A critique on AI Policy and Governance

Aarsh Batra

July, 2018

Abstract

Superhuman Artificial Intelligence in the coming decades/centuries would be a colossal hit to the society.[1] The challenge lies in nudging the present progress away from the potential catastrophic paths that AI might progress into. Another challenge arises from the lack of past 'real world' examples that would serve as case studies in dealing with such a transformative technology. Such transformative technology probably comes around once every several millennia. This limits the amount of precise real-world analogies that we can draw, pre-post scenario evaluations that we can perform.[2] My research would focus on (1) Articulating comprehensive qualitative arguments for each possible alternate future, with the aim of exhaustively addressing the concerns that arise in each alternate time line. This will be complemented by theoretical models that evolves as new evidence arises. This exhaustive approach is time intensive and requires narrowing down the potential alternate future paths to the most probable ones, subject to the constraint that homosapiens thrive, or at the least survive in each one of them. I will discuss in detail (in section 2) as to how this narrowing down process might be implemented. (2) Increasing public awareness about AI and making the arguments of the AI community more accessible to the common man. This would facilitate a relatively smooth transition into the unchartered territories. The first section of this report briefly introduces the above concerns. Next two sections, address the above two points in turn respectively. Finally, the last (fourth) section of the paper concludes the report.

[1] There is a trade-off b/w accuracy and precision in statements making claims about AI progress. The fact that we are 100% 'confident' that AI is the next big mess to be dealt with is a very good news in some sense. This is a very accurate statement. Not knowing/ignoring that this was the case would have been catastrophic.

[2] It is also worth considering the limitations of our analogies. The inter-temporal comparisons we perform in comparing society of the past with the society in the present does not pin down 'causes'. This is because the groups we are comparing are inherently different to begin with.

I Introduction

The Black swan theory is a metaphor that describes events that come as a surprise, has a major effect and is often inappropriately rationalized after the fact with the benefit of hindsight.[3] What maybe a black swan surprise for a turkey might not be a black swan surprise for a butcher, hence the objective should be to avoid being the turkey by identifying areas of vulnerability in order to turn the black swans white.

In preparing for the future with superhuman artificial intelligence, we need to be the butcher, not the turkey. Comprehensively dealing with and preparing for potentially catastrophic scenarios (that will probably arise in an abrupt fashion at a random time) way ahead of time is the only way to be the butcher.[4] The alternative is to deal with the mess once it happens, which is not acceptable.

Superhuman AI will affect everyone in the society at large. It makes sense that everyone should have a good clue as to what is going on in the AI community. To do that, we need to narrow down the potential future timelines into which we evolve as a society. Then, we perform extensive research within these narrow subset (in which we survive, at the least) of potential futures. This research will streamline the information that needs to be disseminated to the general public at large.

The extent to which we can narrow down the possible futures depends on (1) how collaborative we are as a species (2) how collaborative can we possibly be? (3) how much willing are we (as a society) to trade-off the tempting short sighted race to be the best v/s the long-sighted view of avoiding extinction (4) When we think about the question: how should AI research progress, do we think of it on the level of a company, nation, or as a species? (5) the set of concerns in AI research that everybody agrees on (6) the set of concerns in AI research that invite a lot of debate.

At the least, we can immediately start reviewing the implications of the domain of concerns (resulting from the progress of AI) that we agree on as a species. Given just that, we can start breaking new ground by forming basic norms that will eventually evolve into formal treaties and institutions. These ‘new’ types of collaborations will serve as beacons for future AI based collaborations. AI based collaborations are inevitable and unavoidable part of the future. The benefit of collaborating ‘now’ is to make society used to the dynamics of the future world that they are going to live in. This way, the transition to the transformative future will be a relatively smooth one.

II Possibilities v/s Probabilities

We are time constrained. Simple as this may seem, this is the one fact that we may be implicitly ignoring as AI research progresses. Fierce competition with limited market restrictions is a major driving force (in addition to intrinsic motivation) for the AI community.[5]

[3] The theory was developed by Nassim Nicholas Taleb. Black swan theory is discussed extensively in his 2007 book ‘The Black Swan’.

