इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी मंत्रालय
भारत सरकार

Ministry of Electronics & Information Technology
Government of India

REPORT OF COMMITTEE – D ON CYBER SECURITY, SAFETY, LEGAL AND ETHICAL ISSUES

Contents

1. Cyber Security Challenges

In recent years, **cybersecurity threats** have changed in three important ways:

1. The adversarial motivation has changed. Early attack programs were written as a result of an individual's curiosity, more recent attacks are written by well-funded and trained militaries in support of cyberwarfare or by sophisticated criminal organizations. With ransomware, the paradigm has shifted to lockdown what information is valuable to the victims rather than only go after the information that was valuable to the attackers.
2. The breadth and speed of attack adaptation have increased. The earlier attacks exploited software weaknesses found without any automation, were propagated using "sneakernet," and affected single computers or clusters, today's attacks exploit weaknesses found automatically; are automatically propagated over the Internet, packaged even by unsophisticated attackers; and affect computers, tablets, smartphones, and other devices across the globe – including critical information infrastructure.
3. The potential impact of an intrusion has increased substantially. Globally connected devices and people mean that attacks affect not only the digital world as in the past but also the physical world through the Internet of Things (IoT) and the society at large through ubiquitous social media platforms.

The **Internet security tasks can be grouped into two sets**: what humans do well and what computers do well.

1. Traditionally, computers excel at speed, scale, and scope. They can launch attacks in milliseconds and infect millions of computers. They can scan computer code to look for particular kinds of vulnerabilities, and data packets to identify particular kinds of attacks.
2. Humans, conversely, excel at thinking and reasoning. They can look at the data and distinguish a real attack from a false alarm, understand the attack as it's happening, and respond to it. They can find new sorts of vulnerabilities in systems. Humans are creative and adaptive, and can understand context.
3. Computers -- so far, at least -- are bad at what humans do well. They're not creative or adaptive. They don't understand context. They can behave irrationally because of those things. Humans are slow, and not good at repetitive tasks. They're terrible at big data analysis. They use cognitive shortcuts and can only keep a few data points in their head at a time. They can also behave irrationally because of those things.

In this context, cyber security teams are facing many **challenges:**

1. Worldwide shortage in cyber skills - there are not enough people around to follow the security best practice of people, process & technology.
2. Have more compliance mandates to deal with, such as GDPR, which are strengthening data privacy of even the most mature security organisations.

However, for those organisations starting to mature their security capabilities it’s a big challenge, not all organisations have additional resources to support it.

3. Security teams also have to keep up with the ever-increasing pace of business digitalisation.
4. More technology is being deployed and IT teams have to manage more and more diverse devices to ensure security and often even safety.

2. AI and Cyber Security

AI and more specifically, Machine Learning promises to address some of these challenges. AI is composed of different disciplines and Machine Learning is one of them. Machine data is produced by digital interactions within an Enterprise. By making this machine data centrally available and applying machine learning to it, we can radically transform an organisation's Cyber team.

AI and Machine Learning models for cybersecurity are being **applied in two phases**. The first phase involves developing an understanding of the normal historical landscape of network data traffic, extracting actionable insights about threats, and learning to identify anomalies in network traffic. The second phase consists of applying an understanding of “normal” to identify anomalous situations requiring human interaction and action against known threat profiles.

AI will allow computers to take over Internet security tasks from humans, and then do them faster and at scale. These include:

Discovering new vulnerabilities -- and, more importantly, new types of vulnerabilities -- in systems, both by the offense to exploit and by the defense to patch, and then automatically exploiting or patching them.

Reacting and adapting to an adversary's actions, again both on the offense and defense sides. This includes reasoning about those actions and what they mean in the context of the attack and the environment.

Abstracting lessons from individual incidents, generalizing them across systems and networks, and applying those lessons to increase attack and defense effectiveness elsewhere.

Identifying strategic and tactical trends from large datasets and using those trends to adapt attack and defense tactics.

It is near impossible to predict what AI technologies will be capable of. But it's not unreasonable to look at what humans do today and imagine a future where AIs are doing the same things, only at computer speeds, scale, and scope.

However, cybersecurity of AI is as relevant as AI for cybersecurity. The models and data need protection from manipulation. A canonical example of an image classifier, might result in the panda being classified as a gibbon with the introduction of a small perturbation in the training data. Such “adversarial” examples demonstrate that even simple modern algorithms, for both supervised and reinforcement learning, can act in surprising ways, even leading to unintended consequences.

The **new Cyber Security AAA**: **Automation, Analytics and AI**

AI is going to be the next battleground through 2020 and even beyond. The more advanced techniques are moving beyond traditional rule-based algorithms to create systems that understand, learn, predict, adapt and potentially operate autonomously, making smart machines appear "intelligent." *And it will trigger some autonomous systems that will be programmed to operate independently and allow IT departments to go deeper and do more (predicts Gartner).*

As predictive analytics gains ground, mathematics, machine learning and AI will be baked more into security solutions.

Security solutions will learn from the past, and essentially predict attack vectors and behavior based on that historical data. Thus, security solutions will be able to more accurately and intelligently identify and predict attacks by using event data and marrying it to real-world attacks.

- **AI and the Attack/Defense Balance**

  AI technologies have the potential to upend the longstanding advantage that attack has over defense on the Internet. This has to do with the relative strengths and weaknesses of people and computers, how those all interplay in Internet security, and where AI technologies might change things. Defense is currently in a worse position than offense precisely because of the human components.

  Present-day attacks pit the relative advantages of computers and humans against the relative weaknesses of computers and humans thereby creatring asymmetry.

  Both attack and defense will benefit from AI technologies, and we can safely presume that AI has the capability to tip the scales more toward defense. There will be better offensive and defensive AI techniques. Computers moving into what are traditionally human areas will rebalance that equation.

- **Specific Scenarios**
  - The large deployment of **devices** across the globe results in a single vulnerability or failure compromising a significant event, that is usually beyond the ability of human operators to cope with. The necessity is for security artificial intelligence (AI) to act as a force multiplier by augmenting the cybersecurity workforce's ability to defend at scale and speed.
  - The agility created by AI augmentation of a cybersecurity system may however be two sided. Along with a rapid response to both detection and remediation comes the potential for an equally rapid corruption of systems.

- There are already insatnces where attackers are using AI to detect when the malicious activities are being monitored within a "security sandbox," and to alter its behavior accordingly to escape detection and thereby extend the potential damage and surface attack.
- Some companies are employing AI to detect unusual activity on **client networks**. eg. Darktrace, uses AI algorithms created by mathematicians from the University of Cambridge that mimic the human immune system to identify issues on networks of all types and sizes.
- Today, the network consists of human users connected by mobile devices anywhere in the world and autonomous devices broadcasting sensor information from remote locations. This is a very **large and varied attack surface** to manage and defend against adversaries deploying sophisticated cyber attacks aided by AI bots.
- AI can help in a system of **continuous monitoring** as well as take over the more repetitive and time consuming tasks, leaving the technicians with more time to work on damage control. However, attackers will try their best to confuse the AI systems through evasion techniques such as adversarial AI (where the attackers design machine learning models that are intended to confuse the AI model into making a mistake).

3. AI & cyber warfare

Who will win the race to adopt AI for cyber warfare--the defenders of vulnerable networks or the cyber criminals constantly inventing new ways to attack them?

The promise that AI will "transform the world," has given rise to a number of significant races. Most prominent is the race among nations for AI superiority, primarily the U.S. and China, with several European countries (e.g., France) and the European Union attempting to position themselves not too far behind.

As in the case of nuclear arms-race, the use of the same technology for both beneficial and destructive purposes is evident for AI. In a recent paper warning of the potential of AI to "*upend the foundations of nuclear deterrence,*" **RAND researchers** wrote: "*The dual-use nature of many AI algorithms will mean AI research focused on one sector of society can be rapidly modified for use in the security sector as well.*"

For the "sector" cybersecurity, we could talk about "dual use" technology. AI could benefit humanity, but it could also serve as a powerful tool in the hands of cyber criminals and terrorists including nation states just like destructive nuclear weapons.

“Many organizations may be spending too much on technologies that do not minimize the impact of cyber attacks,” the Accenture report states. “*By better analyzing data and applying advanced threat intelligence, organizations can start to anticipate threats and adopt a more proactive approach to defensive strategies*.” One example of the “breakthrough technologies that can make a difference,” according to Accenture, is “*automated orchestration capabilities that use AI, big-data analytics, and machine learning to enable security teams to react and respond in nanoseconds and milliseconds, not minutes, hours or days*.”

While enterprises are slow or reluctant to invest in cyber defenses that can make a difference, cyber criminals are rapidly adopting them. “There are already instances of threat actors and hackers using AI technologies to bolster their attacks and malware”

Recent examples from two smaller software companies:

- AI researchers at Pivotal taught AI to find sensitive information (e.g., passwords) that was accidentally publicly released by looking at the code as if it were a picture—image recognition being one area in which AI has recently become extremely efficient and effective.
- Cybersecurity firm Endgame released an open-source data set, a collection of more than a million representations of benign and malicious Windows-portable executable files, with the aim of assisting cybersecurity researchers and practitioners to train and test new algorithms and improve their malware-hunting capabilities.

4. The Weaponization of AI

At the beginning of 2018, The Malicious Use of Artificial Intelligence Report warned that AI can be exploited by hackers for malicious purposes, possessing the ability to target entire states and alter society as we know it. The authors highlight that globally, we are at "*a critical moment in the co-evolution of AI and cybersecurity, and should proactively prepare for the next wave of attacks*".

The trouble with AI is that it is largely **based on freely available open source software**. In addition, new insights, approaches and successful experiments are widely shared, as AI powerhouses such as Google and Facebook allow their top AI engineers to publish their work, which they need to do in order to stay in the race to attract and keep the best AI minds.

While sharing and collaboration through public access to datasets, algorithms, and new tools are crucial for the success of cyber defenders, there is no question that bad actors could also benefit from it.

Cyber security company Symantec predicted that in 2018, cyber criminals will use Artificial Intelligence (AI) & Machine Learning (ML) to conduct attacks. No cyber security conversation today is complete without a discussion about AI and ML. So far, these conversations have been focused on using these technologies as protection and detection mechanisms. However, this will change in the next year with AI and ML being used by cyber criminals to conduct attacks. It is the first year where we will see AI versus AI in a cyber security context. Cyber criminals will use AI to attack and explore victims' networks, which is typically the most labour-intensive part of compromise after an incursion.

Forrester outlines a number of ways in which cyber criminals will profit from AI in the immediate future, including automating attacks and significantly improving the targeting of victims; better impersonating individuals for more effective social engineering; creating and targeting fake news; better code and better use of attack resources for distributed denial of service (DDoS) attacks; and developing more virulent malware and viruses. In addition, the proliferation of AI-powered systems, such as voice assistant, opens up many new potential vulnerabilities and opportunities for cyber-attacks.

"*Fast-moving and changing attacks will require defenses that move just as quickly, using AI and automation to augment human intelligence*" says Forrester.

AI attackers will use **Adversial AI** where attackers design ML models to confuse the relevant AI model to make a mistake. There are four ways in which adversarial AI can be used by nefarious elements, viz. Intrusion into Target System; Execution of Attacks; Incorporation into Malware; and, Contamination of data.

- An example of Adversial AI is the usage in spear phishing, where carefully targeted digital messages are used to trick people into installing malware or sharing sensitive data. Machine-learning models can now match humans at the art of crafting convincing fake messages, to churn out far more of them without tiring. Similarly, the use of AI to help design malware that is even better at fooling "sandboxes," or security programs that try to spot rogue code before it is deployed in companies' systems or image-based spam where spam content is embedded within an image to evade textual analysis performed by anti-spam filters.
- Another example of adversarial AI is in the realm of crypto jacking that actually went up by 8500 percent in 2017 alone. However, beyond the financial loss on accounting of crypto jacking is the misappropriation and theft of computer processing power that is needed for mining cryptocurrencies. Recent cases have ranged from the hacking of public Wi-Fi in a Starbucks in Argentina to a significant attack on computers at a Russian oil pipeline company. As currency mining activity and the price of certain cryptocurrencies continue to grow, so will hackers' temptation to breach many more computer networks. If they target hospital chains, airports, and other sensitive locations, the potential for collateral damage is deeply worrying.

