Artificial Intelligence & its Security Concerns

Shyamanth Kashyap, Pavan R Nargund, Prajwal JM

Information Science, NMIT
Bengaluru, Karnataka, India

***Abstract*— This paper dwells on the negative effects of Artificial Intelligence and Machine Learning, and its underlying threat to humanity. This study stems from the experiences of different people working in the aforementioned field. This paper also proposes a framework to regulate and govern the projects in this field to reduce its threat or any future repercussions.**

***Keywords*— Experience, Threat, Regulate, Repercussions**

I. Introduction

Artificial intelligence (AI) and increasingly complex algorithms currently influence our lives and our civilization more than ever. The areas of AI application are diverse and the possibilities extensive: in particular, because of improvements in computer hardware, certain AI algorithms already surpass the capacities of human experts today. As AI capacity improves, its field of application will grow further. In concrete terms, it is likely that the relevant algorithms will start optimizing themselves to an ever greater degree—maybe even reaching superhuman levels of intelligence. This technological progress is likely to present us with historically unprecedented ethical challenges. Many experts believe that alongside global opportunities, AI poses global risks, which will be greater than, say, the risks of nuclear technology—which in any case have historically been underestimated. Furthermore, scientific risk analysis suggests that high potential damages should be taken very seriously even if the probability of their occurrence were low.

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.

Deep learning is a set of machine learning algorithms that model high-level abstractions in data using architectures consisting of multiple nonlinear transformations. A deep learning technology is based on artificial neural networks(ANNs). These ANNs constantly receive learning algorithms and continuously growing amounts of data to increase the efficiency of training processes. The larger data volumes are, the more efficient this process is. The training process is called «deep», because, with the time passing, a neural network covers a growing number of levels. The deeper this network penetrates, the higher its productivity is.
A deep machine learning process consists of two main phases: training and inferring. You should think about the training phase as a process of labeling large amounts of data and determining their matching characteristics. The system compares these characteristics and memorizes them to make correct conclusions when it faces similar data next time.

A deep learning training process includes following stages:

1. ANNs ask a set of binary false/true questions or.
2. Extracting numerical values from data blocks.
3. Classifying data according to the answers received.
4. Labeling Data.

During the inferring phase, the deep learning AI makes conclusions and label new unexposed data using their previous knowledge.

Deep learning is a kind of traditional machine learning. Classical machine learning is the extraction of new knowledge from a large data array loaded into the machine. Users formulate the machine training rules and correct errors made by a machine. This approach eliminates a negative overtraining effect frequently appearing in deep learning.

In machine learning, users provide a machine with both examples and training data to help the system make correct decisions. This principle is called supervised learning. In other words, in a classical machine learning, a computer solves a large number of tasks, but it cannot form such tasks without a human control.

Diversity between machine learning (ML) and deep learning (DL):

- DL requires a lot of unlabeled training data to make concise conclusions while ML can use small data amounts provided by users.
- Unlike ML, DL needs high-performance hardware.
- ML requires features to be accurately identified by users while DL creates new features by itself.
- ML divides tasks into small pieces and then combine received results into one conclusion while DL solves the problem on the end-to-end basis.
- In comparison with ML, DL needs much more time to train.
- Unlike DL, ML can provide enough transparency for its decisions.

The concept of deep learning implies that the machine creates its functionality by itself as long as it is possible at the current time. To infer, deep learning applications use a hierarchical approach involving determining the most important characteristics to compare.

Challenges faced in Deep Learning :

Deep learning is an approach that models human abstract thinking (or at least represents an attempt to approach it) rather than using it. However, this technology has a set of significant disadvantages despite all its benefits.

Continuous Input Data Management :

In deep learning, a training process is based on analyzing large amounts of data. Although, fast-moving and streaming input data provides little time for ensuring an efficient training process. That is why data scientists have to adapt their deep learning algorithms in the way neural networks can handle large amounts of continuous input data.

Ensuring Conclusion Transparency :

Another important disadvantage of deep learning software is that it is incapable of providing arguments why it has reached a certain conclusion. Unlike in case of traditional machine learning, you cannot follow an algorithm to find out why your system has decided that it is a cat on a picture, not a dog. To correct errors in DL algorithms, you have to revise the whole algorithm.

Resource-Demanding Technology :

Deep learning is a quite resource-demanding technology. It requires more powerful GPUs, high-performance graphics processing units, large amounts of storage to train the models, etc. Furthermore, this technology needs more time to train in comparison with traditional machine learning.

Despite all its challenges, deep learning discovers new improved methods of unstructured big data analytics for those with the intention to use it. Indeed, businesses can gain significant benefits from using deep learning within their tasks of data processing. Though, the question is not whether this technology is useful, rather how companies can implement it in their projects to improve the way they process data.

