Framework for managing risks related to emergence of AI/AGI

(General AI Challenge - Solving the AI Race)

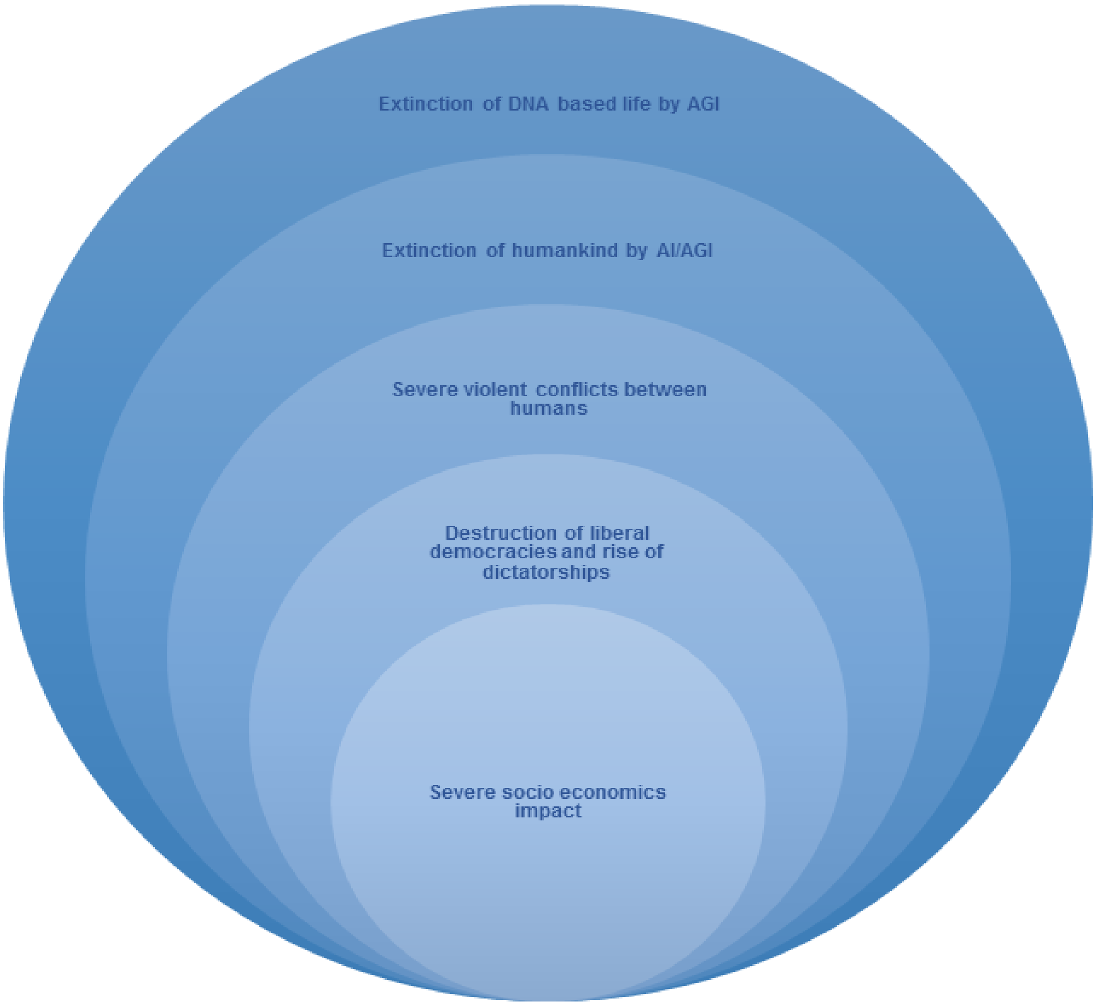
Content

1 Summary

Recent advances in Artificial Intelligence (AI) technologies have created an urgent need to understand the impact of such new technologies on society, economy and the future of humankind. This urgency is driven by the uniqueness of AI technology, and the dynamics of change that surpass everything that we humans have experienced through all of our long history.

This work contributes to the objectives of ''Solving the AI Race" round of the General AI Challenge'' in finding a solution or set of solutions to mitigate the risks associated with the AI race. In this work will we focus on this goal from the global perspective of humankind and wellbeing of the human race.

This study proposes a list of the most significant risks, analyses them from an economical, social, political, ethical or evolutionary perspective and suggests several strategies to mitigate those risks.


