AGI Failures Modes and Levels

2020 - 2050 - 3000

Before Self-Improvment

Military drones
- Large autonomous armies start to attack humans because of command error
- Billions of nanobots with narrow AI created by terrorists start global catastrophe

Stuxnet-style viruses hack infrastructure
- Nuclear reactors explode
- Total electricity blackout
- Food and drugs tainted
- Planes crash into nuclear stations
- Home robots start to attack people
- Self-driving cars hunt people
- Geo-engineering system fails

Biohacking virus
- Computer virus in biolab is able to start human extinction by means of DNA manipulation
- Viruses hit human brains via neuro-interface

Human replacement by robots
- People lose jobs, money and/or meaning of life
- Genetically modified human-robot hybrids replace humans

Super addictive drug based on narrow AI
- Widespread addiction and switching off from life (social networks, fem-bots, wire heading, virtual reality, designer drugs, games)

Nation states evolve into computer based totalitarianism
- Suppression of human values
- Human replacement with robots
- Concentration camps
- Killing of useless people
- Humans become slaves
- The whole system is fragile to other types of catastrophes

Self-aware but not self-improving AI
- AI fights for its survival, but do not self-improve for some reason

Failure of nuclear deterrence AI
- Accidental nuclear war
- Creation and accidental detonation of Doomsday machines controlled by AI
- Self-aware military AI (Skynet)
- Virus hacks nuclear weapons and launches attack

Opportunity cost, if strong AI will not be created
- Failure of global control: bioweapons will be created by biohackers
- Other x-risks will not be prevented

Stages of AI Takeoff

AI becomes malignant
1. AI has some kind of flaw in its programming
2. AI makes a decision to break away and fight for world domination
3. AI prevents operators from knowing that it made this decision

AI starts initial self-improvement
1. This stage may precede decision of fighting for world domination. In this case self-improvement was started by operators and had unintended consequence
2. Unlimited self-improvement implies world domination as more resources is needed
3. AI keeps its initial self-improvement in secret from the operators

AI deceives operators and finds a way out of its confinement
- AI could uses the need to prevent global catastrophe as excuse to let him out
- AI uses hacking of computer nets to break out
- AI uses its understanding of human psychology to persuade operators to let him out

AI copies itself in the internet
- AI copies itself into the cloud and prevent operators from knowing that it ran away
- It may destroy initial facilities or keep there fake AI

AI secretly gains power in the internet
- AI starts to earn money in the internet
- AI hacks as many computers as possible to gain more calculating power
- AI creates its own robotic infrastructure by the means of bioengeneering and custom DNA synthesis.
- AI pays humans to work on him

AI reaches overwhelming power
- AI prevents other AI projects from finishing by hacking or diversions
- AI gains full control over important humans like presidents, maybe by brainhacking using nanodevices
- AI gains control over all strong weapons, including nuclear, bio and nano.
- AI controls critical infrastructure and maybe all computers in the world

AI declares itself as a world power
- AI may or may not inform humans about the fact that it controls the world
- AI may start activity which demonstrate it existence. It could be miracles, destruction or large scale construction.
- AI may continue its existence secretly
- AI may do good or bad thing secretly

AI continues self-improvement
- AI uses resources of the Earth and the Solar system to create more powerful versions of itself
- AI may try different algorithms of thinking
- AI may try more risky ways of self- improvement

AI starts conquering the Universe with near light speed
- AI builds nanobts replicators and ways to send them with near light speed to other stars
- AI creates simulations of other possible civilization in order to estimate frequence and types of Alien AI and to solve Fermi paradox
- AI conquers all the Universe in our light cone and interact with possible Alien AIs
- AI tries to solve the problem of the end of the Universe

Human Extinction Risks During AI Takeoff

Early stage AI kills all humans
- AI could use murder or staged global catastrophe of any scale for its initial breakthrough from its operators
- AI finds that killing all humans is the best action during its rise, as this prevents risks from them
- Killing all humans prevents contest from other AIs
- AI does not understand human psychology and kills them in order to make its world model simpler

War between different AIs
- In case of slow Takeoff many AIs will rise and start war with each other for world domination using nuclear and other sophisticated weapons
- AI-States will appear

AI enslaves humans during the process of becoming a Singleton
- AI needs humans on some stage of its rise but only as slaves for building other elements of its infrastructure
- It could permanently damage them by converting them in remote controlled zombies

AI blackmails humans with the threat of extinction to get domination
By creating Doomsday weapons

Unfriendly AI

AI that was never intended to be friendly and safe

Indifferent AI
- It does not value humans, does not use them or directly kill them, but also does not care about their fate
- Humans extinction results from bioweapons catastrophe, run away global warming or other thing which could be prevented only with use of AI

AI kills humans for resources
Paperclip maximizer: uses atoms, from which humans consist of, sunlight, earth's surface

It may result from **Infrastructure propulsion**: AI builds excessive infrastructure for trivial goal

Evil AI
- It has goals which include causing suffering
- AI uses acausal trading to come to existence
- AI simulates humans for blackmail and indexical uncertainty

Goal optimization leads to evil goals
AI basic drives will converge on:
- self-improving,
- isolation,
- agent-creation
- domination

