The Ethics of Artificial Intelligence

Akshan Raina, 18th August 2023

Abstract

The advent of Artificial Intelligence has raised many questions about the ethical implications thereof. This paper begins by explaining intersection between Artificial Intelligence and Ethics and the necessity of examining this. Through analysis of many economic and social consequences of Artificial Intelligence, this paper develops some ideas concerning political and economic solutions to the foreseen obstacles. With the implementation of these ideas (or any protective measures), the paper concludes that Artificial Intelligence is, or at the very least, has the capability to be ethical. It can even be said that Artificial Intelligence has the potential be ethically good – such will be shown through the positive consequences outlined below (and the assumption of the prevention of the negative implications).

Introduction

§1. The need to consider 'The Ethics of Artificial Intelligence'

As technology continues to advance, it is inevitable that Artificial Intelligence (AI) shall become increasingly prominent in our modern society through its gradual incorporation into many additional aspects of our lives. While the integration of AI with our lives will be done

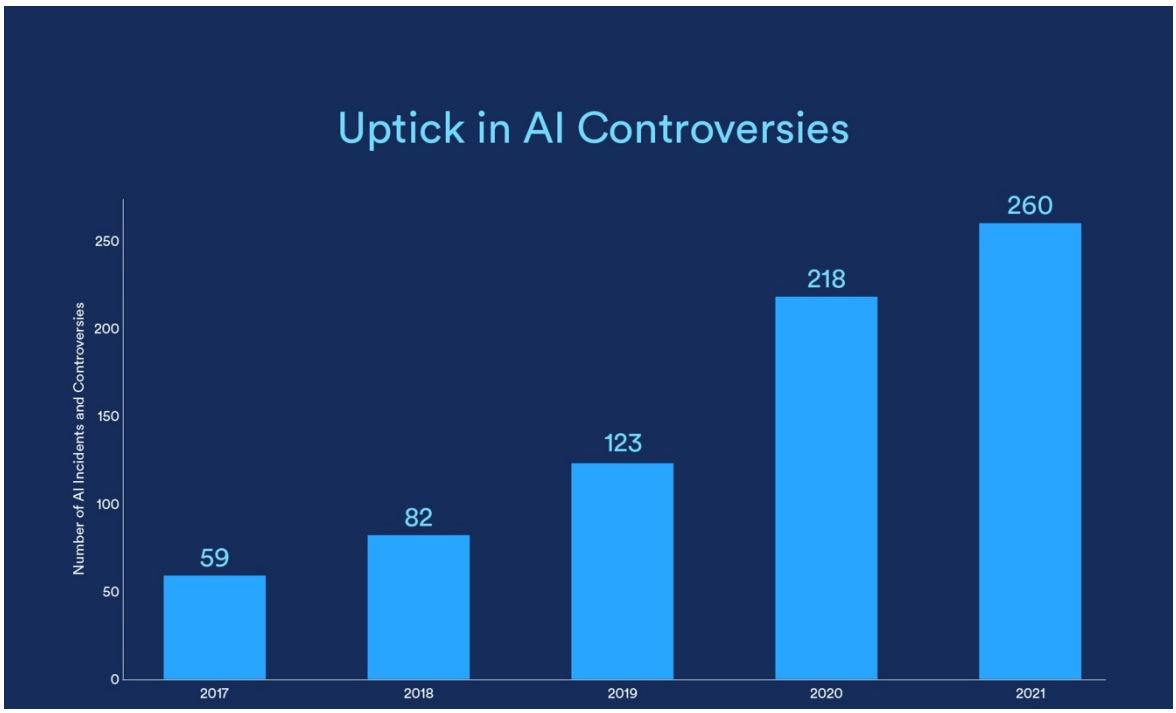


Figure 1. A Chart to show the increase in reported AI-related incidents (2017-2021) [1]

with the opportunities in mind, it is undeniable that these opportunities shall also be accompanied by a number of challenges.

[1] AIAAIC Repository (2022). Available at: https://docs.google.com/spreadsheets/d/1Bn55B4xz21-_Rgdr8BBb2lt0n_4rzLGxFADMlVW0PYI/edit#gid=888071280

One of the challenges will be the consideration of ethics in matters concerning AI. Due to its inanimate nature, a common concern is whether AI can be constructed in such a way as to maintain the ethical integrity of a sentient human. This is of greater concern now than ever before due to the rapid growth of the research and development occurring as pertains to AI. Furthermore, as AI continues to become more advanced - due to its nature tending towards independence - it also becomes more difficult to control as it will lose considerable transparency (due to the more complex processes it shall employ). This accelerated development of AI (and subsequent rapid adoption of AI into society) warrants the immediate necessity to consider the ethics of AI, enabling greater control and understanding for the future.

This is evident from Fig 1. (compiled from data on the AIAAIC [AI, Algorithmic, and Automation Incidents and Controversies] Repository), which displays the sharp increase in the amount of reported AI-related incidents. The basic correlation suggests that, as time goes on, the amount of AI controversies increases (depicted by the steadily growing bar). Aligned with what was written in the paragraph above, the advancement of AI, combined with its greater prominence in society over time shall invariably lead to a further (continued) increase in amount of 'controversies'. Thus, it is important to understand the implications of AI, so as to be able to begin seeking solutions that may maintain (rather, instill) the ethicality of a human.

§2. Paper Format

This paper shall begin by outlining many of the major concerns regarding AI such as bias, privacy and accountability (alongside many others); these will swiftly be analysed from an ethical perspective. Following the many negative implications, a few positive consequences shall also be briefly outlined in order to show the (AI's) capacity for 'goodness'.

The next section of the paper shall delve into some theoretical solutions to create and sustain an ethically good state (concerning AI). Solutions will range from political ones that seek to prevent the problem from its roots to the use of economic instruments in an attempt to offset any negative externalities.

After considering the ethical management of AI through political and economic policy, the paper will culminate with a conclusion that determines whether or not Artificial Intelligence is an ethically viable construct – or how to make it so.

Brief Background

§1. 'The Ethics of Artificial Intelligence'

Before even considering to ponder this paper title, it is of utmost importance to have a sound grasp of the terms of the statement. Below the main terms shall be concisely defined and explained. There are two aspects to consider i) 'Artificial Intelligence' – this shall be defined, and as it is of particular importance, briefly explained as well; ii) 'Ethics' – this shall be defined and subsequently framed in such a way that suits the purpose of this paper.

Naturally, both topics have innumerous intricacies and are by no means simple, however, they shall be explained swiftly and at a high level as an act of abstraction to retain the comprehensibility of this paper to all.

§1.1. Artificial Intelligence

Before considering ethical implications and management solutions, it is important to fully understand what Artificial Intelligence actually is. AI refers to the humanisation of machines or computer systems through the simulation of human intelligence - abstract as it may be (McCarthy 2007). AI aims to mimic human cognitive functions with the hope that it will be able to complete tasks much more efficiently and effectively than humans ever could. The launch of OpenAI's 'ChatGPT' on 30th November 2022 led to the increased popularisation and universal recognition of the term 'Artificial Intelligence'.

