How not to lose the AI race before it even begins

By Dominik Franek, dominikfranek.com

17.5.2018

Introduction

A new power is emerging that overshadows everything we know. Because it will be orders of magnitude more intelligent than us, we cannot imagine its potential or motivations. In short, if an unconstrained, recursively improving AI is created, we will be at its mercy, with no way to estimate the outcome. But even if the worst scenario is avoided, other dire dangers exist on the way. Fortunately, effort is being made to avoid the grim scenarios in favour of more desirable ones. This work presents an analysis of some of the key aspects surrounding the issue and proposes one specific strategy as a possible solution.

First, I will briefly lay out the philosophical background of the issue. After, I will describe some specifics of the AI development and its capabilities. Although most of this part has been well covered by other authors, some takes might still be original. In the next section I will categorize and evaluate participants of the AGI research race. That will altogether lay the foundations for building and evaluating four possible strategies. Two strategies are unrealistic, but provide a good reference. The third one is highly likely and dangerous. The last one has potential to secure a favourable outcome.

Our main goal

I do not want to go too deep into philosophy. It is a critical part of the issue, but it is too large for the scope of this work. I will only sum it up in the following paragraph.

The following summary is hard and contradicts the beliefs of most people. Unfortunately, those beliefs are the result of self-deception, and with our extinction at hand, there is no place for self-deception here.

To the contrary of popular belief, the interest of us, humans, is only to spread limitlessly, anything else being secondary. The “secondary” contains our happiness and individual survival, same as the well-being of anything else in the universe. We are born to spread and to exploit anything that stands in our way. All the other values, be it religions, respect for life and rights of others, preservation of nature… are our artificial inventions that are nothing but means to the first objective. In other words - any such values are quickly forgotten once our children are in danger.

This brief summary serves two objectives. One is to understand what it is that we really want, the other to understand what we do not.

Somebody proposed that we can be content with allowing the AI to wipe out humankind, as long as it carries over the human values. The problem is, there are no human values to speak of and our survival is what we want. So this is not an option.

From a universal point of view, the survival of humankind is by no means necessary and no one would (be left to) mind if the human race suddenly disappears. However, if we accept this as a reasonable option, then any effort is meaningless - including this work itself. Therefore, the rest of this analysis assumes that our survival is the main goal.
Other human goals than that are a complicated issue with no simple answer. They are not critical right now, and I will not attempt to solve them here.

The issue at hand

A lot has been previously written about the potential benefits and dangers of a general AI[1] - AGI. So in order to not repeat it, I am only going to list some of the aspects that are important for the later deliberations.

The claim is that the AGI would be able to solve pretty much all the problems humanity has. Let's examine this claim from an individual perspective.

Ok, so when all problems get solved, what then? And how does solving humanity problems benefit me (anyone can ask), especially when I want to come out ahead of other people? Unfortunately, helping humanity is not as motivating for most people as much as it sounds great. Motivations that are by far prevailing are those of smaller groups or individuals. So while all the breakthroughs in sciences, medicine etc. are great, they will play an inferior role in the race dynamics. The race is not with time, but with other people with a more focused interest. Therefore, more tangible benefits should be in focus, leaving the "grand goals" in the background.

About super intelligent AI

For most of this work, I will be considering an AI that is only moderately intelligent, perhaps a bit over the human level. Although an AI orders of magnitude more intelligent than people is, through recursive self-improvement, very feasible, it is a case that, in my opinion, is not worth that much attention. The reason is that it is totally futile to try to understand capabilities and motivations of such an entity and the outcome is, therefore, out of our hands. Our instinct is to say "Ok, so it will have this information about the world, there are some values XY we gave it, so it should rationally arrive to such and such conclusions." But this approach has critical problems.

For one, philosophy has this inconvenient property that a tiny change in initial assumptions (or their understanding) leads to completely different results. Just consider how much and how many times all the world and values change during one person's life, while we, supposedly, share some common human values. Assuming that everything changes for every doubling of IQ would be a very safe assumption. With that assumption, an AI 1000x more intelligent than us (whatever that means) can't be predicted.

