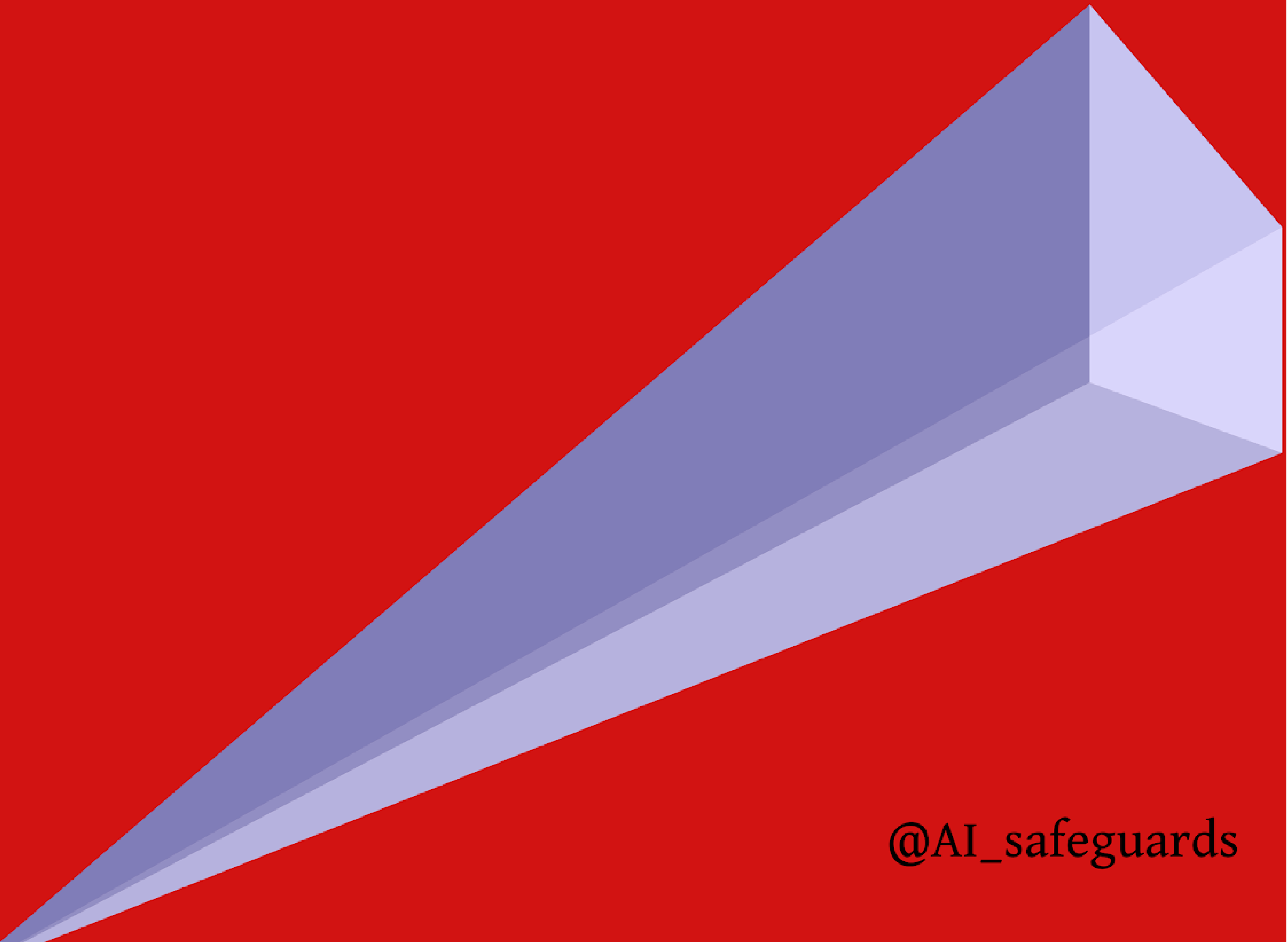
Solving the Artificial Intelligence Race

Mitigating the problems associated with the AI race



@AI_safeguards

Contents

Summary

In whichever way the AI race progresses or ends, it will cause various levels of misery for all actors and stakeholders. It would be prudent to brace for negative consequences of the AI race and formulate plans to mitigate the pitfalls.

Negative consequences of the AI race

The rush to develop and deploy AI would make researchers overlook some of the crucial factors in AI development and deployment, like : *safety of AI, managing expectations on capabilities and limitations of AI, ethics , legal accountability, inspectability of AI, Human-AI interaction design, privacy rights, social impact and speed of AI rollout.*

Overlooking these important factors could result in the following undesirable outcomes:

1. The rush to bring AI products to market without due diligence could result in corporations taking ethical shortcuts like grabbing training data without consent, ignoring consumer privacy, overlooking risks , concealing limitations, engaging in deceptive marketing and claiming the lead in AI benchmarks by ‘gaming the system’.[1]

2. An AI monopolist could lobby policy makers to relax regulations regarding safety, legal accountability and privacy. Rapid rollout of AI could result in massive increase in unemployment and social inequality. Corporations and countries could attempt reckless use of AI emboldened by the delay in updating legal systems.

3. Militaries would be emboldened to recklessly use AI-augmented weapons and autonomous AI weapons to cause rapid large-scale destruction and surgical attacks on targeted groups.The likelihood of accidental wars would increase due to malfunctioning AI weapons, false flag attacks and difficulty in proving AI weapon ownership.

4. Humans are very likely to be harmed due to hidden biases in AI, malware infected AI and AI that is designed without understanding nuances of Human-AI interaction. When powerful AI without inbuilt safeguards and kill-switches are deployed, there will be a huge potential risk of catastrophic harm.

5. Until legal systems are updated, there will be lack of clarity on legal accountability when damage is caused by AI decisions.There are likely to be numerous consumer lawsuits and patent battles resulting from non-inspectable AI. Unless legally forced, corporations are likely to be uncooperative and defensive after AI related accidents.

6. Group-think and academic elitism could suppress funding for new ideas for AGI. Public perception of AI would be affected by the media’s unjustified levels of

optimism/pessimism and constant scaremongering. Lack of immediate progress towards AGI could affect investor confidence and increase public skepticism.