Support of decision makers in determining where the best investments in security would be, to reduce the overall risk to the system, is essential. The development of an optimum attack-defense decision-making algorithm is successful in effectively identifying the optimum attack-defense strategy, thus supporting the decision-making process and plans.

- **Guarding against the weaponization of AI :**To protect against AI-launched attacks, there are three key steps that security teams should take to build a strong defense :
  - **Understand what is being protected.** Teams should lay this out clearly with appropriate solutions implemented for, threat vulnerability management, protection and detection with visibility into the whole environment. It is also important to have the option to rapidly change course when it comes to defense, since the target is always moving.
  - **Having clearly defined processes in place.** Organizations may have the best technology in the world, yet it is only as effective as the process it operates within. Both security teams and the wider organization must understand procedures and it is the responsibility of these security teams to educate employees on cybersecurity best practice.
  - **Knowing exactly what is normal for the environment.** Having context around attacks is crucial and often where companies fail. Possessing a

clear understanding of assets and how they communicate, will allow organizations to correctly isolate events that aren't normal and investigate them. AI/machine learning is an extremely effective tool for providing this context.

5. AI and Privacy

- A cyber terrorist can infiltrate many institutions including banking, medical, education, government, military, and communication and infrastructure systems. The majority of effective malicious cyber-activity has become web-based. Recent trends indicate that hackers are targeting users to steal personal information and moving away from targeting computers by causing system failure.

- AI systems (using big data), filter, sort, score, recommend, personalize, and otherwise shape human experiences. These systems have inherent risks, such as privacy breach, codifying and entrenching biases, reducing accountability and thus increasing the information asymmetry between the developers of such systems and consumers and policymakers.

- Where the sharing of large quantities of personal data is concerned, another issue to be considered is how to make maximum use of this data, with the minimum possible infringement on the privacy of individuals. In practise, it is necessary to use methods such as 'anonymisation' or 'de-identification', whereby datasets are processed in order to remove as much data which relates to individuals as possible, while retaining the usefulness of the dataset for the desired purpose.

Usually, the short-term effects of new technologies are overestimated, but underestimate their long-term effects.

- *AI is notoriously hard to predict and is likely to introduce new asymmetries that we can't foresee.*
- *But it is the most promising technology for bringing defense up to par with offense. For Internet security, that will change everything.*

62% of security experts believe that artificial intelligence (AI) will be weaponized and used for cyberattacks within the next 12 months. - Cylance, August 2017

6. Standards on Cyber Security Using AI & Related Technologies

The Safety and Security of AI Systems mainly depends on overcoming the Complex and uncertain environments; conditions that were never considered during its design, Controlling the system's behavior, Avoiding Goal mis-specification, Maintaining cordial Human-machine interactions & Setting Standards.

Any technology gets adopted and will be accepted only when there is accountability as it goes on to become a part of daily lives. The applications would scale when the concerns are perceived at global level with underlying standards to depend on.

Accordingly, Standards are to be set to address the complete AI development cycle, right from Engineering, Performance, Metrics, Safety, Usability, Interoperability, Usability, Security, Privacy, Traceability and Domains.

Numerous endeavors are underway on developing standards for AI both within the country as well as internationally. Bureau of Indian Standards (BIS), has set up a new committee for standardisation in AI, chaired by the director of IIT Patna. Considering the global nature of this field, it is crucial that India develops a singular national coalition to coordinate and contribute for platforms and institutions where relevant standards are being developed. These include but are not limited to, multilateral entities like ISO, IEC and ITU as well as others like IEEE and IETF, etc.

IEEE has announced three standards for ethics in AI; Standard for ethically driven nudging for robotic, intelligent, and autonomous systems, Standard for fail-safe design of autonomous and semi-autonomous systems and Well-being metrics standard for ethical artificial intelligence and autonomous systems.

If humans are to trust AI, AI -fuelled cyber security must be based on standardized and audited operations. Standards are important as they would chase industry innovation, it is recommended that R&D community, Academic community, industry and standards setting organizations such as IEEE, ISO, BIS and builders of AI systems, and regulators convene standards meetings at regular intervals to articulate the minimum levels of care required in building and operating security AI. Data being the key to the AI security it is necessary to use format of data or meta data to ensure interoperability among organizations with different AI security. Builders and operators of security AI should engage in performing robust self-audit and ensure that security by design principles are in place.

To see a long sustenance of this technology i.e. in order for AI to be safe and trusted enough, it is mandatory to take into account the needs of the global ecosystem and set AI standards through transparent discourse that serves the society at large.

7. Research & Development on Cyber Security & AI

A special committee headed by the National Institution for Transforming India **(NITI) Aayog** Vice chairman Rajiv Kumar was founded with the sole purpose of laying out a roadmap for the country's research and development in the field. Reportedly, the decision to form the committee came after news of China's intense and growing commitment in the field of AI.In this context, it is desirable to prioritize cyber security and privacy research such as quantum cryptography.

Machine learning is a vital part for designing an AI system. Quantum algorithms may be used for designing faster machine learning tools. Security and privacy of data involved in a AI system needs to be addressed. Encryption schemes which are amenable to machine learning techniques are desirable. Mechanisms for use in such AI systems, robust to quantum attacks, is an active area of research.

Recommendations:

Technology Development: There is a need to step up the effort to develop cyber security techniques and tools which use AI to defend against the attacks more effectively. Also, research is needed to identify the new types of vulnerabilities in AI-based applications. In this matter, it is necessary to take the advantage of the experience of other countries. As cyber attacks often involve many countries, international collaboration is necessary to defend cyber attacks.

Cybersecurity Challenges: AI-based cyber security challenges should be organized to select teams for further support for technology development. Such challenges can result in several new ideas for further exploration.

Anonymization Infrastructure: In order to make large sets of data available to the public for development, anonymization infrastructure should be created. The public agencies in possession of useful data should be sensitized and encouraged to share the data with the developers for public good.

Sharing of Best Practices: Often expertise and experience are available within the organizations but not shared. Steps should be taken to encourage the sharing of best practices in this area through various means. Government should use procurement contracts to emphasize on the best practices around security, privacy and other issues. It can make market move towards more responsible use of AI.

National Resource Center: In order to take up the activities listed above, Central Government should establish a National Resource Center for AI in Cyber Security. The Center may also be designated as the nodal agency for other related issues such as safety, ethical and legal issues.

8. Social Impact of AI

With increase in the usage of the Internet, there has been an exponential increase in the use of online social media and networks on the Internet. Websites like Facebook, YouTube, LinkedIn, Twitter, Flickr, Instagram, Google+, FourSquare, Pinterest, Tinder, and the likes have changed the way the Internet is being used. In 60 seconds of time, there are more than 900,000 logins into Facebook, 452,000 tweets posted, and 4.1 Million views on YouTube. Five V's of Social Media Volume, Velocity, Variety, Veracity, and Value are making a big difference in our use of social media for our benefits. Given the proliferations of these platforms, these are used as the primary source to measure the influence of a person, characterize his or her views on a given topic, analyse his or her interests, develop a propaganda, instigate a protest, etc. These platforms have become the real-world experiments for studying AI & cybersecurity / privacy.

In India, these networks have proliferated heavily and India stands in the top 5 countries consuming content from these platforms. Various organisations, entities, and individuals use these platforms effectively to do what they are interested in doing. For example, politicians effectively use it for interacting with citizens, police organisations use it for pushing content to citizens, NGOs use it for collating help, etc. Given the nature and the history of the platforms, Facebook is very popular in India, especially since vernacular languages are being supported on Facebook. Large amount of content from India gets generated on Quora, and Reddit too; there are also repercussions because of sharing information on these platforms. Overall, India is well penetrated in these platforms.

However, widely used, there is a lack of understanding of privacy and cybersecurity issues on online social media. Privacy and security of online social media need to be investigated, studied and characterized from various perspectives (computational, cultural, psychological, etc.). In particular, given the way AI has penetrated into these networks, it will be appropriate to use AI to study the networks and make decisions accordingly.

Here is an (incomplete) list of challenges / topics that are most exciting to work on in AI for Cybersecurity on social media:

- **Fake content / mis-information analysis**: The single most important issue that companies like Facebook, Twitter, Google and Microsoft are all grappling with is the spread of mis information on social media and the Internet. There is an increasing belief that Social Bots [1] control a large proportion of discussions or views on social media. There has been a large body of knowledge created in this space, but still there is lot more ground to cover to be confident on the output from these algorithms, systems, and implementations. Building AI algorithms to detect fake content is one of the biggest challenges in the space of Online Social Media.

- **National security:** Social media is also being used heavily by anti-social elements in the form of propagandas of specific ideologies, recruitment of people for specific anti-social activities, radicalization, fund raising, etc. AI / Machine learning based approaches can be developed to monitor social media content for national security purposes. Crawlers have to be built to monitor different platforms for specific content, triggers can be generated for finding specific content on the platforms. These triggers can be aggregated at various levels.

- **In-organic user behavior**: Given that online influence or network has become a metric to measure the influence of a person in the offline world, it has become more and more relevant for people to start finding ways to increase their online influence, (more the likes and followers, higher the value of the individual). Metrics like Klout [2] measure online influence. Natural behaviour on gaining followers, gaining likes, gaining connections are all organic, i.e. they grow in a specific way depending on your network. While services like Fiverr [3] can help you gain inorganic likes, followers, and connections. Studying the inorganic behaviour using AI methods and measuring the quantum of inorganic behaviour is important to differentiate the real influence vs. bloated influence. This can effectively help in identifying the real social influence of a person or an account.

- **Cybercrime analysis**: Due to increase in use of technology, the Internet, mobile phones, etc. there has been a corresponding increase in cybercrimes and social media has helped in increasing the derivative of these crimes. Research has shown that social information, i.e. one's network on Facebook or Twitter helps in targeting the attack on users and the success rate for these attacks using social information is more than traditional attacks [4]. Online crime can include malware, phishing, scams, terrorism, hate speech, bullying, etc. Understanding how these crimes propagate on the social media, characterizing these crimes can be very useful in building techniques to deter the effect of these cybercrimes. Building these technologies can be very challenging, given the quick changing nature of the cybercrime modus operandi and vectors. Recent progress in AI has helped in analysing large data fairly effectively, so these techniques can be used to analyse large inflowing data into social networks to make decisions on cybercrimes. Anomaly deducting using AI methods is becoming one of the important areas to work on.

- **Privacy on social media**: Given that we cannot live without sharing what we are doing on Facebook / Twitter / Instagram / Linkedin, we sometimes intentionally and sometimes unintentionally share information which can be used / misused against us. There have been many incidents where information posted online has been used againstan individual or people associated with the individual. There is a service www.pleaserobme.com which has been created to showcase how information that one posts on Twitter can be used to say whether one is at home or not and use it to burgle his / her house. There are other pieces of work which use Foursquare information to find out accurately the house of an individual [5]. It is not only problems due to sharing of information by individuals, but also issues and challenges from sharing of information by organizations, what ISPs infer from our online behaviour and how much privacy leaks happen because of this. Studying both the organization side, and the user side of privacy on social media is very necessary. Given the influence / implementation of GDPR, different techniques have to be devised to protect privacy. AI / Machine learning based approaches are being built to protect users' privacy.

In the advent of events like Cambridge Analytica, importance of analysing, and keeping a watch on the social media services has become very essential. Machine learning and AI based methods come in handy for providing better cybersecurity and privacy to users.

Clearly, we see that AI & Machine learning in the context of Cybersecurity is a double edged sword and needs efficient handling. We need appropriate manpower, infrastructure, and motivation to use the strength of these methods for societal good.

References:

[1]. https://cacm.acm.org/magazines/2016/7/204021-the-rise-of-social-bots/fulltext

[2]. https://www.klout.com/

[3]. https://www.fiverr.com/

[4]. https://dl.acm.org/citation.cfm?id=1290968

[5]. http://precog.iiitd.edu.in/Publications_files/TP_lbsn_2012.pdf

9. Safety & AI

In order to be accepted for use by the society, AI systems have to meet high degree of safety standards. Several AI systems such as autonomous vehicles, robots exert forces during the interaction with the environment. It is necessary to design the systems in such a way that it does not harm the people and property during its interactions. There are two primary issues in the context of safety - how much safety is required and how to measure it. Traditionally, regulatory agencies prescribe the safety standards with regard to machines and how to measure these. Such standards or their equivalents should be prescribed for the autonomous machines as well.

