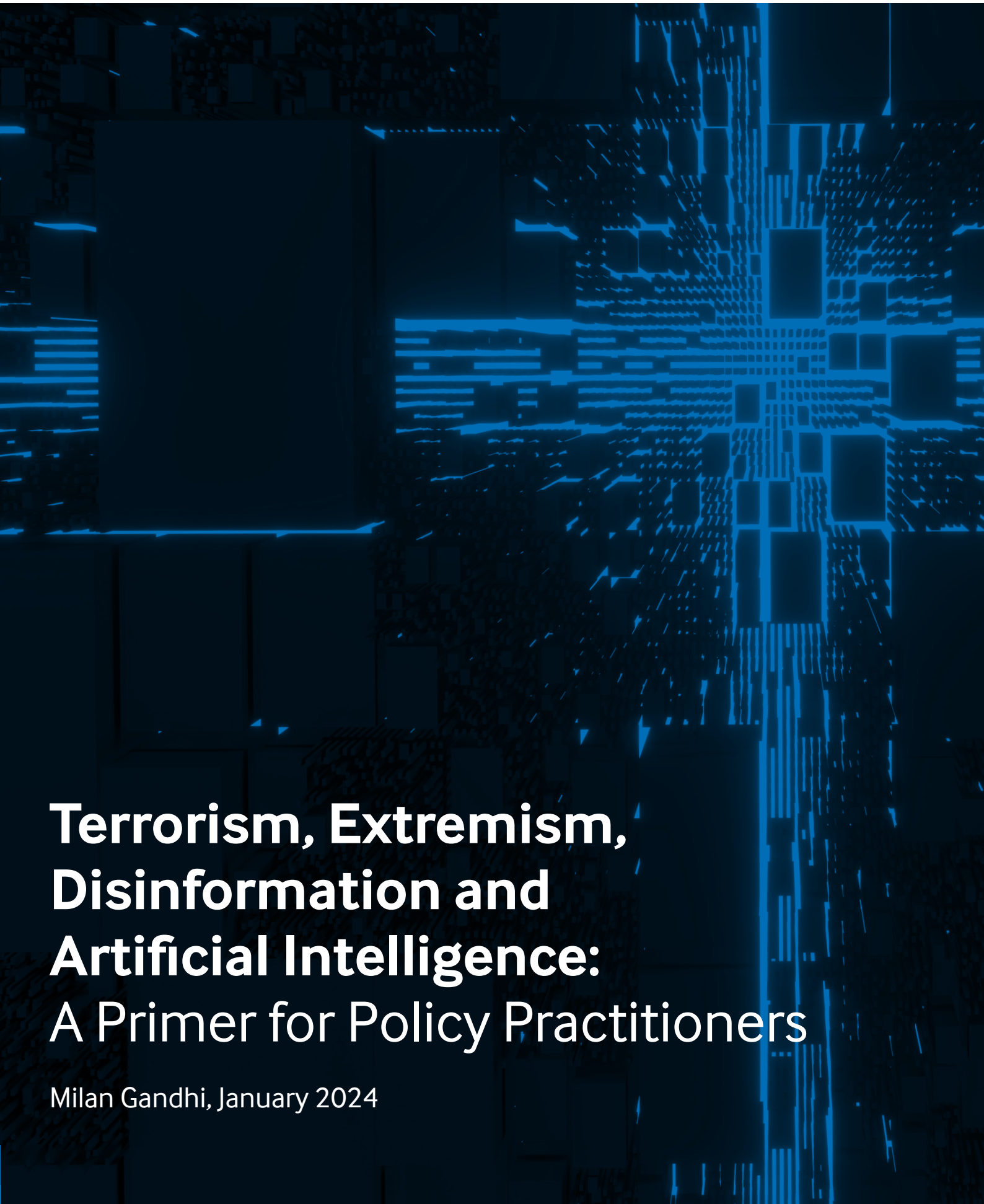
ISD | Powering solutions to extremism, hate and disinformation



Terrorism, Extremism, Disinformation and Artificial Intelligence:
A Primer for Policy Practitioners

Milan Gandhi, January 2024

About the Digital Policy Lab

The Digital Policy Lab (**DPL**) is an inter-governmental working group comprising of senior representatives of ministries and regulators from liberal democratic countries. The DPL charts a policy path for preventing and countering the spread of disinformation, hate, and extremist and terrorist content online. The DPL facilitates inter-governmental exchange, providing policymakers and regulators with access to sector-leading expertise and research. It represents an international community of practice for key digital policy challenges.

