Artificial Intelligence on the Rampage: Speculation on the Need for Guardrails/Regulations

Peter MacKinnon, Senior Research Associate

Faculty of Engineering, University of Ottawa

To the Foresight Synergy Network Via Zoom

August 24, 2023

Agenda

- **A Brief Review** – past comments as context about AI today
- **Generative AI & Integrity**
- **Guardrails & Regulations**
- **Summary**

Why a Rampage?

- **AI has exploded around the globe in recent months because of a new type of commercial AI tool called chatbots**
  - **Chatbots** are derived from a branch of AI called Natural Language Processing (NLP) which manipulates natural languages like English or Mandarin using Large Language Models (LLMs) or components called Transformers
  - **Models** are trained with vast quantities of data, often scrapped from the internet, which may or may not be appropriate data to use, particularly without permission
  - **The global media are full of stories about the power & shenanigans of this in-the-wild technology**, which is clearly causing a rampage around the world
  - **Governments & industry leaders are calling for guardrails & regulation**

In the Sweep of History

- **Human history can be characterised as evolving from a hierarchy based on those who sought to:**
  - Control land (~6,000 years)
  - Then the control of machines (~350 years)
  - Now the collection, manipulation & control of data with AI (<50yrs)
- **These journeys are marked by a transition from:**
  - Manual labour of man and beast
  - To machine assisted labour saving devices
  - To cognitive tools for societal use
    - All in an accelerating rate of return

Lessons from History

- We have learned from history how various technologies sparked the industrial revolution
  - water mills, the cotton gin, steam engines, etc.
- These machines dramatically change our polities, societies, economies, employment patterns, demography & even our cultures
- AI is likely to have a similar if not more profound impact

Now: Suddenly AI

- **AI is not new**, it has been around in academic circles since the mid 1950’s with origins that go back even further
- Since that time **AI has vacillated between periods of research progress with associated business uptake and research failure & consequent business withdrawal**; so-called ‘spring and winter’ periods
- **By the early 2000’s, AI came into full ‘summer bloom’**
- **Where it goes from here depends & technical advancements & social acceptance**
  - Today that path is unclear

Now: Suddenly AI (2)

- Artificial Intelligence has skyrocketed into world news with stories of promise & peril for humanity
- Political leadership is declaring AI a strategic national asset
- Governments around the world are pouring billions of dollars into AI research & related industry development
- Companies are investing billions in applied AI research, product development & enhanced AI-related services
- Competition is heating up between companies & nations
- **Meanwhile companies can't stop competing & the ethics, efficacy & philosophical implications of AI to society are being questioned by a broad community of concerned commentators**

AI in Three Flavours & Timeline

- **Narrow AI** is now
- **General AI** is yet to emerge as a robust capability, perhaps before 2100
- **Superintelligence** is the most difficult and complex challenge in AI and is not, if at all, anticipated in any practical form until well into the future
  - despite media hyperbolae and science fiction apocalypses about the rise of superintelligence
  - Including gross misunderstanding of today's AI in the eyes of the public, including most decision makers be they politicians or business executives

Some Terminology

- **Chatbots** – is a software application or web interface that aims to mimic human conversation through text or voice interaction
- **Generative Artificial Intelligence (GAI)** – a form of AI capable of creating text, images, music & videos using generative models
- **Generative AI Models** – Learn the patterns & structure of their training data & then generate new data that have similar characteristics
- **Large Language Models (LLMs)** – Are neural networks that learn patterns in language through training on vast & diverse data sets (e.g., scraping the internet) & transform that data into new content
- **Neural Networks** – In computer science, a method that teaches computers to process data in a way that is inspired by neural networks of the human brain & constitutes a form of machine learning called deep learning
- **Transformers** – Are a form of deep learning architecture

Current State of AI

- The fact that applications are called “Artificial Intelligence” does not mean they are intelligent
- Today’s AI is based on algorithms that are designed to process massive amounts of training data using sophisticate statistical analysis & presentation techniques
- **We should not apply human qualities such as integrity & ethics to label these applications; rather apply these qualities to the way that they are being developed & the outcomes associated with their usage**

Is AI a Danger?