Absolute safety is often not practical to achieve. Even without using any AI system, we can't ensure 100% safety. Therefore, apart from safety parameters, safety thresholds have to be decided. The thresholds have to be decided for various types of domains under various scenarios. An AI system can be allowed for use by the public if it exceeds the safety thresholds on various safety parameters.

While prescribing the safety measurement parameters, it also has to be prescribed under what circumstances this has to be tested. For instance, it is said that autonomous vehicles should be adopted once these are safer than humans. The question is how to compare to decide whether it is safer. Should a car driven by a human driver be compared with self-driving car or a car driven by human driver assisted with safety devices?

Comprehensive testing is must before releasing any system for use by the public. Government has to establish necessary infrastructure for safety testing and certification. We also need to agree on other relevant points such as who are authorized to test and what tests are to be used.

When AI systems are used in deciding matters with serious implications for the people, it is necessary to take additional precautions. Examples of such decisions include the length of imprisonment for crimes. In order to design systems for making such decisions, certain trade-offs may be needed to satisfy contradicting objectives. For instance, the decision-making has to be just, speedy and inexpensive. It has been reported by the researchers that making the systems more transparent and less biased decreases the overall accuracy and efficiency. A less accurate system may assign more punishment for less sever crimes or vice-versa. A less efficient system may not deliver justice at an acceptable speed. A trade-off should either maximize all the values or should select a combination which is acceptable to the society.

Humans need certain skills to perform a task. They are trained and certified before they are permitted to perform the task. This is especially important in high risk tasks and domains such as aviation, medicine, etc. We need similar certification system for machines. Further, the problem gets more complicated in the case of AI systems which perform the tasks which have never been done by humans. An example is surgical robots which perform the tasks not done by humans so far (at that level of sophistication).

In the context of safety, an issue often debated is whether AI poses existential threat to humanity. This issue needs to be addressed properly else it may distract the policymakers from addressing more immediate challenges. The issue of existential threat has been raised by many thinkers of the time including Elon Musk, Stephen Hawking and Nick Bostrom. In his book Superintelligence, Bostrom has argued that AI being developed would be enormously superior to humans and may even harm its creators. Bostrom has not said that it is inevitable but he has referred to it as a possibility only. The concept of super intelligence has been criticized by many experts of the field. It has been pointed out that first of all, there is no clear way how can one develop it in near future with the current state of technology. So far, we have succeeded in developing intelligent machines for specific tasks only. We have not been able to develop machines which can perform even like a lower intelligence animal. Secondly, even if such a machine is developed, there is no reason to believe that it would be interested in dominating the world as machines don't have intent. If machines with higher intelligence are developed, the ways to control would also be developed in parallel. It is difficult to visualize a situation where powerful intelligent machines have been developed but the ways to control have not been developed.

The only possibility is that an individual or group of individuals with malicious intention may design a machine to harm the humans. For instance, one can imagine the actors who attack our nuclear installations, destabilize the trading market, etc. However, existential threat is a remote possibility. Only time can tell anything. However, most of the people believe that it is not possible in near future. Therefore, this threat should not lead to restrictions on the development of AI technology and its applications.

Recommendations:

Safety Guidelines: All the stakeholders including industry, government agencies and civil society should deliberate to evolve guidelines for safety features for the applications in various domains. Best practices in implementation of safety features should be shared. Government should invest in interdisciplinary research to study the impact of AI on society.

Safety Thresholds: In most of the cases, achieving absolute safety is not practical. Therefore, safety thresholds have to be decided for various domains. When the thresholds involve trade-offs, it should be in the range acceptable to the society.

Human Control: In case of any threat to human life or any other sever implication, humans should be in a position to interrupt or shutdown the system at any point of time. Human checks are necessary before implementing new decision-making strategies in AI systems.

Safety Certification: A mechanism should be established to certify the systems on the safety issues before releasing to the general public. The work should be initiated with the sectors like healthcare, transport, etc. where safety is quite important as it involves human life.

Existential Threat: Though it does not appear to become a serious issue in the near future, it requires deliberations on an ongoing basis. However, it should not restrict the development and deployment of AI systems.

10. Artificial Intelligence Ethics

Artificial Intelligence (AI) is the science of using computers to do things that traditionally required the human mind. It is a technology that will accelerate the digital transformation of industry, and will prove essential to the success of our digital economy in what is an increasingly connected world.

The development, application, and capabilities of AI-based systems are evolving rapidly, leaving largely unanswered a broad range of important short- and long-term questions related to the social impact, governance, and ethical implementations of these technologies and practices. Everyone should think about the ethics of the work they do, and the work they choose not to do. Artificial Intelligence and robots often seem like fun science fiction, but in fact already affect our daily lives. For example, services like Google and Amazon help us find what we want by using AI. They learn both from us and about us when we use them. Many countries and organizations now employ robots in warfare.

For AI to deliver on its promise, however, it will require predictability and trust. These two are interrelated. Predictable treatment of the complex issues AI will throw up, such as accountability and permitted data uses, will encourage investment in and use of AI. Similarly, progress with AI requires consumers to trust the technology and the fairness of how they are affected by it and how their data is used; predictable and transparent treatment facilitates this trust.

We are constantly being used as soft targets for the weaponised AI. Our social media likes and behaviors are being studied by AI systems to serve up deviously accurate ads. At first, we will be unconcerned to see the development of AI algorithms that already exist to optimize computer processes and to clean up systems. As machine learning progresses, these algorithms will inevitably be given growing autonomy to operate across a range of domains, for example to optimize data flows in an integrated supply chain. They will have to be given the ability to hide to operate without interfering with antivirus software. They will also be given the ability to make copies of themselves, just as Trojan viruses already do, and to move autonomously to access complex data sets. So far, this is little different from current machine learning applications. But combine these features, and what has in effect been created is a complex adaptive system: a system made of autonomous agents that replicate themselves, often with variations, leading to self-sustaining evolution. In other words, we build a system that mimics life’s powerful survival instincts.

The more powerful a technology becomes, the more can it be used for nefarious reasons as well as good. This applies not only to robots produced to replace human soldiers, or autonomous weapons, but to AI systems that can cause damage if used maliciously. Because these fights will not be fought on the battleground only, cybersecurity will become even more important. Autonomous machines can be considered artificial intelligences with physical bodies able to interact physically with their surrounding world. As such, from an impact or consequences point of view, they represent the apex of the artificial intelligence discussion. An extreme example could be an autonomous weapons system capable of operating autonomously even up to the point of selecting targets: the combination of that capability with face recognition could create the ultimate assassin. After all, we're dealing with a system that is faster and more capable than us by orders of magnitude.

The more established they will become, like weeds on an old building, the more damage they will do. It is likely that artificial intelligence will be able to develop computer viruses so sophisticated that they can only be stopped by superior AI-created, antivirus programs. A robust legal framework will be needed to deal with those issues too complex or fast changing to be addressed adequately by legislation. Two sets of issues stand out: first, how to address the vulnerabilities of the cyber infrastructure; and second, how to build the necessary safeguards into AI.

The big challenge that the whole security industry and the chief security officers have right now is that they are always chasing yesterday's attack. That is kind of the mindset the whole industry has that if we analyze yesterday's attack on someone else, we can help predict and prevent tomorrow's attack on us. It's flawed, because the attackers keep changing the attack vector. Applying machine learning can help enhance network security defenses and, over time, learn how to automatically detect unusual patterns in encrypted web traffic, cloud, and IoT environments.

Just as humans are given vaccines to train their immune system, the machine learning algorithm can be given mock attacks to learn. In that way, it can predict any potential attack before it happens. In the future, as AIs increase in capability, it is anticipated that they will first reach and then overtake humans in all domains of performance.

If one of today's cybersecurity systems fails, the damage can be unpleasant, but is tolerable in most cases: Someone loses money or privacy. But for human-level AI (or above), the consequences could be catastrophic. A single failure of a superintelligent AI (SAI) system could cause an existential risk event — an event that has the potential to damage human well-being on a global scale.

There is another concept known as cyber resilience. It can be thought of as an organization's ability to withstand or quickly recover from cyber events that disrupt usual business operations. Cyber security aims at preventing stealth of data. The other is meant to prevent anything that aims at disrupting regular business operations.

AI could be used to make non-authentic broadcast more common and more realistic; or make targeted spear-phishing more compelling at the scale of current mass phishing through the misuse or abuse of identity. This affects politics in various ways. The spread of false warning may cause the opposite effect and, namely, a serious peril for collective security. Disinformation and misinformation - to be intended as unfounded news or semi-truths- work often through organized campaigns with the aim to shape national opinion, destabilize social cohesion and affect politics. This will affect both business cyber security (business email compromise could become even more effective than it already is), and national security.

According to Section 70, Information technology Act, 2000 'Critical Information Infrastructure' (CII) as those facilities, systems or functions whose incapacity or destruction would cause a debilitating impact on national security, governance, economy and social well-being of a nation. Once notified as a "protected system" the CII is immediately placed under the ambit of section 66 (F) of the IT Act (Amended) 2008, which defines any cyber attack as an act of Cyber terrorism. Security should be a top priority whether it be in data sharing, video surveillance or even across our country's borders.

Because the Internet is so accessible and contains a wealth of information, it has become a popular resource for public safety professionals to communicate with each other (or neighboring agencies), research topics, and find information about possible suspects. Unfortunately, many people have become so comfortable with the Internet that they may adopt practices that make them vulnerable. For example, although people are typically wary of sharing personal information with strangers, they may not hesitate to post that same information online.

Public safety relies on communications, and the security of that information is critical. Cyber security involves protecting that information by preventing, detecting and responding to attacks. The advent of wireless communications and interagency data sharing has raised concerns about information security. In recent cyber attacks, credit card numbers were being stolen and e-mail viruses were spreading. In one sense, this is good for police officers who are trying to find out more about a suspect's whereabouts or habits. However, due to the nature of the data, it is imperative that the information being passed between agencies is secure and does not get into the wrong hands.

The ever-evolving and uncertain legal landscape in the world of cyber security is a challenge to all management levels in almost every business sector. To help benchmark a company's cyber security policies and programs, here are some continual themes, in no order of importance, that have emerged.

- Encryption of sensitive data at all times.
- Require and enforce robust passwords.
- Ensure vendors/suppliers/third parties have adequate security
- Train the security sector about all possible breaches.
- Practice what you preach.

According to Security Intelligence, there are no mandatory ethical standards that cyber security professionals are obligated to follow. Ethics in cyber security must also extend to consumers. Security Intelligence notes that there can't be any delays in letting customers know that a data breach has occurred and their information may have been stolen. While a deeply integrated code of cyber security ethics and conduct is vital, it is also crucial to cultivate ethical teachings among students and young. By promoting awareness of cyber security ethics at the early stages of learning and professional development, we can help ensure that future representatives stay on the right side of the ethical divide.

AI and ML can help in a system of continuous monitoring as well as take over the more repetitive and time consuming tasks, leaving the technicians with more time to work on damage control. Although it must be kept in mind that AI is not a silver bullet, since attackers will try their best to confuse the AI systems through evasion techniques such as adversarial AI (where the attackers design machine learning models that are intended to confuse the AI model into making a mistake). For instance, spammers and hackers often attempt to evade detection by obfuscating the content of spam emails and malware code. In the evasion setting, malicious samples are modified at test time to evade detection; that is, to be misclassified as legitimate. No influence over the training data is assumed. A clear example of evasion is image-based spam in which the spam content is embedded within an attached image to evade the textual analysis performed by anti-spam filters.

The first adversarial systems merely added random and almost undetectable noise to the input signal and demonstrated that humans could still understand the intended speech perfectly well, but that speech recognizers would come up with substantially different transcriptions. That was interesting but not particularly useful. It was kind of like defacing the Mona Lisa by painting bushy eyebrows a mustache. But the next steps are more interesting. The researchers used machine learning to learn how to deceive machine learning. We need to develop some adversarial AI models itself to counter attack these types of malicious practices.