About this Paper

This paper (**Paper**) is a non-exhaustive introduction to core concepts, public policy challenges and solutions at the intersection of artificial intelligence and focus areas covered by the DPL. It is based on desktop research and informal interviews with colleagues at the Institute for Strategic Dialogue (**ISD**) and the University of Oxford's Blavatnik School of Government (**BSG**). It is not the direct product of any DPL working group meeting(s). The Paper aims, however, to inform policy discussions and thinking among policymakers participating in the DPL and beyond.

About the Author

Milan Gandhi is a Research Fellow at ISD and supports Dr Aaron Maniam to convene the technology policy cluster at BSG. He is currently completing the MSc in Public Policy Research at BSG while conducting policy-relevant research and analysis on issues connecting digital technologies, democracy and geopolitics. Milan holds a Master of Public Policy with Distinction from the University of Oxford and a Bachelor of Laws (1st Class) from the University of Queensland. He is supported by a 2022 John Monash Scholarship and a 2023 BSG Scholarship.

Acknowledgements

Any errors are the author's alone. Henry Tuck, Helena Schwertheim, Terra Rolfe and Ellen Jacobs from ISD provided invaluable project design and editorial assistance. Vincent Zhang and Paola Galvez Callirgos (BSG) offered insightful comments as did others on ISD's DPL team, including Sara Bundtzen and Mauritius Dorn. The following ISD colleagues were generous in contributing insights and expert perspectives by way of informal interviews: Zahed Amanullah, Dr Julia Ebner, Dr Francesca Arcostanzo, Milo Comerford, Dr Tim Squirrell, Jennie King and Melanie Smith.

ISD | Powering solutions to extremism, hate and disinformation

Copyright © Institute for Strategic Dialogue (2024). Institute for Strategic Dialogue (ISD) is a company limited by guarantee, registered office address PO Box 75769, London, SW1P 9ER. ISD is registered in England with company registration number 06581421 and registered charity number 1141069. All Rights Reserved.

www.isdglobal.org

Contents

Glossary

The author acknowledges that terminology is socially constructed and contested. Settling definitional uncertainty in the public policy discourse on artificial intelligence is beyond the scope of this Paper. Nevertheless, definitions are adopted to animate ideas and discussion. Sources are provided by way of endnotes.

Artificial Intelligence (AI) is defined under the subheading titled 'The Concept of AI' within Section 1.

Artificial General Intelligence (AGI) is defined under the subheading titled 'The Concept of AI' within Section 1.

Artificial Narrow Intelligence (ANI) is defined under the subheading titled 'The Concept of AI' within Section 1.

Artificial Neural Networks (ANNs) are defined under the subheading titled 'Specific AI Subfields and Technologies' within Section 1.

Computer Vision is defined under the subheading titled 'Specific AI Subfields and Technologies' within Section 1.

Deep Learning is defined under the subheading titled 'Specific AI Subfields and Technologies' within Section 1.

Disinformation is false, misleading or manipulated content intended to deceive or harm.

Extremism is the advocacy of political and social change in line with a system of belief that claims the superiority and dominance of one identity-based 'in-group' over 'out-groups'. Extremism is rooted in a dehumanising supremacist mind-set which is fundamentally incompatible with pluralism and universal human rights, and can be advanced through violent or non-violent means.