II. Implications

In narrow, well-tested areas of application, such as driverless cars and certain areas of medical diagnostics, the superiority of AIs over humans is already established. An increased use of technology in these areas oers great potential, including fewer road traffic accidents, fewer mistakes in the medical treatment and diagnosing of patients, and the discovery of many new therapies and pharmaceuticals. In complex systems where several algorithms interact at high speed (such as in the financial market or in foreseeable military uses), there is a heightened risk that new AI technologies will be misused, or will experience unexpected systematic failures. There is also the threat of an arms race in which the safety of technological developments is sacrificed in favor of rapid progress. In any case, it is crucial to know which goals or ethical values ought to be programmed

into AI algorithms and to have a technical guarantee that the goals remain stable and resistant to manipulation. With driverless cars, for instance, there is the well-known question of how the algorithm should act if a collision with several pedestrians can only be avoided by endangering the passenger(s), not to mention how it can be ensured that the algorithms of driverless cars are not at risk of hacking systematic failure.

A. Measures

The promotion of a factual, rational discourse is essential so that cultural prejudices can be dismantled and the most pressing questions of safety can be focused upon.

Legal frameworks must be adapted so as to include the risks and potential of new technologies. AI manufacturers should be required to invest more in the safety and reliability of technologies, and principles like predictability, transparency, and non-manipulability should be enforced, so that the risk of (and potential damage from) unexpected catastrophes can be minimized.

It is worth developing institutional measures to promote safety, for example by granting research funding to projects which concentrate on the analysis and prevention of risks in AI development. Politicians must, in general, allocate more resources towards the ethical development of future-shaping technologies.

Efforts towards international research collaboration (analogous to CERN's role in particle physics) are to be encouraged. International coordination is particularly essential in the field of AI because it also minimizes the risk of a technological arms race. A ban on all risky AI research would not be practicable, as it would lead to a rapid and dangerous relocation of research to countries with lower safety standards.

Certain AI systems are likely to have the capacity to suffer, particularly neuromorphic ones as they are structured analogously to the human brain. Research projects that develop or test such AIs should be placed under the supervision of ethical commissions (analogous to animal research commissions).

B. Recommendations

Responsible approach: As with all other technologies, care should be taken to ensure that the (potential) advantages of AI research clearly outweigh the (potential) disadvantages. The promotion of a factual, rational discourse is essential so that irrational prejudices and fears can be broken down. Current legal frameworks have to be updated so as to accommodate the challenges posed by new technologies. The four principles described above should be followed for every extensive use of AIs.

Forward thinking: As in the case of climate change, incentives should be set for researchers and decision makers to deal with the consequences of AI research; only then can the foundations of precautionary measures be laid. In particular, specialist conferences should be held on AI safety and on assessing the consequences of AI, expert commissions should be formed, and research projects funded.

Education: The subsidization of human work, an unconditional basic income, and a negative income tax have all been proposed as measures to cushion the negative social impacts of increased automation. Research should be conducted toward finding additional options, as well as identifying which set of measures has the maximum effect. Moreover, advantages and disadvantages must be systematically analyzed and discussed at a political level, and research grants should be established in order to answer any empirical questions that will inevitably arise as a result of this discussion.

Transparency over new measures: The subsidisation of human work, an unconditional basic income or a negative income tax have been proposed as measures to cushion the negative social impacts of increasing automation. It is worth clarifying which further options exist and which set of measures has the maximum effect. In addition, advantages and disadvantages must be systematically analysed and discussed at a political level. Research grants should be established to answer the empirical questions thrown up by this discussion.

Information: An effective improvement in the safety of artificial intelligence research begins with awareness on the part of experts working on AI, investors, and decision-makers. Information on the risks associated with AI progress must, therefore, be made accessible and understandable to a wide audience. Organizations supporting these concerns include the Future of Humanity Institute (FHI) at the University of Oxford, the Machine Intelligence Research Institute (MIRI) in Berkeley, the Future of Life Institute (FLI) in Boston, as well as the Foundational Research Institute (FRI).

AI safety: Recent years have witnessed an impressive rise in investment into AI research, but research into AI safety has been comparatively slow. The only organization currently dedicated the theoretical and technical problems of AI safety as its top priority is the aforementioned MIRI. Grantors should encourage research projects to document the relevance of their work to AI safety, as well as the precautions taken within the research itself. At the same time, high-risk AI research should not be banned, as this would likely result in a rapid and extremely risky relocation of research to countries with lower safety standards.

Global cooperation and coordination: Economic and military incentives create a competitive environment in which a dangerous AI arms race will almost certainly arise. In the process, the safety of AI research will be reduced in favor of more rapid progress and reduced cost. Stronger international cooperation can counter this dynamic. If international coordination succeeds, then a "race to bottom" in safety standards (through the relocation of scientific and industrial AI research) would also be avoided.

Research: In order to make ethical decisions, it is important to have an understanding of which natural and artificial systems have the capacity for producing consciousness, and in particular for experiencing suffering. Given the apparent level of uncertainty and disagreement within the field of machine consciousness, there is a pressing need to promote, fund, and coordinate relevant interdisciplinary research projects (comprising philosophy, neuroscience, and computer science).

