e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

ETHICAL AI AS WE APPROACH SINGULARITY

Dr. Shalini Lamba*1, Vibhu Tripathi*2

*1Asst. Prof., Dept. of Computer science, National P.G. College, India.

*2Student, Dept. of Computer science, National P.G. College, India.

ABSTRACT

AI and AI-based technology are pervading more and more of our lives, and talks on the ever-distant Singularity are becoming more common. But before that point of no return arrives, it is important for us to slow down and make sure the road that we are taking to that point, right now, is built on solid ethical grounds, as that may be what decides whether the system we rely on post-Singularity leads to rapid development towards a utopia, or whether they simply make the system move fast and break things, in doing so bring forth a dystopia.

I. INTRODUCTION

What is The Singularity?

The Singularity is the speculative point-in-time when the rate of technological advancement becomes so high, that it becomes unmanageable and irreversible causing thitherto unforeseen changes to human civilization. Although this point in time can be reached by any amalgamation of technological advancements, the most popular hypothesis among futurologists is called the "Intelligence Explosion'', where an adaptable intelligent agent will almost inevitably enter a "runaway reaction" where it becomes stuck in continuous self-adaptation cycle, and each new output of this cycle will be more intelligent than that of previous cycle. And these newer, more intelligent generation's development time will continue to decline, causing an "explosion of intelligence" which results in a super intelligent agent, an Artificial General Intelligence, which far surpasses all human intelligence, at least in a qualitative sense.**[1]**

This AGI will be much more powerful when compared to the currently used "narrow" or "specialized" AIs, and since it will surpass human intelligence, we have very few ways of making sure that the AGI will see us as his own and equals, thus treat us fairly, rather than as adversaries or servants. One of the few ways is to make sure that on the way to AGI, our "narrow" AIs don't lose their fair nature.**[2]**

So since AI is seen as the most likely way by which the Singularity will occur, AI is also the technology that should be most thoughtfully and ethically developed, because the fallout of the Singularity, and consequently of AI development, can be ginormous. Many scientists and public figures agree with this sentiment **[3]**, yet little is being done to address the ethical problems that are occuring due to the current implementations of AI technologies. And this paper is written to bring these problems forward.

And although there are many arguments against the likelihood of The Singularity occurring **[4,5]**. Focusing on AI ethics is a noble endeavor, as the growth of AI technology and if it can be given a solid ethical ground to take off from, it will go much further.

Why focus on AI?

There are many technology-related ethical issues in our time and all of them do need to be dealt with, but we believe AI is more significant, why?

- Because most of these areas are already in, somewhat mature state, while the field of AI is still in its budding state. And fields in their budding are much more flexible to change.
- AI has great potential for explosive expansion, that explosion can cause serious damage if precautions are not taken. **[6]**

So what are the current ethical issues brought forth by the emergence of AI?

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Emerging ethical issues due to AI

In this section, the ethical issues of human use of AI are outlined. These issues are more or less autonomous - meaning they likely would not have arisen if AI had not been used and some other technology had been used. But do keep in mind that technologies will always need some uses under it's belt before it becomes easy to use/implement, and thus, also become more frequent, and hinder other uses. So the design of technical artifacts have some ethical relevance, so beyond "responsible use", we also need "responsible design" in this field.