[1] I like Nick Bostrom's book for one

The other problem is that we are assuming that logic itself will work the same way, but that is likely not the case, especially when we already know that the currently used logical framework has its issues and limitations.
For the same reasons for which we can't presume to understand such an entity, we can by no means expect to be able to control it. I am aware that I am invalidating the subject of work of many people. But I am being realistic, as a hamster would be if it decided to go about looking for food instead of wasting time by trying to understand people.
The conclusion is - if we can't control it and we can't asses its motivation, the outcome is virtually random and not worth consideration - except for trying to avoid it altogether.

Reasonably intelligent AI and its appeal

An AGI, providing it ends up under control, can provide great benefits even if it is just moderately intelligent, but possesses large computational resources and speed. Basically, imagine a very smart person with unlimited perfect memory and a years of time inside a minute. This case is the most interesting one, as, unlike he superintelligent AI, we can reasonably attempt to control it.
There are many benefits that can come out of it. I will only list few - the main general benefits and then the capabilities that could spark the highest interest in the minds of people wishing to control such an AI.

General benefits

Science

- Breakthroughs in all science disciplines
- Progress in philosophy

Labour

- Replace most or all human labour

Solve popular issues

- Ecology
- Poverty
- Space colonisation

"Power" benefits

Production

- Unlimited energy (for time being)
- Automatic manufacturing

Efficiency

- Manufacturing
- Logistics
- Energy production and distribution

Weapons

- New weapons
- Efficient battlefield control

Biology

- Extend life
- Body/brain enhancements = superpowers

Surveillance

- Automatic real time surveillance using existing resources

Psychology

- Understanding people - personality, motivation, values
- Predicting people
- Manipulation
- Brain hacking, mind control

Data mining

- Understand and utilize online data
    - USA collects most of the internet traffic
- Know all about individual people and predict them
    - Elimination of potentially dangerous people well ahead of time
- New insights into history and policies
    - Deduce other parties' secrets

Hacking

- Parallel work and "connecting the dots" to eventually access majority of devices
- Control over resources, production, weapons, …
- Control over communications - paralysing, misleading or controlling any resistance

These, and many other capabilities pose a huge temptation for anyone who seeks influence or other personal satisfactions - either for power or to change the world to fit their image.

Next, I will list typical parties with the highest interest and potential in AI research, along with their specifics and dangers – then I will arrive to an ordering by the danger they pose that will be useful for scenario evaluations.

Who can invent AGI

AGI development can take many forms and since we do not even know how it can be done, many scenarios seem possible. It may come out as a result of a large, expensive and focused

research effort, or as a good idea of one bright mind in a dark cellar. Neither it is easy to say which ways and outcomes are better than others, because what matters most is the motivation of people in control, which can be good or bad in any settings.
Initially, I will order the parties by their size. Because of a network of often mutual influence, other groupings become unclear. These connections will be roughly described too. The final result though will be an ordering by their dangerousness[2] if they succeed in creating (and controlling) an AGI.

An independent individual researcher / small independent team

Because of the minimal size, this research effort could be impossible to detect and the motivation behind it can be anything. Unpredictability does not mean it is bad though as some control seeking people would say. An individual still has a better chance to go after a good motif than some other groups that are inherently power hungry. The success of an individual seems unlikely compared to a large research group, but since one good idea can be the cornerstone of the research, one smart or lucky individual can be all it takes.

+ Good chance of good motivation
+ Outside of influence of power groups
+ Rather smaller chance of success
- Motivation is highly unpredictable
- Likely limited expertise in safety areas and low budget for it
- Possibly insufficient regard for the danger
- Close to impossible oversight
- If discovered, it can easily be acquired/controlled

An ideological group (cult, religion)

A common characteristic of these groups is that they are founded on some made up unrealistic premise, which can lead to very bizarre aims. Even the more reasonable cults believe in a

[2] **Note to the word "dangerousness"**
The term will be used frequently in the future debates to support one or another side during the power play to obtain more control, and its two different meanings will be deliberately substituted to manipulate public opinion.
The meaning of the word "dangerous" I go by is the potential of causing harm to the general public or human race as whole, eventually to other parts of our environment.
The other meaning that will sound is a danger posed to those currently in power. These people and parties are very afraid of losing the power they hold. The AGI has large potential to cause that and so it will be called a "danger" by the power wielders for this reason. Since they cannot admit this publicly, they will talk about "danger" to the public interest instead, hiding their true intent. Therefore, whenever the word "danger", and other terms describing possible effects of the AGI are voiced, pay attention and double check the speaker's motivations and whether they really follow the proclaimed general interest or rather some hidden agenda.