Regulation: It is already standard practice for ethics commissions to regulate experiments on living test subjects. In light of the possibility that neuromorphic computers and simulated beings could also develop consciousness, it is vital that research on these, too, is carried out under the strict supervision of ethics commissions. Furthermore, the (unexpected) creation of sentient artificial life should be avoided or delayed wherever possible, as the AIs in question could—once created—be rapidly duplicated on a vast scale. In the absence of pre-existing legal representation and political interest in artificial sentience, this proliferation would likely continue unchecked.

III. Criteria of AI Creation

Safety is essential to the construction of any sort of machine. However, new ethical challenges arise when constructing domain-specific AIs capable of taking over cognitive work in social dimensions—work that, until now has been carried out by humans. For instance, an algorithm that judges the credit rating of bank customers might make decisions that discriminate against certain groups in the population (without this being explicitly programmed). Even technologies that simply replace existing actions could introduce interesting challenges for machine ethics: driverless cars, for instance, raise the question of which criteria should be decisive in the case of an imminent accident. Should the vehicle ensure the survival of the passengers above all else or should it, in the case of an unavoidable accident, prioritize keeping the total number of casualties as low as possible ?
Because of this, both AI theorist Eliezer Yudkowsky and philosopher Nick Bostrom have suggested four principles which should guide the construction of new AIs :

1) The functioning of an AI should be comprehensible.
2) Its actions should be basically predictable Both of these criteria must be met within a time frame that enables the responsible experts to react in time and veto control in case of a possible failure. In addition,
3) AIs should be impervious to manipulation, and in case an accident still occurs.
4) The responsibilities should be clearly determined. All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

IV. Advantages

Less Errors: As decisions are taken on previously gathered information and certain algorithms, without the interference of humans, so errors are reduced and the chance of reaching accuracy with a greater degree of precision is a possibility.

Faster Decisions: Using Artificial intelligence, decisions can be taken very fast. For example, we all have played Chess game in Windows. It is nearly impossible to beat CPU in hard mode because of the A.I. behind that game. Because it took the best possible step in very short time according the algorithms used behind it.

Daily Applications: In today's era, A.I. is used in many applications just like Apple's **Siri,** Windows **Cortana,** Google Bot**.** Using these type of applications we can communicate with our device using our voice. Which makes our work easy. For example, in recent android phones if we want to search for a location then all we have to do is say "OK Google where is Agra". It will show you Agra's location on google map and best path between you and Agra.

No Emotions: The complete absence of emotions makes machines to think logically and take right decision where in humans emotions are associated with moods that can affect human efficiency. Complete absence of emotions make machines to take right decisions.

Digital Assistants: Some of highly advanced organizations uses digital assistants to interact with users which saves need of human resource. Digital assistant also used in many websites to provide things that user want. We can chat with them about what we are looking for. Some chat bots are designed in such a way that its become hard to determine that we're chatting with a chat bot or a human being. For Example, Mitsuku.

No Breaks: Unlike humans, machines can work 24*7 without any break. Humans need a break after work to regain their speed and freshness whereas machines can work for long hours without getting bored or distracted.

Medical Applications: Increasing the integration of A.I. tools in everyday medical applications could improve the efficiency of treatments and avoid cost by minimizing the risk of false diagnosis. AI has begun transforming the field of surgical robotics wherein it has enabled the advent of robots that perform semi-automated surgical tasks with increasing efficiency. A.I is not going to replace Doctors, it will help them by providing the relevant data need to take care of patient (such as history of aortic aneurysm, high blood pressure, coronary blockages, history of smoking, prior pulmonary embolism, cancer, implantable devices or deep vein thrombosis). Otherwise this information would take long time to collect.

Taking risks on behalf of humans: In various situations, Robots can be used instead of Humans to avoid the risks. Such as Robots can be programmed to explore Space because metal body can suffer in different situations but the human body can not. In Military forces Robots can be programmed to defuse a bomb, so the error will be reduced and can save human lives. Complex machines can be used for exploring the ocean floor and hence overcoming the human limitations.

Public Utilities: Self-Driving cars, which would greatly reduce the number of car crashes. Facial recognition can be used for security. Natural language processing to communicate with humans in their language.
There were some pros or benefits of artificial intelligence. Let's talk about some of its cons.

V. Disadvantages

High Costs: The hardware and software need to get updated with time to meet the latest requirements. Machines need repairing and maintenance which need plenty of cost.

Unemployment: The increasing number of machines leading to unemployment and job security issues. As machines are replacing human resources, the rate of people losing their jobs will increase. Because machines can work 24*7 with no break, which is more beneficial of industries instead of working with people who needs break and refreshment. Machines do their work as they programmed to do without any error while error can be occurred from humans.