- The answer is **'It depends'**
  - It is like asking if a match is dangerous, well 'it depends'
- **There are reasons for concern - both technical & socio-economic**
  - Machines are now able to take on less-routine tasks & this transition is occurring during an era in which many workers are already struggling
  - Automation anxiety is made more acute by a labour market that has tilted against workers over the last 30 years, with increasing income inequality & stagnant real wages
- In the past, automation has meant industrial robots & computer hardware & software designed to do predictable, routine & codifiable tasks
- **In the future, AI-enabled robotics & related systems** will be able to perform human-like tasks that **may lead to massive displacement in employment from the shop floor to the executive suite**

Is AI a Danger? (2)

- **These are tasks requiring problem solving, decision making & interaction within a less-than-fully-predictable environment**
  - Automation of this sort includes self-driving cars & diagnosing complex diseases
- Dual use draws both integrity & ethical lines in the sand about where one stands, for example:
  - Are you for or against autonomous weapons?
  - Are you for or against fraud detection?
  - Are you for or against less privacy?
  - Are you for or against fake news?
- **The real answer about AI's danger is still undetermined, time will tell as advances & transformations occur**
  - (e.g., the sudden appearance of GAI-enabled chatbots)

Negative Public Perceptions of AI

- Anticipated massive impact in loss of jobs without replacement
- Perceptions of digital surveillance through loss of privacy & trust
- The rise of superintelligence
- The public fear factor is greater than any expression by political leadership yet some technical & futurist authorities are calling for government imposed guardrails & regulation

AI & Geopolitics

- **AI has enormous potential to be disruptive to the current model of governance of the state & add to geopolitical competition** & by example the use of:
  - Ever increasing echo chambers of social segmentation through social media
  - Hard to discriminate fake news
  - Election meddling through digital maleficence
  - Purposeful disruption of critical infrastructure

Arising Socio-economic Impacts

- Accelerating job losses across multiple business sectors primarily arising as a result of robotics & machine learning
  - This is a serious global public policy issue
- Emergence of Basic Income
- New kinds of jobs, primarily in knowledge intensive areas, often assisted by AI systems
- Dual use prospects are high & broadly worrying

The Future of AI

There are two main views on the medium & long-term future of AI

- **The 1st could be called 'cyclical'**
  - It expects the season-like cycle that has characterized AI so far to continue
- **The 2nd can be seen as an irreversible 'tipping point'** towards an unprecedented (non-human) intelligence explosion
  - It expects unending invention & innovation in AI

The Future of AI (2)

Here is my bifurcation hypothesis, which lies between the previous two within the timeframe of the next 25 years

- **AI bifurcates into dual purpose** streams (~2025)
- A '**White stream**' will explode well into the future with commercial & socio-economic successes across the economy & national fabric
  - This will be primarily the turf of companies
- A **'Dark stream'** will expand in three ways:
  - **Criminal** - expand as fast as possible with caution to the wind
  - **Security** - expand as fast as possible with caution to the Triad
  - **Kinetic Defence** - more slowly & in many respects more cautiously
    - This includes robotics (e.g., autonomous vehicles & weapons) & many kinds of AI apps

Then there are Disruptors

Increasing difficulty of making new breakthroughs

- Progress in science depends not just on funding available & the effort put in, but also on how ‘hard’ progress is

Eventual hardware limitations

- On a related note, it is possible that along with conceptual & software limits, we may also reach fundamental physical limits to our hardware & this will slow progress towards advancing AI

As well as Wild Cards

- A breakthrough in cognitive neuroscience
- New human cognitive enhancement technologies
- A ‘Sputnik event’ – perhaps in retrospect it may turn out to be the release of ChatGPT
- Societal distrust & disinclination
  - public concerns over technological unemployment, machine bias, automated surveillance or digital propaganda will create critical legitimacy problems driving public distrust & societal backlash towards AI

Societal Concerns re AI

- It is important to think about the ethical & legal implications of AI & consider measures for the design & deployment of AI systems
  - Responsible supervision (e.g., integrity & ethics)
  - Governance
  - Guardrails (e.g., voluntary)
  - Regulation (e.g., enforceable by law)

Ethical Direction of AI

- **AI is not moving in a single ethical direction**
- Like all technologies, AI is being pursued by a diverse field of players ranging from:
  - research institutions
  - national and sub-national governments
  - companies & professionals of all descriptions
  - non-state actors such as criminal gangs & terrorist groups
- **Each player anticipating positive outcomes against their goals**

Promising Ethical Applications of AI

- **Infrastructure**
  - Smart Cities
  - Intelligent Transportation Systems
  - Autonomous robots – driverless vehicles
  - Energy ecosystems
- **Manufacturing**
  - Advanced robotics
  - Additive manufacturing
- **Professional Services**
  - Auditors & Accountants
  - Lawyers & Paralegals
  - Healthcare workers
  - Scientists & Engineers

An Emerging AI Race – Why?

