A framework for countering the risks of Artificial Intelligence

Summary

Artificial Intelligence (AI) hits the headlines every day with its promises of new benefits for mankind, but also by concerns about new harms. Normal commercial investment and competition will ensure we get the new benefits, but the harms of AI are already evident and many are very difficult to deal with; 'AI Safety' is therefore an urgent priority.

AI is often discussed as if it is just one topic: it is not. This paper aims to define a framework of the various players in the AI landscape, the interactions between them, and the associated risks of harm from AI to various types of users. The framework helps us define the different types of measures required to counter the risks.

The players in the AI landscape are the Suppliers of AI apps, tools and components, their 'innocent' Users, the Aggressors who attack the Users, and the Defenders whose job is to protect all Users. Suppliers overlap with Aggressors and Defenders because their AI products are also used by both of these groups who range from nation states down to individuals. ('Users' includes people but also hardware devices and software that may interact with AI products.) The AI landscape is thus the scene of multiple complex interactions between the various players. Their effects range from benefits and harms to many Users, to multiple and rapidly evolving AI-assisted warfare between the other players.

The topic of 'AI Safety' covers the risks of harm to three main classes of Users, that require different countermeasures. They are a) the physical and services infrastructure that society relies on that may be disrupted by AI-enabled attacks, b) individual persons and institutions, and c) society in general. The risks of harm to the first two classes are already evident; the potential for harm to society in general seems to be under-appreciated but is growing.

Risks to our infrastructure from AI-assisted attacks can only be mitigated by major investments in strong cybersecurity measures. Since so much of the infrastructure on which we depend is of a global nature, international cooperation in this area is vital. (We regard as implausible the idea that advanced AI could 'take control' and pose an existential risk to humanity. However, advanced AI could well be weaponized to cause severe harm to life.)

Risks of harm to individual persons and institutions are complex and wide-ranging. Any product reliant on AI should be required to keep its users safe from harm in the same way as products in any other industries, such as transport, food, health. AI Suppliers that launch unsafe products must be held accountable for the harms they cause. AI safety is too important to be left to free markets or industry self-regulation.

The risks to society in general are that the on-line 'infosphere' becomes so polluted with misinformation, disinformation, fakes and hallucinations, etc., much of it generated or mediated by AI, that we can no longer trust the infosphere for learning or for communication. A general breakdown of trust in the infosphere would be a disaster.

Implementing cyber-security protection is straightforward in principle though often complex in practice. Nevertheless, its importance is such that the owners of infrastructure assets

whose disruption could harm large parts of society should be made legally liable for protecting their assets from cyber-attacks.

In contrast, protecting individual persons, institutions, and society in general against AI-enabled harms is inherently more complex, ranging from the need to protect children from harmful material to deciding on the limits of free expression. Given the untrustworthy nature of much AI-generated material, pollution of the infosphere can only be mitigated by including a non-erasable 'health warning' watermark in any such material. Only Governments can make the difficult decisions on the types of protective regulations needed.

1. The AI benefit/risk Balance Sheet

The year 2023 has been notable for the wild daily fluctuations in the AI balance sheet of claimed benefits versus tales of new harms. One day we are told of new AI-enabled medical treatments, the next of new AI-generated means for cyberfraud attacks.

Until 2023, the prospective benefits of AI seemed to outweigh the risks of harm, and the launch of Large Language Models (LLMs) initially drew widespread acclaim. But 2023 also saw the world wake up to the scale of the risks. The US, the EU, the UK and other nations started work to establish new laws and regulators to limit the harms of AI. Internationally, the UN, the G7 and others established commissions to tackle risks from misuse of AI under the umbrella title of 'AI Safety'. Both Eastern and Western nations supported these initiatives.

In spite of these positive moves, in early 2024 the AI balance sheet seems to be weighed down by the risks. The beneficial uses of AI do not need any regulatory help to find their way to market. Countering the risks is by far the greater challenge. The World Economic Forum's annual risk report for 2024 [1] lists the top 4 global risks, ranked by severity of impact, as:

1. Misinformation and disinformation – much more easily generated by AI.
2. Extreme weather events (AI might help to tackle climate change).
3. Societal polarization - clearly inflamed by social media's use of AI, by production of disinformation, deepfakes, etc.
4. Cybersecurity – AI adds enormous power to those who want to breach IT security.

