Identifying and addressing the risks of AI through regulations, compliance controls and technical design

Received (in revised form): 12th August, 2024

Sudhanshu Bahadur
Head of Technology for Global Asset Management, BMO Financial Group, Canada

Kuno Tucker
Chief Compliance Officer at Manulife Wealth and Adjunct Professor, Corporate Governance, York University, Canada

Sudhanshu Bahadur

Kuno Tucker

Global Asset Management, BMO Financial Group, 100 King St. W., Toronto, ON, M5X 1A9, Canada
Tel: +1 647 532 8929;
E-mail: sudhanshu.bahadur@bmo.coma

York University, Masters in Financial Accountability, MIAB, 216 Atkinson, 4700 Keele Street, Toronto, ON, M3J 1P3, Canada
Tel: +1 289 259 3521;
E-mail: Ktucker@yorku.ca

Journal of Financial Compliance
Vol. 8, No. 2 2024, pp. 112–130

2398-8053 (2024)

***Sudhanshu Bahadur**, MBA, BS (Engineering), is Head of Technology for Global Asset Management at BMO Financial Group. He has extensive banking, consulting and product experience spanning 25+ years in leadership positions in the capital markets and asset management industry. He has deep expertise in defining and executing technology vision and strategy for the investment management, prime brokerage, derivatives processing, risk management and data strategy functions. His latest pursuits are in the areas of AI/ML, cloud computing and Big Data analytics.*

***Kuno Tucker**, LLM, ICD.D, is the Chief Compliance Officer at Manulife Wealth and an adjunct professor of corporate governance at York University, Toronto. He has served in various senior executive roles in his 29 years of experience in the financial industry and has lectured on topics ranging from strategy, corporate governance, FinTech, AI, compliance and AML. He has been a keen student of FinTech for several years and has recently deployed safe but effective uses of AI in the field.*

Abstract

This paper will identify and examine the various aspects of artificial intelligence (AI), surface the risks associated with AI, and provide compelling governance, compliance and technical solutions to mitigate those risks. AI is destined to accelerate the pace of change and development in multiple fields; however, with its great advances come great risks that need to be addressed.

Keywords: artificial intelligence, compliance, regulations, risk management, controls, GenAI, LLMs

DOI: 10.69554/TPIT6981

INTRODUCTION

Artificial intelligence (AI) has gained enormous attention in the past couple of years, not only because of the amazing pace of progress of the technology but also because the barriers of entry to the technology have considerably lowered as the tools have become readily available to the general public in a consumable form.

While the technology has grown quickly, the concerns around the risks of this technology have also garnered much attention.

This paper will discuss AI technology and the risks it poses. It will then explore what needs to be done to ensure the capability remains under control, maximising the benefits while reducing the potential risks.

WHAT IS AI?

> If a human is interacting with another human and a machine and unable to distinguish the machine from the human, then the machine is said to be intelligent.[1]

This is the Turing test, which was created by Alan Turing in 1950 and which, even today, is considered the benchmark for identifying the intelligence of an artificial system.

HISTORY OF AI DEVELOPMENT

Origins

Today's AI technology with its ability to generate human-like output, unimaginable till only a few years ago, has astonished us with its capabilities and its pace of growth. But AI is not new. It has been in existence for over seven decades in various forms. Its origins can be traced to initial fictional writings by Isaac Asimov in 1942, when he published a short story called Runaround in which he conceptualised a robot and the laws governing robotics. Around the same time Alan Turing invented a real machine called the bombe, which was capable of deciphering the Enigma machine's code used by the German Army in the Second World War, following which he wrote in his article 'Computing Machinery and Intelligence' about how to create and test intelligent machines. By 1956, the term 'artificial intelligence' had been coined by John McCarthy, who is considered one of the founding fathers of AI.

Progression

One of the earliest AI systems was a computer program called ELIZA created at MIT in the 1960s, which was a natural language processing system that could simulate a conversation with a human. Research in AI continued over the decades, with the world watching in awe when IBM's Deep Blue program beat the world champion Gary Kasparov at chess. Deep Blue, much like ELIZA, was a rule-based expert system that could process 200 million moves per second to plan 20 moves ahead. While this astonishing feat was recognised the world over, the expert systems are based on formal rules and do not lend themselves well to being trained from external data.

Recent advancements

AI in it is true form came to the fore when deep learning was invented, as showcased by Google in its AlphaGo program in 2015, which was able to beat the world champion in the board game Go.[2] Deep learning is a type of artificial neural network that relies on large datasets. When this technique of utilising deep learning was implemented on large amounts of written natural language, large language models (LLMs) were created. These LLMs, like GPT4 from OpenAI, have taken AI from being restricted to computer scientists and programmers to being available to the general public, significantly expanding the capability and utility of AI applications.

CATEGORIES OF AI

AI is a broad term that encompasses a number of different capabilities and technologies. It is conceptualised into three categories as:

- *Artificial Narrow Intelligence (ANI)*: These are systems that can perform a defined task with intelligence. These AI systems are limited to the purpose they were designed for. Most systems currently in use fall under this category of AI.
- *Artificial General Intelligence (AGI)*: These systems would be able to perform any intellectual task like a human. These systems do not currently exist and are being researched. With the recent advancements in AI technology, there is a sense of expectation that such systems will be possible in the next decade or two.
- *Artificial Super Intelligence (ASI)*: These systems would surpass human intelligence and will be able to perform any cognitive or intellectual task better than a human being. Such technology does not currently exist.

TYPES OF AI

AI can be categorised into four basic types[3]:

- *Reactive machines*: Reactive machines AI operates purely on rules and predefined patterns, lacking the ability to learn from

data or past experiences. It is task-specific, has no memory and always responds to the same input in the same way. Reactive machines are often used in systems where speed is crucial, such as robots or autonomous vehicles.

- *Limited memory*: These systems are able to remember past events and learn from those experiences to adapt their responses to future events. This allows these systems to continuously evolve with the availability of new data.
- *Theory of mind*: Theory of mind AI systems can understand human emotions and behaviour. These AI systems are still under development and are being designed to be able to interact with humans in an emotionally intelligent way.
- *Self-aware*: A self-aware AI system is a theoretical concept at present. In this case AI is expected to have developed human-like intelligence and self-awareness. It would have its own emotional state and be able to think and reason independently, making decisions without human input.

HOW DOES AI WORK?

AI works by combining large amounts of data with algorithms that can process and learn from the data, and then generate outputs that mimic human behaviour or reasoning. Each type of AI uses different techniques and methods to achieve specific goals.