- The notion that there is an emerging AI race is real
- AI is a disruptive technology in its own right with great promise for wide-scale use
- AI enables other disruptive technologies creating as yet poorly understood synergies that could be of a dual use nature
- Global economic benefits from AI are anticipated to add trillions of dollars to global GDP
- AI capabilities & capacity will define the competitive advantage of nations in the future

Artificial Intelligence & Integrity: Winners & Losers

- AI enabled applications have the potential for societal
- There is a growing concern about the potential for sophisticated use of AI to create deepfake videos, images, audios, and text that manipulates and/or fabricates content
  - with the potential to spread misinformation, deceive individuals & organisations, and manipulate public opinion
- This kind of behaviour can lead to the erosion of trust & credibility
  - such as news, politics, online interactions, and even among friends be they persons or nations, thus leading to a fragmentation of society

Integrity versus Ethics

- **Integrity is an internalization of beliefs such as being honest and fair**
  - It is absolute and lies at the core of the human psyche or anima
- **Integrity manifests itself through human ethical behaviour**
  - which is a cultural set of rules and ideas that have evolved over time against an expanding framework of moral principles
- **Ethics is an externalization shaped by the cultural environment**
- **GAI has a vast potential to create both positive & negative impacts ranging from individuals to society as a whole**
  - The net result would be a segmentation of the world into winners and losers.
  - So, who are they and what are the implications?
- **There are three broad groups of stakeholders at play in the GAI space: product and service suppliers, users, and potentially governments as some kind of regulator**

The Winners

- **Many observers anticipate a significant economic boon to the world economy**
  - with projections of additional trillions of dollars to the world GDP based on primarily & secondary usage of GAI based systems
- **Aside from financial gains, there will be many benefits to society as a whole** as a direct result of new products & services designed to improve the lives of individuals, the value of companies & the efficiency of governments
- **Current biggest winners are the principal platform providers**
  - with their expanded ecosystems of acquisitions of mainly smaller technology-based companies, as well as their strategic & tactical partnerships including with value-added resellers

The Winners (2)

- **The next big winners are the early business & individual adopters of GAI**
  - including business users such as accountants, lawyers, medical practitioners, business & policy analysts as well as others such as writers, researchers, educators & students
- **The third category of "winners" unfortunately are the emerging range of nefarious actors**
  - who use GAI with the intent of influencing others for a wide range of selfish and/or ideologically motivated intents from fraud to espionage
  - Clearly this category operates in a world where integrity is irrelevant and disruption is a purpose.

The Losers

- **Given the potential for dual use & abundance of nefarious actors, cracks in the social fabric are starting to appear**
  - (e.g., lawyers being duped into sighting fake legal cases in court; the creation of faulty financial statements; essay and exam cheating by students & professionals & insidious ramblings from chatbots)
- **In addition, serious concerns about job security are arising**
  - as per the summer of 2023's American writers' and actors' labour dispute, which in part is due to the threat of GAI displacing jobs by machines. Given the potential harm that GAI could rain on civilization, there is a range of threats with impacts that vary from the individual to society at large

The Losers (2)

- **Most AI systems rely on vast quantities of data for training and decision-making, much of which is scrapped from the internet**
  - Leading to a new kind of legal problem with respect to ownership & use of data between AI vendors & owners of the data
  - Moreover, if the training data are intentionally manipulated or biased, it can lead to compromising integrity
  - Adversaries can inject misleading or distorted data into the training process to influence the behaviour or outcomes of AI models, leading to such actions as incorrect decisions, biased results & challenging the integrity of the user with misleading and/or disinformation

What is driving the erosion of trust?

- **AI chatbots can be programmed to manipulate or deceive users**
  - These chatbots can be designed to impersonate humans, spread misinformation, or manipulate emotions to exploit individuals' vulnerabilities including data theft, financial fraud & social engineering attacks such as stalking & bulling
- **AI algorithms can be manipulated or intentionally biased to achieve specific outcomes or objectives**
  - e.g. on social media platforms & online advertising, AI algorithms can be manipulated to amplify certain content, manipulate user behaviour, or reinforce echo chambers
  - This can undermine the integrity of online information, distort public discourse & manipulate user experiences beyond the user's intent
- **AI algorithms employed in decision-making processes can inadvertently perpetuate or amplify existing biases and inequalities in an AI-enabled application**
  - e.g. loan approvals, hiring decisions, or criminal justice applications
  - If these algorithms are not designed with integrity in mind, they can lead to unjust or discriminatory outcomes, consequently undermining the integrity of the decision-making process