Foundation Model is a term that overlaps with the term "general-purpose AI". This Paper adopts the definition proposed by the Ada Lovelace Institute: "[f]oundation models are AI models designed to produce a wide and general variety of outputs. They are capable of a range of possible tasks and applications, such as text, image or audio generation. They can be standalone systems or can be used as a 'base' for many other applications... Some foundation models are capable of taking inputs in a single 'modality' – such as text – while others are 'multimodal' and are capable taking multiple modalities of input at once, for example, text, image, video, etc., and then generating multiple types of output, (such as generating images, summarising text, answering questions) based on those inputs."[1]

Frontier AI Model. In their 2023 paper, *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, Anderljung et al. define frontier AI models as "highly capable foundation models that could exhibit dangerous capabilities. Such harms could take the form of significant physical harm or the disruption of key societal functions on a global scale, resulting from intentional misuse or accident."[2]

Generative AI refers to "to deep-learning models that can take raw data — say, all of Wikipedia or the collected works of Rembrandt — and "learn" to generate statistically probable outputs when prompted. At a high level, generative models encode a simplified representation of their training data and draw from it to create a new work that's similar, but not identical, to the original data. Generative models have been used for years in statistics to analyze numerical data. The rise of deep learning, however, made it possible to extend them to images, speech, and other complex data types."[3]

Large Language Models are defined under the subheading titled 'Specific AI Subfields and Technologies' within Section 1.

Machine Learning is defined under the subheading titled 'Specific AI Subfields and Technologies' within Section 1.

Misinformation is false, misleading or manipulated content shared irrespective of an intent to deceive or harm.

Natural Language Processing is defined under the subheading titled 'Specific AI Subfields and Technologies' within Section 1.

Robotics is defined under the subheading titled 'Specific AI Subfields and Technologies' within Section 1.

Terrorism definitions vary across different national jurisdictions, and there is no universally agreed definition of terrorism. For the purposes of this report, we have chosen to use the shortened version of Schmid's 2011 academic consensus definition, where terrorism is defined as:

> "1. Terrorism refers, on the one hand, to a doctrine about the presumed effectiveness of a special form or tactic of fear-generating, coercive political violence and, on the other hand, to a conspiratorial practice of calculated, demonstrative, direct violent action without legal or moral restraints, targeting mainly civilians and non-combatants, performed for its propagandistic and psychological effects on various audiences and conflict parties; 2. Terrorism as a tactic is employed in three main contexts: (i) illegal state repression; (ii) propagandistic agitation by non-state actors in times of peace or outside zones of conflict; and (iii) as an illicit tactic of irregular warfare employed by state- and non-state actors."[4]

Transformer Model is defined under the subheading titled 'Specific AI Subfields and Technologies' within Section 1.

Executive Summary

Focussing on current and emerging issues, this policy briefing paper (**Paper**) surveys the ways in which technologies under the umbrella of artificial intelligence (**AI**) may interact with democracy and, specifically, extremism, mis/disinformation, and illegal and 'legal but harmful' content online. The Paper considers examples of how AI technologies can be used to mislead and harm citizens and how AI technologies can be used to detect and counter the same or associated harms, exploring risks to democracy and human rights emerging across the spectrum.

The Paper begins by providing a brief primer on AI in Section 1 and outlining general concerns relating to accountability — the "cornerstone" of AI governance[5] — data collection and quality and the opacity of AI models. Special consideration is given to generative AI systems, such as chat bots powered by large language models (**LLMs**), due to their recent popularisation and wide-ranging capabilities.

Section 2 of the Paper categorises AI systems into to those that:

- **generate content**, examining LLMs such as ChatGPT, deepfakes, synthetic media and how LLMs might be leveraged in counter-speech interventions;
- **disseminate and target content**, examining political micro-targeting and AI-powered social bot nets;
- **select and amplify content within online information environments**, examining algorithmic recommender systems;
- **assist in the mitigation of online harms**, examining AI systems that detect harmful content and identify deepfakes; and
- **present a risk to public safety**, examining three novel problems that specifically arise with respect to the unpredictability, deployment safety and proliferation of foundation models.

Responding to risks outlined in Section 2, Section 3 then examines potential mitigations, focusing on ethical principles, public policy and emerging AI regulation.

Given the immense scope and potential impacts of AI on different facets of democracy and human rights, the Paper does not consider every relevant or potential AI use case, nor the long-term horizon. For example, AI-powered kinetic weapons and cyber-attacks are not discussed. Moreover, the Paper is limited in examining questions at the intersection of AI and economics and AI and geopolitics, though both intersections have important implications for democracy in the digital age. Finally, the Paper only briefly discusses how AI and outputs such as deepfakes may exacerbate broader societal concerns relating to political trust and polarisation.