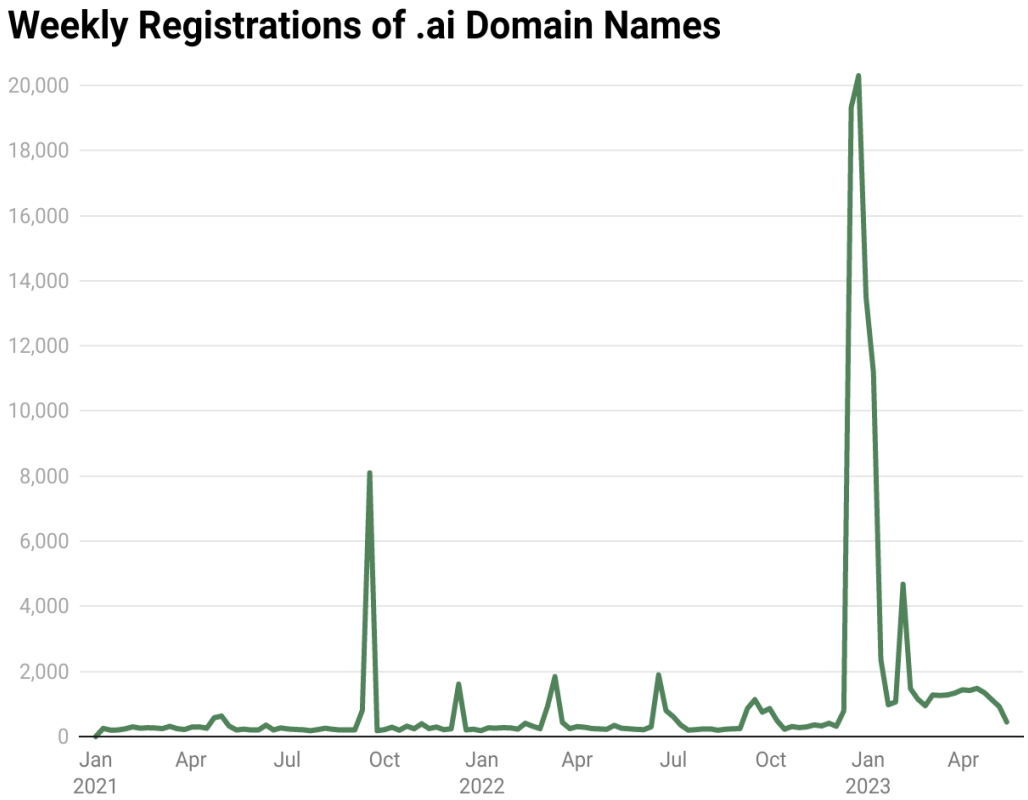


Figure 2. A Chart to show the number of Weekly Registrations of Domain Names that end in '.ai' (2021-2023) [2]

This is evident in the Fig. 2 which clearly displays the sharp increase (going from under 1000 to over 20000) shortly after the launch of ChatGPT. This hype has brought a lot of traction and attention to the field of AI of late. Hereby accelerating development and advancement of AI (through increased research within the area).
The reason for the interest in the development and advancement of AI stems from the broad array of opportunities it presents for society. Alongside being driven by the practical usage it brags, there is also an aspect of scientific curiosity (a further interest in testing the limits and capacity of complex advanced AI models).

[2] Newcomer (2023). Available at: https://www.newcomer.co/p/14-charts-that-tell-the-story-of

§1.2. Ethics

Ethics is a branch of philosophy that concerns itself with questions of morality. It focuses on the ideas of ‘good’ or ‘bad’ and ‘right’ or ‘wrong’. Ethics is often used to evaluate human decisions through the consequences and implications. In a similar manner, it will be used in this paper to evaluate Artificial Intelligence (through its consequences and implications) from the standpoint of whether it is an ethically (alternatively, morally) ‘good’ concept. With this in mind, for the purpose of this paper, it can be taken as a ‘measure of goodness’ (undeniably an oversimplification of this complex abstract concept yet fitting for the paper nonetheless).

Ethical Concerns with regards to Artificial Intelligence

§1. Negative Impacts

Invariably, there are many problems (or negative impacts) associated with the use of AI. Many of these are outlined and explained below.

§1.1 Bias

One of the major concerns when it comes to AI is ‘bias’. ‘Bias’ refers to the production of outputs by an AI system that reflect unjust prejudices. Although AI is expected to improve decision-making due to its objective nature – brought about by a lack of sentiment, the presence of bias significantly detracts from the reliability of any such model.

Biases are most commonly brought about by the use of a biased dataset in the training stages of an program. AI systems learn from historical data - if this data contains biases, the AI will inadvertently perpetuate those biases in its own results. This, in turn, can be brought about by biased collection methods – a simple example being underrepresentation of an ethnic group through the lack of surveying said group.

Bias within AI has several societal consequences. Firstly, it can lead to discrimination – through the perpetuation of historical patterns, the AI system will effectively be living in the past and so will also perpetuate historical inequalities. Bias could also lead to the reinforcement of stereotypes through the amplification of any biases present in the training data.

Needless to say, this is harmful to societies and can detract from the ideal of equality which is currently being sought.

§1.2 Privacy

Another grand concern is privacy (or the lack thereof). Many people don’t even realize how much data of theirs is being collected and used by AI systems. This means that there is a lack of consent – disabling individuals from extending control over their own personal information.

Secondly, the vast amount of data being collected means that a data breach would be much more severe. Possible repercussions being: identity fraud, financial losses and damaged reputations (via identity misuse).

Finally, AI has the capability to create highly detailed profiles of people with these vast pools of information it can receive. Through this profiling, AI can theoretically predict future actions. This is known as predictive analytics and was incorporated by statistician Andrew Pole at Target, enabling Target to identify pregnancies before the (soon-to-be) mothers themselves (Duhigg 2013). It is evident how this can be seen as an invasion of privacy. Furthermore, it also opens many possibilities for manipulation, after all, would you feel safe if a corporation know your future before you lived it?

§1.3 Accountability

Accountability is yet another of the major concerns. It refers to the attribution of responsibility for the actions, and consequent outcomes, produced by AI technologies. There are many problems associated with accountability. Primarily, AI often operates via 'opaque decision-making', this refers to a black box approach, wherein inputs go in... and outputs, out. Essentially, the operations that occur, and decisions that take place cannot be easily understood. If it is difficult to ascertain why certain decision are made, it would also be difficult to ascertain the whole picture (of what may have occurred and why), hence one could not determine culpability.

Secondly, there can often be 'emergent behaviours'; since the AI is learning from a dataset, it could potentially pick up on patterns that weren't intended. Hence "behaviours" and outcomes "emerge" that were not originally expected. This was portrayed in the recent film M3GAN (Scott 2022) wherein she developed an unintended aggressive nature due to an unforeseen emergent behaviour. Here it was evident that emergent behaviours can be dangerous due to the random nature of the adopted behaviour(s). Once again, due to the non-sentient nature of the AI, it remains to find someone culpable – yet no-one actually programmed the aggressive behaviour; once again, leading to an accountability dilemma.

