1: children and programming

when my eldest daughter was about one and a half years old, she could barely speak a few words, yet she liked to mimic her father at work. she did so by climbing onto my chair and then using both of her tiny hands to furiously slam the keyboard while yelling, "DAMMIT, DAMMIT"!

nowadays, when i have to explain the AI risks to someone, i often ask them two questions first: do you have children, and can you program computers?

to parents, i can point out that their children are also part of humanity, and thus defuse arguments of what i call "ersatz humility" – statements like "perhaps humanity does not deserve to survive" or "isn't machine takeover a natural step in the wonderful cosmic evolution?"

parental instincts, i find, always trump the urge to sound clever, so you can use them to call bullshit.

2: programmer intuitions

similarly, a programming career ingrains a set of very useful intuitions that non-programmers don't necessarily share – intuitions about computers that can be used to anchor the AI discussion.

(as an aside, there's at least one intuition that makes it **harder** to discuss superintelligent AI with programmers: the intuition that programs are always amenable to being stopped and debugged.)

still, one very useful programmer intuition gets rubbed in by those "dammit" moments that my daughter unknowingly made fun of: it's the intuition that computers always do exactly what you tell them to do, but almost never what you wanted them to do!

3: constraining intelligent systems

therefore, to borrow an example from professor stuart russell, for programmers it should be entirely uncontroversial to believe that if you program a self driving car to get from point A to point B as quickly as possible, then the passengers would arrive at point B covered in vomit and chased by police helicopters.

moreover, adding the constraint of adhering to the speed limits wouldn't help much, because of the extreme acceleration and braking that would result.

more generally, as stuart points out in his hard-hitting comment on the edge.org AI conversation, whenever the world model of an autonomous agent has free variables (such as acceleration in our example) that are unconstrained by its goal function, those variables will likely end up being pushed to extreme values that are incompatible with human well-being.

from this, it's only one small step to understanding the gist of the AI safety argument: increasing the capabilities of intelligent systems and constraining their behaviour must happen in concert. or, to put it in the language of the open letter many of you signed, we must ensure that capable systems are also robust and beneficial.

4: logic works

another useful programmer intuition is that logic can be trusted: you can have a computer perform a billion steps, yet its output always conforms to what the input and the steps imply.

in fact, it is hard – if not impossible – to get random output using logic alone.

because of that intuition, i take logical arguments seriously – regardless of who makes them, and even if they yield unpleasant or counter-intuitive conclusions.

in other words, if an argument becomes sufficiently precise to be – at least in principle – convertible into computer code, it can and should be evaluated independently of the qualities of the people behind it.

5: meeting with eliezer

therefore, when i stumbled upon the writings of eliezer yudkowsky in 2007 or 2008, and detected the presence of an important logical argument, i had to find out more.

i remember meeting eliezer in march 2009 in a franchise cafe next to a barren californian highway. he wore a badge saying "speak the truth even if your voice trembles" – a token, he later explained, to pick him out from the crowd if necessary. our initial meeting lasted for 4 hours and i came away convinced that his main argument about the default AI outcome not being good was strong enough to be taken seriously, despite him and his "singularity institute" not being part of the industry or academia.

i wasn't the only one to notice them of course.

peter thiel was already their biggest donor, and i remember moshe looks – an AI researcher in google – saying in his talk at one of the first AI conferences that i attended: "you should really listen to what eliezer yudowsky says – even though it's annoying to admit that he's right!"

6: reputational support

it was clear though that when it came to influential people – with the notable exception of peter thiel of course – almost none of them paid any attention to such arguments.

this was where i thought i could help. even though i don't remember formulating an explicit plan to do that, in retrospect there were at least two handles that i could pull.

first, the arguments needed reputational support in order to increase their credibility in the eyes of people who evaluate arguments by the reputation of the sources. towards that end, i started exploiting my own street cred to deliver the AI safety argument to various audiences across the world – a fairly uphill battle.

7: meeting huw, CSER

soon enough i got lucky though by meeting professor huw price and convincing him that these issues warranted immediate action.

in 2013, huw wrote a NYT opinion piece, recounting our first meeting as follows: "In Copenhagen the summer before last, I shared a taxi with a man who thought his chance of dying in an artificial intelligence-related accident was as high as that of heart disease or cancer. No surprise if he'd been the driver, perhaps (never tell a taxi driver that you're a philosopher!), but this was a man who has spent his career with computers."

huw invited me over to cambridge and introduced me to martin rees, a very prominent scientist who back then was the master of trinity college.

together we started CSER – centre for the study of existential risk.

that pretty much took care of the reputational problem: the strategically chosen name of the centre, combined with the credibility of the cambridge university and the distinguished people on our advisory board, makes it very hard to dismiss the topic on purely reputational grounds.

sometimes, i joke that CSER's main contribution to the world has been to give a canonical answer to the question "existential risks? says who?"