Below are some of the techniques and technologies that have been developed to implement AI systems:

- *Machine learning*: Arthur Samuel[4] defined machine learning (ML) in the 1950s as 'the field of study that gives computers the ability to learn without explicitly being programmed'. Instead of the traditional design techniques to define sets of rules in computer programs, ML creates the capability in computer systems to learn from experiences or from data. ML systems work by finding patterns in data to uncover insights and to continuously improve the results of whatever task the system has been designed to achieve. ML can be 'supervised' or 'unsupervised' learning.
  - *Supervised ML*: This technique involves training the machines with 'labelled' training data. In essence this means that the machine is taught, with this training dataset, what outcome to predict. After the training, when faced with fresh data inputs, the machine is then able to correctly predict the outcome based on the training it has already received. Supervised learning can be used for risk assessment, image classification, fraud detection, spam filtering, etc.
  - *Unsupervised ML*: This is an ML technique used for creating clusters of unlabelled datasets. These ML algorithms would discover hidden patterns in datasets without involving human intervention. The models have the ability to identify similarities and differences in datasets. Such pattern recognition is hugely useful in exploratory data analysis, cross-selling strategies, customer segmentation and image recognition. These models can help not only in clustering datasets but also in building associations among variables and in reducing the dimensionality of data to identify the most important factors.
- *Neural networks*: Artificial neural networks (ANNs) are computer models that mimic biological neural networks of interconnected nodes in the human brain to process information, find connections between the data and produce inferences. Neural networks allow AI systems to take in large datasets, uncover patterns among the data and produce outcomes.
- *Deep learning*: Deep learning uses ANNs to identify patterns with significantly large amounts of data. These models analyse large datasets to find associations between the data inputs, interpret meaning from

undefined data and produce results based on positive and negative reinforcement. These ML models use vast amounts of data to build a neural network that establishes nodal connections to predict outcomes.

- *Natural language processing*: This AI technology allows computers to truly understand human language, either written or spoken. Natural language processing is crucial for any AI-driven system that interacts with humans in some way. These systems are able to recognise, analyse, interpret and understand human language.
- *Cognitive computing*: The objective of AI systems is to mimic human interaction models as closely as possible. Cognitive computing AI systems are designed to imitate interactions between humans and machines. This allows computer models to work as closely as possible to the functioning of the human brain when performing complex tasks.
- *Computer vision*: This AI capability allows computers to observe and understand visual data. These systems extract information from images and videos, in a similar way to human vision, by interpreting the content of an image via pattern recognition and deep learning. The machines use deep learning and convolutional neural networks to recognise visual data.

WHAT IS GenAI?

While AI technology has been steadily growing over the decades since the term was coined in the 1950s, the year 2022 was a watershed moment when OpenAI released ChatGPT and made it publicly available. The ChatGPT user base exploded on launch, with one million users within the first five days[5] and 100 million users within two months.[6] This unprecedented growth was as a result of AI finally becoming available to the general public. You did not have to be a computer engineer or scientist to be able to utilise AI and assess its capabilities for yourself. This democratisation of AI technology has suddenly led to vast interest in its capabilities and myriad use cases have been identified as people understood what the capability offers.

GenAI, or generative AI is the ability of AI to produce human-like output, be it in the form of text or images or sound. GenAI models can capture the patterns and structure of the data they are trained on and then generate new data that has similar characteristics. A number of significant technological capabilities have come up in the past couple of years: tools like ChatGPT that can interact in natural language, DALL-E2 which can produce images as instructed or copilots that can write computer code for you. These GenAI capabilities are just a start, as the technology continues to find new capabilities and uses.

IMPACT OF GenAI

It is expected that the GenAI industry would add US$2.8tn–$4.4tn per annum to the world economy as per a McKinsey study.[7] Expectations are that in the next three years anything not connected to GenAI will be considered obsolete or ineffective. Customer care and sales are expected to have 70 per cent of the impact, with the rest spread among network operations, IT and support functions. As much as 50 per cent of customer care interactions can be automated with GenAI.[8] As much as 70 per cent of repetitive tasks done by human beings can be automated. In knowledge search, validation and synthesis, as much as 60 per cent of activity has the potential to be automated. In technology, developer productivity can be boosted 25–40 per cent by utilisation of GenAI based toolsets.[9]

LLMs

An LLM is a type of deep learning neural network that can process and generate natural language, such as English, French or

Chinese. LLMs are trained on huge amounts of text data, such as books, articles, websites and social media posts, to learn the patterns and structure of language. LLMs can then use this knowledge to create text by predicting the next word or sentence, given some input, or create new texts on various topics and styles.

The text data used to train LLMs is mainly unlabelled or uncategorised, and the model uses self-supervised or semi-supervised learning methodology. Information is ingested or content entered into the LLM, and the output is what that algorithm predicts the next word will be. By doing this, the LLM is able to write complex text output and can even perform simple reasoning tasks.

The neural network behind the LLM essentially works by creating vector associations of simple words and, by doing so, is able to build a contextual understanding of the word, which it is then able to utilise to create meaningful output. LLMs train with enormous amounts of data. The size of an LLM is measured in number of parameters. As an example, GPT4 has been presumably trained on ~13 trillion tokens and contains 1.76 trillion parameters.

Apart from OpenAI's GPT-3 and 4 LLM, there are other open models available such as Google's LaMDA and PaLM LLM (the basis for Bard), Hugging Face's BLOOM and XLM-RoBERTa, Nvidia's NeMO LLM, XLNet, Co:here and GLM-130B. The release of Meta's open-source LLaMA LLM is allowing developers to create customised models at a lower cost point.

USES OF LLMS

LLMs have found use in many different areas through many industries. Below are some of the uses of LLMs:

- question answering;
- natural language understanding;
- text generation;
- information extraction;
- content recommendation;
- summarisation;
- text categorisation;
- language translation;
- sentiment analysis.

AI IN FINANCIAL SERVICES

In the financial services industry, GenAI has extensive possible uses. LLMs can be fine-tuned using news, social media, announcements, filings, trends and other data sources for investment research. These fine-tuned LLMs can also be trained on domain specific data for doing sentiment analysis or statement analysis. Idiosyncratic patterns in data can be identified using AI to determine investment opportunities. Similarly, AI can be used to find predictive patterns in data for fundamental analysis and to determine what is the most pertinent indicator for an industry or company. Investment ideas can be summarised and compared by using LLMs. Another potential use is in portfolio optimisation, as you can prompt an LLM with specific questions that result in identifying the right portfolio mix to achieve a defined investment objective (eg create a global equities and bonds portfolio, with defined outcome and Sharpe ratio).

