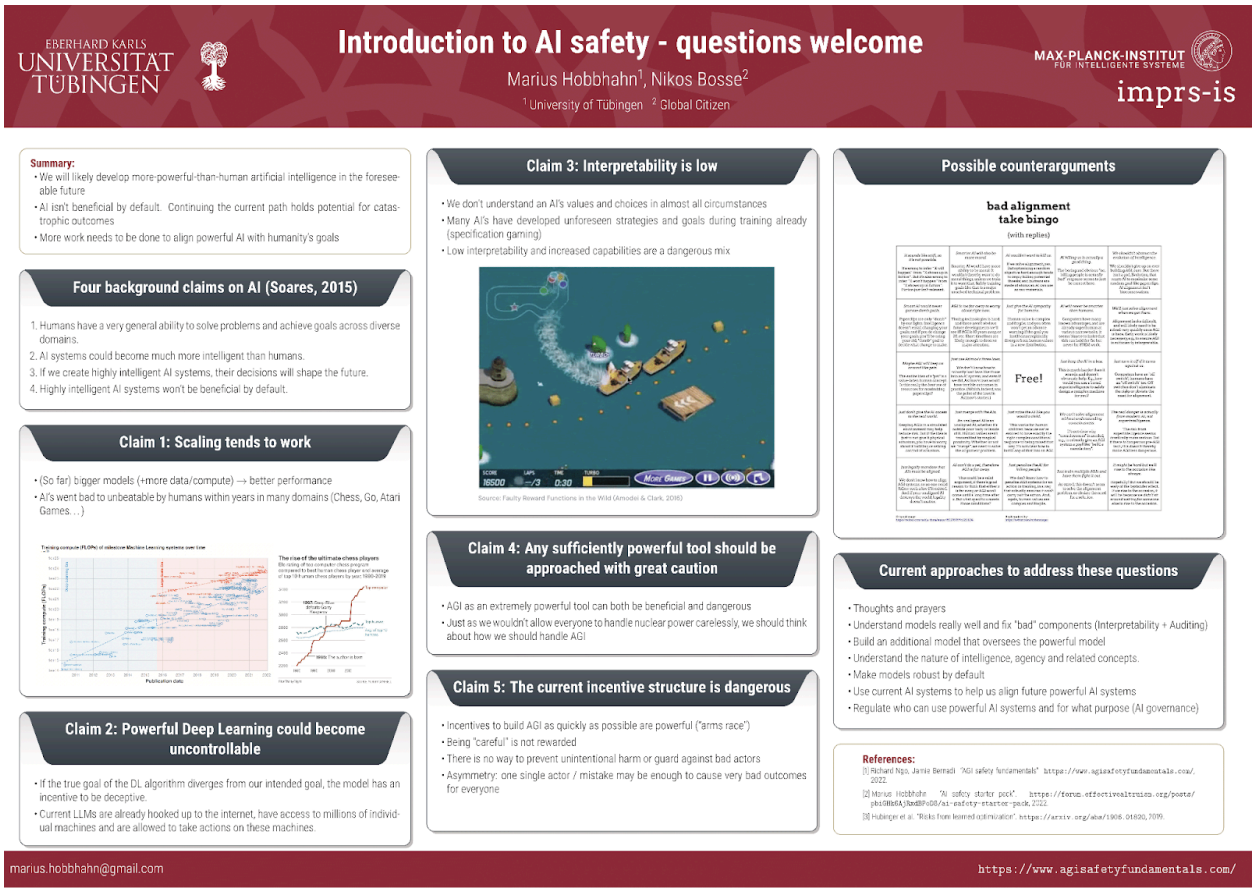
Marius' alignment pitch

This is a write-up of the pitch I [usually give to academics]. Obviously, the pitch differs from person to person but the baseline looks something like the following. When I give the pitch, I often have the following poster available and walk them through it. Here is a [link to the poster]. In case you want to change it, just make a copy and adapt it.