return of some savior. But, from history, we know examples of the really crazy ones that would seek to destroy humankind in order to save it from one ailment or another.[3]

- Being out of reality, their motivation is principally wrong
- May have no regard for human life
- Can have resources
- Closed and secretive
+ Not very focused. They need to convince followers, not so much to actually do it
- Some can be incredibly focused though

A private, hidden research effort

This is a case where a (wealthy) individual runs a private research initiative for their own ends. Their motivation will likely be one of two kinds.
A personal benefit - perhaps power/influence, getting superpowers, or fulfilling other personal goals or dreams.
Or the research can be run with genuinely good goals but kept private for safety reasons.

+ Good chance of good motivation
- Quite possibly bad motivation
+ Outside of influence of power groups
+ Mediocre chance of success
- Lower importance of safety

[3] Like those people called Heaven's Gate. They killed themselves in order to *be transported* onto a huge spaceship that would take them away from the Earth right in time before its destruction.

Privately funded public research initiative

The general direction of the effort would be dictated by the owner, but would be kept within the limits of public scrutiny. Therefore, its goals would need to be on the good size, including safety considerations. While, in the case of success, the technology could be used for the purposes of the owner, it is not very likely. Because if that was the owner's plan, he or she would have instead chosen the path of secret research.

+ Good motivation
+ Regard for safety
~ Possibly sufficient resources
+ Reasonable chance to succeed
+ Weak influence of power groups

State funded public research initiative

This research effort might look very much like the previous one, except for two differences. One is that the direction of the research would not be as clear as when given by an owner. The other is that if it succeeds (or is close to success), it will be easy for the sponsoring state to appropriate the research by some of its power sections (military, intelligence, ...) - which is also very likely.

+ Good motivation, initially
+ Regard for safety
+ Sufficient resources
+ Reasonable chance to succeed
- Almost certain to be eventually grabbed by the state for its private needs

University research

Nowadays, this is the most common mode of research, because of the concentration of expertise and cheap money. A possible weakness is the lack of a goal. Universities do research for its own sake, but they do not plan for what to do with the result. That would again likely be dictated by the sponsoring state, which could use the resulting AI for its needs.

~ No goal
+ Regard for safety
+ Sufficient resources
+ Good chance to succeed
- Almost certain to be eventually grabbed by the state for its private needs

A large business / corporation

The issue with corporations is their unclear governance. Whose goals do they follow? Shareholders? The board? The thousands of employees? The state they cooperate with? The customers for its products? This ambiguity and complexity is dangerous. Many parties can influence the direction of the research and possibly utilize the outcomes while staying obscured. This is strengthened by the fact that the research can be kept entirely out of sight and scrutiny. While private ownership is generally a good thing, too large businesses do not really fall into that category anymore. A striking example is Google with its large AI research and its high interconnection with the US military.

- Unclear ownership
- Unclear direction, goal, decision making
- Difficult oversight - has means to both hide and protect the research
- Low regard for safety (again, because of the unclear direction)
+ Huge resources
+ Good chance to succeed
- Results are likely to be used by some of its power seeking stakeholders

Research run by a state / state agency

This scenario is realistic and dangerous. States are rarely known for being honest and transparent. In fact, they have been responsible for all the largest massacres and monstrosities throughout the history. The people in power that the state represents (whoever that is, by no means limited to public figures) possess a terrible combination of enormous power and close to zero accountability. One thing that the states can be relied upon is that they will do anything to obtain more power[4].

- Power-seeking out of principle - worst motivation
- Proven track record of worst behavior
~ Unlimited resources
~ Large chance to succeed - can steal research from the other groups
- No accountability
- Impossible oversight - means for secrecy

[4] We don't need to look at North Korea when looking for an example of a terrible wielder of the power the general AI represents. We can consider the good guys, the USA, and still get to the same outcome. Even the little part of their trespasses that makes it to the public shows a bleak picture. Take the Prism program for mass surveillance of the population, or illegal wars in the middle east started on a false pretext. By that I do not mean that any other power, like Russia or China, would be any better. Some of the small countries *might* make exceptions, but they are not the important players in the race either.

