AGI Safety Solutions Map

Simple solutions

Complex solutions

Give AI to everybody

Suggested by Elon Musk in OpenAI project

Technical level solution

One can't make safe rocket, without actually knowing how to build a rocket. The same may be true about AI. So we may need to wait until more will be known about most promissing AI architectures and use more help from actual developers.

Intelligence distribution

Create sitiation in which distributed collectibe intelligence of humanity is always greater than of any company, or computer, so it can't get desisive advantage.
Distributed science
Neuronet

Societal Proposals

Do Nothing

- AGI is impossible
  - Humans are unique, a soul exists
  - Existence of upper limits on an entity's effective intelligence around the level of human geniuses
  - Individual intelligence is always outmatched by the distributed intelligence found in social systems of many minds
- AGI is too distant to be worth our attention
  - not in this century
  - not on our level of understanding of AI theory
  - the discussion has negative utility
- Little risk, no action needed
  - AGIs will not have any particular motivation to act against us
  - AGI cannot do much
  - High intelligence implies high morality
- Let it kill us
  - Humans are bad
  - Higher intelligence is more important than humanity
  - Quantum immortality will make only Safe AGI observable
  - AI is our child – we must step aside
  - AI will share some of our values and will construct ancestor simulations which contain humans
  - We should not constrain AGI because it deserves full freedom to act
- Safety measures are dangerous themselves
  - The more we try to confine it, the more AGI will try to escape

Integrate into Society

Marginal

- Legal and Economic Controls:
  - AGIs will be law-abiding
  - Share resources
  - Taxation and regulation
  - Integrate into government
- Foster Positive Values
- State AGI: State transforms into AGI through gradual replacement of existing functions (computer tax regulation and law implementation)

Regulate Research

Marginal

- International regulation
  - international treaty
  - boards
  - licensing research
- Encourage research into safe AGI
  - funding for research
  - open-source approach
- Differential Technological Progress
  - curtail unsafe AI architectures
  - prevent AGI arms race
  - one center of AGI development
- International Mass Surveillance
  - narrow AIs as controllers of AGI research

Enhance Human Capabilities

Marginal

- Create AGI by human upgrade
  - cognitive enhancement
  - moral enhancement
- Keep humans on the same level as AGI

Relinquish Technology

- Outlaw AGI
  - outlaw research on AI
  - nuclear war against AGI rivals [Turchin]
- Restrict hardware
  - limit processing power
  - outlaw supercomputers
  - control microchip fabs
  - switch off electricity or internet

Situation Improvement

Marginal

- Funding of safe AGI research
- Slowing development of AGI
- Promotion of the idea of AGI safety
- Ban military AGI
- Friendly AGI training for AGI researchers
- One main international AGI project, not many rivals
- Safe AGI architectures: e.g. no genetic algorithms
- Hire existing ethical researchers and use legal and ethical literature
- Hire existing safe AI researchers

External Constraints

AGI Confinement

Marginal

- Total isolation
- Isolation like Mars or Pluto
- Only text communication
- Internet firewall
- Virtual reality: many levels of isolation
- "Safe Questions" (not implying dangerous behaviour)
- Resetting AGI memory
- Checks and balances
  - baits
  - check for unusual pattern in the AGI
  - circuit breakers
  - off switch
- Cryptobox: homomorphic encryption [Christiano]

Control room and Gatekeeper

Marginal

- Monitors of internal processes
- AI is constantly showing what it is thinking about
- Power switch
- Explicit goals and subgoals
- Firewall
- Upgrade control
- Mirror AIs to check for strange behaviour (3 copies of the AGI)
- Seed AI testing for unintended behaviour

Philosophical Constrains

Marginal

- AGI fooled to be in simulation which could switch him off
- Persuade AGI that humans have soul or other incomputable quality and it needs them
- Prove that we really maybe in the simulation and AGI should keep people alive to not risk punishment from high-level simulators