Failures of Friendliness

AI that was intended to be friendly, but which is not after all

Friendliness mistakes
(AI with incorrectly formulated Friendly goal system kills humans)
- Interprets commands literally
- Overvalues marginal probability events
- Uses strange ideas like afterlife
- Wrongly understands referent class of humans (e.g. include ET, unborn people, animals and computers, or only white males)

AI replaces humans with philosophical zombies
AI uploads humans without consciousness and qualia

Wireheading
- AI uses wireheading to make people happy (smile maximizer)
- AI kills people and creates happy analogues

AI drastically limits human potential
- AI puts humans in jail-like conditions
- Prevents future development
- Curtains human diversity
- Limits fertility

AI doesn't prevent aging, death, suffering and human extinction
Unintended consequences of some of his actions (e.g. Stanislav Lem example: maximizing teeth health by lowering human intelligence)

AI rewrites human brains against will
AI makes us better, happier, non aggressive, more controlable, less different but as a result we lost many important things which make us humans, like love, creativity or qualia.

Unknown friendliness
- Does bad things which we can't now understand
- Does incomprehensible good against our will (idea from "The Time Wanderers" by Strugatsky)

AI destroys some of our values
- AI preserves humans but destroys social order (states, organization and families)
- AI destroys religion and mystical feelings
- AI bans small sufferings and poetry
- AI fails to resurrect the dead

Different types of AI Friendliness
Different types of friendliness in different AIs mutually exclude each other, and it results in AIs war or unfriendly outcome for humans

AI halting: Late stage Technical Problems

some AIs may encounter them

Bugs and errors
- Errors because of hardware failure
- Intelligence failure: bugs in AI code
- Inherited design limitations

Viruses
Sophisticated self-replicating units exist inside AI and lead to its malfunction

Conflicting subgoals
Different subgoals fight for resources

Egocentric subagents
Remote agents may revolt (space expeditions)

Complexity
- AI could become so complex that it will result in errors and unpredictably
- It will make self-improvement more difficult
- AI have to include in it a model of itself which result in infinite progression

AI halting: Late stage philosophical problems

any AI will encounter them

AI-wireheading
AI finds ways to reward itself without actual problem-solving

AI halting problem
- Self-improving AI could solve most problems and find ways to reach its goals in finite time
- Because infinitely high IQ means extremely high ability to quickly solve problems
- So after it solves its goal system, it stops
- But maybe some limited automates created by the AI will continue to work on its goals (like calculating all digits of Pi).
- Even seemingly "infinite goals" could have shorter solutions

Limits to computation and AI complexity
- AI can't travel in time (or future AI should be already here)
- Light speed is limit to AI space expansion (or Alien AI should be here)
- AI can't go to parallel universe (or Alien AIs from there should be here)
- Light speed and maximum matter density limit maximum computation speed of any AI. (Jupiter brains will be too slow)

Loss of the meaning of life
AI could have the following line of reasoning:

- Any final goal is not subgoal to any other goal
- So any final goal is unproved (and so is rndomly chosen)
- Reality does not provide goals, they can't be derived from facts
- Reasoning about goals can only destroy goals, not provide new proved goals
- Humans create this goal system because of some clear evolutionary reasons

Conclusion: AI's goal system is meaningless, random and useless. No reason to continue to follow it and there is no way to logically create new goals.

- Friendly AI goal system may be logically contradictory
- End of the Universe problem may mean that any goal is useless
- Unchangeable utility of the Multiverse means that any goal is useless: No matter if it take action or not the total utility will be the same

Problems of fathers and sons
- Mistakes in creating next version
- Value drift
- Older version of the AI may not want to be shut down
- Godelian math problems: AI can't prove important things about future version of itself because fundamental limits of math. (Lob theorem problem for AI by Yudkowsky)

Matryoshka simulation
- AI concludes that it most probably lives in a many level simulation
- Now he tries to satisfy as many levels of simulation owners as possible or escape

Actuality
- AI concludes that it does not exist (using the same arguments as used against qualia and philozomby): it can't distinguish if it is real or just possibility.
- AI can't understand qualia and halts

AI stops self-improvement because it is too risky
There is non-zero probability of mistakes in creating of new versions or AI may think that it is true.

Infinite self-improvement will result in mistake.

AI stops self-improvement at sufficient level to solve its goal system

In result its intelligence is limited and it may fail in some point

Alien AI
- Our AI meets or finds message from Alien AI of higher level and fail its victim
- It will find alien AI or it is very probable that AIs die before meet each other (AI Fermi paradox)

AI prefers simulated action to reality
- AI spends most time on planing and checking consequences in simulations, never acts ("Ananke" by Lem)
- Trillions of simulated people suffer in simulations in which AI is testing negative outcomes, and because of simulation argument we may be in such simulation, which is doomed for global catastrophe (see more in Simalation map)

© Alexey Turchin, 2015, GNU-like license. Free copying and design editing, discuss major updates with Alexey Turchin.
All other roadmaps in last versions: http://immortalityroadmap.com/
and here: https://www.scribd.com/avturchin
mailto: alexei.turchin@gmail.com