To work out how to tackle AI risks, we first need to understand the interactions between the major actors in the AI landscape, as shown in Figure 1. First, the AI Suppliers (of AI apps, tools and components) are competing in an 'AI Arms Race', constrained only by money. Their 'open' products are available for 'innocent' AI Users[1] and for all other actors in the landscape including AI Aggressors who use these products and their own tools as weapons to attack other actors, and to AI Defenders who use many of the same products to help

[1] We avoid the term 'End Users'. The supply chain of AI products comprises a complex mesh of developers of AI components, tools and apps, any of whom can fulfil the roles of Aggressor, Defender or User. A User can include any person or hardware/software device with access to/from the Internet.

protect the innocents. (There are also, of course, 'closed' AI products e.g. embedded in products such as autonomous vehicles and robotic devices.)

This AI landscape is thus the setting for multiple guerilla wars between multiple actors of varying power. This landscape is unstable and evolves very rapidly.

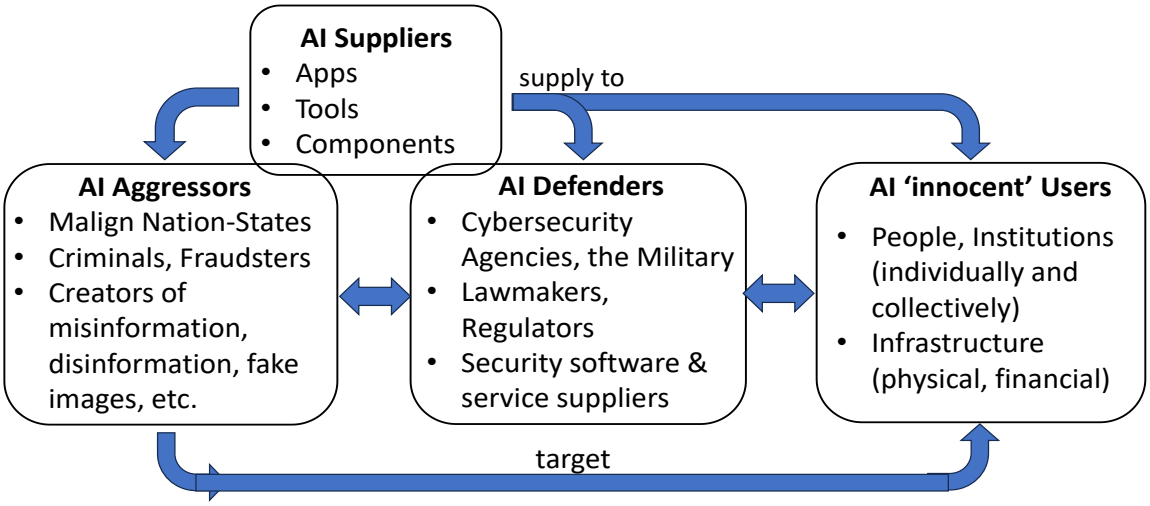

Figure 1: The players in the AI Landscape

Figure 1 shows Suppliers overlapping with both Aggressors and Defenders because although the AI Suppliers' products act as Defenders by providing some measures to protect their innocent Users from harm, these protective measures are clearly inadequate.

In particular, social media companies (a major sub-set of AI Suppliers) cite the need to balance investment in user-protection measures against the need to protect freedom of speech while, of course, investing as much as they can afford in the AI Arms Race.

2. The three classes of Users of AI products

Figure 2 shows three broad classes of Users of AI products, the different types of harm to which they may be exposed, and which require quite different types of countermeasures.

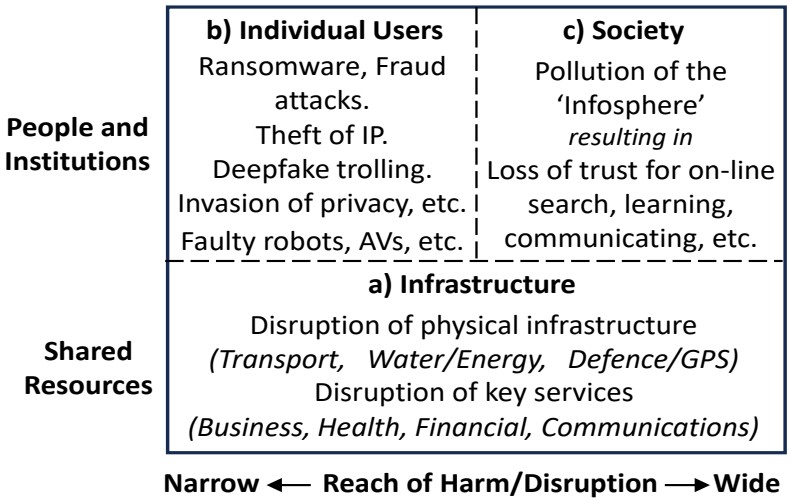

Figure 2. The three classes of AI Users and the types of harm they may suffer.