Although there is a likelihood that aspects of the Paper will be out-of-date the moment it is published given the speed at which new issues, rules and innovations are emerging, the Paper is intended to empower policymakers, especially those working on mis/disinformation, hate, extremism and terrorism specifically, as well as security, democracy and human rights more broadly. It provides explanations of core concerns related to AI and links them to practical examples and possible public policy solutions.

Section 1: Conceptual Foundations

Artificial Intelligence: A Primer

The Concept of AI

Understood as an applied discipline of science and engineering, AI is concerned with "building intelligent entities".[6] An AI system is therefore a system, such as a computer program, that has been designed to carry out a task that was perceived to require intelligence. This begs the question: what counts as intelligent? Professor Joanna Bryson, writing in the *Oxford Handbook of Ethics of AI*, cites the following definition: "[i]ntelligence is the capacity to do the right thing at the right time. It is the ability to respond to the opportunities and challenges presented by a context."[7]

In policy practice, AI is used as an umbrella term for diverse technologies, emerging and evolving on a seemingly daily basis, which "affect profoundly virtually the entire landscape of human activity".[8] The discipline of AI encompasses subfields, "ranging from the general (learning, reasoning, perception and so on) to the specific [or narrow], such as playing chess, proving mathematical theorems, writing poetry, driving a car, or diagnosing a disease."[9] It is analytically useful to distinguish between different sub-types of AI:

- **Artificial Narrow Intelligence (ANI)**. Narrow forms of AI are those systems that perform a singular or narrow set of tasks. Examples include an autonomous vehicle, the recommender system that Netflix utilises, or generative AI tools used to create deepfake images. This Paper focuses on examples of ANI.

- **Artificial General Intelligence (AGI).** Many would argue that other than in science fiction there are no real-world examples of Artificial *General* Intelligence. While WIRED reports that "no concrete definition of the term exists", the intuition is that AGI refers to a system that can perform such a wide array of complex intellectual tasks that it could be said to be generally intelligent.[10] Although human and animal minds are the available benchmark, a hypothetical AGI might be more capable than both in certain arenas. There is disagreement among experts as to whether achieving AGI is possible.[11]

In March 2023, authors from OpenAI and Microsoft claimed that GPT-4, OpenAI's LLM, performs in a manner that is "strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT."[12] They report that GPT-4 can solve a very wide array of "novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology, and more, without needing any special prompting."[13] Their paper is provocatively titled "Sparks of Artificial General Intelligence".[14]

While this Paper explores the implications of LLMs, it does not make a claim about whether such systems exhibit general intelligence. It may be useful, however, to think about LLMs as exhibiting a greater level of general functionality than ANI systems, sitting somewhere between ANI and AGI. The fact that more generally capable systems represent the foundation underpinning a range of narrower consumer-facing systems means that their responsible development is deemed by certain policymakers to be particularly important. This, rather than debates as to whether AGI will be achieved, may be the more important insight for navigating current and emerging policy questions. See, for example, discussion of 'Foundation Models' under 'Specific AI Subfields and Technologies' below.

Policy practitioners should be clear about the kinds of AI systems they are referring to, their design features and the capabilities they exhibit, and their context-specific applications. This will assist in cutting through the definitional morass and identifying more precisely what challenges and opportunities are posed by the AI system(s) in question. As Bryson points out, how we define AI, and specifically intelligence, has practical implications for the laws that purport to regulate it, "[in] order to evade regulation or responsibility, the definition of intelligence is often complicated in manifestos by notions such as sentience, consciousness, intentionality, and so forth... what is essential when considering AI in the context of law is the understanding that no fact of either biology (the study of life) or computer science (the study of what is computable) names a necessary point at which human responsibility should end."[15]

Operationalising the Concept of AI in Public Policy

Lawmakers concerned with defining safety and other standards for AI may have in mind salient and recent examples of systems with the potential to influence society and impact the lives of their citizens or constituents.

Consider, for example, the draft *Proposal for a Regulation laying down harmonised rules for artificial intelligence* (commonly referred to as the "**EU AI Act**"). The EU AI Act adopts a "risk-based" approach, "whereby AI systems are regulated based on the level of risk they pose to the health, safety and fundamental rights of a person". Although the EU AI Act is yet to be finalised and therefore subject to change, the European Parliament proposed that, within the EU AI Act, "AI system" should mean: "...a machine-based system that is designed to operate with *varying levels of autonomy* and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions that influence physical or virtual environments" (emphasis added).