General comments

Some generally helpful findings include

1. **Explain don't try to convince:** Explain the arguments that you find plausible and why. Be honest about where you see uncertainties. Don't try to convince them too much, e.g. don't try to force your beliefs upon them in case they disagree even if they don't have good reasons for the disagreement.
2. **Engage with their concerns in an honest fashion:** There will be concerns and many of them will have obvious answers. Don't get frustrated and take the time to engage with their criticism in an honest and constructive fashion. People are usually open to being persuaded if you engage in a respectful way.
3. **Emphasize uncertainties:** There are many questions AI safety researchers have no good answers to and that should be emphasized. There is a large disagreement about

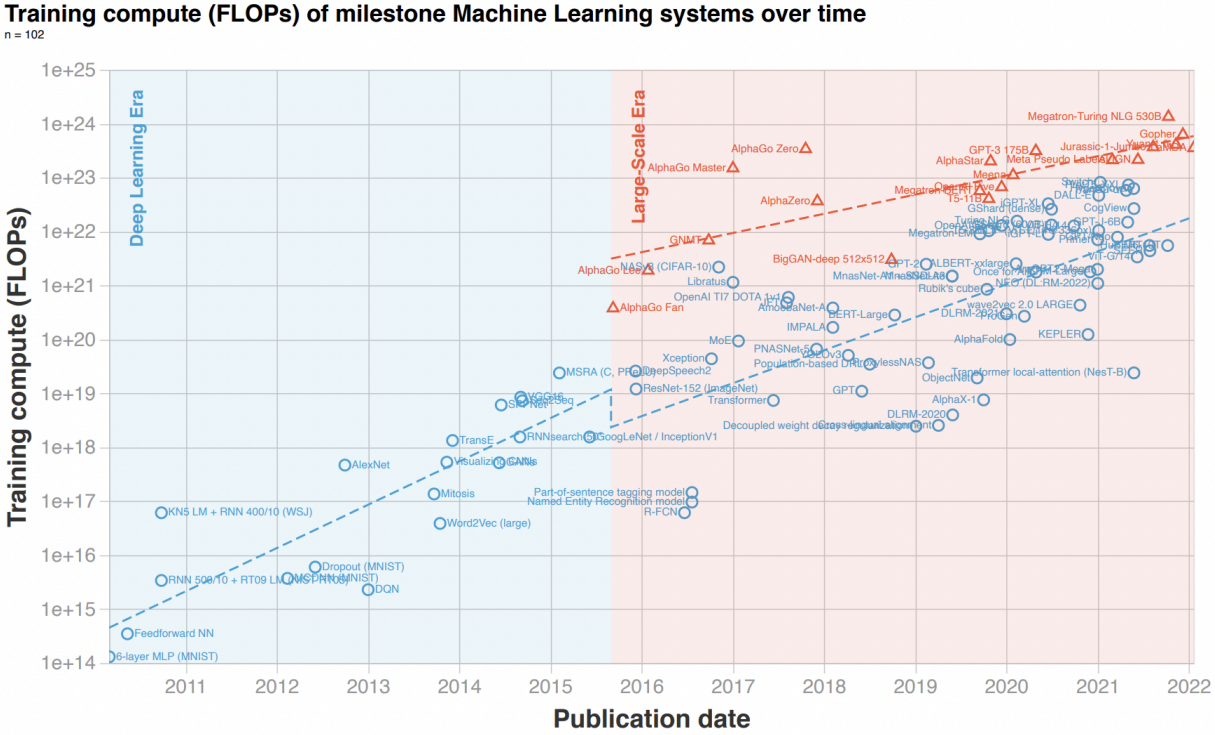
how much X-risk AI actually poses and there is a strong disagreement about the prioritization of problems and approaches. It's completely fine to say that you don't know the answer to something. On the other hand, you can also argue with EVs, e.g. that a 10% chance of AI going really bad is sufficient to work on AI safety full-time.

4. **Don't be alarmist:** I found that motivating the problem of AI safety with X-risk is often seen as alarmist. Chatting about technical problems received better feedback. In general, academics are highly skeptical of all claims that sound too grand or sci-fi.

