intro

my favourite piece of parenting advice is that when you explain something to your kids, treat them as adults and don't dumb down your explanation. it's a good bit of mischievous fun to watch them try to decode what you just said. more importantly though, kids' abilities increase over time, so if you dilute your explanations, you're likely to keep undershooting their intellectual level.

my original plan for this talk was to look back at what has happened in AI safety since the last FLI conference, give positive feedback for the achievements (such as all the technical value alignment work that has been done), take stock of the current challenges and generally congratulate the community for maturing so quickly.

but then it occurred to me.. wait a minute.. wouldn't that parenting advice apply here as well? wouldn't i be undershooting the intellectual level and progress of this community?

so i deleted my draft and started over. new topic: we need a way to figure out what humanity wants.

problem

when i talk about AI value alignment in public, one question i almost always get is "whose values are we talking about?" in response, i usually mumble something about the vast majority of human values being so obvious that we don't even think about them. as stuart russell puts it, "everyone values their right leg" or — my own favourite — "everyone likes our planet to be roughly at room temperature".

that is not to say that we don't have a problem with aggregating values. we do, a massive one.

we don't know what the complete set of humanity's values are, nor can we simply ask people either, since there's a difference between what people say that they value and what they actually value. also, our values are clearly a moving target because they keep evolving over time.

what's more, we don't have a great track record in coordinating on what we already know is valuable. the league of nations failed to prevent WW2, and although the UN has some successes under its belt (such as eradicating the smallpox and fixing the ozone layer), it's widely considered ineffective at its job.

also, free markets and democracy, although better than the alternatives, seem limited in their ability to steer the world towards a bright future.

indeed, i think that the biggest disservice that capitalism has done to the world is that it has created a false sense of security in technological progress!

not to mention the thorny philosophical paradoxes with value aggregation, and even potential dependencies on some unknown aspects of the laws of physics when we get down to the nitty-gritty of how the aggregation should work.

hope

yet there is hope.

first of all, we now know much more about morality, human values and game theory than we did when, say, the UN was established.

second, various new technologies seem to favour global coordination. for example, the internet and mobiles have connected the planet, cheap satellites and other sensors will create an explosion in transparency, and the invention of cryptoeconomics has introduced a new regime: it's now possible to have worldwide consensus about a piece of data without trusting any central authority to maintain it.

just to give one example about what can be done with cryptoeconomics, it's now possible to create global "decentralised courts" on blockchains that resolve conflicts by enlisting random people as a jury, and then game-theoretically incentivising them to produce opinions that society in general would find fair.

not to mention that the continued advances in AI and techniques such as the inverse reinforcement learning and approval-directed agents by paul christiano seem extremely relevant here.

proposal

therefore, i'm proposing that we start designing explicit mechanisms to transparently and robustly aggregate global opinion about what a good future should look like.

the mechanisms have to be open and transparent — blockchain-style — to instill trust that their purpose is to serve everyone in a fair manner.

they have to be robust in the sense of being hard to game, corrupt or otherwise defect against.

this probably requires incorporating philosophical meta-principles, such as veil of ignorance — that is, you could only benefit from the system as a random member of humanity, not as a particular person in a particular position.

basically, i'm advocating extending the technical approach that has been very successful in advancing the frontier in AI safety thinking, to the problem of global preference discovery.

["high-altitude bombardment by thought leaders + ground invasion by technical AI safety researchers"]

pitfalls

of course, all this new technology can also make things harder. one man's coordination is another one's collusion: we have to be careful not to catalyse criminal activity or, worse, paint humanity into a corner by introducing terrible nash equilibria on a global scale.

not to mention the clear and present danger of various AI arms races — both literal and figurative. we absolutely need to avoid these (i know this topic has been on demis's mind for a long time).

finally, the looming AGI limits our time budget. we have lost over half a century since the original warnings about AI value alignment by alan turing and norbert wiener. i certainly hope that we still have another 50 — but i know that several experts in this room are much more optimistic about AI timelines, and thus pessimistic about our remaining time budget.

potential

last month, there was a workshop at the FHI in oxford where one of the sessions was about what to do if the value alignment won't be solved in time. it had this eerie atmosphere of a science fiction story featuring an alien fleet in orbit — aliens who couldn't care less about humanity — and then a roomful of decision theory experts trying to find a philosophical loophole that would allow humanity to keep at least one galaxy out of the 100 billion.

one galaxy as a consolation prize for the losers! that's 50 personal star systems for every human alive today.

with that i want to illustrate a couple of things: 1) even if we mostly screw up, things might still turn out to be pretty okay in the end, and 2) the worst thing we could do is to continue playing our usual political zero sum games while losing 50 galaxies per second!

vision

luckily, a transparent preference discovery mechanism might serve as a ladder for humanity to climb out of the arms races and other bad nash equilibria.

it might also help with a problem that many of you personally feel: society does not trust you with the power you have over the future. granted, they might trust you more than the politicians — but, c'mon, that's a very low standard!

of course there's a valid reason for that mistrust: history is littered with catastrophic tragedies caused by individuals or movements that amassed too much power. i should know that, having personally experienced the tail end of one such tragedy.

now imagine if there was a way to credibly demonstrate that you're working towards a future that not just you personally thought was a good idea, but towards the future indicated by the global preference discovery mechanism.

it's sort of what the open AI people have been talking about — but on steroids :)

finally, having a strong schelling point for humanity's values should be great tool for philanthropists, effective altruists, and politicians who genuinely want to improve the human condition.

and of course, ultimately we want the mechanism to converge into something that can safely guide a superhuman AI.

outro

two years ago standing in front of this conference i compared AI development to launching a rocket: initially, you mostly worry about having enough acceleration, but eventually steering should become your primary concern. to summarise my current talk, i would extend this metaphor to say that now that the AI researchers are producing ever more powerful engines, and the steering systems design by the AI safety researchers is also progressing, it's about time to start plotting our eventual trajectory.

crucially, the trajectory planning must be globally transparent and fair, because everyone — everyone — will be on board.

thank you!