A cyber terrorist can infiltrate many institutions including banking, medical, education, government, military, and communication and infrastructure systems. The majority of effective malicious cyber-activity has become web-based. Recent trends indicate that hackers are targeting users to steal personal information and moving away from targeting computers by causing system failure.

For example, an autonomous vehicle is what is known as a cyber-physical system because it has elements in both the physical and virtual worlds. This makes security particularly challenging. There are security risks to the networks that connect vehicles, whether the financial networks that process payments, roadside sensor networks, electricity infrastructure or traffic control features. It might seem convenient for an autonomous car that gets within 15 minutes of our home to automatically turn on the air conditioner, open the garage and unlock the front door. But any hacker who can breach that vehicle system would be able to walk right in and burglarize the home.

In the energy sector too, cyber threats are a major concern. The rapid decrease in cost of information, communications, and battery storage technologies is enabling more flexible and efficient generation, transmission, distribution, and consumption of energy, notably through smart energy storage solutions. This all leads to a more widespread connection of distributed energy resources, which can increase the number of vulnerabilities in smart devices and electric systems. Smart grid cyber security must address both inadvertent compromises of the electric infrastructure, due to user errors, equipment failures, and natural disasters, and deliberate attacks, such as from disgruntled employees, industrial espionage, and terrorists.

Smart Cities could be similarly susceptible as they rapidly deploy connected devices across municipal domains. For instance, a breach of street light systems could lead to control of the lights, which could lead to servers, in turn leading to data about individual customer behavior, eventually winding up with access to financial information and other personal information about citizens - possibly even their health records.

Interconnected smart cities, driverless cars, smart services, smart homes, and smart utilities providing information in real time, etc sounds appealing. Yet a cyber-safe interconnected utopia includes the right controls with proper implementation to ensure that connected infrastructure is accessible only to the right people at the right time for the right reasons.

As we continuously deploy AI models in the wild we are forced to re-examine the effects of such automation on the conditions of human life. Fueled by big data, these AI systems filter, sort, score, recommend, personalize, and otherwise shape human experiences. Although these systems bring myriad benefits, they also contain inherent risks, such as privacy breach, codifying and entrenching biases, reducing accountability and hindering due process and increasing the information asymmetry between data producers and data holders. This is for certain that algorithms do not exercise their power over us. It is rather the humans who do.

Though some experts use the term AI to refer to software artifacts with equal or superior intelligence to that of a human's, AI actually includes machine learning and other algorithms that supplement or replace traditionally human decisions.

In an increasingly digital world, data analytics offer attractive and competitive opportunities for efficiencies and value creation. Yet the responsibility to make insightful sense of significant patterns in data via algorithm-driven analytics remains ours and not the technical systems we create. For instance, if we detect a pattern that reflects an imbalance in the number of female and male software engineers in a company, and we decide not to perpetuate that pattern, human decision makers should make appropriate changes in hiring practices.

In practice, however, capturing the notion of fairness in an algorithm can be elusive. Everyone has an idea of what fairness means to them, but what is considered to be fair by one individual or group may not easily transfer to others. Using gender information to diagnose breast cancer and prostate cancer, for example, would be highly useful and appropriate. Using gender information to determine the quality of new hires for a computer scientist position, on the other hand, would not. This is the problem with bias in the training set. Bad intentions are not needed to make bad AI. A company might use an AI to search CVs for good job applicants after training it on information about people who rose to the top of the firm. If the culture at the business is healthy, the AI might well spot promising candidates, but if not, it might suggest people for interview who think nothing of trampling on their colleagues for a promotion.

The computer cannot harbor prejudice or stereotypes. While indeed the analysis technique may be completely neutral, given the assumptions, the model, the training data, and so forth, all of these boundary conditions are set by humans, who may reflect their biases in the analysis result, possibly without even intending to do so.

Another issue for fairness is where we have correct results, but they're misleading or they're unfair in some way. Consider a reputation system, a travel system where we are looking at user reviews and using that to choose a hotel to go on a vacation in. Hotel A that gets an average rating of 4.5 out of 5, and it turns out that it has got mostly 4 and 5 out of 5. There is another Hotel B which also gets an average of 4.1 out of 5 but this is a mix of mostly 3 and 5. We would choose Hotel A. Many reputation websites will just focus on the average, rank hotels by it, and order our search results and so on. And this important difference between the two hotels is going to be obscured unless we really look into it. So, hotel A, gets an average of 4.5, based on 20 user reviews. And Hotel B gets an average of 4.1 based on 500 user reviews. We would obviously choose Hotel B. 4.1 is less than 4.5, and so in terms of a sort order, 4.5 might come first. However, we all know that it is too easy on most sites to place a few false positive reviews. But now, if we also know that Hotel A has just 5 rooms, while hotel B has 500 rooms, that again changes our decision and we will probably prefer hotel A, and by a lot. Since hotel A has fewer customers, we should expect it to have fewer reviews. And so the fact that it has fewer reviews is something we should not hold against it, given that it is so much smaller than hotel B. Thus, the false positives and false negatives play a crucial role in defining fairness.

The initial consideration of what fairness means to any organizational culture should be leavened by a consideration of what fairness might mean to the full gamut of affected stakeholders. It therefore becomes very important that they remain transparent. This becomes tricky in case of the so called "black box algorithms"

In a technology context, we do not know the inner workings of the black box algorithms either because they are proprietary or not understandable to the interested parties. The interested parties could be the developers of the system, a third party evaluator or consumers. Some systems are difficult to understand and are quite opaque in the way they make decisions, so achieving transparency technically as well as giving commercial confidentiality may well be difficult.

It's not simply that AI algorithms can make mistakes, but that the whole ecosystem is a closed book to most. The consumers or stakeholders have a little understanding of how decisions that have real-world impacts on people are actually made. Thus the notions of understandability and interpretability become important. If a stakeholder is not able to interpret and understand the workings of a system and its outputs, he will never be able to incorporate fairness. Then, any end-user might forgo the need for complete and transparent access to the underlying algorithm and the dataset if easily understandable information about the system is provided to them by a qualified, trustworthy expert or entity. When people are able to understand how something works, they are more likely to use the system appropriately and to trust those who develop and deploy it.

Trust comes from accountability that means knowing 'things can go wrong'. Accountability for an algorithm and how it is applied begins with those who design and deploy the system that relies on it. Ultimately, designers and deployers share responsibility for the consequences or impact an algorithmic system has on stakeholders and society. They need to analyze if an AI application does exactly what it is designed to do. They need to think about all possible failure modes of an algorithm and actively mitigate the probabilities of high risk failures. Accepting the possibility of unintended consequences is an important element in accepting accountability.

From that starting point, fairness, transparency and accountability issues can be identified and characterized, ideally through a comprehensive internal assessment assisted by an outside, impartial third party. The findings of such an assessment can then be integrated into design and organizational practices, including the involvement of human agency if the identified risks are unacceptable. Such an approach tempers the unrealistic drive to create the "perfect algorithm," and instead, allows us to build an ecosystem of developers, technology and end-users that operates as a fair, transparent, and accountable system.

For example, in existing safety-critical systems like railway software, there is an inspectorate that is only allowed access to commercially sensitive material. It is never made public because, that is a confidential exchange. Only the abstract or high-level details of what went wrong are made public, but not the intellectual property behind the decision. Similarly, we can have a certification system in place before an algorithm is deployed, with a seal that says it's fair and accountable.

Lastly, machines are now able to take on less-routine tasks, and this transition is occurring during an era in which many workers are already struggling. Nonetheless, with the right policies we can get the best of both worlds: automation without rampant unemployment. Automation anxiety is made more acute by a labor market that has tilted against workers over the last 30 years, with increasing income inequality and stagnant real wages. Though there is still much we don't know about how this wave of automation will proceed, there are several areas of action we can identify now. According to a case study in West Virginia, a community of coal minerssaw coal mining evaporate and they were losing all their jobs. A couple of community leaders came up to teach those coal miners to code and write apps. When the short-term pain is inevitable these people were committed to not destroy the community. It is an amazing story of the rebirth of the community from coal mining to being a mini tech hub.Eventually, human ingenuity changes the role of productive work.

While a theoretically perfect AI morality machine is just theoretical, there is hope for using AI to improve our moral decision-making and our overall approach to important, worldly issues. AI could make a big difference when it comes to how society makes and justifies decisions. If we could paint a clearer picture of how our actions will affect people, ranging from everyday decisions to massive social or international programs, we could likely improve humanity and make decisions better rooted in justice and fairness.

Thus, to have a general overview, AI needs ethical strongholds in these areas:

1. Complex adaptive AI systems leading to self sustaining malicious evolution:
    a. AI that can mimic a cancerous growth in human body.
    b. We need to combat such systems using superior evolving AI systems.
2. Cyber-security:
    a. In the age of technology, there will be cyber wars.
    b. Any autonomous system will be used for malicious reasons if hacked.
    c. Such vulnerabilities of AI systems should be checked so that they stay safeguarded against such attacks.
    d. Mock attacking AI systems should be developed that would immunize the existing safeguarding AI system.
    e. There should also be systems that would predict the new type of attacks that can arise.
3. Cyber-resilience:
    a. Any organization must develop systems to quickly recover from cyber events that disrupt usual business operations.
4. Veracity:
    a. AI systems must be developed to check the authenticity of any and every broadcast that is socially made.
5. Privacy:
    a. Critical information that may be maliciously used against an individual should be automatically identified and should be kept private even if people share it publicly.
    b. AI systems can be used to ensure such privacy as well.

6. Consumer ethics:
    a. Cultivation of ethical teachings among young students in schools
    b. Spreading awareness about the ethical standards through media.
7. Transparency:
    a. Letting the end users as well as the stakeholders know that a cyber breach has occurred.
    b. Ensure vendors/suppliers/third parties have adequate security.
    c. The consumers or stakeholders must understand how AI systems work.
8. Certification:
    a. We can have a certification system in place before an algorithm is deployed, with a seal that says it's fair and accountable.
    b. Such a certification or quality check can be done by some AI algorithms as well.
9. Fairness without Bias or Prejudice:
    a. Analysis technique should be completely neutral.
    b. Training data must come from unbiased sampling.
    c. System should automatically detect any bias present in it.
10. Accountability:
    a. Knowing 'things can go wrong'.
    b. Designers and deployers share responsibility for the consequences or impact an algorithmic system has on stakeholders and society.
    c. Analyze if an AI application does exactly what it is designed to do.
    d. All possible failure modes of an algorithm should be thought of
    e. Active mitigation of probable high risk failures
11. Unemployment:
    a. Promote educational opportunity and make more skilled labor.
    b. Transitional Assistance to Needy Families.
    c. Opportunities for re-skilling and up-skilling

Recommendations:

Guidelines on Ethical Issues: There is a need to evolve comprehensive guidelines on ethical issues such as fairness, transparency and accountability in consultation with the stakeholders including civil society. While formulating the guidelines, we should also consider the code of ethics for IT systems developed by professional societies like IEEE. It can't be one-time exercise; the guidelines have to be reviewed and updated periodically.

Infrastructure Development: In order to promote the responsible uses of AI, government should invest in the development of bias-free datasets and techniques/tools for building fairness, transparency and accountability features in the systems.

Testing and Certification: Necessary resources need to be created for testing and certification of AI systems for the desirable ethical features such as fairness, transparency, etc. An independent organization may be entrusted with this responsibility.

Incentives for Compliance: Incentives should be provided for compliance.For instance, when Government procures IT systems to provide different services to the public, compliance to the guidelines on ethics should be made mandatory.

Public Empowerment: Awareness on the ethical issues must be created and information on the working of the AI system should be shared with the public so that people can seek explanation if decision of an AI system makes an adverse impact on them. This will put required pressure on the system designers to incorporate the desirable features.

11. Ethical Issues & AI

Algorithms, particularly with the development of artificial intelligence (AI) and machine learning (ML), are powerful tools that could provide crucial help in advancing science and research, improving access to medical care, and tackling some of humanity's most pressing global challenges in the environment, transportation, and beyond, as well as driving smarter solutions to everyday problems.

It is important to maintain an environment in which innovation is not stifled and users are protected. Instead of rigorous disclosures, it would be more valuable for users to understand the inputs that go into algorithm design and how they help achieve desired outputs. At the same time, it is vital to promote public understanding about the purpose of algorithms and their responsible development and application. Similarly, consensus-driven best practices and self-regulatory bodies that can create flexible and nuanced approaches for regulation are the best way of promoting responsible and ethical development while not inhibiting the potential of rapidly evolving AI technologies to solve major economic, social and environmental challenges.