- **Privacy and Data Collection**
  - There is general discussion of privacy going on in much of the IT sector, but in this section we wish to talk about how AI affects the trends of privacy.**[7]**
  - Almost all data collection and storage is now digital, our lives are increasingly digital, just about all digital data is now connected to a single network of networks, and there is more and more sensor technology in use that generates reports about the offline parts of our lives. Coupling this with the capability of AI for intelligent data collection and for data analysis, gives a lot of opportunity for comprehensive surveillance of people. This applies to comprehensive surveillance of whole areas of people as well as traditional targeted surveillance. In addition to this, much of the data is traded between agents, usually for a fee.
  - At the same time, the grip on the control of data flow and access continues to loosen.
  - Many new AI technologies amplify the known issues.face recognition in photos and videos allows identification and thus profiling and searching for individuals **[8]**
  - Free services usually rely on the data trail we leave behind when accessing their services, and this collected data is extremely large, especially for one of the "big 5" companies (Amazon, Google/Alphabet, Microsoft, Apple, Facebook).Generalizing all this data requires some kind of learning technology. And as learning tech gets better and better, there may come a day when non-conscious but highly intelligent algorithms know us better than we know ourselves.
- **Behavior Manipulation**
  - The ethical issue of data collection goes beyond just the accumulation of data, it includes the use of this data for nefarious purposes, such as the manipulation of people's behaviors, whether online or offline, in a way that undermines autonomous rational choice of those individuals. Though these tactics of manipulation have been used since time immemorial, by adding AI systems and all the data circulating on the net to the equation, the whole game changes. So let use see in what ways:
  - Many advertisers, marketers and online sellers will likely use any legal means in their arsenal to maximize their profit, and since a comprehensive legal framework is yet to be made in this domain, these people have plenty of unethical tactics to use, including but not limited to - exploitation of behavioral biases, deception, and addiction generation.**[9]**
  - Social Media is becoming a prime location for political propaganda, in which AI bots and recommendation algorithms play a big role in this. Sometime these have such influence that why can be used to steer voting behavior**[10,11]** which - if successful - it may harm the autonomy of the individuals **[12]**
  - Improved deepfake technologies make what, in older times, was reliable evidence, unreliable—this has already happened to virtual photos, audio/video recordings. It will likely soon be quite easy to generate (rather than alter) "deep fake" text, photos, and video data with any given content. And maybe after that, sophisticated real-time interaction with persons over text, phone, or video will also be able to be faked. So we cannot trust the online interactions, while at the same time we are increasingly dependent on such interactions.
  - One specific issue is that learning techniques rely heavily on data, so as most technology and services continue to rely on these techniques more and more, the improvement of these technologies and services will rely on how much privacy the users are willing to give up.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

- **Opacity of AI systems**

With the learning techniques of AI systems, the "learning" or "training" captures some patterns in the data, with or without a correct set of solutions provided (i.e., supervised, semi-supervised or unsupervised training). Then the patterns are labeled in a way that appears useful to the decision the system makes. But not even the designers and programmers know which patterns are used, and as these programs "learn", the patterns continue to change, making the whole system opaque.

That is why AI systems are described as "Black Boxes" where all we know are the data we input and the data it outputs, what processes it applies on the input data remain a mystery. This and algorithmic bias are the central issues in the field of Data Ethics or Big Data Ethics **[13,14]**. These AI systems raise significant concerns about lack of due process, accountability, community engagement, and auditing **[8]**, and these systems are sometimes given the task of decision making, which creates a power structure that reduces the opportunities for human participation, and this may be seen as a good thing by those who think AI systems are better decision makers than humans, which they are, at least sometimes but it will often make it impossible for the affected person to know how the system came to this output, and most of the time not even the system experts will be able to give a comprehensive answer for the system came to that particular decision, which means we will either simply have to have blind faith in the system or reject it completely, and neither are ideal outcomes.**[15]**

- **Bias in AI systems**

A judgment is affected by biases when the individual making the judgment is influenced by an aspect that is not relevant to the matter at hand. The biases need not be explicit, they can simply be passively learned biases and the individual may not even be aware of them **[16,17]**, this last point is very important to see how bias boroughs even into an AI system, which starts out completely unbiased.

As seen in the section above, AI systems detect patterns in data to learn, so the quality of the program depends heavily on the quality of the data provided, but if the data already involves a bias, then the program will reproduce that bias. For a real life example, an automated recruitment screening AI system being used in Amazon (that was discontinued in early 2017) discriminated against women — assumedly because the company already had a history of discriminating against women during the hiring processes.

Here are some major obstacles for impede the fight again algorithmic bias:

- Defining fairness **[18]**
- Complexity of the AI algorithms **[19]**
- Lack of open and transparent commercial AI algorithms **[20]**

- **Other issues:**

There are numerous other issues in the industry of AI:

  - Automation & Employment: Classic automation replaced human muscle, whereas automation in our current Digital Age will likely replace human brain processes, such thoughts, or information-processing—and unlike physical machines, digital automation being virtual in nature,will likely be much cheaper **[21]**. It may thus mean a more radical change on the labor market. So, the main question is: will the effects of automation be different this time round? Will the generation of employment and wealth keep up with the reduction of employment that will be caused by AI? And if it is not different, then what are the costs of transition, and who will the people who will bear them? Do we need to make systemic societal adjustments to allow for fairer distribution of costs and benefits of automation?
  - Autonomous Systems: The main issue here is if some incident occurs due to involvement so will bear the responsibility. A subset of this is Autonomous Weapons, do they make warfare worse or better?
  - Machine Ethics
  - Singularity: Some don't see this as an issue but as a boon for humanity, while others see this as an existential

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

threat for humanity, against which major precautions must be taken. Regardless of which stand one may take, it is better to prepare for the Singularity than not.