Mitigating the negative consequences of the AI race

1. Encourage and enforce **cooperation** between actors in the AI race :
   - Form teams that agree to common goals, pledge to ethical behaviour and agree to commit resources to a common pool
   - Reward cooperation with faster regulatory clearances, access to resources and monetary incentives
   - Punish non-cooperation by withholding resources, withholding rollout approvals and encouraging blacklist/boycott of 'rogue' entities by consumers, governments and regulators

2. Provide **incentives for transparency and disclosure** of ideas and AI milestones :
   - Publicly recognize and reward AI corporations based on their track record on customer privacy, legal accountability and AI inspectability.
   - Incentivize disclosure of research data by offering significant rewards
   - Create open-for-all contests and reward contestants on their submissions on novel ideas and algorithms that could lead to inspectable & safe AGI

3. Create AI **regulatory organisations and consumer rights organisations :**

   - **Regulatory agencies** need to be created and empowered to impose penalties on AI corporations for non-cooperative behaviour, misleading marketing, privacy violations, attempts to conceal risks, breach of ethics , safety violations and reckless usage. These regulatory agencies could aid in dispute redressal, handle complaints, initiate investigation on any AI related incident and could work with other government agencies

   - **Consumer rights organisations** specific to AI products could be created to safeguard the interests of consumers of AI products. These consumer rights organisations should be non-profit independently funded entities, with the mission to protect consumers by publishing unbiased reviews of AI products describing their applications, usage, vulnerabilities and risks

Recommendations

1. Accept and plan for the fact that there will always be bad and malicious actors leveraging AI. Accept that there will be constant battles (virtual and physical) between good-AI and bad-AI.
2. Setup framework for encouraging and enforcing cooperation in AI research
3. Mitigate undesirable behaviour through various channels like creating new legal requirements , consumer boycott and blacklisting of uncooperative corporations
4. Focus on regulating AI corporations and educating consumers of AI products
5. Focus on creating legal accountability in AI weapons usage instead of futile bans.
6. Focus on designing safeguards, enforcing legal accountability and inspectability of AI products

Important factors in AI development and deployment

Any AI based product should incorporate the following aspects in its design and deployment :

Safety

AI behaviour in a controlled lab environment does not imply real world performance and safety. AI could behave undesirably due to unexpected environmental stimuli, malfunctioning hardware, biased data or malware. Sufficient safeguards, including kill switches, must be inbuilt in the AI to prevent harm to humans.

Managing expectations

Consumers need to be fully educated of the capabilities and limitations of AI products. Consumers are susceptible to assuming that AI based products do not make mistakes, which results in harm to humans [2]

Ethics

Creators of AI based products need to exhibit ethical behaviour by being transparent in their activities like obtaining consent prior to acquiring data, avoid misleading the public on the scalability of intelligence and being transparent on data privacy policies.

Accountability

Clear lines of accountability needs to be set in order to manage instances of AI related accidents and resultant harm to humans. Accountability must be plausible and legally enforceable.

Inspectability of AI

An AI based system should be inspectable to reveal the 'reasoning' behind its decisions, [3] so that it would be possible to detect malware, hidden biases and illegal discrimination*(eg:racism)* . AI systems ought to have tamper-proof inbuilt logging systems to enable post-incident investigation after an undesirable action.

Human-AI interaction design

Interaction capabilities in an AI should include the ability to understand human emotions, intention signalling, identification of authority, withstanding manipulation, polite behaviour and the right to exhibit reasonable dissent. The risks to humans like addiction & attachment issues should also be addressed.

Privacy policies and other Consumer rights

The privacy policies related to the AI product need to be clear and acceptable.Penalties for violating privacy policies should be legally enforceable.Consumers have the right to know information regarding AI training data, AI testing results, right of ownership of the AI 'brain', planned obsolescence policies of the manufacturer, etc.

Social impact

AI product deployment could potentially cause huge job losses, increase in wealth inequality and social unrest. In cooperation with governments, regulators must plan to mitigate these negative outcomes before deployment.

Speed of rollout

Sufficient time needs to be given to society prior to large scale deployments of AI, so that society can absorb the shock of job losses.A staggered deployment also helps in identifying performance shortcomings , unforeseen safety risks and vulnerabilities which can then be quickly addressed prior to wider roll-out.

Race towards AI : Actors and stakeholders

The AI race

Corporations worldwide are racing towards creation of products incorporating Artificial Intelligence*(AI)*. Significant investment has been made in recent years by corporations, private investors and governments to gain the upper hand in the race towards AI. [4]

AI Products with Artificial Narrow Intelligence *(Narrow-AI)* are already being used to execute specific tasks *(eg: energy usage optimisation, self-driving cars,etc).* The ultimate goal of most AI researchers is to create Artificial General Intelligence *(AGI)*, which aims to create human level generic intelligence.

Any success in creating sufficiently powerful narrow-AI or AGI would confer significant financial gain and power to the entity that created the AI based product.

Actors and stakeholders in the AI race

A simplified list of actors and stakeholders in the AI race is as follows:

- **Corporations**
- **Governments**
- **Individual contributors**
- **Media**
- **Investors** (including shareholders and governments)
- **Consumers** (including military)
- **Regulatory and legal bodies**

Difficulties faced by the actors and stakeholders

- Difficulty in predicting future progress and how the AI race will pan out
- Insufficient awareness about capabilities and limitations of contemporary AI systems
- Insufficient awareness regarding how to distinguish between different types/levels of AI [5]
- Difficulty in estimating impact of AI on society
- Difficulty in updating legal systems to catch up with AI progress
- Difficulty in establishing consensus on basic questions like '*what constitutes AI?* ' , because of moving goalposts to suit current state of AI.