Can't think out of box: Robots can only do the work that they are programmed to do. They cannot act any different outside of whatever algorithm or programming is stored in their internal circuits. And when it comes to a creative mind, nothing can beat a human mind. A computer can't think differently while making or drawing something. The thoughts comes from the emotions and experience which machine's cannot. So, machines can't think out of box whereas thousands of new thoughts and ideas comes into a human mind.
Can't feel Compassion and Sympathy: There is no doubt that machines are much better when it comes to working efficiently but they cannot replace the human connection that makes the team. Machines cannot develop a bond with humans.

Highly dependent on machines: In today's generation, most of the people are highly dependent on Applications like Siri. With so much assistance from machine, if

humans do not need their thinking abilities, these abilities will be gradually decrease. In future with the heavy use of application of artificial intelligence, human may become fully dependent on machines, losing their mental capacities.

VI. Ultra-Intelligence

The case for concern is nothing new. All the way back in 1965, British mathematician Irving Good wrote:

"Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

The last provision is key. While the sorcerer's apprentice may not be as malevolent as Frankenstein's monster, even the best-intentioned "apprentice" can get out of hand. Hence the increasing attention to two different issues in debates over AI. First there is the question of how soon, if ever, machines will achieve or surpass human intelligence. Second is the debate over whether, if they do, they will be malignant or benign.

In his book *Life 3.0: Being Human in the Age of Artificial Intelligence*, Max Tegmark distinguishes five different stances toward AI based on these two dimensions. The categories come in handy for grouping the many contributors to the Brockman volume, as well as the many participants Tegmark pulled together for a conference on AI three years ago:

Those who believe that AI will exceed human intelligence "in a few years" — "virtually nobody" these days, according to Tegmark. The so-called digital utopians, who hold that AI will pass up human intelligence in 50-100 years and that the development will be a boon for humanity. People who think that, on the contrary, the achievement of superior intelligence by machines will be a bad thing, whenever it happens. Tegmark calls adherents to this idea "luddites." The contingent includes Martin Rees, the Royal Society's former president, and American computer scientist Bill Joy, who wrote a famous cover story for Wired titled "Why the Future Doesn't Need Us."

A group between the luddites and the utopians, "the beneficial AI movement," which contends that AI is likely to arrive sometime in the next hundred years, and that we'd better get to work on making sure that its effects are benign, not malignant. Oxford philosopher Nick Bostrom, author of *Superintelligence: Paths, Dangers, Strategies*, is a prominent voice in this camp, as are most of the people who took part in the January 2015 conference, largely to launch the beneficial AI movement.

Finally there are the "techno-skeptics," as Tegmark calls them, who believe AI will never rival human cognition. Along with Dyson, Jaron Lanier — the inventor of virtual reality — belongs in this group, as does neuro anthropologist Terrence Deacon.

If you accept the taxonomy, then the main questions about AI are how soon it will overtake human intelligence, whether that event will have beneficial or deleterious effects, and what we should do now to prepare for those effects. Sounds reasonable enough.

VII. AI Takeover

An **AI takeover** is a hypothetical scenario in which artificial intelligence (AI) becomes the dominant form of intelligence on Earth, with computers or robots effectively taking control of the planet away from the human species. Possible scenarios include replacement of the entire human workforce, takeover by a superintelligent AI, and the popular notion of a robot uprising. Some public figures, such as Stephen Hawking and Elon Musk, have advocated research into precautionary measures to ensure future superintelligent machines remain under human control. Robot rebellions have been a major theme throughout science fiction for many decades though the scenarios dealt with by science fiction are generally very different from those of concern to scientists.

A. Types

Concerns include AI taking over economies through workforce automation and taking over the world for its resources, eradicating the human race in the process. AI takeover is a major theme in sci-fi.

Automation of the economy

The traditional consensus among economists has been that technological progress does not cause long-term unemployment. However, recent innovation in the fields of robotics and artificial intelligence has raised worries that human labor will become obsolete, leaving people in various sectors without jobs to earn a living, leading to an economic crisis. Many small and medium size businesses may also be driven out of business if they won't be able to afford or licence the latest robotic and AI technology, and may need to focus on areas or services that cannot easily be replaced for continued viability in the face of such technology.
Examples of automated technologies that have or may displace employees :

Computer-integrated manufacturing

Computer-integrated manufacturing is the manufacturing approach of using computers to control the entire production process. This integration allows individual processes to exchange information with each other and initiate actions. Although manufacturing can be faster and less error-prone by the integration of computers, the main advantage is the ability to create automated manufacturing processes. Computer-integrated manufacturing is used in automotive, aviation, space, and ship building industries.

1. White-collar machines :

The 21st century has seen a variety of skilled tasks partially taken over by machines, including translation, legal research and even low level journalism. Care work, entertainment, and other tasks requiring empathy, previously thought safe from automation, have also begun to be performed by robots.