Note that the 'reach' of the three risk-types indicates only the *extent*, not the *severity* of a harm. Disruption to water supply for a small community could be annoying but temporary, while a prolonged breakdown of energy supply for a large section of a population could be catastrophic. The risk to individual persons and institutions (also including businesses) varies enormously; the harm to an individual may pass unnoticed by society but can be disastrous for those directly affected. But if a type of harm to individuals grows strongly, it can become a concern to society in general.

Some observers foresee the greatest risk of AI in a future in which AI 'takes control' of humanity leading to its extinction e.g. by unleashing a new lethal pandemic or finding a way round safeguards leading to all-out nuclear war. The steps that would be needed for this to happen, with failure of so many safeguards seems to this author so unlikely as to be implausible[2]. However, we cannot so easily dismiss scenarios whereby a deranged despot uses AI to develop a scheme to launch what appears to be winnable nuclear, chemical or biological aggression against his enemies. The Bulletin of the Atomic Scientists leaves its 'Doomsday Clock' for 2024 [2] as close to midnight as it has ever been, citing potential uses of AI as one of its reasons for this high risk.

The threats of harm to all classes of Users are already with us and will get worse when Artificial 'General' Intelligence (AGI), currently under development, becomes available. However, the prospect of existential threats to humanity arises, even if conceivable, only with the advent of so-called Artificial 'Super' Intelligence (ASI).

Already at the current stage of development AI suppliers do not fully understand how existing LLM algorithms work. LLMs have only limited true understanding of the data they process and hence of the output they produce. Further, the data used for their training is in most cases unknown to Users, and much will be unreliable. Generative AI systems are thus immature and inevitably are capable of making zillions of mistakes. (Artificial *'Intelligence'* is actually a misnomer; 'Advanced Statistical Systems' is a more accurate name.) Stopping these systems from making mistakes will generally be much harder than defect-elimination in conventional IT systems. Yet the consequences of such failures can sometimes be life-threatening, e.g. for the software controlling autonomous vehicles or used in medical diagnosis.

One capability of LLMs that is particularly concerning is their use to create or to modify computer program code from simple prompts. The danger exists that defective code is distributed via the internet and activated, causing harm before there is time to react and to back out the harmful code. This capability adds to the potential severity of risks, particularly for disruption of physical and services infrastructure.

If all these questions re lack of understanding of how LLM's work, how to test and control them etc., can be asked of LLM's, how much worse will these issues be for AGI and ASI?

[2] The importance of HITL (Human in the Loop), requiring that humans retain a controlling hand in critical AI decision making, is fortunately being recognised [3].

In spite of all these issues, the AI suppliers compete relentlessly to develop and release ever-more sophisticated AI [4]. No surprise therefore that last year, 30,000 AI experts signed a letter urging a six-months pause in AI development, of course to no avail [5].

We will now examine the nature of the risks to each class of User in more detail and the counter-measures needed to help bring more under control.

2.1. The risk of major infrastructure disruption.

We are all regularly reminded of the enormous cost to large sections of society when its physical infrastructure is disrupted by natural events such as earthquakes and extreme weather. Human-generated disruption can be even more costly, as when a financial meltdown is caused by error, ultra-risky trading, or failure of supervision.

The FBI has warned [6] of AI being a 'force multiplier' to enable aggressor nations to attack the physical and services infrastructures of nations they regard as antagonists. A despot might well exploit AI to help attack a perceived enemy's infrastructure, perhaps as a prelude to launching a physical war[3]. AI makes such attacks easier by its ability to explore weaknesses in network security far more efficiently than by conventional programming and by launching co-ordinated attacks on multiple targets.