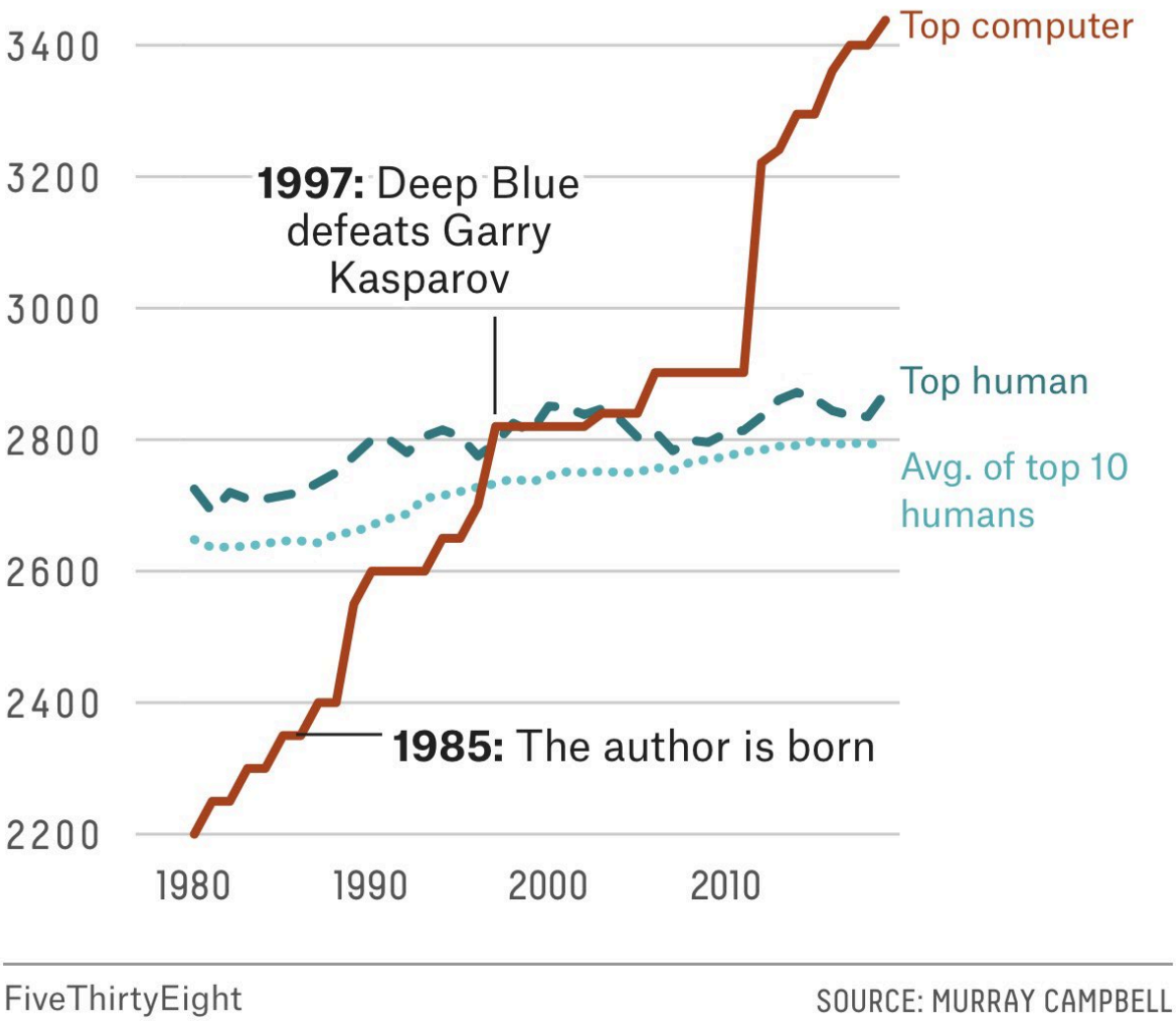
In general, just following some basic norms of conversation like being nice, honest and respectful goes a long way. It shouldn't be necessary to say this but I think many people concerned about AI have tied their identity around the topic which often makes it hard for them to understand when people don't find the reasoning immediately plausible.

Claim 1: Scaling tends to work

Over the past couple of years, there has been a strong trend that more data, more compute and bigger models lead to better performances if combined in the right way. I usually show our graph from Epoch to make this more explicit. There are other resources to show such as even more graphs from Epoch, Danny Hernandez's work, the Chinchilla paper or the bitter lesson. People usually buy this claim with some caveats, e.g. that not all NN architectures can scale that well.



Additionally, I argue that there is no reason to assume that human intelligence poses any sort of magical limit. At least in narrow tasks, human intelligence has been beaten many times, e.g. in Chess, Go, and many Atari games.



Then I point out that it's not clear that something like general intelligence exists or whether humans are just good at learning new tasks but are mostly a collection of narrow skills. For example, humans have a hard time doing math in their heads even though computers don't seem to struggle with it. The task is not hard, humans are just not well-equipped to deal with it. Human intelligence might just be narrow in its own sense.

Claim 2: Powerful Deep Learning could become uncontrollable

The arguments in this section are a bit more technical and I usually don't discuss them in detail. Often my presentation is like "there are some technical arguments that Deep Learning based approaches could be uncontrollable or deceptive. Both of these would be really bad. We can talk about the details if you want to".