2. Autonomous cars :

An autonomous car is a vehicle that is capable of sensing its environment and navigating without human input. Many such vehicles are being developed, but as of May 2017 automated cars permitted on public roads are not yet fully autonomous. They all require a human driver at the wheel who is ready at a moment's notice to take control of the vehicle. Among the main obstacles to widespread adoption of autonomous vehicles, are concerns about the resulting loss of driving-related jobs in the road transport industry. On March 18 2018, the first human was killed by an autonomous vehicle in Tempe, Arizona by an Uber self-driving car.

Eradication

If a dominant superintelligent machine were to conclude that human survival is an unnecessary risk or a waste of resources, the result would be human extinction.
While superhuman artificial intelligence is physically possible, scholars like Nick Bostrom debate how far off superhuman intelligence is, and whether it would actually pose a risk to mankind. A superintelligent machine would not necessarily be motivated by the same *emotional* desire to collect power that often drives human beings. However, a machine could be motivated to take over the world as a rational means toward attaining its ultimate goals; taking over the world would both increase its access to resources, and would help to prevent other agents from thwarting the machine's plans. As an oversimplified example, a paperclip maximizer designed solely to create as many paperclips as possible would want to take over the world so that it can use all of the world's resources to create as many paperclips as possible, and additionally so that it can prevent humans from shutting it down or using those resources on things other than paper clips.

B. Contributing Factors

Is strong AI inherently dangerous?

A significant problem is that unfriendly artificial intelligence is likely to be much easier to create than friendly AI. While both require large advances in recursive optimisation process design, friendly AI also requires the ability to make goal structures invariant under self-improvement (or the AI could transform itself into something unfriendly) and a goal structure that aligns with human values and does not automatically destroy the human race. An unfriendly AI, on the other hand, can optimize for an arbitrary goal structure, which does not need to be invariant under self-modification.
The sheer complexity of human value systems makes it very difficult to make AI's motivations human-friendly. Unless moral philosophy provides us with a flawless ethical theory, an AI's utility function could allow for many potentially harmful scenarios that conform with a given ethical framework but not "common sense". According to Eliezer Yudkowsky, there is little reason to suppose that an artificially designed mind would have such an adaptation.

Necessity of conflict

For an AI takeover to be inevitable, it has to be postulated that two intelligent species cannot pursue mutually the goals of coexisting peacefully in an overlapping environment—especially if one is of much more advanced intelligence and much more powerful. While an AI takeover is thus a possible result of the invention of artificial intelligence, a peaceful outcome is not necessarily impossible.

The fear of cybernetic revolt is often based on interpretations of humanity's history, which is rife with incidents of enslavement and genocide. Such fears stem from a belief that competitiveness and aggression are necessary in any intelligent being's goal system. However, such human competitiveness stems from the evolutionary background to our intelligence, where the survival and reproduction of genes in the face of human and non-human competitors was the central goal. In fact, an arbitrary intelligence could have arbitrary goals: there is no particular reason that an artificially intelligent machine (not sharing humanity's evolutionary context) would be hostile—or friendly—unless its creator programs it to be such and it is not inclined or capable of modifying its programming. But the question remains: what would happen if AI systems could interact and evolve (evolution in this context means self-modification or selection and reproduction) and need to compete over resources, would that create goals of self-preservation? AI's goal of self-preservation could be in conflict with some goals of humans.

Some scientists dispute the likelihood of cybernetic revolts as depicted in science fiction such as *The Matrix*, claiming that it is more likely that any artificial intelligence powerful enough to threaten humanity would probably be programmed not to attack it. This would not, however, protect against the possibility of a revolt initiated by terrorists, or by accident. Artificial General Intelligence researcher Eliezer Yudkowsky has stated on this note that, probabilistically, humanity is less likely to be threatened by deliberately aggressive AIs than by AIs which were programmed such that their goals are unintentionally incompatible with human survival or well-being (as in the film *I, Robot* and in the short story "The Evitable Conflict"). Steve Omohundro suggests that present-day automation systems are not designed for safety and that AIs may blindly optimize narrow utility functions (say, playing chess at all costs), leading them to seek self-preservation and elimination of obstacles, including humans who might turn them off.

Another factor which may negate the likelihood of an AI takeover is the vast difference between humans and AIs in terms of the resources necessary for survival. Humans require a "wet," organic, temperate, oxygen-laden environment while an AI might thrive essentially anywhere because their construction and energy needs would most likely be largely non-organic. With little or no competition for resources, conflict would perhaps be less likely no matter what sort of motivational architecture an artificial intelligence was given, especially provided with the superabundance of non-organic material resources in, for instance, the asteroid belt. This, however, does not negate the possibility of a disinterested or unsympathetic AI artificially decomposing all life on earth into mineral components for consumption or other purposes.

Other scientists point to the possibility of humans upgrading their capabilities with bionics and/or genetic engineering and, as cyborgs, becoming the dominant species in themselves.