This definition is said to align with the definition of AI proposed by the Organization for Economic Cooperation and Development (**OECD**). The OECD definition was itself revised in November 2023 in pursuit of "international alignment of AI definitions".[16] According to the OECD, an "AI system" is "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."[17]

By referring to outputs such as "predictions", "content", "recommendations", and "decisions", the OECD definition calls to mind — and seeks to encompass — certain technologies and computer science subfields that are referred to throughout this Paper. As this Paper explores, many examples of AI systems falling within the OECD definition continue to shape our day-to-day lives.

Specific AI Subfields and Technologies

A non-exhaustive selection of salient subfields and technologies within the field of AI are summarised below in alphabetical order.

Computer Vision

Computer vision is a subfield of AI concerned with enabling computers to interpret and make decisions based on visual data. Applications include but are not limited to autonomous vehicles, facial recognition systems, augmented reality systems and the latest Roomba.

Machine Learning

Machine learning is a subfield of AI concerned with systems that automatically learn and improve from experience.

For example, recommender systems utilised by digital platforms such as Facebook, YouTube, Netflix or Amazon analyse users' previous activity and preferences to recommend online content, movies, products and advertising etc.

Deep Deep Learning and (Artificial) Neural Networks

Deep learning is a subset of machine learning. It employs artificial neural networks (**ANNs**), a methodology inspired by the functioning of a human or animal brain.

ANNs are computational models comprised of node layers, "containing an input layer, one or more hidden layers, and an output layer". They are particularly useful for clustering and classifying information. If a neural network has three or more layers of nodes through which data must pass, it is a deep-learning neural network – the greater number of layers make it literally deeper.

In general, although not always true, the more node layers, the more capable the neural network at handling very large and complicated datasets and discovering patterns within unlabelled and unstructured data.

As IBM explains, "[n]eural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the most well-known neural networks is Google's search algorithm."

A specific kind of ANN, a Transformer Model, is utilised in LLMs (discussed below).

Foundation Models

Foundation models are an emerging category within the field of AI, also sometimes referred to as ‚General-Purpose AI' or ‚GPAI'.

Foundational models refer to systems engineered to perform a broad spectrum of tasks (hence their description as 'general purpose'), encompassing, for example, text synthesis, image manipulation and audio generation. Examples include LLMs such as OpenAI's GPT-3 and GPT-4, which serve as the underlying architecture for ChatGPT.

The modular design of foundation models allows for additional layers of development, enabling the creation of diverse downstream applications. The extensive and sometimes unpredictable capabilities of foundation models differentiate them from narrow AI systems, which are confined to specific or limited tasks.

Large Language Models

Large Language Models (**LLMs**) are statistical models that generate "plausible next words" to a user's prompt. LLMs employ deep learning (discussed above) and are trained on vast datasets, enabling them to produce coherent and contextually relevant responses. As they excel at language-related tasks, they are an applied example of the natural language processing (NLP) subfield (discussed below).

LLMs can produce original text, hold humanlike conversations in multiple languages, pass tertiary exams and generate computer code. Popular examples include GPT-3 and GPT-4, LLaMA 2, and Google Bard.

Natural Language Processing

Natural language processing (**NLP**) is a subfield of AI concerned with enabling computers to understand, interpret and generate human language in meaningful ways. Chatbots, Google Translate and Apple's Siri are all examples of NLP applications. NLP is also utilised for content moderation and detection of unlawful or awful speech including hate speech. For example, the Tensorflow.js library for machine learning in the JavaScript language contains pre-trained models which can assist in automating online content moderation.

Robotics

Robotics is a multifaceted field of science and engineering concerned with designing machines that can carry out a series of tasks automatically. Although distinct from AI, robotics and AI are intertwined inasmuch as AI systems and techniques, as subfields such as machine learning, computer vision and natural language processing, are implemented in certain robots to enhance their "intelligence".

For example, the latest Roomba robot vacuum cleaner is a robot that leverages AI techniques including machine vision.

Tramsformer Models

A transformer model is a type of ANN (discussed above) that comprehends context and thereby grasps significance by observing associations in sequential information, such as the words in a text.