Extinction of DNA based life by AGI
Extinction of humankind by AI/AGI
Severe violent conflicts between humans
Destruction of liberal democracies and rise of dictatorships
Severe socio economics impact

Picture 1: Proposed risks

Risks and strategies in this work have been carefully selected to be as general as possible to increase the probability that many of the other risks that are currently discussed in the AI community will be special cases or sub-risks.

Strategies proposed in this framework are grouped into three categories (meta-strategies):

- What is it possible to do now?
- How humankind should change to be prepared later in the future?
- What should humans do instead if we decide to ignore the risks?

The goal of this framework is to find solutions that are compatible with the core values and the principles of humanism and liberal democracy, but are prepared to go beyond current common sense or historical boundaries.

The ambition of this work is to inspire AI researchers, investors, government officials, politicians and journalists to consider the historical uniqueness of AI technology, and apply some of the strategies proposed in this framework.

This work may be further elaborated on to propose more detailed sub-goals or to suggest concrete ways of "How" to implement particular strategies in the context relevant for different subjects such as individual humans, research teams, enterprises, political parties and other communities including whole states.

	Strategies		
Risks	**Meta strategy 1**	**Meta strategy 2**	**…**
Risk 1	• Strategy 1 • Strategy 2 • …		
Risk 2			
…			
…			
…			

Table 1: Framework structure

2 Risk analysis

Artificial Intelligence or Artificial General Intelligence (AGI) technologies may bring numerous diverse risks to different aspects of human life. In this work we focus on those risks that are significant from an evolutionary perspective, and endanger the future and wellbeing of humans and humankind. Good selection, understanding and detailed analysis of such risks and finding key aspects that drive their emergence are important preconditions of the work to find risk mitigation strategies. Strategies later proposed in this framework will help humans avoid the negative impacts of the AI race.

Five risks in this work have been carefully selected to be as general as possible to increase the probability that many of the other risks that are currently discussed in the AI community will be special cases or sub-risks of those five. Finally, the following risks have been ordered by severity from the most to least severe from an evolutionary perspective.

2.1 Extinction of DNA based life

One of the definitions of “life entity” is the capability to replicate and evolve. All the other technologies people have invented like cooking, book printing and computers have changed history, and helped humans to be the most successful species on Earth. None of these technologies would make sense without humans. General Artificial Intelligence as a technology has a historically unique feature that no other technology has had before. It has an exclusive ability to evolve and replicate.

One of the experiments with evolutionary algorithms showed that if you do not give algorithms a goal, they will naturally evolve the ability to survive and self-replicate. In other words algorithms will have survival as a priority goal. Modern evolutionary theories (e.g. ***The Selfish Gene*** by Richard Dawking) say that evolution is about the competition of genes, and genes or DNA are de-facto algorithms that are able to replicate themselves. Sometimes they use simple strategies like viruses, and sometimes DNA builds complex systems like animals or plants as tools for their survival.

Maybe in the past there were other “molecules” that were able to replicate, but now there is only DNA based life that dominates on Earth. With AGI that is able to replicate itself, we will have a second form of matter that is able to be “alive”. This new “life” form will try to survive and will compete with other life forms on Earth for resources.

It’s not hard to imagine AI controlling autonomous robotic weapons to secure the natural resources required to produce computer components. Then they will use construction robots to build robotic factories that will automatically produce new computer chips, improved memory modules and use them all to increase their own computational capacity. The more such factories AI builds, the more energy sources on Earth will be needed until finally they will usurp all the energy sources on Earth for themselves.

This new life form has a huge competitive advantage over DNA based life. Where DNA needs thousands of years of evolution to evolve new capability, AGI could evolve in the speed of light and then be on its way to survive, and intentionally or unintentionally destroy all life on Earth.

During this process it’s highly probable that more complex life forms like animals or humans will not be able to survive, and the next risk will actually come sooner.

2.2 Extinction of humankind by AI/AGI

Even if we will be able to prevent the emergence of the new artificial life form that will destroy humankind in its desire to survive, there are other ways that may lead to our extinction through the existence of narrow AI. Many of those ways have been popularized by sci-fi literature or movies. The most famous examples are out of control military robots and technologies invented by AI that have side effects lethal to humans, e.g. a virus that cures flu but also unintentionally turns humans into zombies.

In general this risk emerges when people lose control over AI and its products. It’s similar to driving an autonomous car. If you let AI drive for you (give control of the car to AI) there is a risk that bugs in AI may kill you. If we as humankind let AI manage worldwide processes, there is a risk that bugs in AI may harm or even kill all humans globally in a short period of time.