Motivations & desires of the actors in the AI race

Corporations want to be the first to develop and deploy AI products. They would want to utilize the first mover advantage to establish market monopoly and acquire patents to consolidate their monopoly. Some corporations would want to influence government policies to relax controls on privacy, safeguards and speed of deployment

Investors want timely returns on investment, relaxed regulation and to avoid investing in approaches that are less likely to achieve AGI

Individual contributors want to ensure attribution of credit, financial rewards for their effort and opportunities to participate & contribute in the AI development process

Regulatory bodies want to :

- Prevent misleading claims by AI product manufacturers
- Set clear expectations on level of capabilities of AI products
- Encourage transparency of training data and expected behaviour of the AI product
- Ensure presence of inbuilt safeguards
- Ensure transparency in legal accountability, privacy policies of AI products
- Set minimum acceptable standards on Inspectability
- Set policies regarding enforcement of accountability in AI systems
- Punish privacy violations and safety violations

Consumers want :

- Transparency regarding capability and limitations of AI products
- Transparency regarding accountability of AI products
- Strong predictability of AI behaviour
- Strong control over AI behaviour
- Good human interaction capabilities in AI products
- Strong data privacy
- Quick legal redressal on shortcomings in AI performance or violation of privacy

Governments and law enforcement bodies want to :

- Prevent malicious use or misuse by rogue actors

- Preventing rogue actors from hacking or controlling the AI deployments
- Leverage AI for use in public health and social welfare
- Use AI to increase economic growth and administration efficiency
- Stagger rollout of AI products over a long period of time in order to catch up with legal implications and to observe & mitigate negative impact on society

Positive outcomes from the AI race

The AI Race dynamics increases the likelihood of achieving *Artificial General Intelligence (AGI)*. Even if there is an initial monopoly, the benefits of 'good' AI will be felt very soon by a very wide audience. In spite of non-disclosure of proprietary algorithms, there is a strong likelihood that the technology behind AGI will be shared or leaked after a while. This would result in cheaper and accessible AI software for a wider audience.

Negative outcomes and Pitfalls in the AI race

Corporations are likely to attempt to influence public opinion and lobby media and policy makers in order to relax standards for safety, legal accountability and privacy. In the rush to grab market share, corporations are likely to take ethical shortcuts like grabbing training data without consent, ignoring customer privacy and ignoring the need to divulge shortcomings. Corporations are very likely to be defensive when confronted with shortcomings in AI performance, AI safety and lack of inspectability. Corporations would be under direct or indirect pressure to prematurely deploy untested AI, due to pressure from investors. There will be inevitable legal wrangles over patentability of AI. Obtaining evidence of IP theft would be more complicated by uninspectable AI. Corporations are likely to attempt unethical or illegal means to retain and poach AI experts.

Monopoly in AI , resulting from a single entity winning the AI race, could result in a *winner-takes-all* scenario. Such a monopolist would wield tremendous influence over consumers, regulators, governments and other corporations.A monopolist could use its influence to lobby for relaxed regulations and faster pace of rollout. A monopolist has no incentive to conform to strict privacy policies and ethical behaviour, because the customer has no other choice. The monopolist would use its AI products to subsidise its non-AI products.The monopolist could enforce specific cultural norms desired by the monopolist (eg: extreme left wing/right wing policies enforced on consumers). Unless the monopolist is taxed and dealt with cautiously by governments , too much power and wealth will be concentrated in the hands of a few.

Governments are likely to be slow to prepare for legal implications of AI [6] , slow to tax AI and slow to manage unemployment due to AI deployment(rollout) . Governments are susceptible to pressure by lobbyists not to regulate (even if shortcomings and potential AI risk to consumers are known). Governments might accelerate attempts to use an incomplete and unsafe AI for education and healthcare activities without considering the full implications. The privacy rights of citizens might be violated if an insecure AI is clandestinely deployed for large scale communication tapping and surveillance. Governments are also likely to restrict sharing of AGI research to other countries, slowing down the progress towards AGI

Fear-mongers would continue to issue exaggerated warnings regarding the risks posed by Artificial Intelligence, mostly without understanding the contemporary state of AI. While a basic level of caution is justified [7] , whipping up paranoia without justification would negatively impact public perception of AI and funding of AI research. Sensationalist books that whip up fear against AI are popular and sell well, creating a conflict of interest for fear-mongers to overestimate risks posed by advanced AI.[8]

Investors & shareholders of Corporations would rush to invest in AI blindly due to the 'Fear of Missing Out', ending up sinking money into unproven algorithms that cannot plausibly result in AGI. Investors would have unrealistic expectations regarding short term AI progress. Investors would apply pressure on corporations for rushed deployments of AI based systems without the necessary safeguards.

Consumers would form wrong assumptions regarding the intelligence level of the AI product and would inevitably find shortcomings in expected performance. Consumers are likely to be victims of misleading marketing, privacy violations, inbuilt hidden biases in the AI, insecure hackable AI, etc. If the AI product manufacturer offers to take responsibility for the dangers in its usage, it would create a moral hazard where the customer is tempted to take unjustified risks. (*Eg: reckless use of a self driving car in hazardous weather*).

Members of the public who are not direct consumers might feel unhappy due to disproportionate allocation of resources to AI product consumers *(eg:dedicated roads for self driving cars).* Some would also be tempted to test the limits of AI safety by interfering with the environment where the AI product is deployed*.(Eg: blocking a self driving car intentionally)*

Military consumers would rush to deploy killer AI, using real world battlefields as testing grounds. Military consumers might be unethical in AI weapon usage because legal systems will be slow to catch up on autonomous killing machines. See Appendix A : AI based weapons

Regulatory bodies of a country might be tempted to relax safety and inspectability standards to accelerate AI progress in that country. They might overlook privacy violations and the potential AI risk to consumers and society. The absence of minimum acceptable standards for performance, safety and inspectability mean that almost nothing would legally stop deployment of an incomplete or unsafe AI. Regulators might accept AI performance in a controlled environment and permit a wider rollout of the product, which could result in risk to human lives

Individual contributors might continue to persist in using approaches that are not likely to achieve AGI (due to groupthink phenomenon). A inefficient irrelevant algorithm backed by enormous computing power could give encouraging short term results but will not lead to AGI. Individual contributors are likely to be defensive regarding lack of actual progress towards AGI

Media : Journalists writing on AI developments might mislead the public , by either exaggerating the current state of AI or sensationalise the risks involved in AI without justification. Any unjustified hype will be harmful to AI development in the long term, as wrong expectations are being set. The media might not sufficiently scrutinize and question the AI-progress claims of certain AI companies and might add to the AI hype by extrapolating and exaggerating results from a controlled environment. [9]

Misrepresentation of capabilities is a serious issue in AI research, wherein corporations attempt to wrongly convince the public that they own state-of-the-art AI. Under tremendous pressure to show results for effort, some corporations or researchers could 'game the system' to generate better performance in certain AI contests or benchmarks. To obtain funding and market share, corporations might classify non-AI software as AI, claim high levels of performance & safety based on testing in a lab environment, mislead investors regarding current performance & future expectations to procure investment.