Utilising a dynamic set of mathematical strategies, known as attention mechanisms, it discerns the ways in which even separate elements within a data series impact and relate to one another. First introduced by Google in a 2017 paper, transformer models represent one of the most recent and potent models developed thus far, propelling a surge of breakthroughs in machine learning.

Democracy and Artificial Intelligence: A Bird's Eye View

There are numerous ways in which different AI technologies challenge aspects of liberal democratic systems and values, just as there are numerous ways in which AI technologies may assist in strengthening them. To begin to map these touch points, it is helpful to outline the elements of liberal democracies that are the subject of concern. Side-stepping scholarly debates in political science as to the essential characteristics of democratic regimes, one useful starting point is the methodology behind the Economist Intelligence Unit's (**EIU**) annual Democracy Index. Indicators are grouped into five categories or pillars of liberal democracy: (1) functioning of government; (2) electoral process and pluralism; (3) civil liberties; (4) political participation; and (5) democratic political culture.[18] A non-exhaustive description of each pillar is provided below, drawing from EIU's methodology as outlined in its 2022 report[19]:

- **Functioning of government** refers to criteria such as the willingness and ability of the civil service to implement policy (sometimes referred to as state capacity), the pervasiveness of corruption, and public confidence in the government and political parties.
- **The electoral process and pluralism** refer to suffrage, the fairness of elections, and whether citizens can exercise their vote freely.
- **Civil liberties** refer to whether there is freedom of the media, freedom of expression and protest, freedom to access the Internet without political restrictions, judicial independence from government interference, and whether basic human rights are well-respected.
- **Political participation** refers to voter participation and turnout, whether ethnic and other minorities have a voice in the political process, and adult literacy.
- **Democratic political culture** refers to a minimum degree of social cohesion and consensus (contrasted with, for example, polarisation), separation of Church and State and public support for democracy.

To illustrate the challenges that policymakers face when attempting to assess how AI technologies may impact these criteria in practice, the following subsections provide examples of the use of AI technologies in two contexts: counterterrorism and the online information environment.

AI x Counterterrorism

The proper functioning of a government entails its capacity to deliver on the security of the nation state and the safety of its people. The United Kingdom's (**UK**) Strategy for Countering Terrorism (2023) recognises that AI has implications for counterterrorism and "could radically speed up the process of threat detection".[19] Between 2017 and 2020, the European Union invested over €5 million into the development of the Real-time Early Detection and Alert System for Online Terrorist Content (**RED-Alert System**), which harnesses AI technologies to "collect, process, visualize and store online data related to terrorist groups, allowing [law enforcement agencies] to take coordinated action in real-time".[20]

However, bad actors can also exploit AI technologies. In a briefing published in November 2023, Tech Against Terrorism report that they have archived more than 5,000 pieces of AI-generated content produced by terrorist and violent extremist actors.[21]

Further, just as there are concerns about terrorists leveraging AI capabilities, the use by security and law enforcement agencies of biased AI models that may discriminate based on factors such as ethnicity has been flagged by experts as an acute risk to civil liberties, another pillar of liberal democracy. Members of the European Parliament have proposed that the EU AI Act prohibit mass facial recognition programs in public places and predictive policing algorithms.[22] In February 2023, US citizen Alonzo Sawyer was wrongly identified as a criminal suspect by facial recognition software, leading to his arrest.[23] Recently, Porcha Woodruff, a Black woman, was wrongly arrested while eight months pregnant after a false facial recognition match.[24] "Facial recognition systems have the poorest accuracy rates when it comes to identifying people who are Black, female and between the ages of 18 to 30, while false positives "exist broadly", according to a study by the National Institute of Standards and Technology".[25] A joint report by the UN Office of Counter-Terrorism and the UN Interregional Crime and Justice Research Institute recognises "the potential for machine learning algorithms using biased data to compound bias through automated processes and therefore produce discriminatory outputs".[26]