Bias

Bias in algorithms can be a risk—but it should also be remembered that bias is present in many existing non-algorithmic processes and well-designed algorithms could reduce bias and make processes fairer. Best practices in identifying and eliminating potential bias within AI includes ensuring diversity in the sources of data, scrutinising initial data-sets and inputs in order to eliminate bias, continually testing and encouraging feedback, supporting research on practical techniques of promoting fairness and involving multi-stakeholder groups in these feedback processes. Such methods can be formalised using consensus-driven best practices created by self-regulatory bodies allowing for flexible and nuanced approaches.

At the same time, it is important to emphasise that AI technologies should be responsibly designed and implemented—particularly in the public sector. Improvements in cost and efficiency unlocked through algorithmic decision-making should be balanced against the need to ensure equality, accountability, and democratic participation in the creation and provision of services, especially those being provided by the government.

Decision Making

It is clear that algorithmic decision-making will not always be the optimal means of resolving challenges—they aren't a panacea or a catch-all solution. These technologies do not remove the importance of respecting laws, public consent and democratic participation, and in some cases it may well be paramount to preserve the deliberative, context-sensitive decision-making that human involvement can provide. In many cases, feedback loops involving between AI and human decision-makers may be optimal.

As with earlier waves of technological advancement, social values remain important. Convening and supporting robust discussion and collaboration between technology experts and civil society is a key role that government can play as it seeks to fully benefit from these advances. It is important to consider how the ecosystem can maximise the positives of emerging technologies for society whilst minimising potential harms.

Rigorous research can provide a path towards answering the societal questions that increased use of sophisticated algorithms raise. Research should be interdisciplinary, but computer science provides a unique opportunity to develop concrete technical mechanisms and frameworks that can be deployed to diminish or eliminate bias. The degree of scrutiny applied to algorithms should vary depending on the context and use case, taking into account the significance of the decisions that are being made and the existence or absence of any potential harm to individuals. This would be best done sectorally and on a subjective, case to case basis with multi-stakeholder consultations.

There should not be a blanket prohibition on evaluative decisions taken on the basis of automated decision making processes. By imposing onerous restrictions at an early stage of AI's development, India would be limiting the potential of AI technology to bring benefits for a country like India, where a vast number of people still have difficulty accessing ICT services, where AI can help bridge these gaps. According to a recent Accenture report[1], AI is expected to raise India's annual growth rate by 1.3 percentage points—in a scenario where intelligent machines and humans work together to solve the country's most difficult problems. This amounts to an addition of US$957 billion, or 15 percent of current gross value added, to India's economy in 2035 compared with a scenario without AI. Working with the research community to develop concrete tools and practices (including feedback loops between human and AI decision making) to reduce discriminatory outcomes would enable India to reap the benefits of AI while also contributing fair outcomes for all Indians..

12. Understand and Tackle Underlying Factors

The myriad of issues at hand with regard to AI are complex in terms of both their causes and impacts, and thus will not be remedied by one single solution. Therefore, a better understanding of the causal or contributing factors is critical to designing effective methods to reduce or eliminate the bad outcomes algorithms can produce.

Data gaps[2] is one such one factor. Just as many have expressed concern with the digital divide, there is a need to ensure there is not a data divide. As recent cases make clear[3],

[1] https://www.accenture.com/in-en/insight-ai-economic-growth-india

[2] http://www2.datainnovation.org/2014-data-poverty.pdf

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

a lack of good data, or poor quality, incomplete, or biased data sets are problematic in the practices and biases they perpetuate and can potentially produce inequitable results in algorithmic systems. "Growing the artificial intelligence industry in the UK"[4], the independent review carried out by Professor Dame Wendy Hall and Jérôme Pesenti addresses this issue by providing recommendations to improve access to data. These include a proposal that government and industry should deliver a programme to develop Data Trusts—proven and trusted frameworks and agreements—to enable fuller provision of publicly available data and ensure exchanges are secure and mutually beneficial. It is also important to be cognizant of the ways in which bias and unfairness operates in the world today, which data reflects. These data trusts could follow a sectoral, multi-stakeholder and self governance approach within the framework of AI governance.

Establish Accountability Framework

Though algorithms have long been used, their technical sophistication and breadth of use across sectors is ever increasing. A flexible framework for assessing the appropriate scope and technical feasibility of various accountability mechanisms will help , to maximise the benefits algorithms afford and mitigate potential drawbacks. The ideal approach to implementing this is multi-faceted:

- Develop flexible principles that can guide self-regulatory efforts to identify the most appropriate accountability mechanism (matching oversight approach with the appropriate technique) Not all algorithms are created equal, and a given oversight framework, based on self regulation, may be easier or more effective to apply in some cases than others. Decision tree learning, for instance, is a machine learning technique that may lend itself to simpler means of assessing errors than more recent deep learning approaches, albeit at the risk of (sometimes) sacrificing accuracy.
- Consider algorithms in the context in which they are being applied and the potential social and economic impact of their application, with multi-stakeholder input.
- Establish best practices to scrutinise the data sets used to train the algorithms and seek to ensure that:
    - Data sets are not incomplete, inaccurate, biased, or over or under representative of particular communities.
    - Data labels are accurate or accounted for—both on the level of individual data points, as well as for 'metadata' about the dataset (eg: source, date/manner of preparation, etc)
    - Data sets are appropriate for the use case, with sensitive data used appropriately and in line with applicable laws and the principle of minimising the use of personally identifiable information where possible.
- Internally test and audit the algorithms. While precise procedures will depend on context, this research paper "What's your ML test score? A rubric for ML

[4] https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

production systems”[5] can be referred to for some best practices for the robust testing and monitoring of ML systems.

- Engage with civil society. There is a need to emphasise the benefits of leveraging valuable expertise and experience to identify potentially high risk applications or contexts, such as criminal justice, and develop solutions.

Algorithms and data can be used to extraordinary effect in spotting and fightingbias and inequality. Indeed, they already are. For example, the Geena Davis Institute on Gender in Media to help develop the Geena Davis Inclusion Quotient (GD-IQ), a software tool that uses machine learning technology to help researchers analyze gender representation in popular film with unprecedented speed and accuracy and better equip advocates fighting for gender equity. The Geena Davis Institute recently released “The Reel Truth: Women Aren’t Seen or Heard”[6], the first report using findings from the GD-IQ.

Given the broad range of decision making processes that might be labelled “algorithmic”, it is important to recognise that the context in which an algorithm is used is central to determining best practices that should be applied to it. Purposes vary widely: an algorithm might be used to enable someone to find a family photo more easily in their collection, but it also might be leveraged to inform issues of life-changing import, such as housing, education, healthcare, access to finance and issues around criminal justice. Each of these must be governed proportionately, sectorally and keeping the interests of all stakeholders in mind.

Explainability

Algorithms that use modern developments in the field of machine learning sometimes present unique challenges in terms of explainability, due to their complexity. Overcoming the trade-off between interpretability and performance for complex machine learning models are among the most researched area in this space. These efforts will provide us with clearer explanations over time, even if there are limits to what is possible now.

Researchers ability to understand how such systems render decisions is consistently improving year on year, with work on interpretability happening at major technical conferences, such as Conference on Neural Information Processing Systems (NIPS). Advancing this is a priority for the field, not only because it is key to boosting trust in the results of such models, but also because it’s likely to yield insights that lead to further improvements. Examples of research projects in this field include:

[5] https://research.google.com/pubs/pub45742.html

[6] https://seejane.org/wp-content/uploads/gdiq-reel-truth-women-arent-seen-or-heard-automated-analysis.pdf

  - Distill: Open AI and others have established Distill, an independent organisation to support a new open science journal and ecosystem supporting human understanding and clarity in machine learning.
  - Deep Dream: In its earliest incarnation this project was aimed at visualising what different layers within a neural net were learning during training, to make it easier to spot where mistakes in classification arose.
  - Glassbox is a machine learning framework optimised for interpretability. It involves creating mathematical models to smooth out the influence of outliers in a data set, thus helping to make the results more predictable and decipherable.
  - Big Picture is a data visualization group focused on leveraging computational design and information visualization to make complex data sets more accessible, useful and understandable.

Technical Solutions

- Input/output auditing is a technique where you examine an algorithm by selectively submitting different types of input and evaluating the outputs to see if they are as you would expect. This provides a way to test for indicators of whether the model is producing unfair effects (eg: if all inputs were the same except you switched gender from male to female, and it delivered a different output, that would suggest there may be some form of bias that you would want to investigate further). Through systematic variation, it can also help to give a sense of the relative weightings given to different input elements.

- A different situation is when the system has been used to make a decision that you are querying. Here, regardless of what you may know about the general functioning, what is most important is to understand the rationale behind just that particular decision. This is still an area of active research, but examples of techniques being proposed include:
  - Heatmap visualisations show which pixels in an image are the most influential in an image classification decision. Or, if you wanted to apply it to text instead of an image, it can highlight the most influential words.
  - Counterfactual explanations seek to describe the smallest change to the input variables that can be made to obtain a desirable outcome for a decision. Eg: "You were denied a loan because your annual income was Rs. 30,000. If your income has been Rs. 45,000 you would have been offered a loan." In theory, multiple counterfactual explanations could be provided to cover those cases where the desired outcome could be achieved by tweaking multiple variables.

The precise mathematics underpinning the analysis of the most influential factors varies. Two of the most commonly used techniques are:

- *Sensitivity analysis (SA)* - this is the easiest method and assumes that the most relevant inputs are those to which the output is the most sensitive (for mathematicians, it is a form of gradient analysis). BUT one problem is that in practice this may not always be true (eg: imagine a picture of a rooster standing partially hidden by yellow flowers. The pixels with yellow flowers may be deemed very important to determining whether it was of a rooster if varying revealed more of the rooster, even though yellow flowers have no relevance to roosters).
- *Layerwise Relevance Propagation (LRP)* - this approach is more complex and seeks to directly identify those inputs which are the most pivotal in terms of making the classification (rather than just those to which it is the most sensitive). It works by computing a relevance score for each input (eg: in the case of a picture, a separate score for every pixel). There are different ways this score can be mathematically calculated but a common approach is to use a technique called Deep Taylor Decomposition

Increasing Trust

The choice of how best to deliver trust should hinge on the demands of each specific use case, the manner in which algorithms are deployed, and the practicalities of enforcement. Even when a greater degree of transparency is warranted, it is worth noting that there are many ways that it could be implemented. For instance:

- Inputs and Outputs: The functioning of an algorithm might be examined by selectively submitting different types of input and evaluating the outputs in order to provide an indicator of whether it might be producing negative or unfair effects.
- Spotlighting Logic: You could give a visual indication of key metrics relating to an algorithm's functioning without going into the full complexity, akin to the way that car dashboards have gauges for speed, oil pressure, and so on.

Certain transparency designs may also have their limits. Proposals for code or data to be disclosed in its raw form, for instance, are not a meaningful or effective means of creating trust or accountability. Just as privacy notices are encouraged to be written in clear, accessible, and easy to understand language in order to provide effective transparency, so too might a flood of technical details fail to provide adequate notice or understanding about the critical characteristics of a technology.

These "raw" transparency proposals may also generate their own unique problems. Exposing the code, even if just to a small group in a controlled setting, magnifies the risk of gaming and hacking. There is a real risk that transparency could end up hindering more than helping— bringing more error into the system by making it harder to protect. Many of the most critical services run by the government, identity and encryption services for example, would be rendered ineffective if the code were made public. Moreover, some aspects of algorithms and the data underpinning them will, by nature, be proprietary. Mandating disclosure of such materials would conflict with long-standing data protection obligations and legal protections for trade secrets, as discussed below. Allowing unique self-regulatory bodies can also contribute to creating flexible and nuanced approaches

Key stakeholders in AI/ML technology could also help develop better understanding and explainability of algorithms, by funding research on this issue.