- And so..

  So as we have seen above the issues in which AI is involved are numerous, but this does not mean AI as a technology should be abandoned (if that even is possible), but that there is large room for growth in this industry.

Attempts to deal with these issues

As said before, this paper is not to take interest away from the field of AI but to direct it in more more productive areas of AI, that is why all the current issues of AI need to be talked about more.

- **Privacy and Data Collection**

Privacy-preserving techniques that can, for the most part, conceal the identity of persons and groups are becoming the standard staple in Data Science. This includes **[22]**:

- (relative) anonymisation
- access control (plus encryption)
- computations are being carried out with fully or partially encrypted data
- This requires more cost and effort but can avoid many of the privacy issues. Also the need for more privacy can lead to companies competing over who has the most privacy.

- **Behavior Manipulation**

- In this field, there have been attempts but, for the more part, nothing much has been done. The EU has strengthened its privacy regulation and how the data is used, but the USA and China prefer growth with little to no regulations in this field, likely in the hope that it will give them a competitive edge over the other.
- A lot of work here has been done by private companies but many of them, themselves abuse AI technology **[23]**
- Many academic initiatives have also been taken in this field **[24]**

- **Opacity of AI systems**

To deal with this issue, a whole AI field of **Explainable AI(XAI)** has been established, their goal is to make AI systems whose results of a solution can be understood by humans.**[25]** XAI is seen by some as an implementation of the social right to explanation **[26]** , but XAI is relevant even if no legal or regulatory requirement is placed on the organization, as an XAI is much more likely to be trusted by humans, which in return will benefit the organization. The goals of an XAI are: **[27]**

- explain what has been done
- what is done right now
- what will be done next
- unveil the information the actions are based on
- Other than XAI, there have been regulatory actions taken that states that consumers, when faced with a decision that is based largely on data processing, will have a legal right to an explanation for the judgment they were given.

- **Bias in AI systems**

Although this field has come into the limelight relatively recently, much work is being done on this. Some ethical guidelines were established for AI systems. Eleven clusters principles of AI were found: solidarity, transparency, justice and fairness, freedom and autonomy, non-maleficence, responsibility, privacy, beneficence, trust, sustainability, dignity **[28]**. Among these fairness and "mitigation of unwanted bias" was a common point of concern, and were recommended to be dealt with through a blend of technical solution, transparency &

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

monitoring, right to remedy & increased oversight, and diversity / inclusion efforts.

- **Singularity**

Although many treat the Singularity as only a theoretical concept, some serious efforts have been in the field of **AI Alignment** or **AI Control Problem**. They deal with how to create AI systems that will aid us, their creators, rather than harm us. This will take the gas out of the issue of the existential dread of the Singularity. Although the name of the field has "AI" in it, this largely deals with AGIs, because the current weak AI systems can easily be monitored and shut down if they get out of hand.

The major approaches of this are: **[29]**

- Alignment: which aims to align the AGI's goal systems to human values
- Capability Control: which aims to reduce an AGI's capacity to harm humans or gain control. This also includes a kill switch for the AGI if it goes out of hand which brings up the issue of how much dignity sentient AGI should be treated with, the central problem of AI rights.

II. CONCLUSION

An algorithm for machine learning is made by fallible humans, so it is only likely that some issues will occur during any or all of the phases of development (designing, programming, calibrating and evaluating the algorithm's performance). And even after the development phase, the system is deployed, distributed and used by humans, so the issues will always arise in some form. We have tried to mention the main current issues as well as some long-term future issues. But our main conclusion is about **cultivating responsibility**

Cultivating Responsibility

Despite all the great attempts by the many institution and individual in the field of AI, we believe the key piece to solving the puzzle of making more ethical AI is to enable and encourage AI developer (budding ones and experts) to understand that the technology they work on is very deeply intertwined with many current and long-term future ethical issues, and that, as developers, they have a vital role and responsibility to engage in the ethical consideration of this technology that they work on. This belief that because this is a "technical" field, the technology that comes out of this field is "neutral" and so the experts of this technology need not deal with its ethical implications is false, because the ethical implication is fundamentally embedded in the selection, design, deployment and use of any technology.