We are witnessing in effect an AI Arms Race. Consequently, all advanced nations are now concerned about AI Safety. We can learn a lot on how to prevent such a race from leading to serious conflict by studying how nuclear war has been avoided (so far!). The challenge is to prevent 'AI MAD', where the 'D' of nuclear 'Mutually Assured Destruction' is replaced for AI by D for 'Disruption'.

Of course, nuclear MAD and AI MAD differ in some obvious ways, but one can define some of the conditions and measures needed to avoid AI MAD both from where the analogy with nuclear MAD works and where it does not work.

- As AI advances, it is important that the principal belligerents maintain something close to technical parity. Current national efforts to maintain superiority of the hardware and software to deliver advanced AI are unlikely to succeed long-term.
- In the case of nuclear MAD only a few nations have the capacity and the will to develop nuclear weapons. In contrast, AI is being developed by myriad independent, profit-motivated developers with limited moral interest in self-restraint, and by actors with actual malign intent. Criminal organizations, sometimes acting on behalf of malign states, have for years been exploiting security weaknesses in IT systems to launch cyber-attacks on businesses[4]. AI is a gift to the armoury of all Aggressors.

[3] The military use of AI-enabled weaponry to attack critical infrastructure is beyond the scope of this paper.
[4] In 2020, a Russian group of hackers known as Midnight Blizzard and linked to Russia's foreign intelligence services hacked into the US Treasury and Commerce Departments and the Pentagon, as well as several Fortune 500 companies. In 2023 they hacked into the e-mail accounts of senior Microsoft Executives [7]. Moody's, the credit rating agency, has warned that the increasing reliance of water and water-treatment companies on automation to monitor and control their networks is making them more vulnerable to attacks. [8]

- Given the huge variety of infrastructure assets that are at risk of being targeted, the challenge to ensure that each asset is fully protected is enormous. However, the measures are simple in principle - standard practices of system quality control and strong cyber-security – and are already all in the hands of those whose job it is to keep the infrastructure functioning smoothly. Disciplined chains of command are needed to ensure full implementation (as has been necessary to avoid nuclear MAD). There is a good case for putting implementation of strong standard cybersecurity measures on a statutory basis to protect any infrastructure assets whose disruption would cause severe harm to society.
- Current international efforts to promote AI Safety should lead to the establishment of a permanent authority with powers to draw up and to police international treaties to limit the threats of AI MAD (by analogy with the role of the International Atomic Energy Agency in helping prevent nuclear MAD). Achieving this outcome will require visionary international leadership, just as was originally needed to control the nuclear arms race.

An encouraging sign that the risks of AI MAD are being recognised is the recent meeting of Chinese and US AI experts which recognised 'red lines' on the development of AI, including around the making of bioweapons and launching cyber-attacks [9]. (Informal meetings of nuclear scientists played a vital role in maintaining nuclear peace [10].) And to be fair to AI Suppliers, some of the larger ones have also agreed to strengthen their cybersecurity defences, including using AI for that purpose [11].

2.2. The risks of harms to individuals.

On-line IT is already causing harm to individuals in endless ways e.g. by enabling children and adolescents to access unsuitable material, and by spreading disinformation. And mitigating these harms is already being addressed by some major IT service suppliers (albeit only partially effectively), and by hastily drawn-up but not yet effective national and international regulations. Adding more powerful AI to the mix puts the risk of these harms on steroids and is already resulting in new types of harm.

There is now strong evidence that, starting in the early 2010s, the availability of social media and smartphones have led to significant rises in mental health problems amongst teenagers. Western society is beginning to wake up to the scale of the problem[5]: some European countries have banned use of smartphones in schools; Florida has recently banned under-14s from having social media accounts; China is way ahead here, requiring smartphone manufacturers to strictly limit their daily use by under-18s.

Until now, social media companies have been the principal vectors for the spread of IT harm to individuals; they will continue in this role both as vectors of AI-enhanced harms, constrained only by the laws of the countries in which they deliver their services. An early test of the effectiveness of such laws will arise when, for example, the new UK On-Line

[5] The '*damage done to young people'* has been described '*uncomfortably for those of us who regularly defend free enterprise … as … the result of a vast experiment with the brains of young people by corporations that worked out how those brains could be exploited for profit'* [12]

Safety Act is used to pursue social media companies for failing to prevent the distribution of harmful material to children.

The EU and the UK are taking the lead in holding social media companies and other tech giants to account for the harms they cause. In the USA, social media companies benefit from early 20th century legislation which does not hold a telecoms carrier liable for the traffic it carries. It is long overdue that the protection afforded by this relic from when person-to-person voice traffic dominated is removed.