[4] Although, practically, we may not be able to pin point the time to the nearest second at which a human-level AI will pop into existence. This is the randomness I am talking about. But, our research can pin down a relatively narrow interval within which we would definitely expect a breakthrough.

[5] The free market model is a transformative technology that is largely responsible for all the progress that is happening in the world.

But, when it comes to the market for AI research, we need to be cautious in applying the free market model. The free market model is highly generalizable to a lot of fields. But, AI is idiosyncratic in the following sense: The 'unrestricted race to the top' approach may not be the best path for progress when what lies at the top has the capability of fundamentally destroying us.

When we try to address this issue 'now', the natural counter argument that arises from this goes as follows: *This catastrophic event lies in the relatively far distant future. Moreover, given limited resources at our disposal, there are a lot of 'present' problems that need to be dealt with first*. This argument fails (or deliberately ignores) to consider; (1) AI safety research is extremely time intensive. (2) If the world agrees to be revolutionized by the bittersweet candy that is AI, then whatever mess that follows from such a revolution is a 'present' problem, that needs to be dealt with **now**.

It may be suitable to promote AI progress as a guided adventure into the future rather than a wild exploration. There exists a lot of 'possible' futures, but we are interested in the ones in which we survive (at the least). It is important that we do not drown ourselves in the pool of possibilities and focus on the most 'probable' futures.[6] The good news is that we are in a position to nudge the probabilities of different future scenarios in working towards our common good.

The short-sighted view in which various individual actors race to the top creates more vulnerabilities than it fixes. The unlimited degrees of freedom available to each actor in this model makes it difficult for us to narrow down possible futures. On the contrary, as time passes by, such a model explodes the number of possible futures that we need to plan for today.[7]

What is the solution to this problem of narrowing down numerous potential futures to a handful promising candidates? I present two complementary approaches to the solution below.

First approach: Statistically speaking, the larger the number of entities comprising a particular system, the more ways there are (on average) to mess up the system. In context of AI, the analogy is to focus on collaborations of two types. First type of collaboration focusses on grouping country level AI research under a single governing body. This way we can reduce the number of entities to be dealt with from several thousand (individual companies) to a few hundred. Second type of collaboration focusses on establishing a central institution (or adding new divisions to an existing institution) on an international level. This will govern all the individual national institutions. We already have institutions like these e. g. the UN, which has 193 countries as its members. An advantage of forming central governing institutions at the national level is that it allows us to leverage the already built infrastructure like the UN to accelerate new international collaborations (e. g. by adding a new division for AI governance research).

---

[6] What lies at the end of a wild exploration is relatively more difficult to predict. Why? Because relative to a guided exploration, there are many more ways to mess up and we may not be able to deal with (or predict) all of them given our time constraints. The idea is to structure AI progress to reduce the number of possibilities to the ones in which we survive. An e. g. (with potentially no catastrophic consequences) compares a school recess with a school assembly. With the same number of entities (children) in both cases, the latter is much more manageable and predictable. Wild exploration is fun, there is no denying that. Moreover, the AI community may want to keep it that way. But, the specific future catastrophic consequences of AI make it logical to trade-off some amount of fun today with the future survival of our species as a whole. Remember, we cannot have fun once we are dead.

[7] Moreover, if each actor acts in a manner that involves deliberate concealing of important information, it adds a whole new level of uncertainty to the model.