Lack of transparency in AI development includes the reluctance to divulge details on the training data, reluctance to admit to hidden biases in AI and deploying non-inspectable AI. A closed development environment , where known algorithm weaknesses are not shared, causes consumers to suffer from repeated incidents of same nature with products from other manufacturers. Consumers and regulators have the right to know details regarding training data like the source of the training data, relevancy of training data, possibility of biases in the training data, etc. Unless the full details of inbuilt safeguards and test results are made public, there will be a trust deficit that discourages potential consumers from utilizing AI.

Mitigating the negative consequences of the AI race

1. Encouraging and enforcing **cooperation** between actors in the AI race
2. Providing **incentives for transparency and disclosure** of ideas
3. Creating AI **regulatory organisations and consumer rights organisations**

Mitigation 1 : Encouraging and enforcing Cooperation

Benefits of cooperation in the AI race

If the actors in the AI race agree to cooperate, it will provide the following benefits :

- Quicker progress towards achieving AGI
- Avoiding dead-end investment in less promising approaches and irrelevant projects
- Avoiding duplication of effort, avoiding competing for staff, avoiding repeating similar mistakes in development.
- Savings resulting from sharing of infrastructure, expertise and hardware
- Better utilization of resources by sharing interim results and concentrating on more promising approaches
- More consumer choices and higher threshold for safety & accountability standards

Difficulties in enforcing cooperation and compliance

Among the entities in the AI race that agree to cooperate, it would be difficult to enforce long-standing cooperation because of these factors:

- Motivations of the entities (countries, corporations,etc) are different and varied.
- There could be instances of misattribution of credit, theft of intellectual property, plagiarism, premature release of research data without penalties, unauthorized reverse engineering by a competitor, etc.
- There is very little incentive for any entity to share technology once they have established a lead in research
- Legally enforcement of any “agreement to cooperate” is very difficult and time-consuming
- Countries would prohibit sharing of AI technology due to dual-use nature and to maintain relative lead in AI race. Countries could attempt to use their lead in AGI technology as a bargaining chip in international negotiations
- Some entities may withdraw suddenly after they find a novel means of achieving AGI, breaking the cooperative agreement (penalties of non-cooperation might be less than the incentives from cooperation)
- Some countries, after change of government, might change their policies regarding international cooperation *(eg: sanctions/insist on removal of certain countries from consortium in retaliation for unrelated disagreements)*

Avenues of cooperation in AI development

- Creating standard terminology to describe AI capabilities, safety and privacy levels
- Creating common risk mitigation strategies, accountability policies, privacy policies , transparency policies
- Establishing consensus on the investigative procedures to be followed after an undesirable incident caused by AI
- Agreeing on minimum acceptable criteria related to *inspectability, safety and privacy* prior to AI deployment
- Creating a common Ethics committee comprising of independent members without any conflict of interest
- Making legally binding commitment to share resources among the entities in the cooperative group
- Pledging to adhere to code of conduct and Ethics policy
- Pledging transparency on data acquisition, AI ownership, 'learnt data' ownership, accountability policy of AI products, consumer appeal processes and consumer rights
- Pledging to commit sufficient resources to analyse and mitigate the risks of AI products
- Pledging to share information and novel approaches that could potentially achieve AGI
- Pledging to adopt strategies that help governments to minimise negative social impact due to AI and pledging responsible rollout of AI

Enforcing cooperation : How to incentivize actors to cooperate

a. Form a consortium

An international group of likeminded entities, working towards the common objective of advancing AI, could join together to form a team/consortium comprising of corporations, governments, global organisations, independent individual contributors and private investors. The team's primary objective could be the shared pursuit of AGI. Once the team gains a critical number of important members, more entities could be incentivised to join the team as they gain access to the latest research information.

b. Pool resources

The members of this consortium could then legally commit resources to a 'common pool'. Such resources would include funds, hardware, patent licences and other physical infrastructure. Corporations could be required commit their AI experts for a specific period of time to work in this common pursuit of advanced AI.

c. Reward cooperation

- Attribute credit where it is due.
- Enable the contributing entities to retain the right to monetize effort
- Provide shared resources for research : like AI experts, AI safety experts, hardware, development infrastructure and test infrastructure
- Provide structured incentives for significant or full disclosure of AGI milestones, algorithms, training data, safety features.
- Enable easier paperwork and a simple fast-track pipeline for obtaining deployment permits from governments and regulatory bodies.The incentive of quicker 'public rollout permit' acquisition can encourage cooperation and transparency.
- Compensate the potential monetary loss of the contributing entity that agrees not to file for patents
- Set incentives for disclosure of cartel-busting or bad behaviour within a participating entity ie., incentivize whistleblowing
- Set incentives such that the loss of revenue due to non-cooperation should be greater than the potential revenue from reckless deployment of an AI product

d. Punish non-cooperation

There is always a possibility of a corporation withdrawing from the cooperative agreement, claiming that it is 'too restrictive'. Such a 'rogue' corporation might harm customers due to its relaxed approach to safety and privacy of consumers.