What is driving the erosion of trust? (2)

- **Lack of data protection measures**
  - such as unauthorized data sharing, or insufficient transparency about data usage can erode privacy rights and compromise individuals' control over their personal information
- **AI can be employed to invade privacy**
  - by analyzing and mining vast amounts of personal data with AI algorithms that can be used to infer sensitive information about individuals, such as their preferences, habits, or personal details, even without explicit disclosure
  - This intrusion into personal privacy compromises the integrity of individuals' information & can lead to misuse, unauthorized access, and character & extortion attacks on individuals and even organisations from which data were inappropriately acquired
- **AI algorithms are widely used in stock markets for algorithmic trading**
  - however, they also can be manipulated to gain an unfair advantage
  - Malicious actors can employ AI techniques to manipulate stock prices, engage in market shenanigans, or conduct high-frequency trading with the intent of exploiting market vulnerabilities

What is driving the erosion of trust? (3)

- **If integrity is not prioritized in AI development & deployment, society as a whole could suffer**
- **If AI systems are not designed with integrity and ethical considerations, marginalized communities can be disproportionately affected**
- **Biased algorithms or discriminatory practices can reinforce existing disparities & exacerbate social inequalities**
  - Therefore, It is crucial to ensure that AI technologies prioritize fairness, inclusivity, and equal opportunities for all segments of society
- **GAI algorithms can be used to develop automated hacking tools** that exploit vulnerabilities, bypass security measures, as well as conduct targeted attacks
  - These kinds of activities can compromise the integrity of network & computer systems, applications and data, as well as individuals, organisations and infrastructure that become victims of attacks
- **AI can disrupt jobs**, not just manual jobs many of which are being taken over by AI-enabled robotics from shop floors to warehousing and retail; but also many kinds of knowledge worker jobs across virtually all professions
  - **This is leading to workforce displacement and job insecurity, consequently if integrity considerations are not adequately addressed, the impact on affected workers, their organisations and clientele may be disruptive**
  - **Organizations that fail to manage the job displacement transition and provide support for affected individuals can expect social and economic challenges to the organization and disruptions in the community**

What is driving the erosion of trust? (4)

- **The exploitative concerns with GAI arise due to the malicious use or manipulation of AI technology and are not inherently stem from AI itself**
- **These exploitations are caused by the intentions and actions of those who utilize AI for nefarious purposes**
- Responsible development, ethical considerations, and appropriate safeguards can help mitigate these risks and preserve integrity in AI applications
- By prioritizing integrity, responsible AI practices & ethical considerations, stakeholders can strive to create a more equitable & beneficial AI landscape for all
- **The winners and losers in the context of AI and integrity are not fixed or predetermined**
  - The impact can vary based on the actions & decisions taken by stakeholders across different sectors
  - Individuals who lack awareness or understanding of AI risks and the importance of integrity may be at a disadvantage
  - They may unknowingly fall victim to biased or discriminatory AI systems or be affected by privacy breaches
  - Lack of knowledge or control over AI systems can limit one's ability to protect their interests & challenge unfair or harmful AI practices

The Need for Guardrails/Regulation

- **The release of powerful AI-based chatbots in recent months witnessed an unprecedented level of market penetration** reaching over one hundred million internationally distributed users in a matter of weeks from launch
  - Yet these products, plugins & related services are being released without any external safeguards, which can lead to a vast array of illicit uses of GAI.
- **If AI systems are deployed without proper integrity measures, public trust in institutions & organizations utilizing AI can erode**
- Moreover, instances of AI failures, breaches of privacy, or unethical use of AI can undermine trust in those responsible for AI development and deployment
  - Consequently rebuilding trust can be challenging & require significant efforts to re-establish integrity & transparency. Prioritizing integrity in AI development and deployment can help mitigate risks, maximize benefits, and create a more equitable and responsible AI ecosystem for all stakeholders.
- **Implementing & preserving integrity in AI systems points to a need to invoke some kind of guardrails or regulatory framework**
  - Adhering to such rules or frameworks promotes integrity by ensuring that AI systems meet specific standards, respect legal requirements & operate within established boundaries
  - By following these principles, AI can be developed & deployed in a responsible, transparent & beneficial manner.