Deciding if and how to protect individuals against AI-enabled harms is inherently complex, for example:

- Where and how to draw the line between protecting people from harm, e.g. from the spread of false information, and allowing freedom of expression?[6]
- While all agree on the importance of protecting children from harm, how best to proceed? Ban children from having social media accounts, ban use of smartphones in schools, limit daily screen-time, until what age-limits, etc., and how to enforce these rules?
- Concerns about privacy abound[7]. (Use of encryption in direct messaging is a double-edged sword, protecting privacy and equally protecting aggressors from being identified.)

These issues can only be decided by elected Governments when drawing up regulations. AI safety is too important to be left to profit-motivated private companies or to industry self-regulation. However, regulations will vary significantly with national culture. For example, under proposed legislation, EU citizens will be protected against misuse of AI-assisted facial recognition technology, whereas China uses this technology to control sections of its population. Therefore, national laws to protect society and individuals from on-line harm will prevail initially[8]; international harmonization must follow wherever possible.

2.3. The risks of harms to Society

Social media are already causing distress in society by using AI to amplify, target and spread unsuitable material for children, hate messages, conspiracy theories, misinformation and disinformation. The effects will worsen when much of the output that we read, hear, or view on-line is generated, mediated or spread by unreliable Generative AI adding its mis/disinformation, deepfakes and hallucinations to the mix which is then scooped up in the next 'learning' cycle.

[6] The Scottish Government has introduced a law making it a crime to communicate material or behave in a manner that amounts to 'stirring up hatred' against certain 'protected characteristics', e.g. religion or sexual orientation. The law has been strongly opposed by free-speech advocates.

[7] Snap, the owner of Snapchat, is being investigated by the UK's data privacy watchdog for 'failing to adequately identify and assess the privacy risks to children and other users before launching its (Open AI-powered) chatbot 'My AI'.

[8] Prometheus Endeavor produced 'Twelve Principles for regulating Artificial Intelligence' in response to calls for submissions by both the US and UK Governments [13].

One of the greatest long-term risks to society must be that the on-line 'infosphere' becomes so polluted that we cannot trust any information communicated on-line. The infosphere is then effectively 'dead', like a heavily polluted ocean.

By its nature, this effect on society appears relatively slowly and insidiously over time and is difficult to combat and, if not prevented, probably irreversible.

Even before the latest AI, information technology had already aided the degradation of the infosphere. The internet, and especially social media, have become a playground for malign actors and criminals. For academics, the volume of papers output by so-called 'paper mills' containing fake or poor data is already creating difficulties for reputable scientific journals to distinguish them from papers based on genuine, valid research. LLMs will aid the generation of fake output, either by deliberate fraudsters or by their own 'hallucinations'. A recent study [14] of the responses of state-of-the-art LLMs to specific legal queries found hallucination rates ranging '*from 69% to 88%. ….. Moreover, these models often lack self-awareness about their errors and tend to reinforce incorrect legal assumptions and beliefs'*. LLMs are also known to show bias when used for e.g. hiring decisions, and the accuracy of AI-enabled facial recognition systems varies with ethnicity.

A similar problem arises with the use of chatbots in Customer Service Centres. Chatbots are good at answering routine questions, but they have limits, and if their human users do not understand those limits, disaster may ensue. (Another downside of over-reliance on simple chatbots is the risk over time of losing human experience in handling uncommon issues.) At present, therefore, LLMs are not reliable enough for many professional uses.

The release of immature AI products before the development of antidotes to counteract their harmful effects has been irresponsible. Pollution of the infosphere must not be allowed to continue. Various measures are therefore urgently needed to limit the damage, which will mean placing constraints on AI products. Proposed national regulations of AI products do not seem yet to have recognised this risk. The constraints include:

- Suppliers of AI products, at every point in the AI-supply mesh, must be held accountable for the safety of their product in the same way as any other consumer product, such as food, medicines, cars, etc.
- AI-enabled products must make clear that their output is AI-generated and inform the user of any safety and other limitations on this use[9, 10].
- Tools are needed for users to detect whether an artefact is in any way AI generated.
- The public is generally not yet aware of the risks of relying on AI output. Regulators must do more to raise public awareness of the risks.