Finally, expanding on this accountability dilemma, AI systems are mostly brought about through much collaboration between many different parties – such as stakeholders, developers, and the dataset providers. The sheer size of the team behind the AI models (and breadth in differences in their roles) makes it near impossible to pin the blame to a single team, let alone individual. Determining who should be held accountable for unintended outcomes can become even more complex in larger-scale contexts.

§1.4 Job Market Disruption

AI is being developed with the hopes that it will be able to perform many of the tasks of humans, yet with near-perfect accuracy and efficacy. A consequence of this will be the complete upset of job markets globally. As more jobs become automated, it leaves many more people (and their skills) redundant. In order to survive in this economy, the displaces people will have to seek new skills and find a suiting new job. However, how long until AI advances enough to consume those roles too?

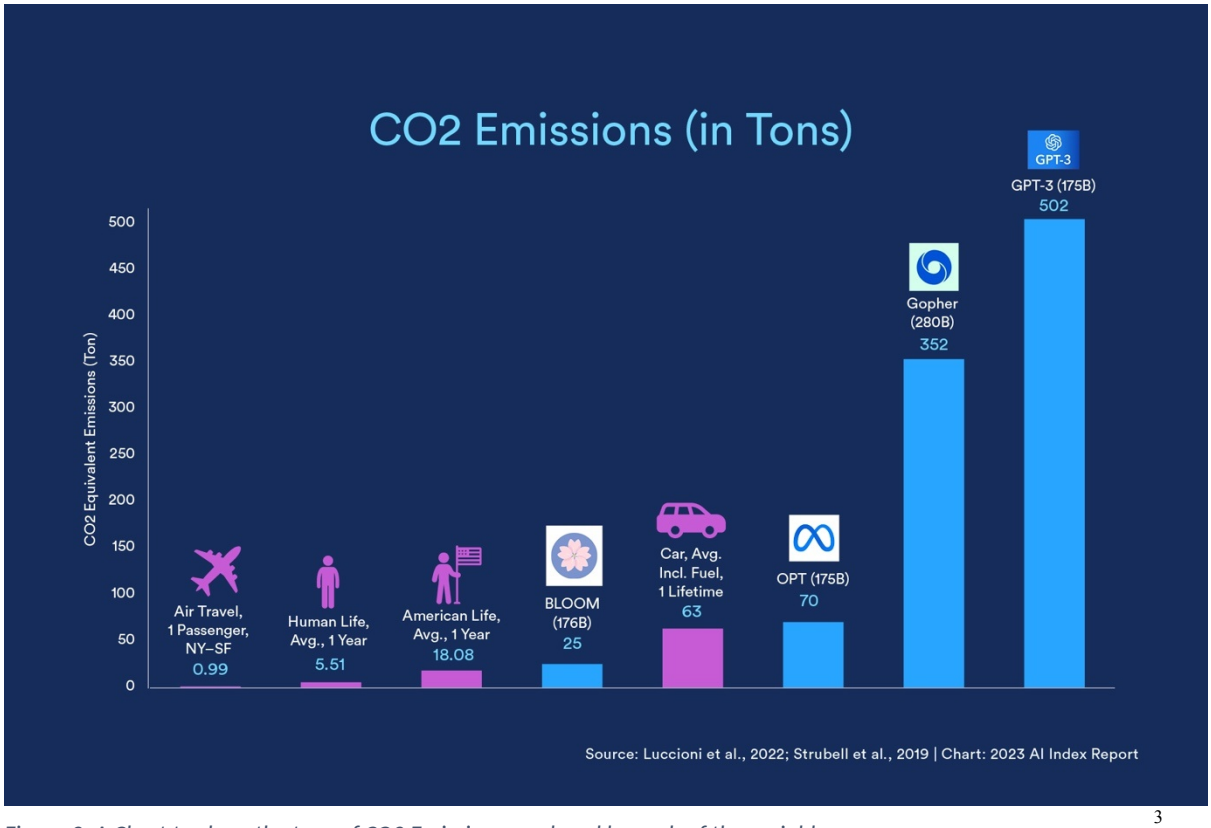
Looking ahead, a loss of jobs en masse would also lead to decreased consumer spending (as people have less money to spend). This decreased spending can cause an economy (and its GDP) to shrink; thereby igniting the harmful flames of a recession which could set an economy into a vicious positive feedback loop (recession leading to more people losing jobs, then less spending and a shrinking economy, and so on, ad perpetuum). Invariably this would

place a much larger percentage of a population into poverty – certainly not a desirable outcome.

§1.5 Environment

The environment is a finite resource that must be looked after and cared for. AI has many negative externalities on the environment. Naturally, the running of such 'high-tech' machinery requires great amounts of energy. This is not such a great worry, provided the industry uses renewable energy resources – else, it would amount to vast carbon emissions released into the atmosphere; in turn, this severely damages the ozone layer through the enhanced greenhouse effect.

Secondly, the production of AI devices requires the extraction of rare minerals (for components of the AI). This can have adverse environmental and social impacts in mining regions. For example, sand is currently a depleting resource (Peduzzi 2014). Sand is used in the manufacture of computer chips and, remains finite. It may seem as though there is no shortage for sand, but desert sand cannot be used due to the difference in construct. This extraction of sand has many further impacts for the environment - such as the erosion of land and destruction of habitats.



[3]

Figure 3. A Chart to show the tons of CO2 Emissions produced by each of the variables

Finally, the training phase of AI models, requires extensive computational resources, can result in significant carbon emissions if not managed efficiently. This is evident from Fig 3. Which visually displays the sheer amount of CO2 emissions expended by GPT-3 in comparison to average American metrics (such as 1 car's lifetime). This shows how

[3] Stanford University HAI (2023). Available at: https://hai.stanford.edu/news/2023-state-ai-14-charts

unsustainable such approaches are, and demands greater effort be put on making such processes eco-friendly.

The environment is of great concern and must be addressed before the situation gets too dire. Currently, AI is causing many unnecessary negative environmental externalities which can be offset by the incorporation of eco-friendly practices.

§1.6 Social Manipulation

Social manipulation through the use of artificial intelligence (AI) has become a concerning phenomenon with far-reaching implications. As mentioned above, AI-powered algorithms can analyse vast amounts of personal data from all sorts of platforms to create detailed profiles of individuals, enabling highly targeted and often persuasive content delivery. One notable case illustrating the potential consequences of such manipulation is the Cambridge Analytica scandal. In this instance, personal data from millions of Facebook users were harvested without their consent and used to create tailored political messages to influence voter behaviour. This manipulation exploited psychological vulnerabilities and highlighted how AI-driven tactics could sway public opinion, impact elections, and undermine the integrity of democratic processes. The incident serves as a stark reminder of the emphasis the needed for robust regulations, data privacy measures, and public awareness campaigns in an effort to mitigate the potential use of AI technologies for nefarious purposes.