- Such research could span a wide range of disciplines and types of outputs, from developing technical tools that automate explainability, to tools that review automated decisions to scan for any harms or biases, to research that highlights key ethical factors for regulators to focus on when reviewing algorithms.
- In time, this could help identify more sophisticated best practices, or even seals and certifications, that regulators could rely on for oversight and that individuals could rely on as a measure of transparency and trust.

Cross Border Flow of Data

According to a 2016 McKinsey report, global data flows now account for a larger share of the increase in global GDP than the global trade in physical goods. All types of flows acting together have raised world GDP by 10.1 percent over what would have resulted in a world without any cross-border flows. This value amounted to some $7.8 trillion in 2014 alone, and data flows account for $2.8 trillion of this impact.

India has been one of the biggest beneficiaries of the global data flows being the world's largest sourcing destination for the IT- BPM (Business Process Management) services. This industry accounts for 55% of the world’s outsourcing market and is worth US$ 173-178 billion. Forty-one percent of the IT exports come from the BFSI sector alone. It is India’s largest private sector employer, employing ~3.97 million people. The IT-BPM industry has a share of 45% in India’s total service exports and contributes ~7.9% to India’s GDP. It isn’t just technology companies who rely on this global cloud network to conduct business. Companies from all industries including agriculture, retail, and banking use Internet technologies to store data globally, so that they can provide faster, more reliable services at lower cost.

This cross border flow of data is essential for the natural and progressive development of the availability of data and the inculcation of skill sets in India’s workforce for technologies such as AI. Limitations on the free and open flow of data can seriously hinder the ability of an economy to remain competitive in the modern globalized world. Therefore, in order to preserve the development taking place on AI technologies in India, it is vital the free flow of data and information be maintained in India.

Data driven innovation and data driven governance, both essential paradigms for the sustainable use of artificial intelligence, are critical for India’s digital economy. A preventing harm principle, as recognized by the APEC Privacy Framework is a framework which helps achieve protection of data on one hand, while balancing innovation on the other. Collection and processing of personal data should be allowed with minimal restrictions provided it is coupled with user transparency, empowerment and control (e.g. easily opting out of a service later) and organizational accountability. This will also increase user trust in AI technologies, which is essential for widespread adoption.

Data-driven innovation for AI use cases and privacy are compatible with each other. Data-driven innovation cannot be scaled without adequate privacy safeguards and gaining users' trust. It is critical to empower the users, without over-regulating data collection, which could negatively impact the development of AI in the space. The framework for this should be outcome driven - legislation alone is not enough unless supported by an adequate implementation ecosystem including an effective grievance redressal system and user awareness. Instead of a prescriptive approach which is weakly enforced, the framework should aim for a light touch outlook which is strongly enforced. There are also intergovernmental arrangements like APEC Cross Border Privacy Rules (CBPRs) which are based on mutual recognition (unlike one way recognition systems like the EU adequacy model), accountability and commonly applicable privacy principles that enable efficient cross border data flows without unnecessary administrative burdens, that can help with the safe adoption of AI technologies.

Regulatory Outlook

AI can impact society in a variety of ways, which is why the government has such an important role to play alongside other stakeholders to ensure positive outcomes. Horizontal regulations on AI technologies would limit innovation, make it difficult for the law to adapt to technology changes and ineffectively govern the use of AI in critical areas such as access to justice, which would require an outlook that a general law cannot provide. Therefore, a sectoral approach to regulating AI would allow for greater flexibility, more effective implementation and targeted approaches that could better govern unique sectors than a general law.

A number of existing generic laws (such as the IT Act SPI Rules) and sectoral regulations, from health care to transportation to communications, already govern technology implementations. Sectoral experts will typically be the best placed to assess context-specific uses, assessing the impact and results of new technologies, but may need support to build AI expertise. As AI advances, the government could expand its technological expertise and explore various multi-stakeholder cooperative frameworks to minimize issues and maximize AI's potential.

Laws and regulations can be an important backstop in ensuring fundamental lines are not crossed. However, because the risks are so specific to the context and use, rather than appointing an overarching entity (like an agency with broad regulatory authority over robotics or machine intelligence), oversight would best provided via existing sector-specific entities. They will be best suited to evaluating whether or not the existing body of rules are sufficient or need to be revised to meet new technological realities (e.g. driverless cars regulated by vehicle-safety bodies; drones by aviation authorities, etc).

Instead of prescribing AI practices in the form of administrative requirements like format of notice and other codes of practice, including assessment/audit standards; the AI framework should define the broad principles and guidelines/requirements and allow organizations to design their own programs in compliance with these principles, with flexibility to adapt as the technology continues to evolve at a rapid pace. The focus should be to define the reasonable security practices and improve internal governance mechanisms in organizations without introducing excessive bureaucracy. While organizations should be allowed to self-regulate, they should be held accountable for any violations. In case of any breach or complaint, the onus to prove due diligence should lie with the organizations.

Recommendations

Civil Liability: As AI systems take independent decisions using the knowledge learned by itself, these are likely to be held responsible for civil liability in long-term. In view of this, stakeholders need to deliberate whether to recognize AI system as a legal person. If legal personhood is conferred, it should be accompanied by an insurance scheme or compensation fund to compensate for the damages. As such systems are to have global presence, these issues have to be discussed at international forums too.

Holistic Approach: A committee of the stakeholders should be constituted to look into all relevant aspects issues in a holistic manner. In order to give a fair opportunity to the technology, a decision on permitting AI systems should be made after considering the increase in the risks as well as decrease in the risks due to adoption of AI.

Review of Existing Laws: The existing laws should be reviewed for any modification which may be necessary for adoption of AI applications in the domain. Preference should be given to modifying the existing provisions in the laws rather than making new provisions altogether. Excessive regulations may be avoided as it may hinder the growth of the technology.

Prioritize Sectors: The review of the laws should be initiated in the sectors where early deployment is expected. These could be transportation, healthcare, finance, etc. The experience gained in these sectors can be used in other domains when the need arises.

Periodical Review: This can't be one time exercise. The laws should be reviewed periodically in view of the development of technology and experience with implementation of the laws.

Other suggestions:

- Commission a sector-by-sector analysis both under the Indian and International laws and best practices to understand how existing regulatory schemes apply to AI-enabled systems, and what gaps (if any) exist, especially in the context of the IT Act.
- Identify any existing constraints which hamper responsible use of AI and seek a solution

    - Eg: inferring race is essential to check that systems aren’t racially biased, but existing laws around discrimination and privacy can make this problematic
    - Eg: on-device AI has different risks and characteristics from cloud AI, and the nuances of this may not be reflected by “one size fits all” data protection rules
    - Eg: copyright rules can restrict data available for use in training AI systems, which may undermine efforts at reducing bias if data from key segments are excluded
- Funding to boost inhouse tech expertise for regulators in the most heavily impacted sectors
- Appoint an advisory committee or lead POC to coordinate on AI governance issues, including representation from the AI research community, industry, and civil society
- Engage with governance bodies around the world to share and learn from experiences
- Encourage industry to share best practices and promote codes of conduct
- Ethics training for government-funded researchers (analogous to research ethics training required)

Illustrative scenarios

Human & Machine Interactions

Description	Threats	Risks
Defining Autonomous Vehicles An autonomous vehicle is one that can guide itself without human conduction or intervention. Amsterdam declaration makes a distinction between connected cars (including cooperative driving(**ENI1)**: communication between vehicles and also with the infrastructure (CITS)) and automated driving (referring to the capability of a vehicle to operate and manoeuvre independently in real traffic situations, using on-board sensors, cameras, associated software, and maps in order to detect its surroundings) Assets present in a typical smart vehicle include but may not be limited to : • **Powertrain control** • **Chassis control** • **Body control** • **Infotainment control** • **Communications** • **Diagnostic and maintenance systems** Modern cars are composed of many embedded Electronic Control Units (ECU) that control mechanical or electronic systems of the vehicle **AI in autonomous vehicles**	A driverless car is a very advanced mode of transportation, possibly without even a readily available steering wheel. They have considerably more electronic components than "traditional" cars, and rely on sensors, radar, GPS mapping, and a variety of artificial intelligence to enable self-driving. These new guidance and safety systems are being integrated into the electronic on-board systems already present in modern-day vehicles, can connect wirelessly to the manufacturer, and probably even offer third-party services via the internet. That's where the problems begin: hackers who manage to **remotely access a vehicle** and compromise one of its on-board systems may be able to engage in a range of criminal activity, from **privacy and commercial data theft**, to imposing actual physical risks to people and property. These vehicles will have to anticipate and defend against a full spectrum of malicious attackers wielding both **traditional cyberattacks** and a new generation of attacks based on so-called **adversarial machine learning** Some of the threats that are likely applicable to the smart cars and autonomous vehicles include but may not be limited to the following :	**Security, safety and privacy Risks** • Compromising powertrain or chassis ECUs and networks may obviously cause a vehicle to behave in an unexpected way, for example if an attacker illegitimately compromises ignition, steering, brakes, speed and gear control, or driving support • Infotainment ECUs and networks may also cause safety issues : incorrect navigation data may lead the car to unsafe areas, and a disturbance of the audio in the entertainment system (such as high volume burst) may distract the driver • Internal networks (for example the CAN bus, but it also includes wireless networks such as Tire Pressure Monitoring Systems (TPMS)): a disruption or integrity breach on these networks may result in a loss of control of a vehicle

Artificial intelligence is an integral component in autonomous vehicles, and responsible for some of the recent technological advancements.

To quote an example, one of the **(Nvi)**capability providers has the DRIVE platform as an in-car AI supercomputer for autonomous driving. This entails numerous **deep neural networks** (DNNs) to be running simultaneously, driving the vehicle autonomously. Several types of DNNs are used to support autonomous driving. One is a **detection and classification network** used not only for object detection (pedestrians, cars, trucks, motorcycles, bicycles, signs, lampposts, and even animals), but also for lane marking detection. Another example is a **segmentation network** which is useful to determine the free space around the vehicle that is available for driving, driving bounds (typically bounded curbs and medians), and blocking objects such as vehicles and pedestrians. A third example is end-to-end networks that **mimic learned driving behaviour**. This can be used as a basic path planner to drive the vehicle in normal or typical circumstances. These networks working together to enable the vehicle's artificial intelligence to drive the vehicle, while keeping the driver, its occupants and pedestrians nearby safe.

V2V Communication

vehicle-to vehicle (V2V) communications, **(NHT)**a system designed to transmit basic safety information between vehicles to

- **Physical Threats**- This category entails Side channel, fault injection, glitching, access to HW debug ports.

- **Unintentional damages**- Unintentional damages may cascade from ill-defined trust relationships: for example, trusting a third party cloud provider with poor data protection, or failing to notify a Tier developer that the data they will store is sensitive

- **Denial of Service**- The denial of service is not only to be understood as a particular form of network outage. A denial of service may also be triggered on internal network by **flooding a CAN bus, or by provoking faults on an ECU via a malicious payload**. The potential impact of such an attack depends on the targeted ECU, but may lead to unexpected behaviours from driving systems

AI based threats- an example

The computer vision and collision avoidance systems under development for autonomous vehicles rely on complex machine-learning algorithms that are not well understood, even by the companies that rely on them.

Researchers have already demonstrated that state-

- Cellular connection of the car may also have adverse impacts on safety, for example in the case of a spoofed firmware update triggered by SMS

- V2X communications, which could lead to accidents, were they disrupted or spoofed

- Data confidentiality and privacy are eventually at risk as well. For example, compromising embedded cameras may lead to privacy issues for the driver and passengers.