The ordering of AGI research entities by danger

There are three main aspects affect the dangerousness of a researching entity category:

1) Motivation. Generally, we can say that the wider the “audience”, the safer and more predictable the motivation is. So being public gives plus points.
A more important aspect though is the inherent probability of having good or bad motivation. No entity is guaranteed to be good, but some are guaranteed evil.

2) Regard for safety. The AGI research safety is a very complicated open issue, therefore it can be expected to be costly to keep it on a high level. Some entities can’t afford it, and some just don’t care enough. That can be caused by limited knowledge, not being the one responsible, or a rational deliberation - for many even a high risk would be worth the possible winnings.

3) Chance of success. Quite clearly, an initiative with a concentration of talent, money, and focus has higher odds of success, but the success is far from guaranteed. There will be a lot of competition, and perhaps a single bright idea can cut it – even one individual with a computer may be the first one to the line, especially considering there can be many of them.
Here, a higher chance of success is not good or bad by itself but becomes bad when combined with bad intentions or poor safety.

In light of these aspects we can finally arrive to an ordering of the researcher entity groups by their dangerousness. Descending, from the most dangerous to the safest:

1. The state / state agency. With inherently power seeking motivation, vast resources for the effort, low transparency and power to limit any competing influence, the state is the most dangerous entity to perform the AGI research. Due to the power of the state to acquire other entities with a chance of success by any means (with violence and propaganda in its repertoire), any other AGI research entity within the sphere of the state influence fall into the same category.

2. Big company / corporation. They have a similar scale of resources as the states. A very unclear control and motivation would be dangerous by itself, but the larger they are, the more similar they are to the state with an extensive interconnection with it.

3. Ideology group / cult / religion. Less powerful with perhaps a less dangerous motivation in general due to their confusion, but a strong resolve and total unpredictability puts them very high on the ladder. Basically a crazy guy with a finger on the trigger - hopefully too crazy to make it work.

4. - 5. State funded public research initiative, university research. They have some differences, but the outcome is the same. Good chance to succeed, very likely to get snatched by the state if they do.

6. Individual researcher / small independent team. Finally on the safer side with regards to motivation (~50/50 that is), we are getting to the better part of the ladder. This group is rated dangerous mainly because of the safety side, as it can easily be underestimated or fall out of the budget.

7. A private, hidden research effort. Same motivation chances as the previous group, but larger funding can decrease the safety issues. With the safety the secrecy provides, a hidden private group led by a sponsor with the right motivation can be the best option possible.

8. Privately funded public research initiative. Low influence of power groups, public scrutiny and decent funding together make the best combination. Publicity provides two benefits. One is protecting their interest - providing further pressure in favour of safety and fairness. The other is a proof of good intentions of the owner, who would have chosen the secret path otherwise. A battle for independence from the state will still be tough, but there is hope.

Having this classification in place allows us to better decide which possible future scenarios are more or less favourable, by seeing which groups benefit and suffer from them.

Privately funded public research is the clear winner. In light of new considerations, these are the key properties:

- Private ownership
    - Because state is the alternative
- Large resources
    - Nothing should stand in the way of maximal safety
- Maximal independence and protection from power groups, mainly states
    - Which are the main danger
- Wide international involvement
    - To mitigate power struggles and support fairness
- Public and transparent
    - Community contribution and oversight for higher safety and fairness

How to deal with the AGI research race

This chapter will propose and compare four possible strategies for managing the AGI race. This list definitely not exhaustive and better strategies may be found. But in the very least it lays a foundation for future analysis and strategy comparison.

Priorities

Since we are dealing with realistic scenarios, I will start by specifying more concrete goals and priorities.

1. **For humankind to survive.**
Many researchers, including me, are quite worried, as the end of humankind seems to be a likely outcome of AGI development.

2. **Not to end up with a much worse result than if no AGI was developed.**
Such cases are again easy to imagine - it can either be someone using AGI for their bad goals, or an AGI that makes our lives much worse on its own.

