Short essays

Aarsh Batra

May, 2018

Abstract

This document is divided into three sections. Each section focusses on a specific research question/idea. For each research question, I briefly discuss the following: my biggest concerns about the idea, why is it worth pursuing? How would I go about pursuing it?

I Human level (and beyond) AI: a potential existential threat

Whenever I wonder about whether human level AI is an existential threat to the society, some very specific questions pop up in my mind. Why is it the case that we visit the zoo to see the animals and not the other way around? Why is it the case that we can put a strap around the neck of a dog and call it 'ours'? We cannot do this with humans, or else it will be called slavery. Do we commit mass genocide when we walk over hundreds of ants each day? Is there a spectrum of species on which there is a threshold, below which we can do as we please and above which the 'law' comes in and take hold?

These questions help me understand the impacts of a significantly more advanced species on those that are less advanced. Why do we strap dogs, crush ants, eat beef? **Because we can!** The fittest species will rule and have its own definition of right and wrong, and the rest of the species will accordingly fall into order.

There is no reason why human level AI cannot pose a similar threat to us, it is just a matter of time. Human level AI is a potential threat because it will put us face to face with a silicon based conscious being.

At the moment, the only characteristic that 'qualitatively' differentiates us from all life on earth is our unique technology called consciousness. When a super-intelligent conscious AI pops into existence, the status-quo will be disrupted and the result will likely be very chaotic. A statistically 'super' significant qualitative shift will occur at the very moment AI becomes

conscious. Why? The silicon-based machines will be able to do everything that humans can, but, significantly faster, more efficiently, and in a fraction of physical space.[1]

Given these concerns, the research problems that interests me are as follows: 'Precisely' pinning down the moment (or interval) of such a breakthrough (This will help in setting timelines within which to make things right). Will consciousness in silicon-based machines manifest itself in similar ways as it does in biological beings? (Developing a better understanding of consciousness in biological beings and how it 'precisely' functions may help us better tackle and prepare for Silicon based consciousness) Will free will (assuming it is a by-product of attaining consciousness) play out differently in silicon-based life compared to biological life? Can consciousness exist independent of free will? Researching other biological species (mainly primates) that may have started to evolve something that is close to, what we call consciousness (e.g. recognizing one's image in the pond) but still far from realizing that they might have free will. Then one can use FMRI, 'physical incentive based' techniques, theoretical/philosophical models to try develop a formal measure for free will. Does free will exist in a spectrum?

II CRISPR and the future of designer babies

CRISPR gene editing technology is a tool that scientists use to change the letters of DNA in cells in precise ways. Jennifer Doudna, co-inventor of CRISPR/Cas9 technology, makes an analogy to the word processor we use in our computers every day. DNA can be thought of as the text of the document which contains instructions that tell the cell, what to do. Then, similar to what we do in a document the CRISPR technology allows scientists to go in and edit the letters of the DNA, just like we might cut or paste text in the document, or even replace a letter, word or even entire chapters.

It is one thing to talk about being able to remove those mutations from the human population that cause genetic diseases (e.g. sickle cell anaemia). But, nothing is stopping us from using the same technology to make advanced human beings tailored to our desires? Will it lead to a new branch (maybe 1500 years from now) in our evolutionary tree (homosapiens ver-2.0)?

Jennifer Doudna describes it as a 'democratizing' technology, i.e. a technology that is easily available, is not very expensive, easily deployable by labs worldwide. For the first time in

[1] Randall Munroe, the author of the fantastic book, what if? compares human computing ability to a mid- range computer processor. In his book, he mentions the following calculation done by the computer scientist, Hans Moravec: A human running computer chip benchmark calculation by hand using pencil and paper can carry out the equivalent of one full instruction every one and a half minute. By this standard, a processor in a mid-range mobile phone could do calculations about 70 times faster than the entire world population (calculations were performed in the year 2014).

human history, we have conjured up something so significant (in terms of its potential impact) that has the ability to challenge evolution by not accepting what it plainly offers.

Such a revolutionizing technology cannot lie dormant for a long time irrespective of the ethical and legal concerns surrounding it. One of the implications being: black market for designer babies with no guidelines in place.

My research would focus on developing an efficient system of protocols that would help guide and regulate the future designer baby markets. The faster we realize that there are potential risks, the faster we can work towards nudging the technology's R&D in the right direction.

III Society's tremendous ability to delude itself

"The really unusual day would be one where nothing unusual happens"
-Persi Diaconis[2]

Most people in their daily lives associate events that they don't (immediately) understand to superficial beliefs and/or entities like 'miracles', 'gods' etc. In many cases they further go ahead and claim that not only is the event associated with a god, it is also 'caused' by it.

The origin of such illogical behaviours lies in the extreme need for an 'immediate closure'. The need for an immediate closure is what I think to be one of the key factors that plague the mind and in doing so it enslaves the subject into believing things for which there isn't any evidence in the scientific world.

The Law of Truly Large Numbers states that given a large enough sample size; any outrageous, unusual thing is likely to happen. Given that, we should expect unusual things more often than we think, such that, a truly unusual day would be the one where nothing unusual happens. So, next time you hear of a person getting hit by lightning thrice in the last 2 months, don't get alarmed (also don't look for closure). It might be that he is a fisherman, working 8 hours a day over the Catatumbo River in Venezuela (the most electric place in the world).

Believing in such delusions is analogous to choosing to walk blindfolded through a busy high-speed lane on a highway in order to get to the other side, rather than taking the walk-over bridge. My concern: most of the world is "plagued" with some version of these delusions (superstitions, religion, …), which results in extremely poor decisions at a colossal scale.

[2] Persi Diaconis is a professor of Statistics and Mathematics at Stanford University. This quote was taken from a book called the Improbability Principle by the British Statistician David Hand. I am mentioning it here because, this book provides great insights into the nature of randomness in via various anecdotes. As a result, it is easily accessible to a very broad audience. Everybody loves good, short, quirky, information packed stories.

My research would focus on performing early childhood interventions (ages 4 - 7) in form of RCT's. These interventions need to be administered in early childhood. Once the child crosses a certain threshold age (e.g. 9 years) it may become very difficult to undo what has been forcefully ingrained in one's mind.[3] A typical intervention would administer 'anecdotes' similar to the one I mentioned above for the Law of Truly Large numbers but dumbed down further to a level that a 5-year old can absorb.[4]

Such anecdotes followed by open discussion sessions for a few months might be the treatment of a RCT. Whether such an RCT would detect an effect on future cognitive test scores/ earnings is a matter of speculation. But, the underlying idea is clear, such quirky stories do stick in people's minds. That is precisely what we need to make these ideas more explicit such that they are at people's disposal later in life whenever they need them.

The idea is to make people realize that these ideas are NOT peripheral. Rather, these ideas play a central role in our lives. Understanding how they function is a necessity.

[3] The threshold age of 9 years that I mentioned above is not an empirical estimate. Figuring this threshold age would be a separate research task. Whether the threshold age is 9 years or 10/11/etc years is debatable, but it is not a bad assumption, if we say that past the age of 25 it is very difficult to undo the delusions ingrained in people's minds. In any case, I think early childhood interventions is the best time to start. The time to build the right foundation is when the foundations are laid. When a building is constructed, it may be too late/ too costly.

[4] A fantastic book that focusses on this anecdote-based information spreading is: The Improbability Principle. I am mentioning it here again (the reason for the redundancy is that, I want the reader to really go and get the book and start reading, it is fantastic).