Second Approach: Forming comprehensive theoretical arguments for various (most probable) future paths for AI (and the society) to progress into. These comprehensive case specific theoretical accounts act as reference 'mental frameworks' for thinking about a problem. This will be implemented graphically in a form similar to a network map and available online. In this map, each node will represent a 'potentially important threshold' (that corresponds to a particular time in a particular timeline). As we go deeper into the map, we move forward in time. A parent node can have multiple children nodes. This means that as we go deeper into the network (i. e. move forward in time) we are able to explore not only a future, but many alternate futures originating from a parent node (which corresponds to a potentially important threshold at a particular time in a particular timeline). A close example of this is node based online graphical system is Wiki Galaxy, which is a 3D visualisation of Wikipedia.[8]

To fill out this map, we need information. Our first move will be to map the current AI literature onto the respective nodes of the network map (Each node is uniquely identified by three parameters; a potentially important threshold, a particular time, a particular timeline). Furthermore, we can standardize the manner in which research papers are written. E. g. a compulsory short section could be added that includes (1) timeline (and corresponding time) addressed in the paper (2) 'potentially important thresholds' addressed in the paper. This type of standardization is not demanding in nature and it will facilitate quick updation and indexing of the AI reference network map.

My research will focus on addressing the above the two approaches of narrowing down (and better managing) the different possible futures to the ones in which we survive with 100% certainty. Next, I will discuss the crucial role of increasing public awareness about the negative and positive implications of AI on the society as a whole.[9]

III Public Awareness and The Wisdom of the Crowd

As far back as the industrial revolution, there have been periodic panics about the impact of automation. Handloom weavers' resistance to new machines earned them a pejorative name---Luddite---that has become a byword for all those who try in vain to stop technological progress.[10] The textile artisans seeking to destroy machines points to the concern of properly addressing the general public as AI progresses.

One of the biggest debates that arises as AI advances is that of mass unemployment. Angst resulting from automation typically focusses on the substitution effect, where the jobs that were once done by humans are now taken over by machines. The anxiety arising out of the future in which the robots will take away jobs is justified to some extent. But, this is not the entire picture. The gloomy view of automation often ignores the fact that, as new technology develops, many jobs become obsolete, but many new types of jobs are created.

David Autor, an Economist at MIT, points out the following example: Between 1980 and 2010, the number of bank clerks in America actually increased despite the rapid spread of the

[8] WikiGalaxy: Explore Wikipedia in 3D. http://wiki.polyfra.me/. (The 'AI reference network map' will function as an online graphically navigable repository (evolving as new research findings arrive) for AI research). It will probably be much more structured and much less densely packed (as 'potentially important thresholds' are sparsely distributed in time) than the 'AI reference network map'.

[9] This topic is worth addressing irrespective of the specific sub-field of AI governance research landscape one is working in.

[10] "Automation Angst: Three new papers examine fears that machines will put humans out of work." The Economist, August, 13, 2015. https://www.economist.com/finance-and-economics/2015/08/13/automation-angst.

'cashpoint'. Although, the IT revolution allowed machines to dispense cash, it also allowed clerks to work out extra financial products that the customers might be interested in and process applications for them.[11] It might be easier to identify disappearing jobs than it is to foresee new ones.

Another important question is: What are the jobs that will remain unaffected by AI (at least for the next few decades)? These may include some of the semi-skilled jobs e. g. the construction worker, gardener, policemen, garbage men, etc.[12]

Other jobs that are still far from the reach of AI are those that demand high levels of creativity and imagination (e. g. a research Mathematician, an Economist, a theoretical Physicist, an Engineer, etc).

Raising public awareness about the various aspects of the argument, avoiding any media sparked lop sided views is one possible solution to the problem of AI Angst.

On an entirely different note, general public can play a very important role in accelerating national and international AI based collaborations. Ethical debates surrounding AI involve asking fundamental unsolved questions about human values. These debates aim to address topics like 'aligning superintelligent AI systems with human values'. These questions have perplexed philosophers for several millennia, but now these debates need a verdict. These open questions are one of the most fundamental hindrances in making more AI based collaborations possible. If we figure out a way to settle this (even to some small but significant extent), we would have solved a very big problem.