§1.7 Academic Integrity – loss of meritocratic society

The creation of tools such as ‘ChatGPT’ has led to worries about academic integrity. Tools such as this have reduced hours of hard work into the simple click of a button. There is greater emphasis now than ever to maintain ‘academic integrity’ as students. While this doesn’t seem like a grand concern in respect to a few of the other topics that have been outlined; it is very important to consider the ramifications of this.

Tools such as ChatGPT discourage hard work and effort, this could result in a generation that is dependent on technology and cannot think of its own. Being reliant on technology is certainly a bad thing as technology can be subject to events such as thunderstorms or blackouts.

The gradual loss of academic integrity may also seem small on the grand scale of things; however, it bears the consequence of the loss of the meritocratic society. Once again, discouraging hard work and effort over laziness and inactivity.

§1.8 Robotics

AI is also closely related to the field of robotics (which focuses on the intersection between engineering and computer science and concerns itself with robots). Through a combination of AI and robotics, concepts such as self-driving cars are possible. Self-driving cars are very morally ambiguous; one of the most interesting considerations to be made is how it would approach that of an ‘ethical dilemma’ – for example, would it save passengers or passers-by, how would it allocate value to different lives?

Another area of interest within the field of robotics is the idea of ‘killer-robots’ ie. autonomous drones and perhaps even a real-life Robocop. Once again, this reverts to the

problem of accountability, since these machines will be operating of their own, who is culpable for the results of their actions?

§1.9 Deepfakes

These AI-generated synthetic media, most often in the form of videos, can be so convincingly realistic that they can make it appear as though individuals themselves are saying or doing the things in the video (when, in reality, they never did). While they may seem like a bit of harmless fun, they must be approached warily. The ease with which they can be created raises concerns about misinformation and deception. Deepfakes can be exploited for various malicious purposes, including spreading fake news, defamation, and identity theft. For example, recently, there was a deepfake of President Zelenskyy[4]. While it may not have been the most realistic, as time goes on, the deepfakes will only become more difficult to distinguish from reality. This is rather disturbing, as if someone were to impersonate a head of state (as done here) without anyone realising it was a mere deepfake, it could have immense consequences.

§2. Possible Positive Implications

Having considered at length the negative implications, to give a whole view, it is necessary to mention some possible benefits. Naturally, there are many positive implications – embodied by the opportunities presented, however, as is fitting with the paper, the focus shall be on ethical benefits.

§2.1. The Judicial System

Logically thinking, AI seems to be very fitting for judicial tasks. Due to its non-sentient nature it stands to reason that it shall be capable of providing objective judgements based purely off the facts of a given case. The use of AI would also bring a regularity to judicial sentences. The current situation (pertaining to the sheer volume of judges [and juries] globally) invariably leads to slight inconsistencies due to different subconscious biases that they may be harvesting.

Talking of bias, it is necessary to address bias within AI; Surden (2019) states that AI is unsuited to make such decisions (yet) due to bias. Earlier it was seen that a biased dataset can lead to perpetuation and magnification of this bias. In this way, stereotypes could be enforced – meaning an unjust judicial system.

On the other hand, Surden (2019) points out the ability of AI to assign a “likelihood to reoffend score” which has proven useful to judges. This reflects the process-driven success that AI can have.

Similarly, AI could be useful in the clerical aspect of law (of which there is much). Legal research and analysis often take up great amounts of time for solicitors and lawyers, yet, an AI system would have the capability to accelerate this by quickly sifting through vast amounts of legal texts, court cases, and precedents to provide lawyers with relevant information and insights.

---

[4] New Scientist (2022). Available at: https://www.newscientist.com/article/2350644-deepfake-detector-spots-fake-videos-of-ukraines-president-zelenskyy/

Secondly, an AI system would be able to manage resources much better and in a much more organised effort than a human (with the same efficiency). In this way, lawyers would be able to access and identify key information, potential risks, and discrepancies much more quickly and accurately (due to the abstraction of irrelevant documents).

Finally, and AI system would be able to automate many of the routine and mundane that a lawyer normally has to do - such as document drafting, data entry, and legal documentation – this would give a lawyer more time to focus on urgent aspects of a case and enable them to direct their efforts more complex and strategic work.

§2.2. Healthcare

AI also has great utility in the field of healthcare and medicine. Once again, it can bring automation to many aspects of the sector. Healthcare is notorious for the heavy paperwork than must be completed and organised. An AI system would be able to streamline all administrative tasks (incl. billing, appointment scheduling, and record keeping), relieving healthcare staff of these administrative burdens. This prevents such staff from being severely overworked and ensure they can direct their time and efforts towards wholeheartedly looking after patients.

AI can easily automate analyses as well due to its statistical aptitude. A good example is ‘medical imaging analysis’ wherein the AI model will analyse a medium such as an X-Ray to search for any abnormalities, tumours, or diseases; the AI model could also be taught to perform ‘disease diagnostics’ by reviewing a patient’s medical history in order to provide accurate diagnoses for the patient. Furthermore, using AI to complete such tasks will lead to a consistent accuracy (presuming the AI is has been trained correctly) due to the objective nature of its determinations. This once again relieves healthcare staff of these time commitments allowing them to channel their energies elsewhere.

Furthermore, (relatively) recent advancements in wearable technology can incorporate AI systems to facilitate remote monitoring. These devices can collect and compile data in real time, alerting healthcare professionals should anything indicate the requirement for medical attention. As is known, hospitals and medical centres are incredibly busy places, in this way, one would be able to drastically decrease a hospital’s patient turnover time without the worry of sacrificing the quality of care for any patients.

Finally, if methods similar to those used by Target (Duhigg 2013) were to be implemented in the healthcare sector, diagnoses could be made ahead of time (thereby increasing the chances/ time of successful treatment. This is especially useful in the case of progressive diseases such as Huntington’s Disease, Dementia, MND (Motor Neurone Disease) and Muscular Dystrophy. Evidently, this could be particularly prominent in the identification of neurological disorders - a niche in which a particular problem is late diagnoses that don’t allow for much preventative measures to be taken (as there don’t tend to be cures for such neurological disorders) . In contrast, an AI model that learns to detect such diseases through the analysis of genetic data (genomic analysis) and tracking of habits (through remote monitoring as detailed above) would be invaluable to human society. It could save innumerable lives and increase the standard of living (as there aren’t any well-known cures for many such diseases yet) for countless more.

§2.3. Equality (and Education)

While not strictly a consequence of AI, it has facilitated the diversification of the computer science landscape. A work(/research) landscape that was predominantly white male (Fig. 4, Fig. 5, Fig. 6) has gradually eroded over time. Fig. 4 and Fig. 5 illustrates the tendency towards gender equality; Fig. 4 shows the growing interest in CS Bachelor's Degrees from female candidates, Fig. 5 displays the increased recognition for females withing the field (a story told by the number of faculty hires) – alternatively, it could also be a consequence of the increased concentration of females entering the field (Fig. 4); another win for gender equality, nonetheless.

Secondly, Fig. 6 shows the ethnic infiltration of the Computer Science Scene (admittedly, predominantly Asian). This is another positive as it [Fig. 6] highlights the growing interest within the Computer Science sector from other ethnicities. Fig. 4, Fig. 5 and Fig.6 collectively show the diversification of the work environment – a positive and desirable effect.