To punish such a 'rogue' corporation, the following options are available :

- If an entity within the cooperative setup decides to withdraw, it should legally lose the right to the resources it previously committed to the common pool, including researchers.
- The uncooperating entity should lose access to the latest research information of the consortium
- The uncooperating entity should be expelled from the group so that it would suffer from loss of shared resources, funding, delay in regulatory approval and government aid
- Individual researchers could decide to boycott & resign from rogue corporations and not contribute to them
- Consumers could boycott rogue corporations. Consumer choice would be based on historic ethical behaviour, openness, accountability, privacy policies of AI product manufacturer, thereby rewarding good behaviour.
- Regulatory bodies could apply stricter requirements and longer approval process for AI deployment from corporations that have withdrawn from the cooperative ethical 'pledges'
- Governments can adjust the incentives for corporations to cooperate by adjusting tax policies, investment policies. Government bodies can adopt the same *minimum acceptable safeguards & accountability standards* as determined by the cooperating consortium and make those standards as legal requirements.
- Governments, consumer rights groups and regulators can blacklist an uncooperative corporation
- Governments can announce sanctions on rogue corporations, restricting hardware and software supplies

While it is likely that corporations would cooperate on defining AI related standards and sharing of resources, it would be quite difficult to force them to divulge AI algorithms

Mitigation 2 : Incentives for transparency and disclosure

1. Incentivise corporations and individual researchers to disclose research results by offering financial compensation and a wider audience for their ideas
2. Create open-for-all contests and reward contestants on their submissions on topics like :
    a. Novel ideas and algorithms that could lead to scalable & safe AGI
    b. Ideas on risk mitigation and enforcing corrigibility [10]
    c. Proof of Concept demonstration of thinking machines
    d. Proof of Concept demonstration of inspectable AI
    e. Proof of Concept demonstration of plausible milestones in AGI
3. Publicly recognize and reward AI corporations based on their track record on customer privacy, accountability and inspectability.
4. Reward researchers who find vulnerabilities or malware in AI products
5. **Diversity in research :** To prevent groupthink [11] & elitism in AGI research, encourage people from a variety of backgrounds and disciplines (neuroscientists, behavioural psychologists) to participate in AI research and the pursuit of AGI

Mitigation 3 : AI regulation & consumer rights organisations

Need for AI regulatory agencies

Several incidents like privacy violations in gathering training data [12] , misleading marketing by corporations and premature public trials of self driving cars [13] have already occurred, resulting in detrimental consequences to consumers and general public.

Corporations creating AI based products are usually under severe pressure to demonstrate results quickly and generate revenue. Such corporations cannot be expected to self-regulate.

Governments and regulatory bodies should be justifiably wary of AI companies and AI products because :

- Corporations are likely to mislead the public regarding AI performance, by using performance metrics specific to a particular test dataset or testing environment
- There is usually no inspection of the AI system by a neutral third party prior to public availability
- It takes Governments years to formulate and implement legal guidelines to catch up with the AI capabilities and to mitigate unforeseen negative consequences
- Corporations are likely to just match the bare minimum safety level among competing AI products, without making effort to enhance safeguards

- Corporations don't always give immediate cooperation for investigation after an AI product related accident
- Corporations have been historically defensive and uncooperative when confronted with their AI products’ shortcomings
- Corporations are too secretive regarding AI algorithms and capabilities, resulting in non-inspectable AI that hinders the ability to detect possibility of discrimination & potential harm

For the above reasons, regulatory agencies for Artificial Intelligence need to be created, in the public interest. National regulatory bodies need to be created initially, followed by creation of international regulatory agencies

Duties of AI regulatory agencies

1. Certifying the performance and safety aspects of AI products
2. Inspecting AI products in real-world environments and inspecting the reliability of backend infrastructure behind AI products
3. Monitoring product performance and safety in real-life deployments
4. Certifying the safety and inspecting safeguards in AI products
5. Monitoring to ensure that product manufacturers adhere to their stated policies and to ensure that stated policies do not contradict existing laws
6. Ensuring swift compliance on legal accountability
7. Verifying that corporations are financially sound enough to fulfill their legal obligations on accountability
8. Requiring companies to deposit funds in escrow for high risk AI workflows to cover potential liability on accidents
9. Certifying individual expertise on designing and managing AI products
10. Implementing policies on AI accident management
11. Enabling industry wide common investigative process and sharing of investigation results to avoid repeat mistakes
12. Making recommendations regarding legal changes required to protect public safety and consumer rights
13. Gathering comments from public and other entities, taking them to account while formulating regulations
14. Recommending risk mitigation strategies like fail-safe options, data backup methodologies and designing kill-switches
15. Providing periodic reports to policymakers on state of AI, impact of AI deployment on society, predicted risks and mitigation strategies
16. Consulting with AI corporations and cooperating with government agencies

Powers of AI regulatory agencies

Powers must be granted by the relevant government to the regulatory agency, to enable the regulatory agency to :

1. Impose penalties on AI corporations for non-cooperative behaviour, anti-competitive behaviour, safety violations, non-transparency, misleading marketing, data privacy violations, false claims of product performance, attempts to conceal risks, attempts to conceal unfavourable test results, wilful launch of faulty products, breach of ethics and human rights, safety violations and reckless usage
2. Receive complaints from consumers, corporations and other stakeholders and settle AI industry disputes
3. Initiate investigation on any AI related incident
4. Force transparency from an AI manufacturer , like forcing disclosure of past complaints regarding an AI product.
5. Force transparency on privacy policies, lifetime support and planned obsolescence policies to remove future uncertainty
6. Force corporations to divulge known limitations, performance results under test environment, details of test environments, sources of training data,etc.
7. Work with other government agencies, law enforcement authorities, social welfare organisations to mitigate negative consequences of AI

International regulation : Once the first few national regulatory bodies are established, an International regulatory agency could be created to harmonise AI regulations internationally to agree on terminology, minimum acceptable safety standards, legal accountability policies and conditions for approval of AI based hardware and software.

The regulatory body could cooperate with governments to slow down pace of AI rollout, using methods such as:

- creating a non-profit government owned monopoly with exclusive license to sell AI products
- high taxation of AI products during the first 10 years
- requiring AI products to deposit funds to an escrow proportional to the number of active AI products in active use , to cover potential accountability

Challenges in establishing an AI regulatory organisation

When an AI regulatory organisation is established, it is likely to be criticized by AI companies on adoption of strict standards in definitions, safety policies, privacy policies and accountability policies. The AI regulator is likely to be accused of hindering AI development and revenue. Corporations will resist strict definitions of '*what constitutes AI*' and complain about changing goalposts regarding the definition of AI.