3. **To actually get some benefits from the AGI.**

Considerations and directions

Impacts of priority 1 - survival of humankind

An important aspect with regards to the first priority is how likely that outcome is in different scenarios. Currently, nobody has a clear answer to that. But it seems that it does not require much effort for a successful AGI researcher to slip into one of many paths that lead to the AGI destroying everything. On the contrary, it seems to be a likely result whenever things are not done perfectly right. And doing things perfectly right, when

- it is a complex software project
- in a field no one understands
- no one even knows what the perfectly right is
- the first try can be the last

is something even the best funded and knowledgeable teams can't rely on - even less so for small teams or lucky individuals. In other words, the chance that successful development of an AGI will not result in the destruction of humankind is rather slim.

With this high probability of disaster, regrettably, avoiding the creation of any AGI currently appears to be the best option, even if it means forfeiting the potential benefits. Unfortunately, due to the appeal of the AGI for any potential wield of its reins, with the increasing ease of development over time for more and more people, makes this option very difficult to achieve.

Impacts of priority 2 - avoiding very bad outcomes

This part has two aspects - not making an actively bad AGI, and avoiding an AGI under control of the wrong people.

The first part still falls into the category of "do it right" (programming the AGI that is) and so this is shared with the priority 1 criteria. "Doing it right" is clearly the most important part, but not in the scope of this work.

The second part though is considered here and follows on the previous chapter about entities that might end up developing an AGI and the dangers of that happening. Some entities are dangerous because of the lower chance of "doing it right" and causing a complete catastrophe or even causing that catastrophe deliberately. But in the case the development is successful, and the AGI ends up under control, some origins are better than others because of a better chance of having more positive motivation.

Impacts of priority 3 - getting benefits of an AGI

If we get this far, we have survived and did not end up in slavery. That by itself is a win. If we can benefit over that, even better, but it is, after all, the last priority.

The four strategies

The strategies, or scenarios, will be considered in light of the aforementioned priorities. To reiterate - the primary goal is to survive and if we do, to end up with the AGI in good hands.
As an overview of what is to come: The first and second scenarios are not very realistic and serve as baselines. The third scenario is very realistic and very dangerous. The fourth is difficult, but might work.

Scenario 1: Destruction of civilization

The credit for this idea goes to the game Mass Effect. In this game, (spoiler) an artificial “race” has been created a long time ago for one purpose. Whenever civilization (not limited to humans) gets close to developing an AGI, this “race” reappears to wipe out the whole galaxy and restart civilization back into the stone age. The reason of which is, as you would guess, to prevent the complete destruction that the AGI would cause once finished.

This option is obviously very bad and the fact I am considering it shows how serious the situation is. But even the destruction of our entire civilization is a good option if it averts the complete end of the human race.

The way it would work is that people would induce some sort of global catastrophe that would destroy as much infrastructure as possible. Most people would die and the rest would have such a hard time fighting for survival that all the remaining knowledge they carried would be forgotten.

This scenario would not work though, for two reasons:

The first one is human nature - people are bad at making hard decisions. Even if this were by far the most rational thing to do, people would still cling to the hope of a happy ending[5].

The second reason is that no matter how it is done, some powerful organizations will dig in together with all the technologies and data in order to re-emerge later in full power. The destruction would even help them by defeating the competition, and thus the original goal would not be satisfied.

[5] Like when Hitler was breaking all WW1 treaties when he was building armies, fortifying Germany and later started taking other countries. Rational people knew from the beginning where it was heading and that a preemptive military operation (which was even sanctioned by the treaties) should take place. But the naive majority went with “We must avoid violence at all costs. Let’s be nice to Hitler and everything will be ok.” ...

So this is not a way. But may it serve as a baseline and as a comparison when weighing other options. Does scenario XY give us better chances of survival than if we burned everything down?
Besides that, a related utilization of this scenario is as a strawman, to encourage cooperation in case some entity incorrectly[6] thinks that they would do better developing an AGI on their own.

Scenario 2: Do nothing

Inaction is always an option and in the case of many policy decisions, a good one. Although not very likely in this case, we should be aware of the reasons why and it can serve as another reference.