8: reaching out to demis, deepmind

the other handle i pulled was to reach out to a couple of AI companies that i thought should be aware of the AI safety argument.

as michael vassar – a good friend and former president of the singularity institute – says, a nice side effect of skype is that gives me a reason to be in the room – a fact that i've been exploiting by sending cold emails or simply walking up to people i want to talk to.

in that spirit, in early 2011 i walked up to demis hassabis after seeing him present his AI research at a conference, and we started a conversation that eventually lead to my investment in deepmind, followed by joining their board of directors.

as a quick aside, my career in deepmind exposed me to several moments of internal turmoil where the technologist in me was fascinated, while the AI-safety activist was concerned.

for instance, i remember a lunch with demis where he brought his ipad in order to "show me something cool". while we were waiting for our pasta, he shifted it over and, with a smirk, started a video: it was the video of deepmind's atari-playing AI having figured out how to beat the breakout game by deliberately sending the ball behind the wall of bricks. while the technologist in me marvelled at the achievement, the other thought i had was that i was witnessing a toy model of how an AI disaster would begin – a sudden demonstration of an unexpected intellectual capability.

with that said, i was glad to see that the safety concerns we promoted together with luke nosek, who represented the founders fund, did indeed resonate with the deepmind team. even the current conference is an evidence for that, given how instrumental deepmind has been in organising it.

by the way, a hilarious case can be made that if this conference ends up significantly pulling the AI research trajectory towards safety, sponsoring it might have been the best way humanity has spent money, ever; and helping to organise it would also be a strong outlier on the scale of effective altruism.

9: vicarious

later, and to a lesser degree, i repeated my "invest-to-be-heard" approach with vicarious.

effectively, my strategy with AI startups has been to invest in them so i could hang around in their kitchen and talk about AI safety to anyone who listens – and i'm happy to report that, as far as i can tell, it seems to work: for example, both deepmind and vicarious are co-signatories to the open letter.

in fact, let me take a moment to thank all of you who joined the letter – i truly think that the world is now a safer place because of you, and i hope that the letter will function as a corner stone for the official collaboration between the AI-safety and AI-research communities.

10: building the bridge: x-risk code

yet when it comes to strengthening the bridge between the AI-safety community and the AI-research community, there's a long way still to go, and both sides need to invest and evolve.

speaking of the AI-safety community (or x-risk organisations in general), i think the biggest improvement they can make is to ground their research more strongly in computer science and mathematics. they need to write code and design formal models in order to demonstrate that their arguments and ideas are not just some post-modernist babbling they are often mistaken for. not to mention that working on the math and code would be conducive to integrating their results with the on-going AI research.

so i'm glad to see that MIRI has been recently pursuing that strategy in earnest, and can only endorse it to other x-risk organisations.

to quote daniel dennett, computers keep the philosophy honest.

11: building the bridge: industry arguments

now turning to the AI researchers, one fairly low hanging fruit is to make an effort to retire the low quality arguments that infest the AI safety discourse: there are about a dozen knee-jerk statements that your colleagues whip out again and again when they feel the need to signal their alliances.

some of them don't even bother to be logically consistent, frustrating those who want a genuine discussion about this important issue. stuart armstrong from FHI recently wrote about one such argument as follows: "I want a big stick. On the stick, will be written 'we don't know' does not mean 'we are safe'. Then I would hit anyone who made that kind of argument, with that stick. Then I would feel better, and they would feel wiser."

you can examine a generous sample of these arguments at the edge.org AI conversation that i mentioned earlier.

12: building the bridge: doomsayers

however, allow me to – just for public display – hang and bury one of these arguments here, and then stomp on its grave.

i'm sure you've heard of the "doomsayers argument": sometimes otherwise intelligent people point out that throughout history there have been doomsayers predicting the end of the world in order to draw attention to themselves and their cause.

the AI safety argument, they maintain, is no different.

now, leaving aside the basic anthropics 101 error such arguers are making, there's a simple way to demonstrate that this argument is not only flawed heuristic and sloppy thinking, but advancing it in the context of new technology is actively harmful to humanity.

consider the report labelled LA-602.

prepared by physicists konopinski, marvin and teller 6 months before the first test of the nuclear bomb, it investigated the possibility of the nuclear detonation igniting the atmosphere and thereby destroying earth. as far as i know, that was the first existential risk research project humanity has undertaken.

the problem with the doomsayers argument is that it lumps together religious nuts who pull up reasons why the end is near from their hind quarters, and people like those manhattan project scientists promoting a sober scientific analysis. the argument lowers the credibility of people who advocate such analyses, and makes their concerns harder to hear.

i could therefore make the case that every time one wields the doomsayers argument against scientists or engineers, he kills a thousand children by increasing the probability of an existential catastrophe.

please, help me to stop people from doing that.