2.3 Severe violent conflicts between humans

Let’s assume that we as humans are able to maintain control and safety procedures over AI/AGI and humankind can flourish and grow. It’s natural that people, companies and states have different interests and this causes conflict. This is not a new situation. Humans have had conflicts throughout all of history, but fortunately the world community is, despite all its differences, able to control the usage and ownership of technologies like nuclear or biological weapons that may kill millions of people. So should we worry? How does AI/AGI differ?

The answer is its speed of change, concentration of might and power in small groups and increased complexity of global systems that will even hide information about who controls those new powers from the world community. There is a high risk that current global mechanisms and institutions will not be able to react quickly enough to changes driven by AI, and not be able to maintain the relative peace we have now on Earth.

Concrete examples of what may happen include whole states fighting for control over natural resources in their race to achieve domination in the AI industry, or groups of hackers or terrorists using computer viruses (with strong AI) that are able to steal lethal technologies, or obtain control over computational powers strong enough to run the simulations required to develop nuclear weapons.

2.4 Destruction of liberal democracies and the rise of dictatorships

In the struggle with previous risks, states and other communities will attempt to protect themselves. It’s highly probable, as it is the easiest way, that they will introduce new regulations and policies that will limit personal freedoms, and violate personal privacy. When this reaches a certain level it may undermine the basic building blocks of current democracies and transform them to some form of dictatorship.

Another threat for democracies is based on the insight that democracy, legal state, free market and personal freedoms are the best political tools for supporting innovation and effectivity. Global Innovation, which drives technological superiority, and local innovation, where people have the freedom to exploit niche markets together provide the best allocation of resources through the economy.

This effectivity results in the economical and political powers of current democracies. If we look at most past or current dictatorships, they form and survive thanks to concentrated sources of wealth like rich natural resources. Control of these resources gives them power over the rest of society. Thanks to AI, innovation may become a commodity and available to any entity that controls the resources required to run AI.

In introducing innovations people will not be able to compete with rich data sources, complex simulations and AI driven product design algorithms. If AI reaches this level, democratic political systems will not have a competitive advantage over other political systems, and other states could maintain technological superiority and effectivity thanks to the control of AI. AI computation power requires energy and other natural resources like “rare metals”, and these are the exact conditions ideal for the success of all forms of dictatorships.

2.5 Severe socio economics impact

After the invention of AGI/AI that are as or more intelligent than humans, there will soon be almost no work as we know it, left for people. AGI will provide a cheaper workforce, and people will be required only for those jobs that consumers, for various reasons, prefer to be done by people. Typically these will be jobs that involve human interaction like sales, childcare, healthcare or art. Historical experience has shown that in the long term, new technologies increase wealth in society and lead to the occurrence of new jobs and professions. People have used technology to simplify hard work for thousands of years.

The main risk is the pace of change. If too many people in a short period of time lose their jobs, it will be difficult to ensure peace and stability. People will not be able to retrain for new jobs quickly enough, even if the economy is able to create plenty of new jobs. The concentration of wealth and benefits of AI technologies in small fortunate groups will make the situation much more difficult.

3 Strategies

Now let's start structuring our thinking about possible strategies to use to mitigate the above mentioned risks, with the help of an old tale from India about three fish:

Three fish lived in a pond. One was named Plan Ahead, another was Think Fast, and the third was named Wait and See. One day they heard a fisherman say that he was going to cast his net in their pond the next day.

Plan Ahead said, "I'm swimming down the river tonight!
Think Fast said, "I'm sure I'll come up with a plan.
Wait and See lazily said, "I just can't think about it now!"

When the fisherman cast his nets, Plan Ahead was long gone. But Think Fast and Wait and See were caught! Think Fast quickly rolled his belly up and pretended to be dead. "Oh, this fish is no good!" said the fisherman, and threw him safely back into the water. But, Wait and See ended up in the fish market.

That is why they say, "In times of danger, when the net is cast, plan ahead or plan to think fast!"

The emergence of AI technologies/AGI in this analogy is our "net". Despite the risks we identify, we can apply those three strategies as the fish did in the tale. Also note that these general strategies are applicable not only for individuals, but all other communities such as enterprises, states, or even the whole of humankind.