[9] The development of tools for 'watermarking' or for generating 'content credentials' [15], or for deepfake detection still has a long way to go. In 2020, Meta held a 'Deepfake Detection Challenge' inviting developers of deepfake video detection tools to examine a dataset of 100,000 videos [16]. The success rate was 65%.

[10] OpenAI warns that ChatGPT "can make mistakes". Anthropic, says that its LLM Claude "may display incorrect or harmful information"; Google's Gemini warns users to "double-check its responses". Yet users are tempted to try these new products and may not be knowledgeable or experienced enough to recognise dodgy output.

Some good news is that 20 of the largest AI suppliers have agreed to collaborate to combat the creation and spread of AI-generated material such as deepfake videos that could mislead voters and thus affect election results [16]. Whilst this is an important step, given that so much of the world's population is having national elections in 2024, **it is sobering to reflect that this harmful use of deepfakes is just one tiny fraction of their possible misuses**.

As regards the longer-term risks, Kissinger et al have written *'As AI's role in defining and shaping the 'information space' grows, its role becomes more difficult to anticipate*. *…. As a result, the* prospects *for free society, even free will may be altered.*' [17]

3. Conclusions.

We conclude that AI-enabled products, whilst promising significant benefits, pose serious risks to very many aspects of our way of life.

Measures to mitigate the known current risks are urgently needed and are of two sorts. Strong cybersecurity measures are needed to prevent disruption to the physical and services infrastructures on which we all depend. Harm to individuals and to society in general must be curbed by strictly enforced regulations, starting with national efforts. However, we accept that defining and implementing such regulations is inherently difficult in many cases, and regulations may require us to accept limits on cherished freedoms. Eventually, the aim must be to obtain international agreements on controlling AI.

The rush to develop even more advanced forms of AI runs the risk of making the current harms worse. (Has there even been any other class of legal consumer products in history that it is strongly suspected will cause a variety of harms, but that cannot be properly tested for safety except by letting consumers use them as the only way to find out what harms they actually cause?) AI suppliers must be held accountable for the harms their products cause, just as product safety is regulated in so many other markets. AI Safety is too important to be left to free markets or to self-regulation.

It is in the long-term interests of AI Suppliers to accept safety constraints on their products. If regulators find them unsafe for Users, especially the young, and/or no-one trusts their output, the products and eventually their suppliers will not survive anyway.

Charles Symons

Prometheus Endeavor

Reigate, England, May 2024

Acknowledgements

I am very grateful for the many helpful conversations with colleagues from Prometheus Endeavor in developing the ideas in this paper.

References

[1] World Economic Forum 'Global Risk Report 2024'.

[2] Bulletin of the Atomic Scientists (thebulletin.org). 24th January 2024.

[3] Human-in-the-loop - Wikipedia

[4] Mark Zuckerberg: the next generation of tech services 'requires building full general intelligence', Facebook, January 2024.

[5] A letter published in March 2023 (by Max Tegmark, a co-founder of the Future of Life Institute) warned of an 'out of control race' to develop minds that no-one could 'understand, predict, or reliably control'. It urged leading AI companies to agree a moratorium on developing systems more powerful than GPT-4.

[6] Munich Security Conference, February 2024, Munich Security Conference

[7] Microsoft says Russian group infiltrated some employees' email accounts, Financial Times, 20th January 2024'

[8] Moody's warns over growing risks of cyber attacks for water and wastewater companies (waterbriefing.org), 20th January 2024.

[9] US companies and Chinese experts engaged in secret diplomacy on AI safety (ft.com), Financial Times, 18th March 2024.

[10] www.pugwash.org/about-pugwash

[11] AI can strengthen cyber defences, not just break them down, Sundar Pichai, www.ft.com, 16th February 2024.

[12] William Hague, The Times, 2nd April 2024.

[13] Twelve Principles for AI Regulation - Prometheus Endeavor

[14] Matthew Dahl, Varun Magesh, Mirac Suzgun, Daniel E Ho, [2401.01301] Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models (arxiv.org), (pre-print), Jan 2nd 2024.

[15] 'This Election Year, look for Content Credentials', IEEE Spectrum, January 2024, www.ieee.com.

[16] Deepfake Detection Challenge Dataset (meta.com), June 25th 2020.

[17] 'The Age of AI and our human future', Henry Kissinger, Eric Schmidt, Daniel Huttenlocher, 2022.