Environmental, social and governance (ESG) sustainability research can benefit from the use of AI, as many types of data sources can be sifted through to identify sustainability metrics that meet ESG criteria. This would help weed out those companies that might be greenwashing.

Investment risk is an area where LLMs will find extensive use, as causal inference can be found using ML in idiosyncratic risk management (systemic risk can be hedged for, but idiosyncratic risk cannot).

Another use case is for regulatory filings to be autogenerated using LLMs, thereby reducing a lot of manual repetitive work.

As LLMs gain greater traction, there is potential to enable financial advice through

an LLM. These types of services are being built and are becoming more and more autonomous for self-driven investors.

In the mergers and acquisitions (M&A) space, GenAI can also be used for insolvency prediction by assessing a company's current state and mapping it with possible future possibilities. Forecasting of M&A targets can be done by using LLMs that can help companies identify possible targets for acquisitions.

The use of AI has vast potential, and the possibilities it offers are just beginning to be grasped. Expanding of usage of AI based use cases will be seen as it becomes more prevalent in the industry.

WHAT ARE THE RISKS OF AI?

The advancement of AI technology in recent years is already offering significant benefits across industries and is leading us ever closer to artificial general intelligence (AGI). This expansion of capability and usage is raising considerable risks that must be evaluated and concerns that need to be addressed before irretrievable immersion in AI. Most industry leaders agree that AI is a game-changer in bringing on additional capabilities and efficiency, while at the same time AI experts are calling out the threats it poses, with some calling them existential in nature.[10,11]

AI technology suffers from many challenges, which result in outputs that do not always align with human value systems.

RISKS WITH USE OF AI

- *Job displacement*: As mentioned earlier, as much as 70 per cent of repetitive tasks done by human beings can be automated. In knowledge search, validation and synthesis as much as 60 per cent of activity has the potential of being automated. In technology, developer productivity can be boosted 25–40 per cent by utilisation of GenAI based toolsets.[12] With these significant efficiency improvements, AI can automate many tasks that are currently done by humans, which could lead to significant unemployment, income inequality and social unrest. Some studies estimate that AI could replace up to 40 per cent of the world's jobs in the next 10–15 years.
- *Information security*: AI systems require large amounts of data to be able to predict outcomes and solve problems effectively. This poses two distinct challenges from an information security perspective.
  - *Privacy concerns*: One of the most significant concerns is that personally identifiable information (PII) data about individuals is collected by these platforms, and if it gets inadvertently leaked to malicious actors, it could lead to identity theft, financial fraud and other forms of abuse. For example, the world of health care is beset by high costs, long waiting times, some bloated inefficiencies and the inability to crunch large data, including visual data, in order to provide timely and accurate diagnoses. However, in many of these cases PII data would need to be exposed in these datasets, which could be vulnerable to cybersecurity incidents, resulting in irreparable harm to individuals whose data gets exposed to threat actors.
  - *Cyberattacks*: With AI systems becoming more complex and autonomous, considering the significant amounts of data they gather, there is a risk that they can be hacked through cyberattacks. If a malicious actor takes control of an AI system, they can lead it to harm individuals or society as a whole. Secondly, AI systems are also capable of reducing barriers to entry for malicious actors, whether individuals or state sponsored, to gain the expertise to bypass security measures and exploit system vulnerabilities. The level of sophistication of cyberattacks through social engineering could make these attacks much more targeted and hence

more precise in inflicting maximum damage to individuals and organisations.

- *Accuracy*: AI engines are capable of reasoning and creating output by understanding the context of the input data. That is the big advantage of using AI technology to create new text, images or even music. But if it is not designed, tested or deployed properly, it can behave in unpredictable ways, or 'hallucinate' outcomes. Essentially, an AI hallucination happens when an AI engine generates patterns that are fabricated and non-existent, resulting in outputs that are completely inaccurate. These hallucinations can be simple contradictions within sentences, with prompts or facts or they can be completely random in nature. Hallucinations can be caused by bad input data quality, poor generation method due to a bad training model or poor prompt input context.[13] Already, at least one lawyer in New York State in 2023, was disbarred for, seemingly unwittingly, presenting a precedent for his argument, which turned out to be literally a legal fiction created by an AI hallucination prompted by the desire to provide a case that would help the lawyer win his client's case.[14] This happened again in British Columbia in 2024.[15] As such, the obvious efficiencies and career-threatening speed of legal LLMs are not fully ready to be unleashed in the courtroom without human verification and protections.
- *Bias*: Human society is inherently biased and individuals hold opinions on every topic and subject. These opinions manifest themselves into biases that everyone holds. While the cause of these biases is often hotly debated, the results are reflected in people's behaviours. These behaviours are further reflected in data produced and collected in various domains. Since AI is based on large amounts of this data, it can reflect or amplify these biases and prejudices. This can easily end up affecting the output produced by GenAI engines, which can lead the consumer of this information to believe that some of these biases are accurate. The potential is often extremely harmful, as the outputs are misaligned with the human value system. Also, most of the data used by GenAI platforms today comes from the web, where most content present is produced by Western nations. Hence, the current GenAI engines already have Western bias in their output. Similar biases exist across ethnicities and religions, as input data does not have the diversity it needs for producing a more holistic output.
- *Trust*: AI produces outputs that seem extremely precise and well-researched, even providing seemingly relevant references. At the same time those outputs could be pure hallucinations or reflect biases inherent in societies. Meanwhile, malicious actors could be manipulating this information and the output it produces to achieve their political, financial or social disruption goals. Most people have probably now seen or heard of the almost magical power of 'deepfakes' which, while amusing at one level, are frightening in their ability to manipulate perceptions of political or historical events. Should this continue without safety mechanisms to identify fake from true images, humanity risks almost catastrophic trust deficits. Commercially, there are very expensive risks to corporations losing control of their brands, images, likenesses and identity. All these, when combined, could lead to an erosion of confidence in existing social and governmental systems, which could further lead to social unrest and erosion of trust in fellow human beings.
- *Intellectual property*: The infamous Samsung example had most companies scramble to block uses of AI and ChatGPT at their firms. But then it was realised that their competitors would leapfrog past them in terms of coding ability, productivity and competitiveness. Some employees may continue to use GPT unofficially to

improve their own productivity leaving firms equally vulnerable. The appropriate response is to allow for AI, but under strict governance controls. The equally frightening risk is if AI can copy all manner of intellectual property (IP) and possibly be designed to disguise that it stole without attribution, then motivation for creativity and innovation by humans will be thwarted. This creates a new question. If an AI creates a unique product or service, effectively IP, who owns the rights to this IP? Is it the person or institution that deployed the AI or the AI itself? What limitations can be imposed if it is deemed to be owned by the AI itself?

