanced quantum technology and avoiding dangerous world-lines
- Search for potential terrorists using new scanning technologies
- Special AI to predict and prevent new x-risks

Manipulation of the extinction probability using Doomsday argument

- Decision to create more observers in case unfavorable event X starts to happen, thereby lowering its probability (method UN++ by Bostrom)
- Lowering the birth density to get more time for civilization

Control of the simulation (if we are in it)

- Live an interesting life so our simulation isn't switched off
- Don't let them know that we know we live in simulation
- Hack the simulation and control it
- Negotiation with the simulators or pray for help

Bad plans

Prevent x-risk research because it only increases risk

- **Do not advertise** the idea of man-made global catastrophe
- **Don't try to control** risks as it would only give rise to them
- As we **can't measure the probability** of global catastrophe, it may be unreasonable to try to change the probability
- **Do nothing**

Controlled regression

- Use a **small catastrophe** to prevent large one (Willard Wells)
- **Luddism** (Kaczynski): relinquishment of dangerous science
- Creation of **ecological civilization** without technology ("World made by hand", anarcho-primitivism)
- **Limitation** of personal and collective **intelligence** to prevent dangerous science
- Antiglobalism and diversification into **multipolar world** (may raise probabilities of wars)

Depopulation

Could provide resource preservation and make control simpler

- **Natural causes**: pandemics, war, hunger (Malthus)
- Extreme **birth control**
- **Deliberate small catastrophe** (bio-weapons)

Computerised totalitarian control

- **Mind-shield** – control of dangerous ideation by the means of brain implants
- **Secret police which use mind control** to find potential terrorists to stop them

Choosing the way of extinction: UFAI

- Quick die off is better
- Any super AI will have some **memory** of humanity
- It will use **simulations** of human civilization to study the probability of its own existence
- It may share some **human values** and distribute them throughout the Universe

Attracting good outcome by positive thinking

- Start **partying now**
- Preventing **negative thoughts** about the end of the world and about violence
- Maximum **positive attitude** "to attract" positive outcome

Global catastrophe

Plans A1-A4 represent the best possible ways to prevent global catastrophe.
These plans could be realized simultaneously until the later stages.
Plans B, C and D are subsequently worse and give only a small chance of survival.

(c) Alexey Turchin, 2014 - 2016