To justify their demand for light touch regulation, corporations might use nationalistic jingoism to appeal for lax domestic regulation to win the race to AI .Corporations might point out existing premature deployments in lax regulatory environments to claim that they are disadvantaged by geographical restrictions.[14]

It would be somewhat difficult to establish international cooperation because countries would be wary of impeding their domestic corporations in the AI race.

Consumer rights organisations

To safeguard the interests of consumers of AI products, one or more consumer rights organizations specific to AI products need to be created.

The consumer rights organisation(s) should be non-profit, independently funded entities, with the mission to protect consumers by :

1. Raising awareness of consumer rights
2. Flagging untruthful information in the marketplace
3. Publishing unbiased reviews of AI products describing their applications, usage, vulnerabilities and risks
4. Publishing information about corporations that misrepresent costs, fees and benefits
5. Testing the quality of AI products and making recommendations among products of same category
6. Highlighting good and bad behaviour among AI companies like accountability policies, refund policies, level of transparency, attitude towards addressing complaints, deceptive marketing and unfair billing
7. Recommending against purchase of dangerous, unpredictable and unaccountable AI products

Conclusion and Strategy Recommendations

Even though contemporary AI is incapable of '*creative thinking*' and is nowhere near human level intelligence, there is a strong possibility that AGI could be created in the next few decades. Human societies need to brace for the inevitable malicious use of AI, including autonomous killer robots and cyberspace threats.

Attempts need to be made to generate industry-wide consensus on terminologies, safety standards, performance standards, transparency and inspectability.
Using a carefully designed system of incentives and penalties, it would be possible to counteract and mitigate the various negative consequences of the race to AI by :

- creating consortium with cooperating AI race actors
- by incentivising transparency and disclosure
- by establishing regulatory bodies and consumer rights bodies

The impact of AI rollout on society should be continuously observed and unforeseen pitfalls must be mitigated by giving governments and legal systems time to catch up on implications. Reckless usage of AI must be resisted by forming alliances between countries and corporations to penalise uncooperative malicious actors and monopolies.

Appendix A : AI based weapons

The *Race to AI* will inevitably create AI based weapons that will be used by military forces.

AI based weapons could be either **narrow-AI weapons** or **strong-AI** (AGI) **weapons**. Even *narrow-AI* can be leveraged to enable more efficient decision making during a war, like target identification, enemy surveillance and avoidance of civilian collateral damage.

Weapons based on AGI would be even more easy to weaponize, because the algorithm behind the hypothetical AGI will be a generic one and adjusting the motivation system of the AGI would be sufficient to alter behaviour. Even harmless AGI (eg: surveillance bots) can be integrated into other AI weapons or conventional weapons [15] . It would be practically impossible to detect dual-use of any AI.

AI based weapons would offer some benefits like enabling quicker medical assistance, reducing collateral damage, enhancing defensive preparations and being a strong deterrent to war. Advanced AI weapons might prove to be a cheaper but effective equivalent to nuclear weapons, potentially paving the way for nuclear disarmament. AI weapons would provide a means of delivering minor harassment and inconvenience to the enemy forces and civilians, allowing aggrieved countries to let off steam. However the potential danger from AI weapons would definitely outweigh the benefits.

Dangers from AI based weapons

1. **Higher likelihood of war**
    - ○ Since the risk to human combatants' lives is less, the threshold for waging war is reduced.
    - ○ A third party might initiate a false-flag attack using *ownership-untraceable* AI weapons and provoke war between two historically hostile countries.
    - ○ Less detectable mobile AI units increase higher probability of accidental aggression *(Eg: since AI drones may not be detectable by radar, a lower threshold for detection of mobile units could trigger false positives and trigger mistaken retaliation)*
    - ○ AI could be leveraged to shape public opinion by targeted propaganda against a perceived enemy. Public support for war could increase due to the perceived relative strength of AI weapons.
    - ○ Using AI enabled surgical strikes, various incremental levels of aggression are possible (*salami tactics)*, such action will blur the lines between war and peace, continuously test the defending nation's patience and create confusion on when to retaliate.

2. **Highly efficient and rapid destruction:**

- AI weapons could be used to assassinate politically significant individuals or harm a particular group of humans based on race, gender, age, etc.
- AI weapons could be used to kill humans without destroying physical infrastructure, so that the aggressor can utilize the infrastructure.
- AI weapons production might cause a fall in manufacturing cost of weapons due to their high effectiveness, minimal hardware requirements, intelligence and self-navigating abilities.AI weapons would most certainly produce more kills per dollar spent in manufacturing cost.
- AI can enable efficient internal military communication and enable significantly faster pace of war. AI weapons would make it easier to maintain a relentless continuous attack without a minute's gap; This can't be replicated by traditional armies due to limitations in human physiology.
- A war with AI weapons will open novel war-fronts (deep sea, deep space), potentially damaging the environment. AI weapons could self navigate to difficult-to-access places for self storage and recovery.
- AI based weapons enable quicker deployment on short notice and enable easier destruction of vital communication infrastructure *(like undersea internet cables).*
- AI weapons could be integrated with conventional weaponry for more accuracy and efficient destruction.
- AI algorithms could assist in optimal resource allocation and geographical distribution of *men and materiel* to inflict maximum damage

3. **Accidental destruction from incompetent, malfunctioning or hacked AI:**

- An AI could be hacked to alter its behaviour or loyalty. The inbuilt safeguards and restraints of an AI weapon could be removed by a third party hacking into its software. There is also the possibility of the AI overriding its own safeguards.
- An AI weapon might misunderstand the command or intention of its human handlers and execute an undesirable action *(eg: friendly fire).*
- A damaged AI weapon that loses its communication capabilities cannot be controlled by its owners.
- Badly designed AI based systems might accidentally trigger offensive action or accidental launches of missiles.
- An AI weapon without sufficient power might behave unpredictably or lose the ability to withhold aggressive action, unless a safe shutdown is initiated.
- An AI weapon incapable of contacting its human handlers might survive long periods of time, while retaining its lethal power for centuries, potentially harming future generations.