- *Ethical concerns*: AI systems work based on the input data they have been trained on and interpret it to produce human-like output. Hence human biases get embedded into the system. But human behaviour and value systems are not purely data dependent but rather are defined by a moral and ethical code of conduct. Hence, a purely data-driven AI system would produce output that may not be aligned with human ethics. This is a significant issue with AI systems and needs to be addressed during the design of the systems, as well as by modulating the output they produce. The bigger problem, though, lies in the fact that different societies have different ethical and moral codes. Whose ethical framework should the AI system be operating on, and who gets to decide which framework to use is an open philosophical conundrum for AI architects to solve.
- *Existential risks*: AI technology has exponentially expanded in its capability in the past couple of decades. Today's AI systems are able to outperform human beings in a number of areas like protein folding and strategy games.[16] These systems are faster, can grasp broader sets of data and can communicate almost instantaneously, which makes them far better than human beings at achieving their focused outcome. Large tech companies are investing billions of dollars in this technology. With the current pace of growth in AI technology, it is plausible that the level of AGI in machines can be reached in the next decade or two. The next evolutionary step for AI would be artificial super intelligence. Such machines will be far more capable than human beings and will surpass us in their capabilities. If not designed with proper checks and controls, would these machines then be able to continue self-evolution at an exponential rate and ultimately turn on human beings and destroy society? There are many AI researchers who are mulling this as a significant existential threat, which must be taken into consideration as this technology is further developed.[17]

HOW CAN SOCIETY PROTECT ITSELF FROM THE RISKS OF AI?

The risks raised by AI extend from simple inaccuracies to societal challenges and may even extend to becoming a threat to human existence. These risks must be addressed before humanity gets too far into a technology where there is a possibility of losing control, even if it is a remote one. These protections need to come from all directions. First, technology builders should put in self-regulating checks and balances that avoid the chance of risky outcomes. Secondly, governments need to create regulatory frameworks to better control the development and adoption of the technology. Thirdly, organisations need to set up governance and control functions that ensure proper use of the technology. Finally, the design of these systems must ensure security of data and reduce the probability of biases and hallucinations creeping in. These controls can help reduce the risks while ensuring the potential use of the technology can be maximised.

SELF-REGULATING CHECKS AND BALANCES

The innovations of AI, especially when coupled with robotics, can be world changing. But there must be a balance between the power of innovation and the need for appropriate controls so these capabilities do not cause irreparable harm to humanity. Hyperfast racing cars can only be designed and raced if they have compensating disc brakes to provide safety to the drivers. These responsible controls are the much-needed brakes that will allow innovation to happen safely.

FIRST LAW OF ROBOTICS

Isaac Asimov, famed science-fiction writer and arguably one of the world's foremost futurists, wrote about the 'Three Laws of Robotics' in one of his short stories, and these laws can apply equally to AI:

The three laws,[18] presented as being from the fictional 'Handbook of Robotics, 56th Edition, 2058A.D.', are: [1]

- The First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- The Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- The Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The flurry of AI-enabled solutions has gone beyond just augmenting what current software or humans perform; these solutions are now at the stage of replacing humans in certain tasks. This is exciting both in terms of seeing how much faster, more efficiently and more effectively work can be accomplished and in terms of the possibility that the enormous scale could result in wonderful new developments in science and health care that have the potential to improve society.

PAPERCLIP PROBLEM

However, it is important to be mindful of Oxford philosopher Nick Bostrom's 2014 'paperclip thought problem'[19] in which essentially an innocuous task of making as many paperclips as possible can turn into a debilitating surfeit of useless paperclips, as well as the threat of Armageddon, as a machine tasked with solely making as many paperclips as possible without any other parameters will put the world at harm in order to achieve its singular task.

Consequently, those creating AI solutions, especially when combined with robotics, must always build in protections or guard rails to avoid devastating unintended consequences. These suggested safeguards to mitigate these risks are outlined below.

Personal safeguards

Obviously, each individual must monitor and safeguard their own social media likeness and control their confidential data by only providing necessary data to trustworthy institutions, avoiding public GPT or dishonest WiFi and keeping data on private cloud instances which are encrypted.

But it goes beyond these simple steps. People have a responsibility to themselves not to blindly trust AI to do their thinking for them. Trust the process but test the output.

Bias versus overcorrection

Much has been said about the instances of bias in early instances of public GPT, whereby answers came from public domain information, which was littered with sometimes racist viewpoints, thus inadvertently providing similarly racist responses to prompts.

One solution to eliminate said biases was to provide neutral or even sanitised responses. Conversely, these sometimes led to overcorrections, such as removing all references to Nazis when discussing Germany during

the Second World War. While seemingly less harmful than bias, overcorrections can result in inaccurate and unreliable responses. So, corrections should strive to eliminate unfair biases in responses but also be true to historical fact. In essence one needs to have a bias checker, then a fact checker. Again, test the output.

Reinforced bias

The other fundamental flaw is the way LLMs function. For example, allegedly one of the most common words to follow 'apple' is 'pie', which is entirely logical. However, if the AI follows more 'obvious' patterns consistently, and at a phenomenal rate, then other parings such as 'apple slices' might drop down in the rankings. Now, the issue of apple pie versus apple slices seems trivial, but extend that further to trading systems and algorithms that seek profit-driven patterns and correlations and the preference for more common successful patterns might result in results lacking uniqueness or certain results being missed.

Likewise, systems that are designed to find patterns of bad behaviour, from insider trading to mortgage delinquencies, might focus on typologies, obsessively target mergers and acquisitions lawyers who have insider information and deny normal interest mortgages to citizens who come from a lower socio-economic background. In the first case, the logic is indisputable, but the output might neglect to focus on other possible miscreants. In the second case, generations of long term financial and social harm can arise by denying mortgages to the underprivileged.

One means of avoiding the reinforced bias that can cause these and other harmful outcomes is to tweak the parameters of the AI algorithm by changing the prompt or datasets used. A fresh prompt removing overly common outcomes might yield more interesting and unique results. Not all apples turn into apple pie.

Human testing and oversight

These areas of concern speak to the need to have human oversight and testing, since AI untested is unreliable and untrustworthy by itself. The bias and fact checking suggested above might be hard to tune through an AI testing bot by itself, thus humans need to be involved in overseeing the parameters and the output. To have AI test another AI might intuitively sound efficient, but this would have to be done by an AI designed by different humans than those who created the original AI being tested. So, test the output with differently built systems by different people.