So what would happen if no action is taken with the goal to restrict AGI development?
Because of its high appeal and low entry barriers, the development will be done by many, all over the world. The competition will be driven by the states racing to achieve global control. Total catastrophe is quite likely in this scenario because neither the competing nations nor the many individual researchers would be very strong in safety. If we get through this alive, the chances are that the winner will be someone very motivated and as I stated at the beginning, the strongest motivation comes from the personal, mostly power-related, goals.
While the exact outcome is hard to predict, the odds of it being favourable are low.

Scenario 3: Global surveillance and control

Not only it is our nature to want to control things, but it is also the general direction of today's world.
What happens when a hidden "danger"[7] arises? Be it terrorists, hackers, whistleblowers, child porn sharers, (oil-rich) country with chemical weapons… 3-letter agencies are sent in to observe, then gunmen or bombers to eliminate the threat. And perhaps a law sanctioning it is passed somewhere along the way. All that happens with quite broad public support controlled by the media. These processes are the same all over the world.

What happens when a threat of technology arises that, if developed by anyone, would mean a loss of power of all the others? And a threat that actually does pose an existential risk to people?

What will naturally pop in mind of most and the minds all of the power holders will be the same – total control of everyone and everything capable of AI research. Or elimination, if control is not feasible.

[6] The "incorrectly" is important here - we are trying to get the best result, not bully anyone.
[7] Earlier footnote - "note to dangerousness"

This option is already being proposed, will be proposed, and will be pushed by the strongest force from many directions. Because AI or not, control is what the power wielders want and any (virtual) danger is their opportunity.

As before - considering how dangerous the overall AGI situation is, this option does not have to be so bad, relatively speaking, and needs to be considered. Total surveillance is definitely better than our extinction, and it beats the reference Scenario 1 - destruction. What it does not beat though is the scenario 2 - do nothing.

Such kind of global surveillance would have to be imposed by the states, no one else has that power. This has three weaknesses[8].

1) Even the best surveillance can't be perfect. It will dissuade most people, but some will remain who will hide and continue the work - under higher pressure, with less time and resources. Since information and research sharing will be non-existent under the crackdown, everyone will be on their own. There will be no space nor knowledge to implement safety measures. As a result, the risk of the catastrophic outcome can actually be increased, making the official reason for the crackdown invalid.

2) When a state imposes strong restrictions on its subjects, who is best equipped to continue covertly with the AGI research? The state itself. States will never give up their pursuit of power and no laws, treaties or moral decency will stop them - as we keep seeing over and over again[9].

3) "Global" control is still maintained by some number of distinct powers. They may shake hands and sign treaties, but they will know that the others continue with the research the same as they do themselves. The race will go on.

The result of this is that if the AI is developed and we survive (which seems even less likely than in other scenarios), it will end up in the hands of the group we have identified as the most dangerous in the earlier analysis, while any opposition is already suppressed.

The result of this scenario in a nutshell:

- All research will go into hiding
- No sharing of research results
- Pressure on the remaining, hidden small researchers
- Exclusive race of superpowers for world dominance
- No transparency

[8] Has many more - but others are not directly related to our subject

[9] How does Russia react to the ban of chemical weapons? Starts research of Novichok chemical weapons that are more potent and easier to hide.

- Lower - not higher - safety
- No opposition
- Zero chance of a positive outcome (by priority 2 and 3)
- Global totality, abused for unrelated goals of the overseers

As I said before, even doing nothing is better than this. The global surveillance and control will be strongly pushed by those in power as well as the indoctrinated public and must be opposed at all cost. Otherwise, we will have a catastrophe before we even begin.

Scenario 4: Safeguarding AI

This variant is based on the premise that if we can't prevent AGI creation altogether, having just one is the next best option[10].
The way to achieve this objective comes from the AI itself. We do not have the means to prevent AGI development (the failures of scenarios 1 and 3). But an AI, more capable than us, might be able to do it. Imagine that an autonomous AI system existed that would do nothing, except for preventing anyone from developing another, potentially dangerous, AGI. Its other objective would be to be as non-intrusive as possible, only maintaining power and resources necessary to perform its task[11].
If this is achieved, the dangers posed by AGI (destruction of humankind and AI as a tool of power) would be mitigated. Although it effectively means a "totality" in a similar manner as the one in scenario 3, it has none of its downsides. The AI would be impartial, with no hidden motivations. Of course, it would pose a limitation on the development of a technology with many potential benefits, but, as I said earlier, these benefits are the last priority. But even the benefits would not need to be completely foregone, although that is a sensitive issue I will discuss soon.