- Trade secrets may be at risk in several systems: TCU/ECU firmware, which might be sensitive with regard to the competition. Some industry actors, in particular, may be wary of the possibility of device cloning (for example the cloning of aftermarket products)

facilitate warnings to drivers concerning impending crashes V2V communications use on-board dedicated short-range radio communication devices to transmit messages about a vehicle’s speed, heading, brake status, and other information to other vehicles and receive the same information from the messages, with range and “line-of sight” capabilities that exceed current and near-term “vehicle-resident” systems -- in some cases, nearly twice the range. **Example of HMI for an autonomousvehicle (working) (GM)** Customers will begin their interaction with self-driving vehicle before they get in the vehicle by using a mobile application to request a ride. Once inside the vehicle, the customers will use touchscreen tablets with an intuitive interface allowing riders to control the HVAC and radio, access general information about the vehicle, and receive real-time status information pertinent to the current ride. Before the ride begins, the tablets will provide helpful safety reminders, such as to close all doors and fasten seat belts. After customers enter the vehicle and meet all preconditions, such as closing the doors and pressing the begin ride button, the vehicle will start to move. At any point, a customer having an emergency may end the ride by making a stop request, and the vehicle will pull to the side of the road at the next available safe place. If the vehicle	of-the-art face recognition algorithms could be defeated by wearing a pair of clear glasses with a funky pattern printed on their frames. Something about the pattern tipped the algorithm in just the right way, and it thought it saw what wasn’t there	

has a malfunction, it will provide explanatory information to the passengers, as well as offer communications with a remote operator		

Machine Based Decision Making and software vulnerabilities

Description	Threats	Risks
With the current rapid economic growth, vehicle ownership is fast increasing, accompanied by more than one million traffic accidents per year worldwide. According to statistics, about 89.8% of accidents are caused by driver's wrong decision-making. In order to alleviate traffic accidents, autonomous vehicles have been the world's special attention for its non-driver's participation but it's no secret that developers of automated vehicles face a host of complex issuesto be solved before self-driving cars can hit the road, from building the necessary infrastructure and defining legal issues to safety testing and coping with the vagaries of weather and urban environments. Along with these issues they neglect the vital	Cybersecurity threats unique to automated vehicles, including hackers who would try to take control over or shut-down a vehicle, criminals who could try to ransom a vehicle or its passengers and thieves who would direct a self-driving car to relocate itself to the local chop-shop but what kind of threats would likely to hit automated vehicle. Following cyber threats classification taken into consideration (Uni1) **Internal versus external** Internal attacker threat is part of network who is capable to communicate with system internally, external threats is considered by network as an	With advance in artificial intelligence, the risk of hackers using such technologies to launch malicious attacks are increasing. They could use such AI to turn autonomous cars into potential weapon against human. For example self-driving cars could be tricked into misinterpreting a stop sign that might cause road accidents. Intelligent machines (AI equipped bad bots) could be used against the AI based decision making systems. Potential risk to the automated vehicles may be considered in following three scenarios • **Industry Risk**

issue of cybersecurity in automated vehicles.

Cybersecurity is an overlooked area of research in the development of driverless vehicles, even though many threats and vulnerabilities exist, and more are likely to emerge as the technology progresses to higher levels of automated mobility. Several advance AI -deep learning models are developed to obtain the inherent complexity of driving decision but while adopting machine based decision making models it is mandatory to study the cyber threats, risk involved in the overall AI- based system.

The Threats to automated vehicles can come through any of the system connect to the vehicle's sensors, communications applications, processors, and control systems as well as external inputs from the other vehicles, roadways, infrastructure and, mapping and GPS data systems.

In addition, Threats to the autonomous cars must reviewed by their motivations and capabilities. For Example. While two different attacker might focus on a vehicle's self -parking capabilities, for example, the threat of loan car thief trying to steal single vehicle would be significantly different from an organized group of dedicated hacktivists looing to hurt a manufacturer by disabling a huge number of vehicles. Along with motivations and capabilities, must to address the possible attack vectors, related threats and associated risk in autonomous cars.

intruder and limited in the diversity of cyberattacks.

Malicious versus rational

Malicious threats are not for personal benefits from attackers, aims to harm functionality of the system. Rational threats seek personal profit and, hence more predictable in terms of attack means.

Active versus passive

An Active attacker threats can generate the signals or any other mechanism to attack the autonomous car system where as passive attacker threats just eavesdrop on communication channels or system network**.**

Local versus extended

An attacker threat can be limited in scope, even it can control entities(vehicles or base stations), which make that attacker threat a local.

Extended attacker threats can control several entities attached with vehicles system.

Attack vectors

Autonomous automated vehicles may integrate more components, all components

- **Consumer Risk**

Industry Risk

The safety of driverless vehicles should be the paramount concern of the auto and insurance industries, if for no other reason than flaws and failures in automated vehicle systems will impose potentially enormous, even catastrophic liability upon hardware and software manufacturers in the event their products cause harm, and lead to more, and more costly, insurance claims. It is important for developers and manufacturers to understand the current threat landscape, what is currently being developed by criminals, and who among them have an interest in attacking self-driving vehicles and for what purpose.

Consumer Risk

Security Standpoint, vehicles connected technologies -including laser rage finders, cameras, ultrasonic devices, wheel sensors, and inertial measurement systems will be access points for hackers. If even one of these potential points of compromise is not properly secured, then entire decision making system cloud may crash down and may leads to following consequences (ENI2)

The core components of an autonomous vehicle are the computer, sensors and programs. Array of electronic devices will be accountable for detecting the external signal and conditions to navigate vehicle. The onboard computer system collects data from various inputs, process the information and make decisions that are presently executed by humans. Software analyse the sensors and communication data flow and instruct vehicles how to navigate. Particularly critical functions of the software will be replace the judgement of human motorist. Challenge will be not to just avoid collision or drive on correct path but also comply with traffic laws and rules. It will be therefore necessary that Software is bug free and not contain any vulnerability that may escalate into cyber attack**(MDP)**	will be interconnected and dependent on each other to make driving decision. We consider following attack vectors. **Machine Vision** Video signs (static and dynamic), Pictures used for direction (Road, Obstacles, stop signs, red lights etc.)**GPS** Global Positioning system used for driving directions. It direct the vehicles in specified direction. Attacks on Map Navigation system leads to destination change or very severe consequences. **In-Vehicle Devices** It concludes hand-held devices brought by user or car's interior devices connected with infotainment system via Bluetooth or Wi-Fi**.** **Sensors and Radar** Sensors are to sense the known signals and can intimate to take proper action based on the signal. (E.g. crash sound get detected) and then inform to open airbags. Radar system make use of microwave radiations to detect objects. Attacking Sensors and Radar System create complications in car operations and reactive response system. **Lidar**	**Physical Damage Risk** • Side Channel, fault injection, escalation to access embedded systems might lead to various type of risks such as nefarious activity / consumer abuse or eavesdropping or consumer hijacking. • Vehicle sensors attacks -GPS, light detection and ranging sensors and ultrasonic sensors are vulnerable to jamming and spoofing, these attacks can disable or otherwise affect connected cars controls. It can potentially lead to crashes and other safety concerns • Key fob done techniques, where vulnerability in key less entry system allow adversaries to eavesdrop on a signal sent by a remote control and gain unauthorized access which can affect consumer safety. **Automated Ransomware for vehicles** Through the advent of ransomware for connected cars, locking owners and even technicians out of vehicles until a ransom is paid. Or through the information theft and leakage, for resale into the

	Active system that that uses return of infrared(IR) or visible light instead of radio waves to detect the object **Specific Threats** AI based decision making systems play significant role in driving car, taking right or left, direction to follow, stop signs and many more. Decision making system is at the centre of autonomous cars. Following threats are to be considered for it. **Physical Threats** Certain attacks could be carried out by those with physical access to the vehicle. Vehicular systems that are exposed to passengers such as USB ports or OBD-2 ports might provide mechanisms to allow for malicious use or exploitation. As with other technological systems, physical access often bypasses controls that are specifically in place to prevent remote exploitation. **Sensor jamming, spoofing & blinding** Current approaches to self-driving automation leverage a variety of cameras, lasers, GPS, radar and other sensors to give the vehicle the environmental and situational awareness it needs. Each of these types of sensors can be blinded or jammed, thereby hindering the vehicle's ability to retain full awareness of environmental	thriving data economy. In addition to that hacktivist or terrorist groups taking advantage of software vulnerabilities to weapinise individual or even fleets of vehicles for use in attacks. **AI Against the AI Algorithms** With the rise of AI, AI is teaching bots to be more like humans. Mobile bots or smart AI bots implemented against the legitimate AI based systems. Adversaries are continuously implementing sophisticated attacking techniques. AI based attacks are faster, well executed and are more difficult to differentiate between real user and non-user, and human versus bots. AI based Bots might be used to conquer the ECU's which may leads to disruption of the Autonomous car operations. For Example Automatically acquiring controls over Map or GPS system might change the consumer destination and hijack the car. **AI based Network Attacks (Network Attacks)** Bugs and Vulnerabilities in network may leads to network outrages, directly affects availability – results in a denial of service for sensitive operations. Vehicles are relying on network connectivity for its operations, in case of network

	conditions or potential obstructions. **Forged vehicle communications** Another risk involving communications would be the forging of vehicle communications to spoof hazards that don’t exist or attempts to cause a vehicle to behave in ways it wasn’t designed or intended to. One potential problem revolves around protocols that lack cryptographically sound integrity checks. These protocols may be vulnerable to spoofing depending on implementation & communication methods. **Invasive Attacks** (Int) • **Code Modification** Through remote attacking methods bad guys makes unauthorized modifications to code or data, attacking its integrity. These attacks can take many different forms and have a variety of consequences. • **Code Injection** Code Injection, or Remote Code Execution (RCE) attack where in an attacker is able to execute malicious code as a result of an injection attack. When Software is not validated or tested, possibility of code injection attack is relatively high. • **Packet Sniffing** Packet sniffing, in which an application or device can,	failure or outrage entire vehicle fleet may face potential issues. **Automated Distributed Denial of Service:** Distributed denial of service not only on external but on internal network by flooding a CAN bus, or by provoking faults on an ECU via a malicious payload. The potential impact of such an attack depends on the targeted ECU, may lead to unexpected behaviours from driving systems. Driving system is at centre of autonomous vehicle system, unexpected behaviour from driving system can take control of entire car by sending unusual commands, car hijacking, traffic rule violations, mass accidents etc. are possible due to driving system compromise. **Privacy Risk** Modern cars have become computers on wheels, collecting significant amounts of data about the vehicle and the habits of the motorist that drives them. For example some insurance companies have installed “black boxes” in the vehicles they insure to track vehicle location, speed and other metrics. Automation technologies will collect, process and communicate vast amounts of information. The recipients of the data stream will include, eventually, other vehicles and likely the government agencies that operate the intelligent transportation grid. The data

	monitor, and capture network data exchanges and read network packets. If the packets are not encrypted, a packet sniffing provides a full view of the data inside the packet. - **Packet fuzzing** Automobiles are vulnerable to the attacks through the network-CAN which connects the ECUs (Electrical Control Units) embedded in the automobiles. Attack packets are constructed in a completely random manner without any previous information such as CAN IDs. The packets are injected into the network via Bluetooth, a wireless channel. **Non-Invasive Attacks** - **Side -Channel Attacks** Side-channel attacks are little different than usual cyber-attacks, Side Channel attacks are based on information gained from the implementation of a computer system, rather than weaknesses or vulnerabilities in implemented software or algorithm itself. Timing information, power consumption, electromagnetic leaks or sound can provide an extra source of information, which can be exploited.	is extremely valuable to software and hardware manufacturers and insurance companies, but could prove costly for consumers when it comes to profiling and usage of personal data without consent.