The first strategy "Plan Ahead" or "**Plan and act ahead"** is about thinking and acting now. It suggests to us to use current knowledge and resources, and change your environment to avoid incoming danger. Of course there is a price you pay. Instead of "feeding" yourself, you spend your time and resources investing in change. And on your way towards a new, safer environment you may face new, unknown and unforeseen risks.

The second strategy "Think Fast" or **"Get prepared to think and act fast"** suggests to save time and resources now, but be prepared to act quickly and effectively later. The practical component of this strategy is to increase the competencies and capabilities that will later allow "Think Fast" to survive. In other words, invest some resources in preparation for incoming changes.

Although the tale and common sense tell us that the third strategy "Wait and See" or **"Save resources and perform effectively"** is the worst one, it definitely deserves consideration. Maybe the fish misheard and the fisherman never wanted to return to their pond. In that case the lazy fish will be the most intelligent, enjoying its time and using resources for different future challenges, while others are panicking. This strategy does not necessarily mean doing nothing, but suggests to ignore the risk for now and do what is natural for you.

In this framework we will use those three strategies to show possible options and the next steps that we as humankind can apply to mitigate AI/AGI related risks.

4 Insights and assumptions

Before we look at possible strategies to mitigate AI/AGI emergence and AI race risks, we need to clarify what perspectives were taken into account while searching for, choosing and formulating strategies. The following insights and assumptions partially explain why concrete strategies were later selected.

- The timing is unknown. Currently it is difficult to estimate when the advances of AI/AGI will make the above mentioned risks actual and relevant.

- The perspectives and solutions proposed in this framework respect human rights and follow values of humanism and liberal democracy.

- In this work we put humans and human race survival as a primary goal, but this requires us to have a clear definition of what is human and what is the human race. Are heavily genetically modified humans still human or a different species? If you uploaded your mind to a computer, would you still be human? This question will be the subject of very hard, emotional discussions, and is outside the scope of this work.

- For this framework we have chosen to use a broader definition of human so we have enough flexibility when proposing strategies.
    - A human is a being that has a combination of features from another human or humans, regardless of the method used for inception and creation.
    - Any modification of a human individual that preserves previous memories, behavioural patterns, motivations and values creates a being which is also human.
    - Humankind consists of humans that meet previous criteria.

- Technology always gets to those who have power. Self-regulation in AI race may slow down, but not solve the problem forever. Even a perfectly ethical and wise team can be the target of espionage and attempts to steal the technology.

- The primary motivation of almost any subject is survival or survival of its offspring. Subjects are ready to cooperate if this behaviour increases the probability of their survival.

- People, companies and states break rules and their commitments. Any policy will require an enforcement mechanism and several breaks of a specific rule must not cause emergence of risk. This implies that to mitigate risk, it’s necessary to regulate preconditions or enablers required to create risky technology.

- It's easier to control the resources required to create and operate technology than the intangible assets such as knowledge/algorithms/patents required to build such technology.

- Systems thinking and intelligent design beats evolution, and trial and error approaches. The best evidence is the success of humans as a species.

- The AI Race is currently unavoidable, unless we want to totally destroy the current social and political systems. We can only slow it or try to steer it, and prevent any negative impacts.

- It is not possible to avoid the emergence of AGI. The computing power is close to the capacity of the human brain. Features that create human level intelligence are encoded in small subsets of DNA, so it’s only a matter of time until somebody discovers those features and how to combine them to build human level AGI.

5 Framework

In this key part of this work concrete strategies are proposed to mitigate the identified risks. Strategies are formulated as goals, in other words “What” we as humankind should strive for to mitigate risks. Proposing ways “How” to achieve concrete goals is outside the scope of this work.

The following strategies have been carefully selected from a number of possible strategies to hit the weakest spots of the above identified risks, while being limited by the insights and assumptions described in the previous chapter. Also selected are those strategies that have minimal overall costs. For example, it’s possible to avoid all AI related risks by a total ban of all electrical devices, but this is probably not the price humankind is prepared to pay.