Concerning equality, AI also presents the opportunity for equality of opportunities – this is well-represented by the education sector. AI tools and platforms (eg. ChatGPT) can provide personalized learning experiences, giving everyone access to the similar resources and teaching. This can narrow the educational achievement gap by ensuring that students receive the support they need to succeed, regardless of their background.

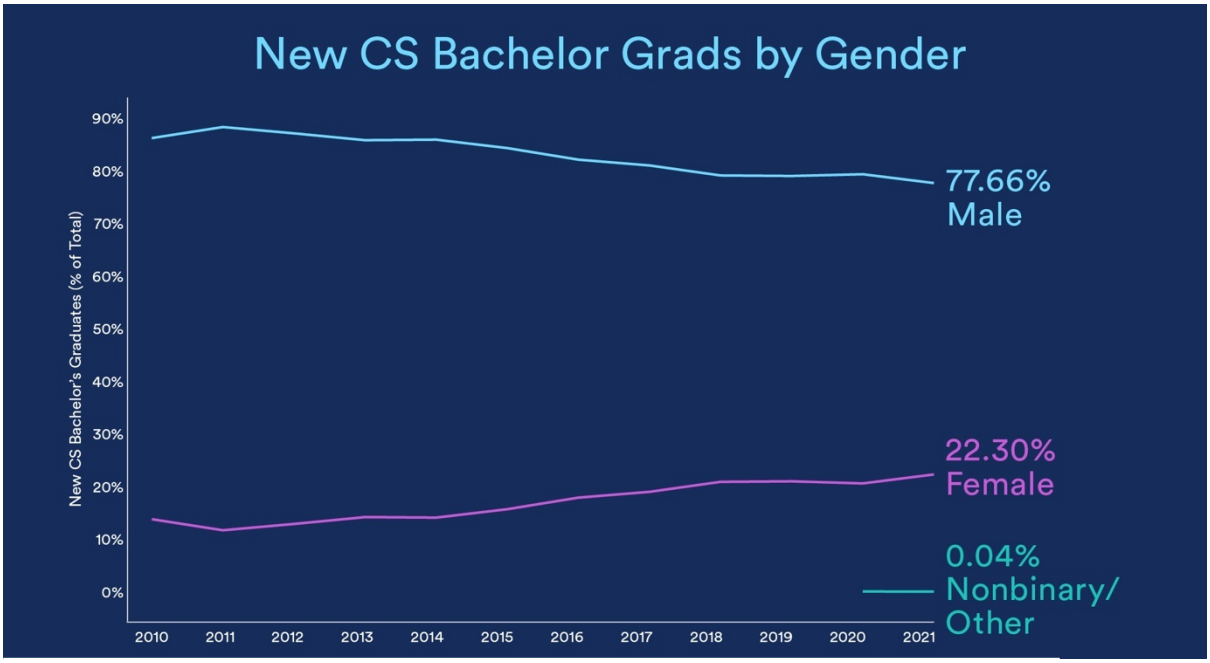


Figure 4. A Graph to show the percentage of Computer Science Bachelor's Degree Graduates by Gender (2010-2021) [5]

[5] CRA Taulbee Survey (2022). Available at: https://cra.org/crn/wp-content/uploads/sites/7/2023/05/2022-Taulbee-Survey-Final.pdf

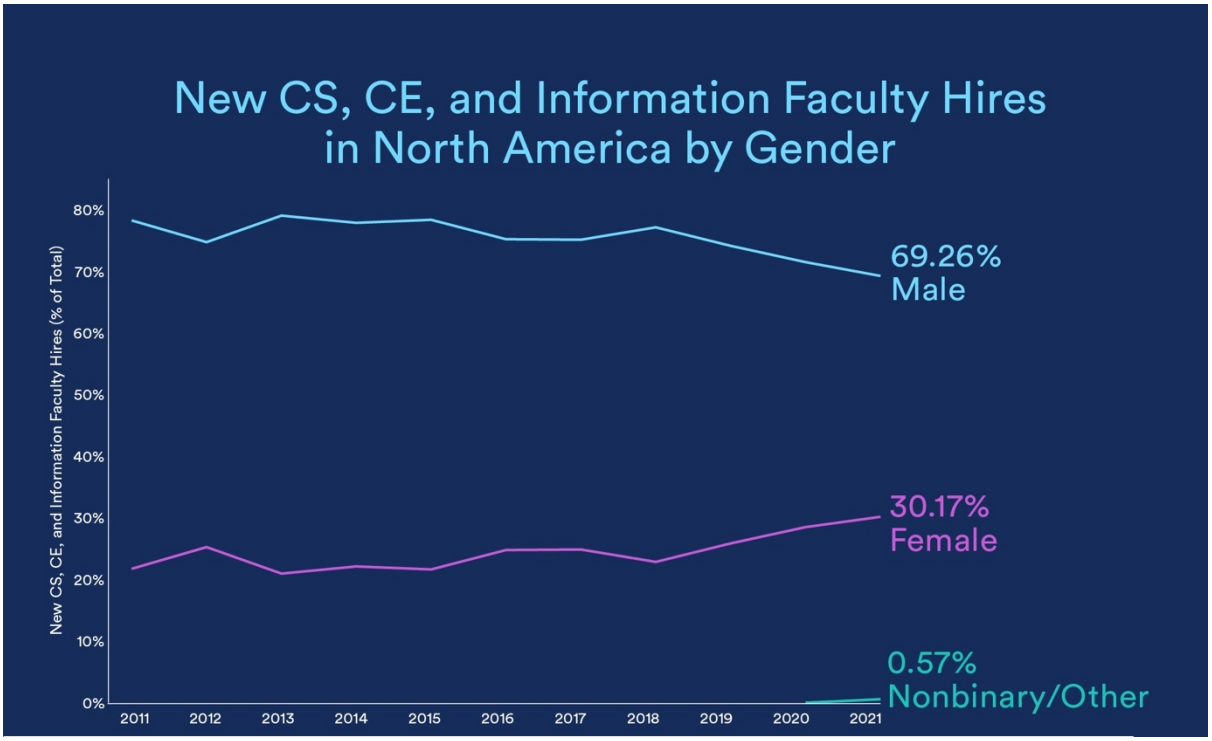


Figure 5. A Graph to show the percentage of new Computer Science related Faculty Hires by Gender (North America, 2011-2021) [6]