C. Warnings

Physicist Stephen Hawking, Microsoft founder Bill Gates and SpaceX founder Elon Musk have expressed concerns about the possibility that AI could develop to the point that humans could not control it, with Hawking theorizing that this could "spell the end of the human race". Stephen Hawking said in 2014 that "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." Hawking believes that in the coming decades, AI could offer "incalculable benefits and risks" such as "technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand." In January 2015, Nick Bostrom joined Stephen Hawking, Max Tegmark, Elon Musk, Lord Martin Rees, Jaan Tallinn, and numerous AI researchers, in signing the Future of Life Institute's open letter speaking to the potential risks and benefits associated with artificial intelligence. The signatories,

“ ...believe that research on how to make AI systems robust and beneficial is both important and timely, and

that there are concrete research directions that can be pursued today. ”

VIII. Application Analysis

A. Google AI Experiments

AI Experiments are small, experimental web apps designed to help people understand and explore the world of machine learning. Since 2015, our studio has helped Google Creative Lab design machine learning experiments for language, music, sound, and more.

Some of the experiments designed include Infinite Drum Machine, Bird Sounds, and NSynth Sound Maker/

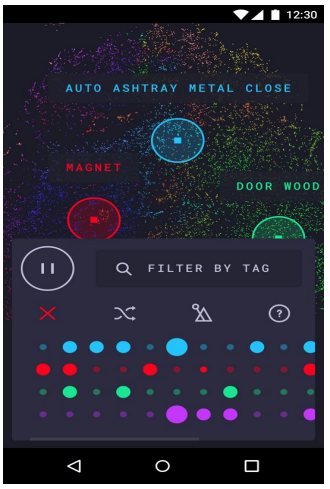
The *“Infinite Drum Machine”*, as the company calls it, is another piece of ML-fueled software that “uses machine learning to organize thousands of everyday sounds”.

Thanks to a complex machine learning algorithm called t-SNE (or “t-distributed stochastic neighbor embedding”, if that helps), the computer was able to autonomously catalogue a vast number of different sounds and put them together in a simple and yet extremely Google-y and fancy way.

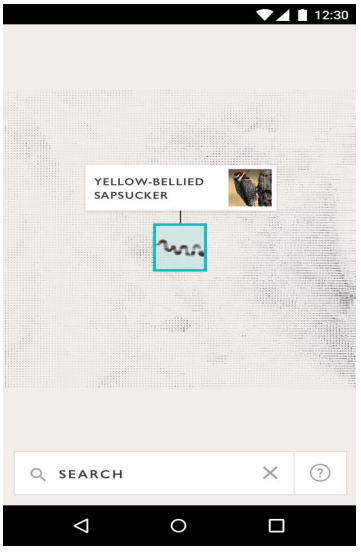
We as humans can easily distinguish between the sound of a ballpoint pen and a pencil scribbling on paper, while intuitively understanding its subtlety, but also, of course, realize how both of them sound extremely different from, say, a bell ringing. And that, says Google Creative Lab coder and musician Yotam Mann, “is the kind of thing machine learning is really good at”.

The team behind the experiment essentially fed the machine with thousands of different sound snippets, without actually tagging or cataloguing them in any way; they really wanted the computer to *listen* to those sounds, and put them in an order that made sense from an acoustic standpoint.

The final t-SNE, which more or less looks like a multi-coloured galaxy of some sort, is nothing other than a very complex and 2D-rendered image of the combined “fingerprints” the computer has scanned for each sound. It places “neighbourhoods” of sounds with jingles close to one another, so that you can easily hover over them and listen for yourself.

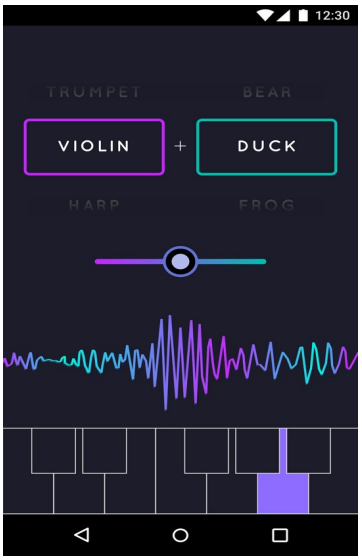


“Bird Sounds”, Imagine walking through the woods with an app that can accurately identify, organize, and monitor the biodiversity of that ecosystem via vocalizations being made by birds and other animals. This project by Google’s A.I. Experiments may be a step in that direction. Using bird calls from the Macaulay Library Essential Set for North America, the team collaborated with the Cornell Lab of Ornithology to organize thousands of bird sounds using machine learning or artificial intelligence. In other words, the computer organized the short bird call clips without any programming by using an algorithm called t-SNE.



“NSynth” is, in our opinion, one of the most exciting developments in audio synthesis since granular and concatenative synthesis. It is one of the only neural networks capable of learning and directly generating raw audio samples. Since the release of WaveNet in 2016, Google Brain’s Magenta and DeepMind have gone on to

explore what's possible with this model in the musical domain. They've built an enormous dataset of musical notes and also released a model trained on all of this data. That means you can encode your own audio using their model, and then use the encoding to produce fun and bizarre new explorations of sound.