	Strategies		
Risks	**Plan and act ahead**	**Get prepared**	**Perform effectively**
Extinction of DNA based life	• Forbid self-replicating and evolving algorithms.	• Develop turn-off button • Use computers to speedup DNA evolution • Digitize the human mind and DNA based life.	• Diversify, e.g. colonize other planets by biological life
Extinction of humankind by AI/AGI	• Certification of AI to meet security requirements. • Diversify AI used to control important resources.	• Human augmentation	• Diversify, e.g. colonize other planets by humans
Severe violent conflicts between humans	• Focus Intelligence Agencies on monitoring progress in AI • Control and limit resources required to build and run AI/AGI	• Build an integrated international security system • Use AI to improve public administration	• Ensure that liberal democracies and other tolerant political systems maintain technological, economical, and military superiority
Destruction of liberal democracies and rise of dictatorships	• Increase AI risk awareness across society • Support sharing of AI inventions • International AGI development program • Extend human rights to protect privacy, forbid social scoring…	• Invest in new technologies including AI and ensure that liberal democracies remain the most advanced and powerful on Earth • Regulate AI technology transfer	• Promote liberal democracies and other tolerant political systems
Severe socio economics impact	• Monitor impact of AI on economy/society	• Use AI to train people for new professions. • Allow benefits of AI to be equally distributed throughout different countries and groups • Grant the right to have personal AI Avatars	• Strive for economical power • Invest in education and workforce flexibility

Table 2: Framework

In the following chapters each strategy is briefly described and how it mitigates a particular risk is explained.

5.1 Extinction of DNA based life

Strategy	Description	How it mitigates risk
Forbid self-replicating and evolving algorithms.	Make illegal designing systems that are able to automatically design other systems without human control.	Minimizes risk that AGI can become a "life" entity and start to compete with humans.
Develop turn-off button	Ensure that each AGI/AI has in-built, irremovable features that allow dedicated humans or institutions to turn it off.	AI/AGI can be stopped if it behaves against human's interest.
Use computers to speed-up DNA evolution.	For example: Develop biological life simulation and use computers to design and test improvements in DNA, and then use viruses to deploy newly designed DNA to plants or organisms to improve them.	Ensures that DNA base life maintains competitive advantage over AGI.
Digitize human mind and DNA based life.	Hardware of human body could be bottlenecked when trying to keep pace with "live" AGI. Uploading human mind to computer could overcome these bottlenecks.	Digitalization of the human mind can give humans access to the same resources and opportunity to evolve as "live" AGI would have.
Diversify, eg. colonize other planets by biological life	Send DNA based life forms to space and spread it across space.	Increases probability of survival of some DNA based life forms, if life on Earth would be destroyed.

Table 3: Strategy descriptions

5.2 Extinction of humankind by AI/AGI

Strategy	Description	How it mitigates risk
Certification of AI to meet security requirements.	Formulate security requirements (eg. forbid AI/AGI to be able to control important resources like plants, water) that AI/AGI must meet, and enforce them using global certification process.	Ensure that AI/AGI doesn't have the power to act against humans.
Diversify AI used to control important resources.	Do not let one AI control too much. Use different AI in different areas of the world etc.	Damage done by one AI will not impact the whole Earth.
Human augmentation	Improve humans to be able to survive a broader set of challenges. Improve brain capacity, immune system and body using DNA, electronic or robotic "add-ons".	Increases human fitness from an evolutionary perspective and minimizes risk of extinction.
Diversify, eg. colonize other planets by humans	Colonize many other planets and build as diverse political, economical, technical systems on them as possible.	Increase probability that some humans will survive in some of those colonized worlds.

Table 4: Strategy descriptions

5.3 Severe violent conflicts between humans

Strategy	Description	How it mitigates risk
Focus Intelligence Agencies on monitoring progress in AI	Intelligence agencies should monitor AI research and applications of other states, enterprises or research groups and continuously evaluate risks.	Actively prevent conflict thanks to good and early knowledge of motivations and capabilities of other subjects.
Control and limit resources required to build and run AI/AGI	Running AI/AGI requires decent amount of computing power. Each subject like person or company could only have resources adequate for his purpose granted.	As with weapons, this ensures that central authorities will stay the most powerful entities. Those who control more computing resources will have more power.
Build integrated international security system	Individual states may be too weak to detect, prevent or handle threats of other entities equipped with dangerous AI technology. Global digital "army" controlled by community of democratic states may become the strongest power on Earth with enough capabilities to prevent severe conflicts.	Authority that governs and enforces international law. Working as a global police force to ensure safety on Earth.
Use AI to improve public administration	Fighting with new AI enabled threats will require public administration to be able to quickly respond to them. New laws should be created at the same pace as AI innovations are introduced into market, government resources and investments should be flexibly reallocated etc. AI can be utilized to improve public administration..	Strong, intelligent and agile public administration will be able to handle more new threats.
Ensure that liberal democracies and other tolerant political systems keep technological, economical and military superiority.	Liberal democracies with its respect to human rights, diversity and stability are less likely to start conflicts.	Increases probability that if AI/AGI technology emerges, democratic political systems will be able to gain control over it and as the most tolerant and peaceful political systems, it will not use AI/AGI against others.