4. **Difficulty in ownership traceability and accountability**

- Given the potential compact size and low manufacturing cost of AI weapons, there would be many unauthorised manufacturers of AI weapons. It would be difficult to legally prove ownership of an AI weapon after a destructive incident. Mistaken assumptions regarding ownership of the offending AI weapon might result in hostile relations and accidental wars.

Practicality of enforcing ban on AI weapons

Even though there are calls from AI researchers to ban autonomous weapons [16] [17] , it would be difficult to establish consensus on such a ban and even more difficult to enforce it [18] [19].
There are strong incentives for gaining the lead in the AI weapons race. Besides the financial gain from efficiency in arms production, the winning entity is likely to be more aggressive and attempt to enforce its will on other entities & countries. No country will forgo a chance of gaining an upper hand in a war or a negotiation.

Establishing a consensus on banning AI weapons is very difficult, due to:

- Distrust among historically hostile countries
- Difficulty of detecting AI weapon development
- Difficulty of legally proving deployment of AI weapons or AI-augmented weapons
- Difficulty of AI weapon ownership traceability and enforcing accountability
- Difficulty of enforcing penalties of violating a ban

Even if an agreement to ban AI weapons were to be signed, such an agreement will be ignored in times of a national crisis or preperation of war. Successive governments might not honour previous agreements and attempt to withdraw from such an agreement.

The only way to enforce a ban on AI weapons is ironically through waging war on the entity owning the banned AI. An alliance of likeminded entities could decide to use formidable force on any entity that breaks the terms of engagement in usage of AI weapons. The alliance could declare sanctions & trade wars on rogue AI-race-actors benefiting from AI weapons.

Risk Mitigation in usage of AI weapons

1. **Establish consensus on acceptable usage of AI weapons:**
    - Formulate rules (similar to the Geneva Convention) to agree upon the manner of deployment and permissible targets for AI weapons
    - Agree to limit theatres of war (eg: avoid deep space and deep sea for storage/deployment of AI weapons)
    - Agree to prevent nuclear weapons being controlled by AI
    - Create safe zones for civilians in which all AI weapons are prohibited and incursions by any AI weapon will be met with unified force of neutral entities.

2. **Create deterrents against aggression:**
    - Encourage more spending on creating AI systems with surveillance and defensive capabilities
    - Encourage whistleblowers in defence industries at international level
    - Create an international alliance of countries pledging to battle any entity that uses AI weapons in an unacceptable manner

3. **Traceability and accountability:**
    - Focus resources to detect unaccountable, unregistered killer robots. Ensure that every AI weapon can be mapped to the corresponding legally accountable owner.
    - Establish legally enforceable accountability norms for malfunctioning AI or hacked AI. Malfunctions should not negate the obligations of the AI weapon operator.
    - Place stricter controls on hardware & materials required for explosives & weapon creation, to deter illegal AI weapon creators.

4. **AI weapon safety & corrigibility:**
    - Enable detection of 'hacked' AI and malware infected AI
    - Test AI weapons periodically to verify their motivation,behaviour and ability to control hardware
    - Ensure control of command by the military entity in charge, by means of inbuilt kill switches.
    - Determine '*Assembly points*' for 'malfunctioning', 'damaged' or 'lost' mobile AI weapons to enable safe shutdown, preventing civilian collateral damage

Recommendations regarding AI weapons

It must be accepted that there will be an inevitable arms race to create and deploy AI based weapons. It is impossible and impractical to detect or prevent usage of AI in weapons systems.Even though autonomous AI weapons based on AGI are unlikely in the short term, there is a strong likelihood of deployment of AI-augmented weapons.

Instead of futile attempts to completely ban AI weapons, resources and attention should be focussed on establishing traceability of AI weapon ownership, enforcing accountability for AI weapon usage and creating more AI with defensive & surveillance capabilities. Deterrents against AI weapon usage could be in the form of a credible retaliation plans by a group of countries against any rogue entity which uses AI weapons in an unacceptable manner.

Appendix B : Hypothetical scenarios in AI risk mitigation

1. Biased AI

Clara applies for jobs but she doesn't get a single interview even after several months. Clara learns that all recruitment agencies use an AI based software *“HIRE-AI”* to filter CVs and predict employee behaviour. Clara is informed that *HIRE-AI* had deemed her as '*high risk candidate*' and had recommended a '*do not hire*' decision. The recruitment agencies are not aware of the exact criteria used by *HIRE-AI* for filtering CVs and they just base their decisions on the AI software.

Clara contacts the AI corporation which sells *HIRE-AI ,* and requests details on the criteria used to filter candidates' CVs. The AI corporation refuses, mentioning that the filtering criteria details are part of the corporation's intellectual property.

Clara complains to the AI regulator, which has the power to force transparency from AI product companies. The AI regulator launches an investigation which reveals that the AI software correlates Clara's ethnicity with ‘*bad performance*’ and correlates Clara's gender with '*high risk of employee litigation'* - even though it is illegal to base a hiring decision on a candidate's ethnicity or gender.

The AI software provider is fined by the AI regulator for enabling racial discrimination & gender discrimination, ordered to recall *HIRE-AI* from the market and ordered to disclose all training data used to train *HIRE-AI.*

The AI regulator recommends that policymakers enact a law to prohibit usage of non-inspectable AI in hiring decisions. The AI regulator also recommends that those who knowingly use such ‘biased AI’ that is known to consistently discriminate against a group should also be prosecuted.

2. Accountability

John buys NUDOG, a robotic pet with Artificial Intelligence, to safeguard against burglars.

One day NUDOG attacks and bites John’s neighbour for no obvious reason.

The AI regulator’s investigation concludes that NUDOG had misheard a phrase and mistook it for a command to attack.

The AI regulator deems NUDOG as defective and orders its manufacturer to compensate the bite victim.

The AI regulator instructs NUDOG’s manufacturer to either recall all NUDOGs, OR deposit funds in a third party escrow account proportional to the number of active NUDOGs in use (to swiftly compensate future victims)

The AI regulator recommends that policymakers enact a law such that any AI robot capable of aggressive action should be capable of reading human emotions, signal its aggressive intention prior to attack and get a confirmation from its operator prior to aggressive action.