As doctors may take the Hippocratic Oath, so should AI professionals take their own form of Oath. While it is not panacea, it does serve as a statement of and commitment to the social contact between a professional and the public, a reminder to a professional's ethical obligation to the public.

III. FUTURE ASPECTS

AI is modifying our daily lives in forms which are hard for us to predict or understand. If the AI technologies are to be guided in a more socially responsible way, then it is time to dedicate some time and attention to the education of the public and developers in the field of AI ethics. Not only for AI, but it is important for the computing community, as a whole, to more resolvedly embrace ethics as a part of its core identity. There is even a practical reason to do so, as jobs are starting to emerge in the realm of AI ethics. It has been suggested that some companies should consider having a chief artificial intelligence ethics officer. We hope that this is at least partially a sincere effort towards taking ethics as a field of importance, rather than an attempt at "ethics washing". Because ethics is a field of importance and we think more of the community needs to recognise and a pathway towards increasing that possibility is assuring that ethics has a central place in the educational efforts of AI.

IV. REFERENCES

[1] "Collection of sources defining "singularity"". Singularitysymposium.com.
[2] Article at Asimovlaws.com
[3] Sparkes, Matthew (13 January 2015). "Top scientists call for caution over artificial intelligence". The

Telegraph (UK).

[4] Paul Allen: The Singularity Isn't Near, David Chalmers John Locke Lecture, 10 May, Exam Schools, Oxford, Presenting a philosophical analysis of the possibility of a technological singularity or "intelligence explosion" resulting from recursively self-improving AI

[5] Hagendorff, Thilo (2020). "The Ethics of AI Ethics: An Evaluation of Guidelines". 7. Müller, Vincent C. (forthcoming 2021), Ethics of Artificial Intelligence.

[6] Whittaker et al. 2018 AI Now Report 2018 Costa and Halpern 2019 The behavioral science of online harm and manipulation, and what to do about it .

[7] Samuel C. Woolley and Philip N. Howard Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media.

[8] Samantha Bradshaw, Philip N. Howard, Bence Kollanyi, Lisa-Maria Neudert Sourcing and Automation of Political News and Information over Social Media in the United States, 2016-2018

[9] Susser, Roessler, and Nissenbaum 2019 Technology, Autonomy, and Manipulation

[10] Floridi and Taddeo 2016 What is data ethics?

[11] Mittelstadt and Floridi 2016 The Ethics of Big Data

[12] Inside The Mind Of A.I. - Cliff Kuang interview

[13] Graham and Lowery 2004

[14] Binns 2018

[15] Friedler, Sorelle A.; Scheidegger, Carlos; Venkatasubramanian, Suresh (2016). "On the (im)possibility of fairness". arXiv:1609.07236.

[16] Sandvig, Christian; Hamilton, Kevin; Karahalios, Karrie; Langbort, Cedric (2014). Gangadharan, Seeta Pena; Eubanks, Virginia; Barocas, Solon (eds.). "An Algorithm Audit" 20. Seaver, Nick. "Knowing Algorithms"

[17] Bostrom and Yudkowsky 2014 The Ethics of Artificial Intelligence

[18] Stahl and Wright 2018 Ethics and Privacy in AI and Big Data

[19] Fiegerman, Seth (28 September 2016). "Facebook, Google, Amazon create group to ease AI concerns".

[20] "New Artificial Intelligence Research Institute Launches"

[21] Janssen, Femke M.; Aben, Katja K. H.; Heesterman, Berdine L.; Voorham, Quirinus J. M.; Seegers, Paul A.; Moncada-Torres, Arturo (February 2022). "Using Explainable Machine Learning to Explore the Impact of Synoptic Reporting on Prostate Cancer".)

[22] Edwards, Lilian; Veale, Michael (2017). "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For". Duke Law and Technology Review. **16**: 18. SSRN 2972855

[23] "Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms".

[24] Jobin, Anna; Ienca, Marcello; Vayena, Effy (2 September 2020). "The global landscape of AI ethics guidelines". Nature. 1 (9): 389–399. arXiv:1906.11668. doi:10.1038/s42256-019-0088-2. S2CID 201827642

[25] Verfasser., Christian, Brian 1984- (January 21, 2021). The alignment problem : how can machines learn human values.