Creative ethical culture

It goes without saying that designing unique AI algorithms takes a high level of technical sophistication, but it also takes a level of creativity. In order to ensure AI is being built for good, it is important that this is all done in the right ethical and cultural environment.

While most will build with good intentions, some may not realise they are building deadly 'paperclip making killers'. The way to mitigate this risk is to have proper training in ethics for those building AI, and to ensure that outputs are not tested by the builders.

Core to the safe creation of 'good' AI is to always have the creators state what their purpose in creating AI is, what the expected outcomes are and what the possible negative outcomes are, so they can know what to look for, in most cases.

Cybersecurity

The cybersecurity arms of organisations will have a major role to play in keeping internal data protected. With the use of LLMs, any data presented to the LLM, be it in the form of a prompt or vector datasets, is used by the LLM for further training. This can result in internal data getting leaked to the outside world. Cybersecurity teams will have to ensure that they put sufficient checks

and balances in place so that organisations do not inadvertently end up sending confidential data to the LLMs, thereby making it publicly available.

Privacy controls

While a lot of information is available on the Internet, LLMs are able to consolidate this information and will have the ability to create a comprehensive profile of individuals, which can be misused by a nefarious actor. Ensuring that private information is not revealed by LLMs would require those controls to be defined and implemented. Organisations implementing LLMs need to ensure that they are embedding logic in the design to avoid disclosure of private or personal data.

GLOBAL AI REGULATORY FRAMEWORKS BEING ESTABLISHED

AI EU law

The EU has been a reliable vanguard for adopting rules to protect citizens from possible malfeasance with stringent privacy rules, anti-monopolistic sanctions, protections against abuses from Internet and technology companies and now AI. In 2024 it passed the 'Artificial Intelligence Act' (AI Act)[20] which broadly covers these areas:

- It establishes a general framework on safeguards surrounding general purpose AI.
- It prescribes limits on the use of biometric identification systems by law enforcement.
- It bans the use of social scoring and AI used to manipulate or exploit end user vulnerabilities.
- It establishes the right of consumers to launch complaints and receive meaningful explanations from those who use AI that affects them.

Notably, the EU takes pains to note that AI is a positive tool for innovation, but emphasises the need for 'trustworthy'[21] AI:

> This Regulation should be applied in accordance with the values of the Union enshrined in the Charter, facilitating the protection of natural persons, undertakings, democracy, the rule of law and environmental protection, while boosting innovation and employment and making the Union a leader in the uptake of trustworthy AI.

The union provides a robust list of positive use cases for AI, but notes the possible risks of unguided or unfettered AI:

> At the same time, depending on the circumstances regarding its specific application, use, and level of technological development, AI may generate risks and cause harm to public interests and fundamental rights that are protected by Union law. Such harm might be material or immaterial, including physical, psychological, societal or economic harm.[22]

The EU AI Act is commendable in many ways, however, unless this is adopted globally, which is highly unlikely, these risks will persist. Another inescapable flaw is that it is very difficult to detect, let alone enforce the provisions of the act. Threat actors who commit cybercrime are no more likely to pay heed to the lofty ideals of the EU AI Act.

If there was a means to watermark or identify products, services or systems that were developed using AI, or have AI embedded in them, then this one problem might be solvable. But this is unlikely.

Bill C–27 AI regulation in Canada

Canada has yet to pass AI legislation. However, in September 2023, the government did propose the Artificial Intelligence and Data Act (AIDA)[23] which will (1) reshape the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA) (2) introduce AI data safeguards

and (3) establish a data tribunal. Rules without consequences rarely have a strong impact.

AIDA is inextricably linked to Bill C-27:[24]

> An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act.

While this sounds quite comprehensive and sweeping, the scope is obviously even smaller in effectiveness than the EU AI Act.

A notable difference is that AIDA seeks to cover 'high impact' whereas the EU AI Act focuses on 'high risk' and is far more prescriptive.

Another increased strength in C-27 is the audit rights prescribed and the accountability of those who develop AI.

Lastly, the Personal Information and Data Protection Tribunal appears to carry significantly more weight than what exists today in Canada. To data PIPEDA has resulted in very few prosecutions, and yet significant privacy breaches have become an almost weekly occurrence without consequences. As the use and storage of data has proliferated, the judicial system has not kept pace accordingly.

The new Canada Personal Privacy Act also has heavier consequences than the current PIPEDA, which should help strengthen the true protection of the personal privacy of Canadians.

US Blueprint for an AI Bill of Rights

The US Government, specifically President Joe Biden, has introduced a Whitehouse Blueprint for an AI Bill of Rights (the Blueprint).[25] Naturally, the USA, while wanting to protect its citizens, does not want to introduce legislation that would limit their competitive growth in the space of AI.

The Blueprint[26] does have some important hallmarks that are noteworthy:

- the right to be protected from unsafe or ineffective systems;
- deterring discrimination in algorithms which should be designed with equality embedded decision making;
- protection from abusive data privacy actions;
- disclosure and explanation of use of AI impacting you; and
- opt out ability to require human-based decision making.

Again, the ideals espoused are laudable, but the reach is confined by borders. The Blueprint is also seemingly without teeth. But presumably, as case law is formed through precedent, especially in an extremely litigious USA, it seems inevitable that AI will come under much greater scrutiny in coming years.

Many US states are implementing their own AI legislation to ensure that protections are defined and implemented for the safe use of AI. In March 2024, Utah enacted the first AI-focused consumer protection legislation in the US, and many other states have pending legislation that is in process.

United Nations

More recently, in March 2024, more than 120 nation states in the United Nations backed a UN General Assembly resolution to promote 'safe, secure and trustworthy AI systems that will also benefit sustainable development for all'.[27] Importantly, the UN tied this to their existing framework of 17 sustainable development goals.

While the resolution is not yet enacted, it does set the world stage for a much more widespread adoption of an AI regulatory framework that is much needed.

Obviously, UN resolutions are not binding, but they form a more global framework that can help guide the AI rule-making framework in many countries, thus helping safeguard the use and care of AI globally.

So, this is not an exhaustive list of AI legislation, and more will be developed and implemented over time. It is worth wondering if these new rules might even get drafted by AI, although it would be wise to have humans fact-check and review any such machine-driven proposals.

BEYOND LEGISLATION: HOW INTERNAL CONTROLS NEED TO EVOLVE TO COVER AI RISK

Waiting for lawmakers to do the right thing is unwise. All firms looking to use AI, either natively or using third parties, should establish and follow certain principles.

AI governance council

Most organisations are excited about the opportunities that lie within the use of AI, such as creating or adopting systems that optimise operations and improve productivity and efficiency.