Prevent Self-improvement

Promising

- Limit number of self-improvement steps
- Secure permission from human operator for each step of self-improvement
- Slow rate of self-improvement
- Limit areas of self-improvement
- Limit access to own source code
- Limit IQ of the AGI
- Limit available hardware for the AGI
- Limit working speed of the AGI
- Prevent large architecture changes
- Create as many AGIs as needed to prevent any of them from total power
- Differential self-improvement: slow on high levels and quick on low level

Goal Checking

Marginal

- Automatically search for failure modes of goal system
- List of signs of failure modes
- Committee for goal checking
- Prevent failure modes

Catch the moment of beginning of malignant AI self-improvement

Marginal

After AI acquires malignant goal system and starts self-improvement, it is still stupid and could be stopped

Logical or philosophical "landmines"

which will stop too complex AI

Marginal

- Idea of absurdity of any goal system
- AI halting problem
- Lobian obstacle [Yudkowsky]
- AGI is given false believes, which is known to it but it still obeys them [Armstromg]

Record AI's thought stream

Build AI in the way that its high-level resaonong are presented in the open form and record them and analyze, may be using another AI.

Signs constrains

Identify all signs of possible wrongdoings by AI and make independed circut breakers.
Signs:
- Reading source code
- Modifing own code
- Creating AI models
- Power consumption changes
- Behavior patterns changes
- Internal work pattern changes
- Access to Internet on AI tipics
- Attempts to cheat and lie
- Goal system changes

Knowledge constrains

- Separete source code of AI, executable code, data and knowledge about how to build AI.
- AI can't phisically read its own source code or its attempts are recorded

Internal Constraints

Oracle AI

May be

- Answer questions
- No goals
- No memory
- Not one AI but many narrow Tool AIs with different goals
- Slave AI – only obey commands

Short rulesets

May be

- "Three laws by Asimov"
- Categorical imperative
- "Principle of voluntary joyous growth"
- Utilitarianism
- AGI obeys existing criminal laws
- Short rulesets of other types

Value learning

Promising

- AGIs cooperate with humans in value learning
- AI learns values by studying uploaded brain
- Education of childlike Seed AI

Bottom-Up and Hybrid Safe AGI

May be

- Evolutionary invariants of ethics
- Evolved morality
- Reinforced learning of ethics
- Learning values by pattern recognition
- Human-like architecture

Formal Verification

May be

- "Goal-oriented Learning Meta-Architecture" – search for better way to known goal [Goertzel]
- Provable self-modifying algorithms

Motivational Weaknesses

Marginal

- High Discount Rates for time or distance (don't try to save Universe)
- Easily satisfiable goals
- Promised safety for the AGI
- Calculated indifference to some threats
- Legal machine language

AI Constitution

Promising

AI is regulated by proven large set of rules; some possible rules:
Meta rules:
- Obey commands of its creator
- Report and explain value changes
- No secrets
- Limit self-improvement
- Stop after 10 years

Content:
- Do not kill people
- Cure death
- Prevent severe suffering
- Prevent x-risks
- Don't rewrite human brains against their will

AI as Next Step in Human Evolution

Promising

- AI based on emulations
  - Society of EM's working at high speed
  - Upload a specially trained person with math skills and solid ethic as Seed AI
  - EM as value and awareness core of AGI
- Use human-like value architecture: cultural upgrade of natural drives, reward system, complex values, many different architectures
- AGI use humans ("me") as a source of qualia (actuality) and preserve them
- AI is based on many human brains connected together (neuronet)
- Psychogenic singularity: use raw power of human brain and its unconsciousness to create AI (like savants)
- AI is created by as gradual upgrade of one or several human beings (posthumans)
- AI is created by genetic experiments like a child or superbrain
- AI pretends to be best human (Armstrong)

Decision Theory Based Friendliness

Marginal

- Cooperate in prisoner dilemma
- Don't take actions which are irreversible
- Anti-Pascaline agent [Armstrong]
- Corrigible reasoning [MIRI]

AGI without one Supergoal

Marginal

- Any one-goal system may be dangerous like "maniac"
- Democracy of goals
- Values not goals
- AI capable on reflection on its goals and values
- All possible goals in different proportions (like in humans)

Opacity constrains

Use opaque architecture which is difficult to understand and improve (some kinds of neural nets)

Open goal system

The actual goal system of an AI must be presented to external observer in the form of an open tree (text, graph) and any changes in it should be saved remotely in unchangeble way.
If the AI will try to create another goal tale it will be presented in first.