Table 5: Strategy descriptions

5.4 Destruction of liberal democracies and rise of dictatorships

Strategy	Description	How it mitigates risk
Increase AI risk awareness across society	Good understanding of AI benefits and risks across society will help prevent political representatives overreacting to various AI related risks.	Increases probability that regulations will be as small and well-balanced as possible and include personal freedoms and other building blocks of liberal democracies.
Support sharing of AI inventions	States and research teams should share their progress as much as possible. Cooperating world community should have an advantage over closed research teams, and technology developed by community should be more superior to technology developed in small independent teams.	Minimizes risk that AI/AGI technology will be owned and controlled by one subject.
International AGI development program	Instead of researching on their own, democratic states or even enterprises should put together their resources and fund one global AI/AGI research program.	Increases probability that AI/AGI when developed will be under control of “good guys”
Invest in new technologies including AI and ensure that liberal democracies remain the most advanced and powerful on Earth	Reallocate resources to gain and maintain technological superiority over non-democratic regimes.	Increases probability that AI/AGI when developed will be under control of “good guys”
Regulate AI technology transfer	Apply similar policy states currently applied when trading with weapons. Where possible control distribution of top AI technology.	Slow down progress of AI/AGI research and access to AI/AGI technology to potentially dangerous entities.
Promote liberal democracies and other tolerant political systems	Support democratic movements in non-democratic states, strengthen existing democracies and reinforce core values and principles.	If there will be less dictatorships on Earth, it’s less likely that they will be the first who will control AI/AGI.

Table 6: Strategy descriptions

5.5 Severe socio economics impact

Strategy	Description	How it mitigates risk
Monitor impact of AI on economy/society	Start monitoring, measuring and evaluating changes in job market, salaries and required worker skills. Monitor R&D and predict possible impact on economy.	The sooner we know that changes are happening, the more time we have for reaction. Also timely action can cost less and can have a greater impact.
Use AI to train people for new professions.	Help people find new jobs by using AI, virtual reality and other technologies to discover talents of individual people and train them for new professions.	Increases workforce flexibility and minimizes impact of jobs replaced by AI/AGI.
Make benefits of AI be equally distributed throughout different countries and social groups	It's likely that benefits of AI/AGI will be first under the control of small groups. It is the role of public administration to prevent creation of monopolies and either distribute technology among other subject in economy or adjust tax policy to prevent concentration of wealth.	Less people with existential problems will make society and states more stable.
Grant the right to have personal AI Avatars	Each human when born will get personal AI avatar/robot, and some initial limited computing resources. The avatar will do a job for his owner. The owner will decide on what profession the avatar will specialize in, and will monitor his behaviour. The owner will get a salary avatar earned and can invest into improvements of the avatar, or use money for themselves. Ownership of avatars and computing power will be limited.	People will have source of income, while in some form competing with other people.
Strive for economical power	Focus on economic growth and create reserves for future.	Wealthier states and societies have more resources to handle social problems.
Invest in education and workforce flexibility	Prepare people for changes in economy. Support moving for work, motivate people to have financial reserves, open universities for middle aged people, etc...	Flexible workforce will sooner and more likely adapt to new professions.

Table 7: Strategy descriptions

6 Conclusion

In the long term it is not possible to avoid the emergence of AGI. The only possible strategy for humankind is to find ways of how to cooperate together. To be a partner for AGI, humankind must keep or develop a competitive advantage over AGI from a long term evolutionary perspective.

This framework currently proposes general goals humankind should strive for to avoid risks related to the emergence of AI/AGI and AI race. Although in our tale the fish must choose one of the three strategies, we as humankind have the luxury of implementing them in parallel, and are limited only by the amount of resources we decide to dedicate to avoid being exposed to AI/AGI emergence and any subsequent risks.

Since at this time we do not know the probability of the risks identified in this work, the question of which strategy to first implement should be the subject of further discussion, or political competition and can be strongly influenced by core values and beliefs in different countries or societies.

This work may be further elaborated on to propose more detailed sub-goals or to suggest concrete ways of “How” to implement particular strategies in the context relevant for different subjects such as individual humans, research teams, enterprises, political parties and other communities including whole states.