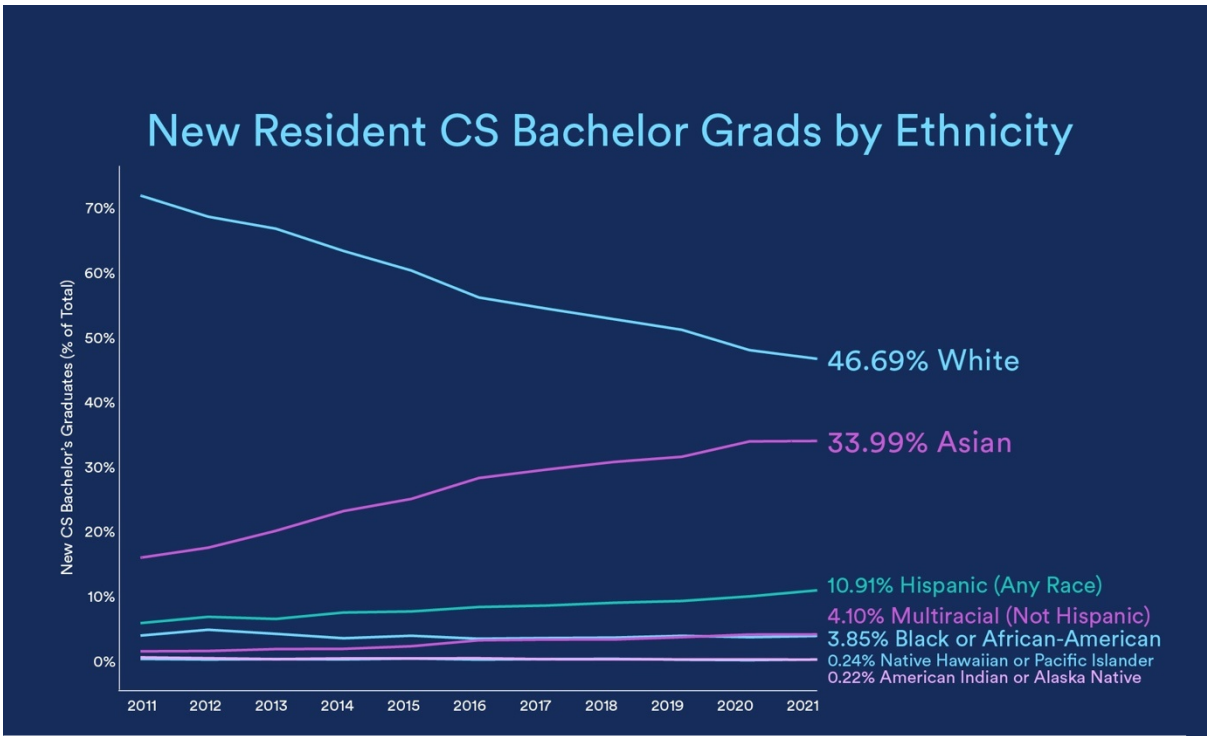


Figure 6. A Graph to show the percentage of Computer Science Bachelor's Degree Graduates by Ethnicity (2011-2021) [7]

[6] CRA Taulbee Survey (2022). Available at: https://cra.org/crn/wp-content/uploads/sites/7/2023/05/2022-Taulbee-Survey-Final.pdf

[7] CRA Taulbee Survey (2022). Available at: https://cra.org/crn/wp-content/uploads/sites/7/2023/05/2022-Taulbee-Survey-Final.pdf

Ethical Management

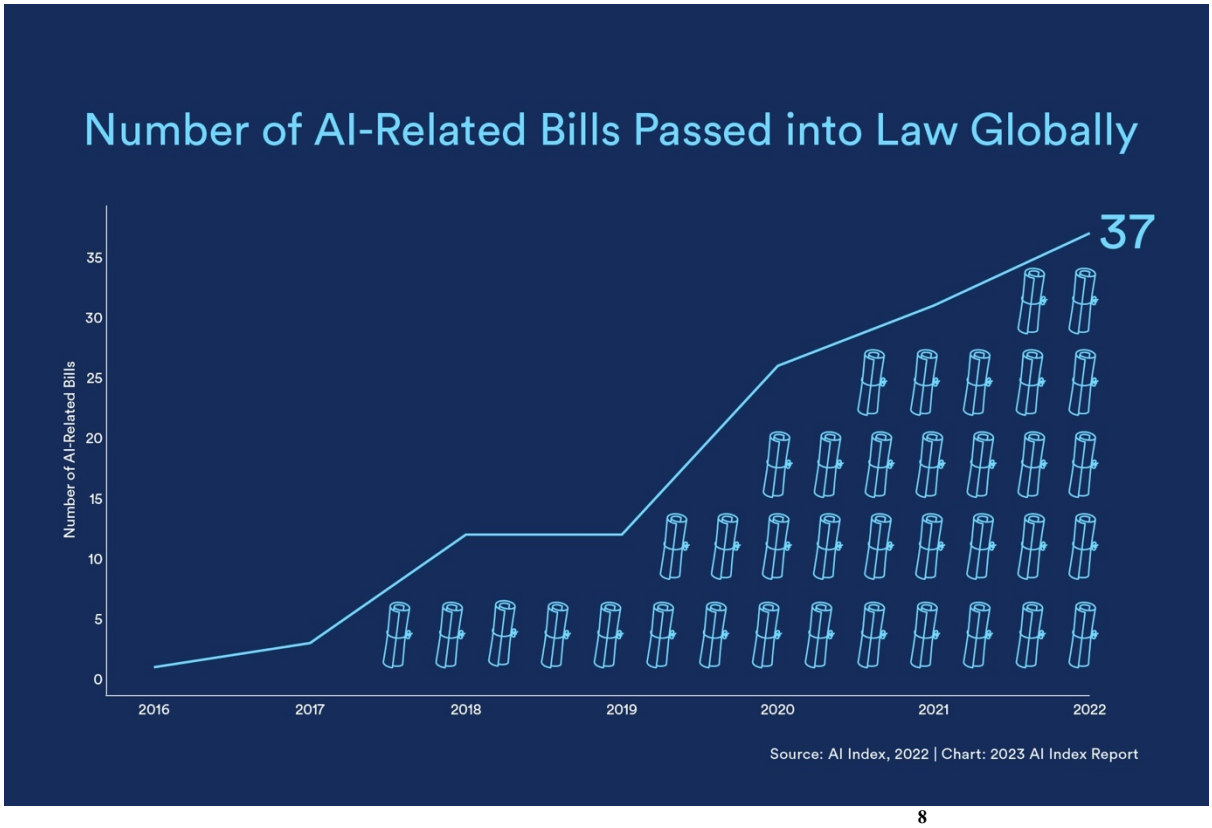
Having looked at many of the negative implications produced by considering the ethics of AI, this paper shall now suggest possible solutions that aim to maintain the ethical integrity of AI while allowing it to progress. Solutions shall be split into two basic sections i) political, these are regarding AI policy and making AI more ethical with a forward-looking angle; ii) economic, these focus on addressing current problems and offsetting these ad infinitum, thus these can be considered comparatively more retrospective.

§1. Political Solutions

AI has quickly become a political matter due to its far-reaching implications for society (a few of which were outlined above). Since it has rapidly become a matter of national importance (due to privacy concerns, potential further economic impact, and ethical concerns alongside other factors), it is only right to consider political solutions.

§1.1. Regulations

Perhaps the most sure-fire way to manage AI is the introduction of regulations and laws. Such regulations will ensure that all AI produced and evolved will abide by the desired set of rules lest they risk a lawsuit.