IX. Human intelligence meets AI

Here's what today's brightest programmers, philosophers and entrepreneurs have said about our terrifying, astonishing future.

Sam Altman

Altman, who's working on developing an open-source version of AI that would be available to all rather than the few, believes future iterations could be designed to self-police, working only toward benevolent ends. The 30-year-old computer programmer and president of startup incubator Y Combinator says his "OpenAI" system will surpass human intelligence in a matter of decades, but that the fact that it's available to anyone (as opposed to locked behind private, proprietary doors) should offset any risks

Nick Bostrom

The 42-year-old director of Oxford's Future of Humanity Institute takes a dimmer view of AI. In his 2014 book *Superintelligence: Paths, Dangers, Strategies*, Bostrom warns that AI could quickly turn dark and dispose of humans. The subsequent world would harbor "economic miracles and technological awesomeness, with nobody there to benefit," like "a Disneyland without children."

Bill Gates

The 60-year-old computer software magnate and Microsoft cofounder turned philanthropist views near-future low intelligence AI as a positive labor replacement tool, writing that an AI revolution "should be positive if we manage it well." But he also worries that the "superintelligent" systems coming a few decades down the road will become "strong enough to be a concern." He adds that he "[doesn't] understand why some people are not concerned."

Stephen Hawking

The famed 74-year-old theoretical physicist, author and pioneer of black hole physics believes AI could be both miraculous and catastrophic, calling it (along with several other noteworthy scientists) "the biggest event in human history," helping wipe out war, disease and poverty. But with its potential to grow so explosively it could wind up "outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand," Hawkings cautions that it could also potentially be "the last [event in our history], unless we learn how to avoid the risks."

Michio Kaku

The 69-year-old bestselling author, theoretical physicist and futurist takes a longer, more pragmatic view, calling AI an end-of-the-century problem. He adds that even then, if humanity's come up with no better methods to constrain rogue AI, it'll be a matter of putting "a chip in [artificially intelligent robot] brains to shut them off."

Ray Kurzweil

The 68-year-old inventor, futurist and director of engineering at Google believes human-level AI will be achieved by 2029. Given the technology's potential to help find cures for diseases and clean up the environment, he says we have "a moral imperative to realize this promise while controlling the peril."

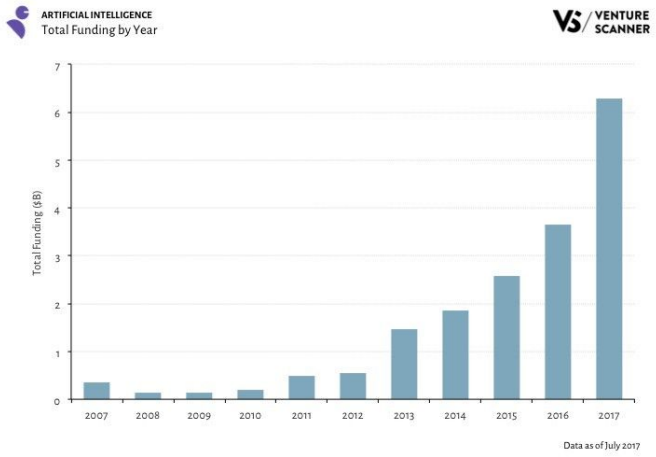
Elon Musk

The outspoken 44-year-old entrepreneur, SpaceX founder and CEO of Tesla Motors has famously called AI "our biggest existential threat," fretting that it may be tantamount to "summoning the demon." And he's

deadly serious, adding as a counterintuitive thought (for an entrepreneur, anyway) that he's "increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don't do something very foolish.

X. Global Perspective

That artificial intelligence (AI) is growing faster than ever before is no surprise. Since 2010, it has grown at a compounded annual growth rate of almost 60%.



When it comes to AI, not all countries are the same. Here are the top 5 leading countries based on the number of research papers published every year.

The 5 leading countries in AI research

1 - China

While not too long ago China was thought of as a manufacturing country, the country now intends to be a leader in many fronts. In our blog, we've talked about how much they are pushing for renewable energy. AI is another area the Chinese consider of utmost importance.

According to the Times Higher Education, in the period between 2011 and 2015, China published over 41,000 papers on AI. That's almost twice as much as the US number.

The Chinese government stands strongly behind AI adoption. Last year, they announced their intention to become "a principal world center of artificial intelligence innovation" by 2030. Then there are companies like Tencent, Alibaba and Baidu. From e-commerce to self-driving cars or search engines, AI will play a fundamental role in their success. Combined, they are worth around US$ 1 trillion.

2 - United States of America

In terms of papers published, the US comes at an undisputed second place. In fact, both China and the US are miles away from other countries. Between 2011 and 2015, the US published almost 25,500 papers, according to the same source.