The Need for Guardrails/Regulation (2)

- **Governments around the world are taking notice of emerging issues around AI** & invoking discussions on what kind of rules are needed and how to coordinate internationally
  - This includes consideration for the emergence of automated decision-making driven by AI
- **An example of a major concern is that the training data used in GAI systems can replicate intentional & unintentional biases & discriminations created by designers & programmers leading to tainted content that favours certain groups at the expense of others**
- **At issue is there is generally no specific remedies for addressing AI maleficence** other than a broad set of existing legislation associated with human rights, privacy law, tort law & intellectual property law

What are Guardrails & Who is taking Action?

- **A guardrail is** a strong fence that protects people & things from falling off a precipice
- **Currently AI Guardrails amount to voluntary rules** to meet minimum certification & performance requirements with the intent that users are not deceived in what they request
- **China** introduced Guidelines for Large Language Models in April 2023
- **Canada** introduced *Canadian Guardrails for Generative AI – Code of Practice* in August 2023
  - These guardrails are based on:
    - Adopting safety measures
    - Testing codes
    - Disclosure practices

Guardrails

- **Ontario** introduced legislation on its own use of AI with a *Trustworthy Artificial Intelligence (AI) Framework* back in 2021
- **The USA** introduced voluntary guardrails for AI with six major US AI companies signing on to have 3rd party inspection of the firms' algorithms
  - Google, OpenAI, Microsoft & Anthropic form coalition for Responsible AI following meeting with the US President
- **European Union** – is on the verge of passing the first major legislation to regulate AI based on risk & mandates various development & use requirements

Regulations

- **Regulation are rules made by government or other authority** in order to control the way something is done or the way people behave
- **To date the introduction of regulations is mostly a debate** in legislatures and corporate Board Rooms
- **However, the use of AI within the European Union soon will be regulated** by the *EU Artificial Intelligence Act*, the 1st of its kind in the world

Precautionary Note

- The promise of AI may yet fizzle due to:
  - **Technical barriers**
    - Limitations in computing resources, inability to scale training models, lack of explainability in models

  or

  - **Hit a wall brought on by social resistance**
    - Triggered by rapid changes & fragmentation as nations, companies & individuals try to grapple with profound change AI will bring to the order of things

AI & Integrity in Summary

- **If AI systems are developed and deployed with integrity as a priority of the providers, then society can benefit from improved services, enhanced decision-making & innovative solutions**
- **Fair and unbiased AI applications can ensure equal opportunities**, reduce discrimination, and enhance access to resources & services
- **Ethical AI can also empower individuals** with greater control over their personal data and foster transparency in decision-making processes
- **By integrating integrity into the development, deployment, and use of AI systems, organizations can ensure that AI technologies are aligned with ethical principles, societal values & human well-being**
- **Emphasizing integrity in AI fosters trust, reduces risks, and paves the way for the responsible and sustainable integration of AI into various aspects of society**

AI & Integrity in Summary (2)

- **Embedding integrity into AI systems requires a multidimensional approach**, encompassing technical, organizational & ethical considerations on the part of the constructors as well as the users
  - It also involves a commitment to ethical principles, transparency, fairness, accountability, and responsible governance to ensure that AI systems benefit society while upholding human values
- Without adhering to developing GAI systems designed & developed by people with integrity, an approach casted against an ultimately international regulatory framework, **society will become the biggest loser**
- **The results will be unpredictable and could even lead to the demise of ethical principles** that have been established over many centuries to enable world order as we know it
- **By prioritizing integrity in AI deployment and decision-making, society can harness the potential of AI while upholding ethical standards and ensuring a fair and just future**

General Summary

- Artificial Intelligence has evolved over the past 60 years in fits & starts
- AI is now an emergent disruptive technology
- AI is dual purpose (White Box - Black Box)
- The limits to applying AI are wide open
- **AI wars are possible** among commercial, economic & criminal players as well between states depending on dual use developments
- **All publicly facing AI applications, be they public or private sector, should adopt the use of explainable AI**
    - A concept where the logic of an AI decision is explainable as opposed to a common form of AI that is 'black box' whereby the logic associated with a given output cannot be explained
- **AI will likely have a profound effect on society with its overall net benefits still unclear**

Merci – Thank You

Peter MacKinnon

Senior Research Associate

Faculty of Engineering, University of Ottawa

mackinnon.peter@gmail.com

Resources

- **Elements of AI**: a free online course from the University of Helsinki https://www.elementsofai.com
- **Artificial Intelligence and Integrity: Winners and Losers**: op-ed by Eli Fathi C.M. & Peter K. MacKinnon, Medium website, published August 9, 2023 https://www.medium.co/mackinnon.peter/artificial-intelligence-integrity-winners-loosers-95e66ea8c945