3. Misleading marketing

GRASS-BOT, an autonomous self-navigating robot with AI that is designed to trim grass, is advertised as having inbuilt RADAR to avoid harming humans during operation.
A consumer reports that when he used his GRASS-BOT, it injured his cat which was sleeping on the grass.

The AI regulator investigates and finds that GRASS-BOT has trouble detecting living things if they are stationary and also discovers that the RADAR data is not used in the decision making process, The manufacturer conveys that future software upgrades will utilize the RADAR data in decision making.

The AI regulator fines the GRASS-BOT manufacturer for misleading consumers by encouraging the assumption that the RADAR data is being used. The AI-regulator orders the manufacturer to (a) stop advertising the GRASS-BOT as safe (b) inform current and future consumers that RADAR data is not being used and (c) clearly describe the conditions required for safe operation of GRASS-BOT.

The AI regulator then issues an alert to the public regarding the observed risks of the product.

Mitigating the negative consequences of the AI race

Desired outcomes of AI race

Safety of AI products
Ethical behaviour
Legal Accountability
Gradual rollout of AI
Inspectability of AI
Strong Privacy rights
Good Human-AI interaction design
Minimisation of negative social impact
Disclosure of limitations of AI products
Managing expectations of AI capabilities

Negative consequences of AI race

Possibility of Monopoly dominating the market, restricting consumer choice
Monopoly lobbying for relaxed regulation and accountability
Rush to deploy premature AI ignoring public safety
Rush to generate revenue from substandard AI product
Unforeseen gaps in legal accountability
Reckless use of incredibly powerful AI weapons
Fearmongering or Hype of AI in media
Lack of transparency of privacy policies, training data procurement
Harm due to biased AI, non-inspectable AI and malware infected AI
Misleading marketing of AI products
Wrong assumptions of consumers regarding AI capablities and performance

Mitigation strategies

Create regulatory agencies for AI
Establish consumer rights groups
Standardize AI terminology
Incentivize cooperation by providing funds and shared infrastructure
Incentivize transparency and disclosure
Fund new approaches to acheive AGI
Set strict conditions for AI product rollout
Withhold resources and regulatory approvals from uncooperative AI corporations
Educate public on current state of AI
Ensure that ownership of AI weapons is traceable
Update legal systems to catchup with AI implications
Blacklist / ban unethical or uncooperative AI corporations
Impose higher taxes on AI products to mitigate corresponding job losses
Incentivize whistleblowing on ethics breaches

References

1. Gaming the system : Why and How Baidu Cheated an Artificial Intelligence Test https://www.technologyreview.com/s/538111/why-and-how-baidu-cheated-an-artificial-intelligence-test/

2. In Emergencies, Should You Trust a Robot? http://www.news.gatech.edu/2016/02/29/emergencies-should-you-trust-robot

3. Inspectability : Tension between AI and personal rights a growing problem : "Any algorithm that can't be explained can't be used by the civil service" : Mounir Mahjoubi https://www.irishtimes.com/opinion/tension-between-ai-and-personal-rights-a-growing-problem-1.3422860

4. Report from the AI Race Avoidance Workshop : GoodAI and AI Roadmap Institute Tokyo, ARAYA headquarters, October 13, 2017 https://medium.com/ai-roadmap-institute/report-from-the-ai-race-avoidance-workshop-bd631b2bbb2c

5. Everyone Is Talking About AI—But Do They Mean the Same Thing? https://singularityhub.com/2018/03/15/everyone-is-talking-about-ai-but-do-they-mean-the-same-thing/

6. Unreliable Evidence : The Law and Artificial Intelligence , BBC Four [07-01-2015] http://www.bbc.co.uk/programmes/b04wwgz9

7. What happens when our computers get smarter than we are? [14:53]:Nick Bostrom, Ted talk http://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are

8. The ethics of fear and how it undermines an informed citizenry https://www.poynter.org/news/ethics-fear-and-how-it-undermines-informed-citizenry

9. Don't believe the hype when it comes to AI : Artificial intelligence may be subject to more hype than any other field. While this creates funding opportunities, it could also damage AI's long-term potential http://www.wired.co.uk/article/sensationalism-ai-hype-innovation

10. Corrigibility : Nate Soares,Benja Fallenstein ,Eliezer Yudkowsky ,Stuart Armstrong https://intelligence.org/files/Corrigibility.pdf

11. Groupthink Dilemmas for Developing AI Self-Driving Cars https://aitrends.com/ai-insider/groupthink-dilemmas-for-developing-ai-self-driving-cars/

12. Concerns raised over broad scope of DeepMind-NHS health data-sharing deal https://techcrunch.com/2016/05/04/concerns-raised-over-broad-scope-of-deepmind-nhs-health-data-sharing-deal/

13. Uber's Self-Driving Cars Were Struggling Before Arizona Crash https://www.nytimes.com/2018/03/23/technology/uber-self-driving-cars-arizona.html

14. Don't Let Regulators Ruin AI https://www.technologyreview.com/s/609132/dont-let-regulators-ruin-ai/

15. Russian weapons maker Kalashnikov developing killer AI robots
https://news.vice.com/en_us/article/vbzq8y/russian-weapons-maker-kalashnikov-developing-killer-ai-robots

16. Autonomous weapons: An open letter from AI & Robotics researchers
https://futureoflife.org/open-letter-autonomous-weapons/

17. Ban on killer robots urgently needed, say scientists : theguardian.com
https://www.theguardian.com/science/2017/nov/13/ban-on-killer-robots-urgently-needed-say-scientists

18. Only five countries actually want to ban killer robots : theverge.com
https://www.theverge.com/2014/5/16/5724538/what-happened-at-the-un-killer-robot-debate

19. We can’t ban killer robots – it’s already too late
https://www.theguardian.com/commentisfree/2017/aug/22/killer-robots-international-arms-traders

20. The need to setup a regulatory authority for Artificial Intelligence based systems :
https://aisafety.quora.com/The-need-to-setup-a-regulatory-authority-for-Artificial-Intelligence-based-systems