In order to address this problem, I was inspired by the idea called 'The Wisdom of the Crowd'.[13] It proposes that the collective opinion of the crowd (aggregated using an averaging technique, e. g. Arithmetic mean, Geometric mean, etc.) is closer to the truth than an individual opinion, especially in questions concerning quantity estimation, general world knowledge, spatial reasoning. Why? Given a large enough sample size, when we record the individual responses of many individuals (e. g. estimating the total number of candies in a jar), we will find that many people overestimate, and many under estimate the actual number of candies in the jar. Once we take the average, the noise resulting from the extremes cancels out and what we are left with is an estimate that is very close to the 'true' (actual number of candies in the jar) answer.[14]

---

[11] David H. Autor. "Why are there still so many jobs? The history and future of workplace automation". Journal of Economic Perspectives, Volume 29, Number 3, Summer 2015, Pages 3-30. https://economics.mit.edu/files/11563.

[12] Pentagon sponsored DARPA challenge to clean the Fukushima Daiichi reactor (which melted multiple times in 2011). These results hints at the current capabilities of robotics in performing some semi-skilled jobs. "Seeking Robots to go where First responders can't." The New York times. April, 9, 2012. https://www.nytimes.com/2012/04/10/science/pentagon-contest-to-develop-robots-to-work-in-disaster-areas.html;
"The Pentagon's fleet of robots may not be so menacing after all" The Washington Post. June, 7, 2015. https://www.washingtonpost.com/news/the-switch/wp/2015/06/07/the-pentagons-fleet-of-robots-may-not-be-so-menacing-after-all/

[13] Aristotle was the first person to be credited with this idea in his famous work "The Politics". But, the classic wisdom-of-the-crowd finding involves point estimation of a continuous quantity. At a 1906 country fair in Plymouth, 800 people participated in a contest to estimate the weight of a slaughtered and dressed ox. Statistician Francis Galton observed that the median guess, 1207 pounds, was accurate within 1% of the true weight of 1198 pounds.

[14] Professor Marcus Du Sautoy. "The wisdom of the crowd" The Royal Society Science events and lectures, November, 29, 2017. In this lecture Professor Sautoy give a live demonstration of the 'count the candies in the jar' example. https://royalsociety.org/science-events-and-lectures/2017/11/wisdom-crowd/. Starting at "10:00

My research would involve working on figuring out ways to leverage the 'wisdom of the crowd' idea in answering the fundamental questions concerning human values. Aggregating human values is not as simple as aggregating the number of candies in a jar. In the candies example we have a 'correct' answer (i. e. we can count the number of candies in the jar) with which we can compare our average results received from the crowd. In case of human values, we do not have the 'correct answer'.

But, given the following: (1) the sample size is large enough (2) we have a robust statistic that captures an individual's value system reasonably well. (3) A suitable averaging method, we can possibly **pin down the 'correct but unobserved answer'** to the various human value problems. Another advantage of this approach is that, the basic premise of the idea is easily demonstrable and understood by everyone.[15]

In my research I will work on developing a robust index for capturing a human's value system (mentioned in point 2).[16] Besides that, I will work on developing suitable averaging methods to aggregate various 'Human Value Index' values for various individuals.

IV Conclusion

Governance of AI is crucial for the successful growth of AI. A structured and guided exploration into the unknown that takes a long-sighted view in which we survive as a society is better than a short-sighted dive into the oblivion. Accepting a guided approach does not imply a complete ban on free exploration, rather it involves trading off some of the free exploration scenarios (and as a result narrowing down possible futures) with the future survival of our species. In all of this, public awareness plays a critical role. An informed public is in a position to choose for itself and collectively ensure a relatively smooth transition into the unchartered territories. Finally, the public's wisdom can be utilized in solving open problems (solutions to which will accelerate international collaborations) that require verdict 'now' and no later.

minutes" into the video, he describes the problem. Towards the end he displays the results. He also presents a wide variety of problems in which this the wisdom of the crowd can be leveraged.

[15] This idea, to me, is exciting and terrifying at the same time. It is easier said than done. But, I think we do not have the luxury of choice here. If we think that a particular approach 'reasonably' addresses a problem that is wide open for a long time and needs immediate resolving, we should at the least start trying it out and see how it goes.

[16] By index I mean a 'measure' of an underlying quantity (as in the Body Mass Index). But there is no 'standardized unit' of human values. Given that, it would require researching other indirect methods to extract out the 'Human Value Index'.