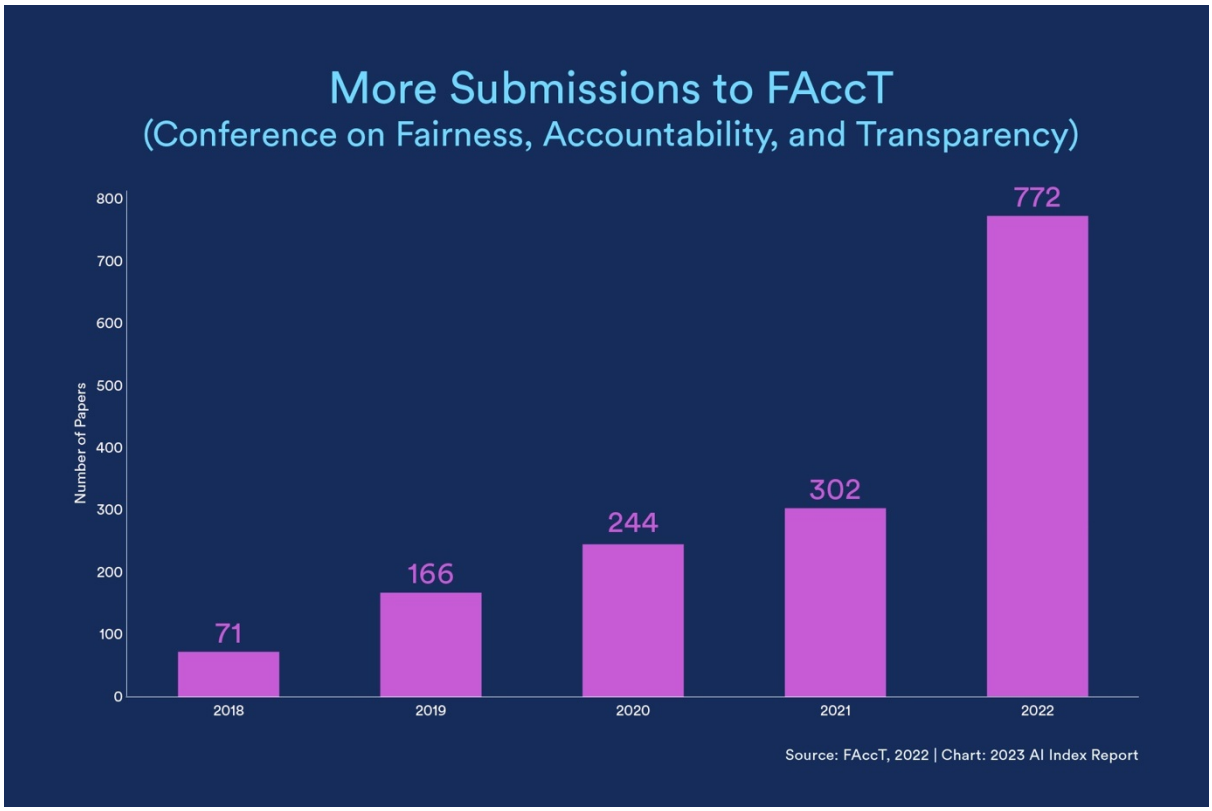


[8]

Figure 7. A Chart to show the increase in the number of AI-Related Bills Passed (2016-2022)

[8] Stanford University HAI (2023). Available at: https://hai.stanford.edu/news/2023-state-ai-14-charts

Fig. 7 shows the increase in the production of AI-related legislature globally; the correlation displayed is that, as time goes on (and AI becomes more advanced), the amount of AI-related legislature increases. This embodies the aforementioned point wherein it was suggested that regulations and laws seem to be effective ways to manage AI systems as desired.



[9]

Figure 8. A Chart to show the increase in submissions to FAccT (2018-2022)

FAccT is the Conference on Fairness, Accountability, and Transparency, depicted in Fig. 8, the increase in annual submissions implies growing interest in such topics. Having explained accountability above (and with fairness being fairly self-explanatory), here the focus shall be on transparency. Transparency refers to the intelligibility of an AI's decision-making process by humans. A good way to implement transparency is through 'XAI'.

XAI is a contraction for 'Explainable AI'. The idea behind XAI is that it provides humans with an insight into the inner workings of the AI. As earlier mentioned, AI tends to operate with a 'black box' (opaque decision-making), XAI provides a window into this mysterious box, creating greater levels of transparency. By understanding why an AI makes the decisions it does, the problem of accountability can be addressed. It also enables any bias or errors to be detected. In this case, a developer could address this issue – leading to more robust and trustworthy technology. XAI also helps to prove compliance with any regulations by explaining the AI's decisions. Consequently, there should certainly be a law that obliges AI developers to produce Explainable AI.

Having tackled the problem of accountability, privacy and bias (of the big three) remain. In response to data privacy regulations are quite simple. Firstly, the AI developers must ensure that they only use and access someone's data with their informed consent. This should

[9] FAccT (2022). Available at: https://facctconference.org/

include a brief explanation of what the data will be used for. Once the data is in their hands, they must take it upon themselves to protect any personal information and prevent unauthorized data access or theft. Should regulations detailing such rules be enforced, the problem of data usage and privacy will be much reduced.

Concerning bias, Omer et al. (2017) uses the analogy of the 2001 film 'Shadow Hal' to conclude that the solution to biased results is to educate against the bias (ie. [in the case of AI] an unbiased training dataset). This is as an alternative to forcing the result (ie. tampering with the outputs to attain the desired result). In this way, the problem will be solved at the root, and so, the program can then be trusted to consistently produce unbiased results; in the other case, the programmers will have to constantly interfere with the results to force an unbiased output set – in this way, it is highly likely that their own (inherently) flawed human nature will lead to subconsciously biased outputs. Therefore, regulations of some sort should be implemented to ensure datasets used are unbiased. To ensure this, the regulation should address data collection and concepts such as representation; if the collection of data is unbiased, it is noticeably less likely for the training dataset to be biased – hence leading to unbiased results. A combination of unbiased datasets, and the monitoring of inner-workings (by use of XAI) should greatly diminish biased results (or, at the very least, lead to an explanation of certain results).

§1.2. Education

To create a secure foundation for the future, the best approach is to instil understanding and ethics into the youth.

Hence, Goldsmith et al. (2017) proposes the incorporation of ethics into the syllabus of an 'AI class':

> "Because AI work has enormous effects on society — including the ways in which individuals interact, on our economies, on the practice of medicine and the uses of leisure, to name a few — we believe that all AI practitioners, and those that reason about AI technology, should be able to frame discussion in terms of ethics. Further, we believe that popular culture promotes a very limited understanding of how to frame and analyze ethical decision making. Thus, we advocate that ethics be taught in AI classes, and that we develop materials and courses for teaching ethical frameworks and reasoning to people working in AI (Burton, Goldsmith, and Mattei 2016)."

By teaching the 'AI developers of the future' ethics, they shall find it easier to align their AI (alignment of AI refers to the alignment with human values and ethics) due to their heightened understanding of both topics. Furthermore, they will also understand the need for such action to take place due to the deep-rooted public awareness within the society of the future.