13: building the bridge: culture of care

another obvious way how AI researchers could help is by promoting a culture of care within their organisations. considering the dramatic changes that you might be bringing about, i would say that you are not in the business of creating artificial people. instead, you are in the business of rewriting the laws of physics – so the level of required care should go way beyond even what they had in the manhattan project.

14: building the bridge: media, politics

finally, both the AI-safety and AI-research communities need to advance dialogue about the correct level of involvement from media and politics.

in case of the media, as we see at CSER, reporters have a lot of appetite for these topics. we should tap into that interest to advance correct arguments and draw in people who could contribute to good outcomes – yet we definitely want to avoid fuelling public hysteria or endangering the trust between the AI-safety and AI-research communities.

the same is true for politics. in my view, politics is a very blunt instrument whose side-effects usually dominate the intended effects. therefore, extreme care is needed to build consensus and only promote policies with predictably positive side-effects.

15: x-risk research is fun!

before i conclude, allow me to make the case for why more capable people should spend their brain cycles on AI safety research (and x-risk research in general).

no, i'm not going to talk about how these topics are under-appreciated or how contributing to the x-risk reduction is the easiest way to maximise one's positive impact. while both points are true, myself and many others (not to mention the entire effective altruism movement!) are working hard to make more resources and reputation available to x-risk researchers – thereby weakening those arguments over time.

instead – and perhaps somewhat surprisingly – i can recommend x-risk research because it's just so much fun!

you get to routinely combine hard technical problems with deep questions from cutting-edge philosophy.

you get to talk to some of the smartest people on the planet – people like nick bostrom and many others here.

also, despite how gloomy the topic seems, some x-risk community members are the most cheerful and charming people i know. anders sandberg from FHI, or the host of this conference, max tegmark, come to mind.

personally though, my own favourite side-effect from being active in the x-risk community is that you get regularly exposed to controversial hypotheses that – if true – would turn your world upside down. of course, the idea of technological existential risks itself is a prime example of such ideas, but there's a lot more where that came from!

i sometimes joke about michael vassar that he should get a guinness record for his ability to produce crazy-sounding statements that are not obviously false. one of my tongue-in-cheek talks is based on my conversations with him and other x-risk researchers. feel free to google it: it's called "why now? a quest in metaphysics" and it tries to address the weird coincidence of being born into a seemingly mind-blowingly important moment in the history of the universe.

i find that based on the exposure to such topics alone, one often gets to be the most interesting person in the room. for instance, i remember once being approached by a stranger in a crowded restaurant – he was on his way out, yet wanted to first thank me and michael for the opportunity to overhear our conversation!

last but not least, there's a strong sense of camaraderie within the x-risk community, because of the important common goal that transcends the organisational boundaries. since the FHI is located in oxford, my original hope with CSER was to play on the oxford-cambridge rivalry to ratchet up the intensity of the x-risk research – but no, there is just too much sense of community between the x-risk researchers for that to work.

so let me repeat: saving the world is a lot of fun!

16: conclusion: moloch

in conclusion, allow me to propose a "rocket metaphor" for AI development: just like when launching a rocket, in AI development you should first maximise acceleration, but in order to avoid catastrophe, steering should gradually become your primary concern.

of course, since AI development is a distributed global effort, steering it implies a coordination challenge.

earlier this year, the young doctor and essayist scott siskind published an essay called "meditations on moloch", which i can heartily recommend. the gist of the essay is that most global problems humanity faces – such as global warming, arms races, intelligence explosion – can be reduced to coordination problems where, just like in prisoner's dilemma or tragedy of the commons, local incentives conspire to yield a suboptimal outcome globally.

using poetry and mysticism, scott paints a picture of us voluntarily feeding our children to an emergent game theoretic monster named moloch – with the most perverse thing being that the moloch does not even care about eating children, nor anything else really.

it's just a horrible, horrible piece of cold mathematics.

that little girl at the start of my story who climbed my chair to yell DAMMIT at the computer is now a young lady. when i look back at the last few years of my life straddling the x-risk reduction and AI research communities, i see many wonderful conversations, great people, technical breakthroughs, and philosophical insights that shook my world..

yet when i squint my eyes, i can also tell the shape of that game theoretic monster, waiting patiently for the AI rocket to miss its target, so it could devour the girl who yelled at the computer.. as well as her 5 siblings.. and your children.. and you.. and everyone.. and anything anyone ever cared about or hoped for.. to only yield back the silent indifference of a dead universe.

please, let's do all we can to steer the course of AI away from that future.

thank you!