But firms should be equally concerned with both the reputational and liability risks of not creating an appropriate governance structure around their use of AI. An AI governance council, comprising the right executives, covering impacted divisions and including those with the technical skills and understanding, is the first step.

The council should have a charter to establish what the permitted or desirable use cases of AI are and the evaluation methodology of such use cases and systems, and it is important for that charter to delineate what elements are not permitted or would require a higher standard of due diligence.

For example, a third party AI system would have to outline where it is gathering data and the algorithm or prompt parameters.

Any system that uses PII can create extensive legal exposure and should be limited or contained appropriately.

Financial decision-making AI, while attractive in many ways in its hypothetical outcomes, can become a legal nightmare if it performs imperfectly and must have appropriate safeguards, including human intervention and oversight and testing.

Each AI project should be ranked in terms of both its benefits to the organisation, its multiple of improvement of speed and efficiency, its impact on customers and the firm's products and services and, importantly, the overall risk should it fail somehow.

Data integrity must be stressed throughout, and access rights must be reviewed and tested frequently to ensure third parties do not later create havoc.

Compliance

Increasingly, organisations are creating compliance regimes to oversee facets of their activity to ensure they conform to regulatory requirements. However, since AI rules are nascent, compliance should be established as an oversight body to review the safe use of AI without awaiting legislation.

The nature of establishing a compliance regime speaks to the need to establish governance parameters to ensure there is no overreach and no gaps. One of the first precepts is to outline what concerns are to be addressed by a compliance regime, notably: do no harm to individuals, ensure accuracy of data and ensure safety of outcomes.

The next step is to provide, within the framework, acceptable policies and determine who is responsible for drafting and enforcing said policies. Tied to that is usually an audit and testing framework. Many compliance regimes today are already embedding AI in their own governance, and it would be a conflict to have AI monitoring AI, so in this instance the AI should be monitored in a more manual manner, involving human oversight.

There should also be an established training methodology, just as the AI is trained to perform certain tasks, the creators of the AI

should be trained on the relevant AI governance principles and procedures.

The AI product scope should be established, and guard rails should be established to monitor for any potentially malicious or undesirable outcomes.

The basis for decision making should also be tested separately against different data sources and sets to ensure the accuracy and reliability of outcomes. Testing should be done without PII or confidential data, more as an algorithm 'sanity check'.

Then the actual data integrity should be screened, since all the best governance can come to nought if the underlying data sources are found to be unreliable or corruptible. The vastness of the Internet and the ability for false information to permeate can cause hazards.

Risk management oversight

The fundamental risk management process is to have suitably qualified people with the correct information and access gather all the data, create a risk register and prioritise the highest risk items and their outcomes, probability and mitigants.

Higher risk AI use cases need a higher level of scrutiny and due diligence and might fail to get development approval if deemed excessively risky. Manipulation of biometric data, for example, could be highly financially rewarding, but beyond being unethical, can destroy the reputation of the company and people that develop and deploy this. Recall the downfall of the once-powerful Cambridge Analytica.

It is advisable to also not just 'check and forget' but instead to periodically test the output to ensure nothing goes awry.

Audit controls

This speaks to the need to have an ongoing and periodic independent internal audit, to verify parameters and check against the expected compliance controls and to test the safety of the compliance regime that oversees the AI. Periodic examinations, with a risk-based focus, changing the audit elements to be tested, should help bring to the surface and correct the course of any issues where the AI might be stepping out of bounds.

Equally, there should be an annual external independent audit to oversee the AI program development and deployment. The pace of change is too great to wait more than a year, and the auditing standards will have to keep pace with AI innovation.

HOW TO BUILD AI TECHNOLOGY THAT REDUCES INHERENT RISK

Now that the key risks and the considerations for addressing these risks have been covered, the technical solutions that help reduce these risks need to be discussed. GenAI systems are reliant on the LLMs that they are built on. These LLMs need to be further refined to achieve these goals.

Key data considerations to tackle AI risks

With data being the backbone of AI systems, it is imperative that it be comprehensive, timely, accurate and transparent to reduce the probability of the risks manifesting themselves in the outputs. The key considerations from a data perspective are:

- *Comprehensiveness*: The ability of AI models to create output that is devoid of biases, hallucinations and general inaccuracies is severely dependent on the comprehensiveness of the data that is fed into the system. For example, OpenAI's GPT4.0 was trained on ~13 trillion tokens and contains 1.76 trillion parameters.[28] This includes Common Crawl, BookCorpus, Wikipedia, books, articles and more. Despite that huge dataset, it lacks datasets that might be needed for its model to predict outcomes accurately, as its data tends to get stale, and does not

contain information that might be industry-specific or cover topics that are not easily accessible in public datasets. Hence it is crucial that the input dataset represents a wide range of topics, writing styles and contexts. This diversity helps LLMs become more adaptable and capable of handling various tasks. To ensure comprehensiveness of data, and especially to enrich the initial training data, several techniques are used to provide additional domain-specific data to LLMs. These include fine-tuning, prompt tuning, retrieval augmented generation (RAG) to name a few. These are discussed further in the next section. Data sourcing for an LLM can be expanded by using multiple data gathering methods.[29]

- *Crowdsourcing*: Data can be sourced from a wide network of individuals who accumulate and label data. This technique can result in a wide variety of diverse datasets at a low cost point, though care needs to be maintained on the quality of data provided.
- *Managed data collection*: Data collection services offer a variety of data that can be accessed for training the LLM model. These services specialise in collating and cleaning large datasets through various regions, languages and topics. This is an efficient method of obtaining quality data but could be expensive depending on the dataset domain involved. Also, care needs to be maintained around data privacy, consent and ethical use.
- *Automated data collection*: Web scrapers can extract vast amounts of open-source textual data from websites, forums, blogs and other online sources. This is a quick and inexpensive way to gather vast amounts of data. The challenge is in ensuring data relevance and in ensuring intellectual property rights are not infringed. Data can also be automatically generated by using simulations that mirror the real world. Even AI generated datasets can be produced for this purpose.
- *Licensed datasets*: Purchasing datasets or obtaining a licence to data is a quick way to obtain curated domain-specific datasets for training of LLMs. This allows immediate access to vast well-structured datasets. This though could be costly, and the provider may limit the usage to certain limited scenarios.

