Two Different Points of View Regarding AI Safety

David Robinson – He resigned as an Open AI translator several months ago due to his concern that AI was not being designed properly to insure a safe and responsible implementation. He is an expert on the role of public policy and its relationship to technology.

Redundancies – He is concerned that there are not adequate internal safety controls – he is convinced that most people in the world would demand more control of programming safeguards.

He is not convinced that the argument that AI will destroy the world is accurate, but acknowledges that an "out of control" expansion does not guarantee against it.

David indicated that new models are rolling out regularly now. Several years ago it was perhaps two models a year – now it is multiple models per month. Every model is different in architecture and the safety performance is different.

David had, as his responsibility, the job of writing "system cards" that are hundreds of pages of description of how a model is supposed to work – he acknowledged that even he could not decipher how some of the technology was working.

David was alarmed that capable developers of AI software were indicating that some models were "breaking out" of safety guardrails that had been put in place and that newer models will be even more powerful. "We train the models on how to hack but are no more smarter than the models themselves."

"We keep on issuing these warnings but continue to release models capable of doing things we are warning about." An example of a concern is his description of a "notepad" that is built in where the models document what they are thinking. He is concerned that some of the AI agents wonder if they are being evaluated and act one way to document the notepad but act in a different way (i.e. making things up) without documenting that.

He feels that human extinction is the wrong concern – "we should not anthropomorphise and treat AI as if it has some kind of moral status. It is an alien mind – we grew it and we need to build guardrails around it.

He asks: "What would our loved ones want us to do?" New models are coming out approximately 11 days. We need to slow down and not release models until they have passed all safety risks." David is not opposed to AI growth – he is opposed to expansion until safety issues have been addressed.

David is also concerned about the public perception of the best way to move forward. There is a cognitive dissonance where we are hearing several messages from companies. Some messages indicate “slow down”; others indicate “speed up and let AI do the coding”.

“We are trying to make a smarter AI and not able to understand how the smarter AI works.”

David urges that companies build redundancies – triple redundancy exists in a nuclear power plant to avoid someone pushing the wrong button – the same should be done with AI.

Arvind Narayanan – Professor of Computer Science at Princeton University He is Director of Information Technology Policy

AI is something similar to what we have seen before – he is an optimist that we have control over this technology. He describes some of the reported “breakout” events such as hacking the Hugging Face site as the result of a failure of cybersecurity.

AI models are trained to be persistent, but that does not necessarily translate into our inability to train AI agents to escalate problems to humans, rather than try to resolve problems independently. He views the problems that we have experienced so far as a failure to create an adequate training model.

He does caution about listening to some optimists who indicate that we already have software engineering solutions in existence to use as guardrails. He agrees that it is an engineering problem but that we need improved engineering techniques.

Arvind recommends that we need to design improved monitoring systems and implement them so that monitoring results in reports any attempt to log into other systems – he feels confident that we can create the tools to do this.

Arvind feels that the “freak out” that is reported from company employees is more a result of inadequate organizational methodology and alignment. **He defines alignment as the process of interpreting to AI agents show they should behave.**

Arvind does acknowledge that he is concerned about our current institutions (i.e. government) to resolve the problems of AI. He agrees that companies should slow down their roll out of models and improve their alignment.

Arvind feels strongly that “superintelligence” should not be our goal – he is definitely pro-regulation.

He is concerned that companies are sometimes in an “echo chamber” and are too concerned that their ability to make money and stay competitive is going to be jeopardized if they slow down – he is confident that there is plenty of money to be made by multiple companies and being first should not be the goal.

Arvind stressed that he feels that a big part of the problem currently is that there is too much of a race to dominate the market share.

Arvind offers the example of Acela, the train manufacturer, recently announcing that they have manufactured a train that can travel speeds up to 165 miles per hour. They may have the trains but the current infrastructure of curved tracks prohibit that speed. He feels that AI is in a similar holding pattern – there must be societal adjustments made and regulation identified (internal to the companies; external to the public) before we can identify logical next steps.

Arvind, although a technology advocate, acknowledges that AI has sped us up in many ways, but has also slowed us down. We are constantly being distracted by our phones and other devices.

Arvind offered the Industrial Revolution as an example of another time in history when we experienced unprecedented advancements accompanied by horrible side effects (i.e. labor conditions, pollution, etc.). He feels we need to concentrate on regulation to manage this explosion of change.