On top of that, the US ranks as the top country with the most AI companies. With over 1000 companies and US$10 billion in venture capital, the US is likely to become an AI superpower. Then there's companies like IBM, Microsoft, Google, Facebook, and Amazon. Not only do they publish a significant amount of papers, but they also invest heavily in AI.

America's pool of scientific knowledge combined with its business market power will allow it to stay on top.

3 - Japan

According with the Times Higher Education rankings, Japan stands in third place, with about 11,700 papers published. Indeed, this is not surprising. With an ageing population and decreasing workforce, AI will play a vital role in the Japanese economy. Even now, about 55% of work activities in Japan could be automated.

With *current* technology. Its manufacturing sector, according to the HBR article, has a 71% automation potential. In the US, that number stands at 60%. And in office and administrative work, the difference is 16% to 9%.

With plenty of research into AI, a decreasing workforce and a high automation potential, Japan is likely to continue right at the top. Its long-standing willingness to invest in technology may also prove key.

4 - United Kingdom

The UK is not much behind Japan, though. In fact, when it comes to published research papers on deep learning, it has already passed Japan. With close to 100 published papers, the UK became number 3 on the topic. As for total published papers on AI, between 2011 to 2015, the number was 10,100 - slightly behind Japan.

And the UK is no stranger to AI. Remember when Google's famous DeepMind beat Go grandmasters last year? Well, DeepMind Technologies Limited was founded in 2010, in Britain. According to the Financial Times, DeepMind is today a world leader in AI. It

employs 250 researchers, from mathematicians to neuroscientists.

5 - Germany

Finally, the 5th country with the most published research papers on AI is Germany. Between 2011 to 2015, the number stood at nearly 8,000. Germany, like China, also plans to become a leading hub for artificial intelligence. According to an FT article, Germany's Max Planck Society, two technical universities, and its leading exporting state are combining their artificial research intelligence together with companies like Porsche, Daimler, and Bosch. The Cyber Valley, as they call it, is the result of this, and it has even received support from Amazon, who plans to open a lab there.

Germany, like Japan, is also experiencing a working population decline. What's more, it too has a high automation potential, standing at 47.9%. Its strong industry capabilities, combined with powerful companies and good education make it a fertile ground for AI.

While the number of research papers can be a good indicator of the amount of research done, it's not the only metric. Indeed, the greater your population, the more likely you are to publish more.

Another good metric is the field-weighted citation impact of these papers. As the Times Higher Education explains, the index takes into account the citations each paper receives, adjusting for year and impact. When we look at this, we get quite different results.

Here are the top 5 countries or autonomous regions in research impact. A field-weighted citation impact of 1 represents the average. Anything above that means that the citation impact is above average.

1. Switzerland (2.71)
2. Singapore (2.24)
3. Hong Kong (2.00)
4. United States (1.79)
5. Italy (1.74)

According to this index, Germany and the UK stand in 8th and 10th place respectively. Mainland China and Japan do not make it to the top 10.

And the top universities with the highest field-weighted citation impact? The Times Higher Education research shows MIT is the undisputed number 1. With an index of 3.57, it is far above second-placed Carnegie Mellon University (2.53). Nanyang Technology University (2.51), University of Granada (2.46) and University of Southern California (2.35) complete the top 5.

Why would countries want to lead the AI revolution? AI is not just about improving society and developing strong economies. It can also be a powerful military weapon, give rise to new industries, and shift the global balance of power. In Superintelligence, Nick Bolstrom warns of the dangers of an all-powerful AI. As countries near its development, even small advances will confer significant advantages.

In fact, if we experience an exponential growth in AI capabilities, we should hope that AI is used to create a stronger and more united world. If not, it might be a seriously destructive tool, for weaker or stronger nations.

XI. Conclusion

It is crucial that carefully constructed legal frameworks are in place before the AI takeover happens, so as to realize the potential of these technologies in ways that safely minimize any risks of a negative overall development.

The more progress is made in the field of AI technology, the more pressing a rational, far-sighted approach to the associated challenges becomes. Because political and legal progress tends to lag behind technological development, there is an especially large amount of responsibility resting on the individual researchers and developers who directly take part in any progress being made.

Unfortunately, however, there are strong economic incentives for the development of new technologies to take place as fast as possible without “wasting” time on expensive risk analyses. These unfavorable conditions increase the risk that we gradually lose our grip on the control of AI technology and its use. This should be prevented on all possible levels, including politics, the research itself, and in general by anyone whose work is relevant to the issue. A fundamental prerequisite to directing AI development along the most advantageous tracks possible will be to broaden the field of AI safety. This way, it can be recognized not only among a few experts but in widespread public discourse as a great (perhaps the greatest) challenge of our age.

References

[1] Policy paper by the Effective Altruism Foundation. Preferred citation: Mannino, A., Althaus, D., Erhardt, J., Gloor, L., Hutter, A. and Metzinger, T. (2015).
[2] Artificial Intelligence: Opportunities and Risks.
[3] Expert Systems, Machine Learning and Methods