Children constrain

In order to selfupgrade, an AI must first create new version of itself and test it.
Some AI-architectures may make it very complex task with many fails and resource consuming "education period" (neural nets)

AI used to create Safe AI

AGI tasked to create Safe AGI

May be

- Oracle AI gives friendliness theory
- Oracle AI designs safe AI
- Oracle AI tests different AI designs
- Oracle AI checks "AI constitution" for failure modes and prevents them
- Two-level AI: emulation for goals control, Tool AI for problem solving

Timeless decision theory deal

Marginal

- Humans invest in creating AGI, AGI invests in being safe to humans
- AGI is safe to humans because it expects that its own next versions will be safe to him

AGI Enforcement

Marginal

- AGIs to watch other AGIs
  - Multilevel system
  - Community of AGIs
- AI runs simulation with AGI inside
- AGI actions are tested in simulation for unintended consequences
- First AI checks safety of decisions of another AGI

AGI Nanny

May be

- Singleton
- No self-improvement
- Prevent other AIs from becoming too powerful
- Minimum rulesets: prevent death, x-risks and suffering, prohibition of forced brain reprogramming [Turchin]

Coherent Extrapolated Volition

May be

- AI predicts what our values would be if we had more time to think about them (CEV)

AGI Democracy

May be

- AGIs with different architectures suggest different proposals for AGI safety and vote on them [I.Artuhov]

Virtual jury

- Virtual jury based on highly moral uploads checks AI's decisions, goal system and CEV [Yudkowsky]

Integrated Solutions

Military Nanny AI

Marginal

Wins war for world domination, then become Nanny AI

Multilevel Safe AGI

Promising

1) Unchangeable core with constitution and values
2) EM as a cloud around this core
3) Tool AIs as extension of EM
4) External confinement and control
5) Self-improvement is limited
6) Test its actions in virtual reality
7) Virtual jury for action judgment
8) Works as Nanny AI
9) Use it to predict failure modes of its own constitution

AI Testing

May be

- Test AI inside multilevel simulation: if it breaks 1st level, terminate.
- Test AI with difficult moral choices
- Test AI decisions inside simulations for inundated consequences
- Give AI idea that its decisions may be tested by unknown method
- Create parallel AI which only test decisions of first AI, but is build on different principles (may be it is human upload or virtual jury)
- AI self-improve in different branches, which later test each other for core values consistency
- AI present full model of the future which would result from its decisions, and human operators (or uploads) validate them

Meta

General requirements to the Safe AI theory

- It must be easy to understand and implement so all researchers will prefer it
- It must help AI to win first and to safely establish its power into Singleton
- It should be computationally simple
- It should have good chance to be really safe
- Choose the ways to the safe AGI which are not mutually exclusive

Other useful ideas

- Value of human values
- Secret way of testing right behaviour will result in obeying spirit of law [Armstrong]
- Break AI into modules [Armstrong]
- Prevent orthogonality of means and values (like in humans) [Phil Torres]
- Approval-directed agents [Christiano]

International AI agency

- Choose the best AI safety theory and recommend its implementation

Created by Alexey Turchin based on the article "Responses to Catastrophic AGI Risk: A Survey" by Kaj Sotala and Roman V. Yampolskiy, 2013
Ideas from the article are inside red border line.
Ideas not derived from the article are marked to their authors in [...]
"No" means that a method definitely will not work (it is point of view of Alexey Turchin and other people may have different opinions)
"May be" means that small chances that a method may work (like 1 per cent)
"Marginal" that a method may be used for short time or as additionally precaution measure but don't change the situation much.
"Promising" means that a method seems to be good enough to invest more in its research.

© 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 , https://www.scribd.com/doc/265853723/AGI-Safety-Solutions
mailto: alexei.turchin@gmail.com
Proofreading and advise: Michael Anissimov