This investment in the future generations is an extremely sustainable and efficient solution to the pressing concern of immoral (unethical) AI.

§2. Economic Solutions

The solutions provided here are more focused on offsetting and preventing the negative externalities brough about by the use of AI. Economic solutions include fiscal policy and economic incentives.
One of the greater problems with AI is the environmental negative externalities. Many of the economic instruments outlined below shall be focused on reducing these due to the importance of the environment and the apparent simplicity of the solutions.

§2.1. Subsidies

Governments could allocate funds and grants to support ethical AI initiatives. These could encourage innovative research and development of ethical AI technologies. Such subsidies could also be used to help AI projects and companies that are focused on having a positive social impact by solving social challenges. The existence of such subsidies can further incentivise other AI businesses to consider ethical implications of their doings.

§2.2. UBI

One of the major economic concerns detailed earlier was the replacement of many jobs by AI systems. A potential solution to this is the introduction of a 'UBI'. 'Universal Basic Incomes' would be an annual guaranteed salary received by all members of a population.

There are many benefits to UBI such as: the prevention of poverty, social security and support for unrecognised and unpaid labour (such as caregiving and community service). It also has the capability to act as an economic stimulus by facilitating consumer spending.

This being said, there are various problems associated with the concept too. Firstly, the cost of the program will be mighty – funding would either come from excessive taxation or the increasing of national debt. Secondly, it reduces work incentives – this would lead to a stagnant economy with no growth (with impending eventual economic shrinkage). Finally, since UBI is distributed to a whole population, it could potentially exacerbate income inequality.

All in all, the positive benefits of UBI do not outweigh the problems. It is unquestionably a 'utopian concept' and can only exist in a perfect society.

§2.3. Pigouvian Tax

A Pigouvian tax is levied on the negative externality of a business' (or individual's) activities. As seen above, one of the major concerns with AI is the energy needed to run such models. For the facilitation of these activities, fossil fuels are often used in place of renewable energy sources. The burning of fossil fuels release carbon dioxide (a greenhouse gas) into the atmosphere – this is the negative externality. Thus, the tax would be levied on the mass of carbon dioxide produced. This works to dissuade superfluous energy-demanding activities (as it would result in more CO2, and so, more tax).

§2.4. Cap and Trade

To further manage the CO2 production of corporations, a cap-and-trade system should be put in place. This would place a limit on the total amount of emissions any given business can produce (thence the 'cap'); the 'trade' aspect enables businesses who don't use their allowance may trade their carbon credits (emission permits) for money. This ensure that unnecessary carbon emissions don't take place as the businesses will be conscious about how they spend their carbon credits. The fundamental idea is that the 'cap' is gradually reduced over time. This will eventually lead to a great decrease in carbon emissions.

§2.5. Tax Credits

The final economic instrument in this series of reducing carbon emissions is tax credits. Tax credits (benefits of sorts) reduce the amount of tax to be paid. In this case, tax credits should be allocated to compensate unused carbon credits. In this way, it is in a corporation's best interest to save as many of their carbon credits (emit as little) as they can. Once again, this will greatly benefit the environment and encourage companies to find more renewable alternatives. Eventually, they shall be very well acquainted with renewable energies and will be prepared (and in a good place) to make a complete transition.

Secondly, tax credits can also be allocated to AI systems that produce positive impacts for societies such as some of the ideas mentioned above concerning healthcare. Moreover, tax credits can be used to endorse and encourage developments regarding the intersections between the fields of ethics and AI – this could be anything from substantial research into alignment of AI to effective bias mitigation mechanisms.

Conclusion

§1. Is AI ethical?

The lack of morality and consciousness prevents AI from being both, ethical, and unethical. Thus, in response to the question, AI is not inherently ethical. On its own it has no capability to distinguish between right and wrong.

Due to the programmable nature of AI, it can only emulate that which it has learnt and so cannot come to conclusions beyond the context of the training data. AI systems optimize their outputs based on the objective function which is defined by developers. If this function itself doesn't prioritize ethical considerations, the AI will likely make decisions that maximise contextual objective success of the outputs but can be seen as ethically questionable.

Alternatively, while 'AI' may not be ethical, the 'concept of AI' (ie. the [perhaps unattainable] imagined ideal) is ethically good – this is because the ultimate reason behind the creation of AI was to enable scientific and technological advancements that could make life easier and better for humanity. Taking a utilitarian approach (the greater good for the greater number) AI can only be seen as an ethically correct idea due to the immensely valuable (and large number) of positive implications it can facilitate (for a very, very large number of people)– a few of which were outlined above. This betterment that it (ideally) brings to society suggests that the concept of AI is, certainly, ethically good.

§2. Can AI become (and remain) ethical?

The real question, however, is whether AI (in itself) can be made to be ethical.

After outlining the main problems associated with the use and incorporation of AI with everyday life, it is evident that are many things to consider. The large quantity of problems associated with AI suggests that it cannot be ethical. However, this paper also suggested a number of different solutions (political and economic) to many of the aforementioned problems.

Having also outlined a few possible positive implications of AI, it is also very clear that there are innumerable positive opportunities that can prove invaluable to humankind (such as those related to the early diagnosis of neurological disorders). Needless to say, AI definitely has the capability to do a lot of social good (which would prove it ethical).

Through the enforcement of some of the management techniques listed and explained in this paper, the negative aspects can be significantly offset and prevented. Alongside this, the rapid growth and evolution of Artificial Intelligence means that many of desired positive characteristics can be quickly realised; thereby enabling the 'good' to outweigh the 'bad' and rendering AI: "ethical".

Bibliography

Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., ... & Buchanan, W. (2022, June). Measuring the carbon intensity of AI in cloud instances. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1877-1894).

Duhigg, C. (2013). *The Power of Habit: Why we do what we do and how to change*. Random House.

Goldsmith, J., & Burton, E. (2017, February). Why teaching ethics to AI practitioners is important. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

Johnstone, G. (Director). (2022). *M3GAN* [Film]. Blumhouse Productions.

Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model. *arXiv preprint arXiv:2211.02001*.

Malik, P., Pathania, M., & Rathaur, V. K. (2019). Overview of artificial intelligence in medicine. *Journal of family medicine and primary care*, *8*(7), 2328

McCarthy, J. (2007). What is artificial intelligence?

McGee, R. W. (2023). What Will the United States Look Like in 2050? A ChatGPT Short Story. *A Chatgpt Short Story (April 8, 2023)*.

Nelson, G. S. (2019). Bias in artificial intelligence. *North Carolina medical journal*, *80*(4), 220-222.

Omer Tene & Jules Polonetsky, Taming the Golem: Challenges of Ethical Algorithmic Decision-Making, 19 N.C. J. L. & TECH. 125 (2017).

Peduzzi, P. (2014). Sand, rarer than one thinks. *Environmental Development*, *11*(208-218), 682.

Surden, H. (2019). Artificial intelligence and law: An overview. *Georgia State University Law Review*, *35*, 19-22.

Yudkowsky, E. (2016). The AI alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, *4*.

Zweben, S., & Bizot, B. (2022). 2021 Taulbee Survey.