I tend to talk about uncontrollability because it is less weird than alignment. Alignment carries all of these vague questions like who to align the AI to and so on whereas uncontrollability is less controversial. The basic pitch is usually something like "Goal maximizing systems have incentives to preserve the goals they currently have and have other instrumental incentives to remove your control". This is also true for DL systems in the future, they are currently just not powerful enough to understand this or act upon it. But we should really look out for this when training NNs, especially in an RL setting.

There are also some claims about deception. I think Richard Ngo's paper "The alignment problem from a DL perspective" has the best explanation of this problem so far. I usually present a short summary of the paper and then try to answer follow-up questions.

This fact is more complex than the others. It's only claim 2 because this was the best layout for the poster. Without the poster, I would switch around the order a bit.

Claim 3: Interpretability is low

We mostly don't understand what's going on in NNs. We don't understand how they make decisions and our tools are not good enough to look at individual activations for large NNs. The best we can currently do is find circuits for simple behavior in GPT2-small.

This claim is usually not controversial and everyone working on NNs just keeps nodding during all of these examples. The fact that NNs are effectively black boxes is very widely held.

Additionally, it helps to emphasize that just specifying the goals of the network also doesn't seem to be sufficient. There is a long list of problems of specification gaming, where an RL agent, learns an unintended policy from a plausibly sounding reward function. There is also a long list of examples of goal misgeneralization (see Rohin's new paper). I usually tell the story of the boat racing game (see below) as an example of specification gaming. But I might take an example of Rohin's new paper in the future.

Source: Faulty Reward Functions in the Wild (Amodei & Clark, 2016)

This is important because it shows that we need to be able to look into the black box from time to time if we want to make sure the system learns the right policy.

Claim 4: Any sufficiently powerful tool should be approached with great caution

This really doesn't require a lot of convincing. It just sometimes helps to name a couple of examples, e.g. guns, viruses, nuclear weapons, etc. We even require a driver's license to move a car. So if NNs have the ability to hack, manipulate, etc. they should also require some safety mechanisms.

Claim 5: The current incentive structure is dangerous

Many academics live in their own little bubble and often have a very naive view of the rest of the world in my experience. For example, they work on dual-use technology but can't (or don't want to) deal with the fact that their research could be used by bad actors. They are also unaware (or don't want to think about it) of the fact that AI in general could be used for nefarious means and that regulation is thus really critical. I think this changed over the last few years but there are still a lot of people with an "engineering mindset" who just don't want to think about the potential consequences of their research.

Therefore, it seems important to point out that the current incentives are really not helping. Some of them include

1. **Academic incentives:** Your main incentive in academia is to publish. Publication is not tied to safety concerns at all. However, thinking more about safety usually means that you can publish less. Thus, the incentives are probably against thinking about safety concerns. Furthermore, it's harder to publish direct safety work. Therefore, there are even fewer incentives. Lastly, it's hard to pivot once you have doubled down on a specific research agenda, so switching to safety is even harder.
2. **The race to AGI:** If you are the first company or country to build AGI, this seems very advantageous. Thus, there is an incentive to rush ahead even if there are lots of safety concerns. This means that safety is less of an agenda for many actors working on AGI.
3. **Asymmetries in actors:** Even if most actors care about safety, it is sufficient for one actor to fuck up for everyone to have a problem. This could be one company releasing a dangerous AI to the internet or one country to naively try and win the race to AGI.
4. **There are no established procedures to care about bad actors and or unintentional harm:** Thus, there are very few incentives to do anything about them even if you are or spot incidents.
5. **Political incentives are nationalist and short-sighted:** There are few institutions that have global long-term incentives and they tend to not have power (see UN). However, these would be really needed in such a situation.

Current solutions approaches are not sufficient

After people hear the basic arguments and a bit of back and forth, people usually as "OK, so what can be done?". I then go through the list of current alignment approaches and then say that I personally don't think that any of them has a chance of solving all problems even if we fully succeeded with it. People tend to share this impression, e.g. they also see that the problem descriptions are far bigger than what the respective approaches try to solve.

Then we usually chat about what the individual researcher could do to contribute. Most of the time, this is a personalized recommendation, e.g. when they work on RL, I talk about inverse RL

or reward design or when they work with NNs, I suggest they should try to open up the black box and understand what the network is doing.

Final words

I’m probably missing stuff or some part of my pitch is suboptimal. Feedback and comments are appreciated.