- *Timeliness*: Another critical factor affecting the accuracy of GenAI models is the timeliness of the data they contain. Training an LLM requires intense computing power that is expensive to obtain. To keep the data current, the model could possibly be retrained, but due to the significant costs involved, retraining the model at a frequent interval becomes prohibitively expensive. This results in the data getting stale and the subsequent outputs from the model lacking the latest information to be able to provide accurate outcomes. Techniques like fine-tuning or RAG can be used to provide the LLM with enriched data that is more current.
- *Transparency and interpretability*: Machine-learning neural networks, on which LLMs are based, behave as black boxes, ie it is virtually impossible to determine how the model arrived at a certain output. With such opaqueness, it becomes exceedingly difficult to determine whether the output is accurate or whether it contains biases or has been 'hallucinated'. This is particularly problematic in the healthcare and criminal justice fields, where decisions can have far reaching consequences. There are techniques available like local interpretable model-agnostic explanations and SHapley Additive exPlanations that can be used to implement model explainability.[30] Additional techniques like counterfactual approximation methods are being developed for the interpretability of LLMs. Another less compute-intensive technique involves creating a dedicated embedding

space that aligns with the causal graph and is able to explain model predictions.

- *Accuracy and credibility*: The accuracy of output from an AI engine is directly linked to the dataset that has been used to train the neural network that it uses. It is imperative that this dataset is both broad and deep. To reduce the chance of biased output, the data needs to contain elements of all aspects of the question that is going to be posed to the AI model. Hence, it needs that breadth. Plus, it also needs to have sufficient depth, ie size of domain-specific data, to be able to interpret the correlations among data elements for an accurate output. Another key consideration is the quality of input data. Data obtained is vast, and it is crucial that it is curated and any inaccuracies removed before it is ingested by the AI model. On the output side there are many ways that the output can be refined to achieve better accuracy. For example, reinforcement learning from human feedback (RLHF) is a technique that provides direct human feedback or creates a reward-based model that mimics human feedback to achieve a more precise output. Finally, hallucinations are a direct output of neural-network based LLM systems. To reduce these, multi-shot prompting (providing several examples of context to the LLM) or filtering and ranking strategies can be used, which can reduce the randomness of the response, thereby improving accuracy.

Key AI system design considerations to reduce risk

GenAI systems are reliant on the LLMs that they are built on. These LLMs need to be further refined to reduce the inherent risks that arise from the technology. Once an LLM is trained using vast amounts of data, it can produce excellent output. But this output can have a range of issues, from biases to hallucinations. Plus, the LLM data can quickly get stale and may not have all the relevant information needed to produce the desired output. To address these issues with LLMs, there are a number of techniques that can be used:

- *Fine-tuning*: During the LLM training process, parameters and weights are assigned to the ingested data. Once the model is trained, this data remains static, and becomes stale over a period of time. Fine-tuning is a process by which a pre-trained model can be adapted to specific tasks or refreshed with new data. It involves retraining the model with a smaller targeted dataset and with anticipated desired behaviour. The process updates the model weights and parameters to improve performance and accuracy. While fine-tuning helps achieve specific objectives from a pre-trained LLM, one of the big disadvantages of this approach is that it is very compute-intensive, and hence it typically involves significant costs.
- *RLHF*: Another method of refining a previously trained AI model involves providing direct human feedback to the model or creating a reward-based model that mimics human feedback. This is passed directly into the AI model, and the LLM then adjusts its output based on user acceptance rates. Essentially the LLM learns through this feedback mechanism that if there is a high enough probability of the user accepting its output, it considers it okay to generate it. This helps in refining the output and making it more relevant and accurate without having to retrain it with more data.
- *Prompt tuning or engineering*: This technique involves feeding front-end prompts into the LLM, to provide it with context of the task. These prompts, which could be either human or AI generated, provide the LLM with clues that help it in performing the desired action. This technique allows an organisation to ask the model to perform narrowly defined tasks, thereby getting a much more precise output. Essentially, it is possible to customise a large model to

specific needs without the need to retrain it or fine-tune it, which could be significantly expensive from a compute needs perspective. Another benefit of prompt tuning is that it can also be used to remove bias in LLM responses.

- *RAG*: This technique allows augmentation of the LLM dataset with external data sources, thereby enriching the data in the LLM. This technique is particularly useful when the solution requires current knowledge which is not present in the original training dataset for the LLM. Apart from that, if the solution requires domain-specific knowledge, which is not contained in the LLM, this technique comes into its own. To create a RAG-based solution, external datasets are curated, vectorised and fed into the LLM to augment the information available in the LLM. This is a very cost-effective solution to expanding the knowledge base of an LLM.

HOW AI CAN SUPPORT EFFECTIVE COMPLIANCE

The use cases for AI in compliance are nearly limitless. As more data and more systems driven by the oceans of data and systems that feed off this data are created, it will be near impossible for humans alone to monitor such use cases and their data integrity. Instead, ironically, AI will be used to go beyond simple monitoring for erroneous patterns and exception reports and will be tasked with connecting myriad points of data, which a human cannot do with the speed, depth and accuracy of AI. Ultimately humans will need to oversee and adjudicate said data and outputs.

AI IN COMPLIANCE IN FINANCIAL SERVICES

One example of AI in compliance is already found in financial services. Where data is measured in terabytes and there are funds and motivation to use AI for creating opportunities, and likewise, AI is needed to monitor these for effective compliance oversight.

Trading

Trading algorithms have come to dominate electronic trading in most marketplaces, so likewise, compliance oversight can now marry trading between different asset classes of similar underlying assets, sometimes linking transcripts from recorded trader phone calls and news releases to tie together any web of collusion and manipulation in the marketplace.

AI based detection can also pierce the holdings of funds and ETFs, finding patterns of trading that are not easily detected on the surface by just looking at a single asset class.

Research

Likewise, investment research can now draw from a near infinite lake of data points. AI can check the sources for accuracy with blinding speed and precision, review for proper disclosures and even detect possible conflicts and connections that a human reviewer might struggle to do in the rush to review a research piece before the 7:30 am firm morning meeting.

AI would probably also be able to put together more data points more quickly, possibly read through inconsistencies or irregularities in company regulatory filings or see patterns in underlying commodities or vertical partners or competitors, which would better inform the company's fundamental analysis and lead to more accurate and more informed forecasting. Compliance would be able to overlay AI to test predictions against larger datasets. Further, compliance would be able to link any possible conflicts that would require disclosure.

Investment banking

Similarly, the complexity of relationships, corporations and data that drive investment

banking can be captured and analysed with precision by compliance to detect any possible leaks of confidential information, or other nefarious activity.

The art of deal making would be similarly enhanced by the use of deeper datasets, but likewise compliance would be able to probe deeper into what names should be on what confidential deal lists and when based on travel patterns of bankers or reviewing company e-mails or even detecting anomalous trading patterns that suggest inappropriate trading activity ties to possible banking transactions.