Scenario 5: Smart City

Description	Threats	Risks (EY, 2016)
A smart city consists of blocks such as smart governance, smart	Cyber security in the context of Smart Cities is a hot topic. The	**Cascading Impact**: The reliance on large automated

mobility, smart utilities, smart buildings, smart environment, smart surveillance etc. The smart city is based on a dense and heterogeneous set of IoT devices deployed over the urban area. It generates different types of data and augmented actions to deliver citizen-centric services. The components of an IoT architecture, would be as follows but not limited to:**(Direct, 2015)** • Urban IoT infrastructure capable of integrating with multi-technology model • Accessibility of data of IoT devices for the authorities and citizens • Responsiveness of IoT devices for solving smart city problems • Availability of end-to-end ICT infrastructure for the smart city to function • Web services convergence with IoT devices in a smart city • Availability of interoperable and open standards for communication of multiple IoT devices operating in a smart city • Sufficient gateways and peripheral nodes for the efficient functioning of a smart city • Integration of ICT architecture and IoT devices • Centralised command and control center for harnessing holistic data for assimilation and decision making	objective of Smart Cities is to optimize the city in a dynamic way in order to offer a better quality of life to the citizens through the application of information and communication technology (ICT). The range of areas where cities can become smarter is extensive: it is an evolution of "Connected Cities" with the prevalence of data exchange at a larger scale. The increase of data exchange controls multiple services and assets leads to a higher degree of automation in the city. As several critical services become interconnected and automated, the need for cyber security surges to protect data exchanges, privacy as well as the security and safety of citizens. The cyber **threat categories** taken into account with respect to smart cities are • Availability threats • Integrity threats • Authenticity threats Confidentiality threats • Non-repudiation/accountability threats **AI Specific Threats**(ENISA, 2015) **AI based Eavesdropping/wiretapping** is a deliberate act of capturing network traffic and listening to communications between two or more parties without authorisation or consent with help of automated tools. Recent experience has shown that wireless and cellular networks are vulnerable to eavesdropping	central servers to control systems, especially to control utilities such as water or power supply, can have a cascading impact in case of disruption. The efficacy of back-ups, measures to contain damage and quick response mechanisms, which are being discussed and implemented, will actually be tested only in a live situation. With large interdependent automated networks, security and risk-mitigation are big worries as is the time required for recovery. **Complexity in Face Recognition and Investigation**: What should be the policy on using publicly available information for training image recognition algorithms based on machine learning used in CCTVs of smart cities? The combination of face recognition with camera-equipped drones and easily accessible tagged photos makes a very powerful surveillance system available for everyone. The supervision of face recognition for intelligence purpose may become difficult in AI based smart city scenario. **Natural Language Processing Risks:** Natural language processing opens the possibility of processing unstructured data, we may see an erosion on purpose limitation for personal data

	equipment based on standard components. **Machine learning based Theft** refers to the unauthorised appropriation of information/data or technology. Theft may affect availability and confidentiality. Theft of cryptographic keys to decentralised ticketing systems, for instance, can cause serious financial and reputational loss. **Backdoor Tampering/alteration** aims at altering information/data, applications or technology with direct and potentially significant effect on availability and integrity. It is also relevant from the perspective of nonrepudiation/accountability **Automated Unauthorized use/access** can be at the source of other threats. Apart from eavesdropping/wiretapping, theft and tampering/alteration, it may also be that information/data, applications or technology are used/accessed in an unauthorised way with help of machine learning based Trojans. This includes unauthorised connection to a network, data leaks, and browsing files, acquiring private data, controlling field components and using resources for personal use. **Automated Distributed Denial of Service (DDoS)** consists in the usage of several sources connecting simultaneously to one destination, with the objective of overflowing the connection	stored in (old) documents. - Natural language processing uses like question answering and customer like interactions, may foster automated decision which may lead to complexities in regulatory oversight w.r.t consumer protection. **Reverse Engineering of Machine learning models used in smart cities**, much like any piece of software, are prone to theft and subsequent reverse-engineering. In late 2016, researchers at Cornell Tech, the Swiss Institute EPFL, and the University of North Carolina reverse-engineered a sophisticated Amazon AI used in transportation system by analysing its responses to only a few thousand queries; their clone replicated the original model’s output with nearly perfect accuracy. The process is not difficult to execute, and once completed, hackers will have effectively “copied” the entire machine learning algorithm—which its creators presumably spent generously to develop. Machine learning also faces the risk of **adversarial “injection”**—sending malicious data that disrupts a neural network’s functionality. In 2017, for instance, researchers from four top universities confused image recognition systems by adding small stickers onto a photo,

	leveraging IoT devices with help of AI cyber weapons. With the increase of IP-connected devices, DDoS are a main threat to smart city systems, in particular for devices and services relying on Internet connectivity. **Automated Malware Attacks** What happens when devices across a smart city (or the equipment at the power grid) are infected with AI based malware and instead of turning off the power, they increase power usage instead on their own. In the future, these types of cyber-attacks could affect street lighting, city management, traffic control, power, water grid, surveillance, public transportation, location-based services, and much more with no human intervention as required. **AI Device hijacking**: The AI system as an attacker hijacks and effectively assumes control of a device. These attacks can be difficult to detect because in many cases, the AI system does not alter the basic functionality of the device. In the context of a smart city, a malicious AI system could exploit hijacked smart meters to launch ransomware attacks on Energy Management Systems (EMS) or stealthily siphon energy from a municipality	through what they termed Robust Physical Perturbation (RP2) attacks; the networks in question then misclassified the image. Another team at NYU showed a similar attack against a facial recognition system, which would allow a suspect individual to easily escape detection system. **Privacy Risks**: With AI based integrated circuits (ICs) and microphones becoming cheaper, smaller, more powerful and easy to embed into all devices, snooping and an end to privacy is a reality even for kids—if for nothing else, to be able to sell you goods and services. Smart city technologies capture data relating to all forms of privacy and drastically expand the volume, range and granularity of the data being generated about people and places. Privacy can be threatened and breached by a number of practices which are normally treated as unacceptable, however are part of operations in a smart city eco system. **Surveillance**: Watching, tracking, listening to or recording a person's activities **Aggregation**: Combination of various aspects of data

		about a person to identify a trend or pattern of activities **Data leakage**: lack of data protection policies can lead to leakage or improper access of sensitive information **Extended usage:** use of data collected for period longer than stated or for purposes other than **Big data, Profiling and automatic decision making**: Complexities for data protection authorities (DPAs) support the right to information from the data subject when confronted with big data, artificial intelligence and machine learning. Evaluating the bias in automated decisions when artificial intelligence and machine learning may become an arduous tasksfor the regulatory authorities. **Simple Bugs with Huge Impact**: a simple software bug can have huge impact. As Smart Cities will run on hundreds of systems and devices managing critical services, a simple software bug can have huge impact. For instance November 2013 Bay Area Rapid Transit (BART): major software glitch, service was shut down by a technical problem involving track switching, it

		affected 19 trains with about 500 to 1,000 passengers on board. The system was based on machine learning techniques to manage trains. **Bandwidth consumption**: Thousands of AI based sensors, or actuators, trying to communicate to a single server without prioritizing their ICT needs will create a flood of data traffic which can bring down the server. The bandwidth consumption from billions of AI based devices will put a strain on the spectrum of other wireless communications, which also operate on the megahertz frequency.

Smart Energy

Description(MDPI, 2018)	Threats	Risks
Over the last decade, Smart Grids have led the revolution of the electrical grid, transforming itinto a set of automated and efficiently controlled processes by the incorporation of ICTs. Smart Grids promote electrical energy management in a distributed and flexible manner. However, the current management systems are (i) centralized regarding their management; (ii)	**Automated bot attacks**: With the rise of AI, AI is teaching bots to be more human-like. Mobile bot farms where bots are implemented on many thousands of devices to appear more human-like are just one	**Privacy and the Smart Grid** The Smart Grid brings with it many new data collection, communication, and information sharing capabilities related to energy usage, and these technologies in turn

located in independent locations between them; and (iii) managed by fragmented applications, without integration between them and only intercommunicated thanks to specific communication channels, which are generally proprietary The main goal of Smart Grids is to provide better services and features (also known as smart functions), for both consumers and for producers and prosumers. In addition, the increased use of distributed and renewable energy generation requires changes in the electricity management system. It is necessary to improve automation systems, distributed intelligence, real-time data mining and management to improve network control functions, simplify configuration and also reduce system recovery and self-healing times. Smart Grid architecture is composed of three main independent modules as follows. **Hybrid Cloud Data Management**: This module provides a data storage and processing system that intrinsically adapts to the Smart Grid's topology in a scalable and flexible way. Also, it implements an algorithm (also referred to as orchestrator) to decide whether collected data should be stored at the private cloud or could be placed at the public cloud. Such decisions are taken considering the smart functions' requirements (e.g., reliability, delay, and cybersecurity) associated with the collected data.	example – making it more difficult to differentiate between real users and non-users, and humans versus bots. Bots are increasingly being used in smart grid systems to respond to consumer queries. **AI Advance Persistent Threats**: A Stealthy, Long-term AI presence on your network will have ample time to learn what your working style is and how this differs depending on grid types, based on operator duties and the distinctions based on the intrusions fed into the grid systems.	introduce concerns about privacy. Four dimensions of privacy are considered: (1) personal information—any information relating to an individual, who can be identified, directly or indirectly, by that information and in particular by reference to an identification number or to one or more factors specific to their physical, physiological, mental, economic, cultural, locational, or social identity; (2) personal privacy—the right to control the integrity of one's own body; (3) behavioural privacy—the right of individuals to make their own choices about what they do and to keep certain personal behaviours from being shared with others; and (4) personal communications privacy—the right to communicate without undue surveillance, monitoring, or censorship.(NIST, 2010) **Algorithms Complexities:** Greater complexity in self-healing grid systems increases exposure to potential attackers and unintentional errors in grid system. **Automated Connected Networks:** Networks that link more frequently to other networks introduce common vulnerabilities that may now span multiple Smart Grid domainsand increase the potential for cascading failures. **AI increases the power of DDoS**: More interconnections present increased opportunities for "denial of service" attacks,

Web of Energy: This module provides a ubiquitous (i.e., web based) monitoring interface that enables a seamless management of the whole IoT architecture of the Smart Grid. In addition to providing a mechanism to communicate humans and machines, it also enables the interactions Sensors among those IoT resource-constrained and small devices (i.e., machine to machine) through the HTTP protocol. Note that the below mentioned context-aware security module might decide to add an extra layer of security by switching from HTTP to HTTPS in accordance with the device features, network status, and smart function under execution demands. This is done through an open API that both couples and decouples all the modules. Hence, this module also acts as a bridge between the distributed storage layer—that takes care of all the Smart Grid's Big Data concerns—and the context-aware security layer—that gives the necessary access control and cybersecurity mechanism. **Context-aware security**: This module aims to individually provide the needed security level for the proper operation of every smart function. For instance, for the use case of Smart Metering, this module can update all the encryption keys of the Smart Grid once an unauthorized access to the metering infrastructure has been detected.	**AI Malware**: An automated tool that learns how to mask a malicious file from anti-virus engines, by changing just a few bytes of its code in a way that maintains malicious capacity. This allows it to evade common security measures deployed on grid systems, which typically rely on file signatures – much like a fingerprint – to detect a malicious file.	introduction of automated malicious code (in software/firmware) or compromised hardware, and related types of attacks and intrusions. **Creation of grid nodes based on AI**: As the number of network nodes increases based on automated learning via smart grid requirements, the number of entry points and paths that potential adversaries might exploit also increases.

References for this Scenarios section

Smart City Architecture [Report] / auth. Direct Science. - [s.l.] : https://www.sciencedirect.com/science/article/pii/S1877050915009229, 2015.

Cyber Security : A necessary pillar of Smart Cities [Report] / auth. EY. - [s.l.] : EY, 2016.

Cyber security for Smart Cities [Report] / auth. ENISA. - [s.l.] : https://www.enisa.europa.eu/publications/smart-cities-architecture-model/at_download/fullReport, 2015.

Smart Energy Architecture [Report] / auth. MDPI. - [s.l.] : http://www.mdpi.com/1424-8220/18/2/400/pdf, 2018.

Guidelines for Smart Grid Cyber Security [Report] / auth. NIST. - [s.l.] : https://www.nist.gov/sites/default/files/documents/smartgrid/nistir-7628_total.pdf, 2010.

ENISA [Online]. - https://www.enisa.europa.eu/publications/cyber-security-and-resilience-of-smart-cars.

Nvidia [Online]. - http://www.semiconwest.org/programs-catalog/smart-automotive-future-smart-connected-self-driving-cars/artificial-intelligence-ai-and-self-driving-cars.

NHTSA [Online]. - file:///C:/Users/aditya.bhatia/Downloads/Readiness-of-V2V-Technology-for-Application-812014.pdf.

GM [Online]. - https://www.gm.com/content/dam/gm/en_us/english/selfdriving/gmsafetyreport.pdf.

MDPI [Online]. - http://www.mdpi.com/2076-3417/8/1/13.

University of Machigan [Online]. - https://mcity.umich.edu/wp-content/uploads/2017/12/Mcity-white-paper_cybersecurity.pdf.

Intelligent Tranportation System -2014 [Online]. - https://www.researchgate.net/publication/266780575 .

ENISA [Online]. - https://www.enisa.europa.eu/publications/cyber-security-and-resilience-of-smart-cars.

University of Michigan [Online]. - Potential Cyberattacks on Automated Vehicles” https://www.researchgate.net/publication/266780575 .