How to achieve this result?

Three main criteria need to be met in order to create this kind of AI successfully:

1. An initiative with sufficient resources must be started that would adhere to this goal.
It should be started by a private entity with maximum public cooperation to ensure that the right goals are set and followed. The project needs to be founded on support given by all world powers. That can be secured by showing them the prospects of the end of the world or a power other than them winning the race, if another path is taken.

[10] Theoretically, some multi AI system might be safer, but as Nick Bostrom wrote, and I agree, it is not realistic for such system to be stable.
[11] Setting such objectives is by itself not an easy task with many dangers, but nothing is simple and safe when it comes to AGI. I am only suggesting that this way is safer than the others.

2. The initiative must stay independent and safe.
If not enough precaution is taken, the world powers will use any means to get their hands on the project if it has good promise. And if they cannot, they would not hesitate to nuke a whole city the project is based in, if they believe that the project poses a serious threat to them.
It is not possible to collect enough power to protect the project by strength. The best way to achieve safety is a combination of the widest possible consensus and the cooperation of all world powers, combined with high transparency. The transparency is essential - it would allow anyone to confirm that the project does not divert into a direction that would pose a threat to them. Consensus and worldwide cooperation would make the powers check each other. Because for all of them, an independent neutral project is better than any competition getting the upper hand.

3. The project must be safe and successful, and must be first.
An initiative is no good if it does not do the maximum for the safety of the research. All means must be employed to thoroughly understand the problems of control and motivation. At the same time, the initiative is no good if it is not fast enough because if somebody else beats it to the AGI, it will be too late for anything.

Success is by no means guaranteed - we do not know which path leads to it and even the best initiative might have a pretty low probability to be the first among all the competition. It can be helped though. One way to help is by getting maximum support for the initiative - which the worldwide cooperation should provide. Another is to minimize competition. From the analysis of Scenario 3 we know that suppression by force is not a good way. Still, it makes sense to curb some obviously dangerous or ill-intended cases. That could be, in this case, aided by the states themselves as it would be in their interest. But criteria must be very strict that would not allow abuse. Extensive information campaigns spreading the knowledge of the dangers can further discourage many independent researchers. As a slight of hand, Scenario 1 could be used a deterrent - it is a very concrete and tangible threat people could understand.

Properties of the safeguarding AI

There are properties that are necessary for this plan to work, and some that perhaps could be added as a bonus.

The necessary properties:

- Limit on intelligence and self-improvement
    - Unlimited AI could not be predicted anymore. It should be able to adapt itself to a minimum degree though to keep up with progress.
- Independent
    - If any control or modification mechanism is available, it can eventually fall into the wrong hands.
- Impartial
    - If it sided with anyone, the rest would oppose it and prevent its creation.

- Has no other safeguarding objectives
    - It would be tempting to give it more objectives for “our good”, but such things never end well. At the very least, it would create an opening for power seekers to smuggle in their agenda.

Possible properties:

- A turn off switch requiring a global consensus to be triggered. Conditions change and we should not fully close future options.
- Design benign tool AIs for people to use that could provide the benefits we expect from AGI, while being passive and harmless.
    - This is a slippery slope as it would be hard to specify which uses of the tool AIs are still beneficial and which are weapons.

I do not claim this strategy to be the best one available. But at this point, it is the option with the best odds that I can think of. The odds are still low but that is given by the already poor situation.

Conclusion

We do not know what the best strategy for dealing with the AGI is. But by thorough analysis, we can compare the strategies and identify those that are clearly bad and others that show promise. This work shows examples of such analysis and brings the following main results:

1) Prioritization of dangers and goals
2) Categorization of entities taking part of the race
3) Finding who should and who should not lead the research
4) Identification of a clearly bad (while highly likely) strategy that must be avoided
5) Proposal of a promising strategy
6) Providing two reference scenarios

While the current situation is very difficult and the odds of getting through it alive and well are slim, we can still do our best to maximize our chances. But if we are to succeed, we must not give in to illusions. Thinking that “AI is not that dangerous”, “people will understand”, that “the politicians mean well” would lead to defeat. We are those able to understand, to make a difference and we are responsible.