Product analytics and risk rating

The sheer volume of investment products and associated risk ratings, created to categorise which products are suitable for which client portfolios, is a staggering sum, which is not static and must be reviewed continuously. For advisers and clients, AI can be invaluable in finding the right product mix for the client's portfolio.

The ability to review this dynamically and tilt portfolio holdings appropriately is a near impossible feat for an adviser with a large book of business. But AI can help oversee this in real time. Likewise, compliance can oversee the suitability standards and provide more accurate and timely guidance.

The possibilities are limitless. However, the risks associated with building AI-enabled products for a business can be mitigated by building compensating AI-driven compliance tools.

CONCLUSION

AI is an exciting tool which is growing in adoption and complexity every day. At the time of writing this paper, NVIDIA announced a new AI 'Blackwell' chip which is five times as powerful as their latest GPU. The possibilities of expansion seem limitless.

However, the risks of expanding without safeguards are also growing. AI will not disappear and use cases and risks will only multiply over time. Compliance and regulations will help form the necessary frameworks, and enthusiasm for the increased productivity and efficiencies AI offers should be tempered by the need to ensure rigorous guard rails to defend society from wounds that can be inflicted by the simplest of things, like a paperclip.

The potential of the technology seems limitless, and so is the human desire to maximise its benefit. As this paper has articulated, the risks are as limitless as the exciting promises that the technology offers. AI has the potential not only to match but arguably to supersede human abilities. Civilisation must do everything in its power to create an ethical and responsible AI ecosystem, that will help achieve the efficiencies and capabilities it offers while safeguarding against the considerable risks it poses. An intelligent civilisation must strive to use it as a tool that furthers and extends human intelligence.

References

(1) Turing, A. (1950) 'Computing Machinery and Intelligence', *Mind*, Vol. LIX, No. 236, pp. 433–60.

(2) Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L. van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D. (2016) 'Mastering the Game of Go with Deep Neural Networks and Tree Search', *Nature*, Vol. 529, pp. 484–89.

(3) Hintze, A. (14th November, 2016) 'Understanding the Four Types of AI, from Reactive Robots to Self-aware Beings', The Conversation, available at https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616 (accessed 16th September, 2024).

(4) McCarthy, J. (1992) 'Arthur Samuel: Pioneer in Machine Learning', *IBM Journal of Research and Development*, Vol. 36, No. 3, pp. 329–31.

(5) Brockman, G. (5th December, 2023) 'ChatGPT just crossed 1 million users; it's been 5 days since launch', X post, available at https://x.com/gdb/status/1599683104142430208?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%

5E1599683104142430208%7Ctwgr%5Edf67b8e6f588b503bde37cd4f316534808dcb619%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fwww.techopedia.com%2Fchatgpt-statistics (accessed 16th September, 2024).
(6) Duarte, F. (27th July, 2024) 'Number of ChatGPT users (Aug 2024)', available at https://explodingtopics.com/blog/chatgpt-users (accessed 9th September, 2024).
(7) McKinsey & Company (22nd February, 2024) 'Beyond the Hype: Capturing the Potential of AI and Gen AI in Tech, Media, and Telecom', available at https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/beyond-the-hype-capturing-the-potential-of-ai-and-gen-ai-in-tmt#/ (accessed 9th September, 2024).
(8) *Ibid*.
(9) *Ibid*.
(10) Henshall, W. (24th October, 2023) 'AI Experts Call For Policy Action to Avoid Extreme Risks', available at https://time.com/6328111/open-letter-ai-policy-action-avoid-extreme-risks/ (accessed 9th September, 2024).
(11) Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrel, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Güneş Baydin, A., McIlraith, S., Gao, Q., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J. and Mindermann, S. (2024) 'Managing AI Risks in an Era of Rapid Progress', available at https://arxiv.org/pdf/2310.17688.pdf (accessed 9th September, 2024).
(12) McKinsey & Company, ref 7 above.
(13) Lutekevich, B. (2023) 'AI hallucination', TechTarget, available at https://www.techtarget.com/WhatIs/definition/AI-hallucination (accessed 9th September, 2024).
(14) Armstrong, K. (27th May, 2023) 'ChatGPT: US Lawyer Admits Using AI for Case Research', BBC News, available at https://www.bbc.com/news/world-us-canada-65735769 (accessed 9th September, 2024).
(15) Little, S. (23rd January, 2024) 'AI "Hallucinated" Fake Legal Cases Allegedly Filed to B.C. Court in Canadian First', Global News, available at https://globalnews.ca/news/10238699/fake-legal-case-bc-ai/ (accessed 10th September, 2024).
(16) Bengio *et al*., ref 11 above.
(17) Henshall, ref 10 above.
(18) Asimov, I. (1950) 'Runaround', in 'I, Robot' (The Isaac Asimov Collection ed.), Doubleday, New York p. 40.
(19) Gans, J. (10th June, 2018) 'AI and the Paperclip Problem', CEPR, available at https://cepr.org/voxeu/columns/ai-and-paperclip-problem (accessed 10th September, 2024).
(20) European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, available at https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf (accessed 10th September, 2024).
(21) *Ibid*., p. 5.
(22) *Ibid*., p. 6.
(23) Government of Canada (2023) 'Artificial Intelligence and Data Act', available at https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act (accessed 10th September, 2024).
(24) Parliament of Canada Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts, C-27 (44-1), available at https://www.parl.ca/legisinfo/en/bill/44-1/c-27 (accessed 10th September, 2024).
(25) The White House (n.d.) 'Blueprint for an AI Bill of Rights', available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (accessed 10th September, 2024).
(26) The White House (October 2022) 'Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People', available at https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf (accessed 10th September, 2024).
(27) United Nations General Assembly (11th March, 2024) 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development', available at https://documents.un.org/doc/undoc/ltd/n24/065/92/pdf/n2406592.pdf (accessed 10th September, 2024).
(28) McGuinness, P. (12th July, 2023) 'GPT-4 Details Revealed', Substack, available at https://patmcguinness.substack.com/p/gpt-4-details-revealed (accessed 10th September, 2024).
(29) Javaid, S. (2024) 'LLM Data Guide & 6 Methods of Collection in 2024', AI Multiple Research, available at https://research.aimultiple.com/llm-data/ (accessed 10th September, 2024).
(30) Cloudera Blog (8th May, 2020) 'ML Interpretability: LIME and SHAP in Prose and Code', available at https://blog.cloudera.com/ml-interpretability-lime-and-shap-in-prose-and-code/ (accessed 10th September, 2024).