AI x Online Information Environment

It is important to the health of liberal democracy that citizens have access to diverse, but ideally accurate, sources of information about politics and policy issues. Citizens should know about the decisions their governments and politicians make in order to hold them accountable. Moreover, the quality of information that is widely shared, whether through traditional media sources, across social media platforms or elsewhere on the internet, may shape citizens' preferences at the ballot box. Consequently, disinformation has been characterised by the United Nations Secretary General as a threat to free and fair elections.[27] UNESCO member states have declared that information and the quality of collective public debate are public goods, expressing concern "at the increasing proliferation, amplification and promotion, through human and automated systems, of potentially harmful content".[28] The same declaration expresses concern at laws that would unduly restrict freedom of expression under the guise of opposing false information.[29]

The implications of AI technologies for information, the quality of public debate, and free and fair elections are a concern for policymakers across the globe. Scholars hypothesise that algorithmic recommender systems, which structure how information is seen and shared by users of social media, may preference "misleading, hateful, conspiratorial, or extremist" content (see subsection 'Recommender Systems and Data Access' for more a more detailed examination).[30] Consequently, a goal of the EU Digital Services Act is to address "manipulative" algorithmic systems that "amplify the spread of disinformation".[31] Members of the European Parliament have proposed that the EU AI Act designate recommender systems as "high risk",[32] requiring that they be registered on a publicly-accessible database and meet stringent requirements under a "conformity assessment" process.[33] An emerging complication, discussed under 'Content Generation' below, is that LLMs, such as ChatGPT, may be exploited to increase the scale and effectiveness of disinformation campaigns and lower the barriers to entry for actors who would perpetrate them.[34] Other scholars suggest that such fears are overblown.[35]

On the other hand, there are also opportunities for AI to enhance the quality and safety of the online information environment. For example, ISD and CASM Technology leverage AI systems in digital research projects, identifying hate speech and discovering links between different forms of extremism through automated analyses of social media data. Meta utilised a multilingual AI system called "Few-Shot Learner" to identify Facebook and Instagram content that shared "misleading or sensationalized information discouraging COVID-19 vaccinations".[36] This is one of many examples of social media platforms utilising such technology to assist in content moderation. Further, researchers are at the early stages of considering the ethics, implications and feasibility of harnessing LLMs in counter-speech efforts to neutralise the harmful effects of online hate speech and de-bunk and pre-bunk misleading information.[37]

The Balancing Act

In the domains of counterterrorism and policing, using AI may be a boon to security or intelligence agencies concerned with safeguarding citizens, just as it may erode the rights of citizens. Similarly, different technologies that fall under the umbrella of AI, including LLMs, can negatively affect the health of the online information environment. However, they might also be deployed to safeguard the same communal good and push back against disinformation, extremism, hate speech, illegal content and "legal but harmful" content, with positive ramifications for democracy and public safety. These interactions between AI and democracy highlight a familiar complication for policymakers, namely, that there are trade-offs at each stage of adopting and regulating AI technologies. Some mirror longstanding contestations between policy goals such as security and minority rights; as well as between individual rights, such as freedom of expression, and communal goods, such as the health of citizens' information environment and public discourse.

The breadth of possible AI use cases means that there is sense in revisiting the ethical principles that should govern the development and operation of AI systems, irrespective of their application (see Section 3). In essence, the difficulty is to establish individual (or corporate) responsibility for harmful outputs and shortcomings of

AI systems as such systems are opaque, unpredictable, and involve a multitude of actors and resources.[38] Accountability, therefore, is the "cornerstone" of AI governance.[39] By holding those who develop, deploy and use AI systems accountable, policymakers can incentivise more transparent and trustworthy AI systems that advance, rather than deteriorate, democratic and other important societal values.

A related and significant challenge lies in the sourcing and quality of data used to train AI systems. Concerns relate to privacy, intellectual property and the potential for systems trained on incomplete or biased data to produce biased results. As Manyika, Silberg and Presten explain, "AI systems learn to make decisions based on training data, which can include biased human decisions or reflect historical or social inequities, even if sensitive variables such as gender, race, or sexual orientation are removed."[40]

Risks to public safety may stem from highly capable foundation models, including LLMs, which evolve unpredictably. Therefore, systems trained on vast quantities of data with general-purpose functionality, such as LLMs, require special consideration. A computer scientist specialising in AI safety recently cautioned, "[t]here are no reliable techniques for steering the behaviour of LLMs".[41] Experts warned in July 2023 that certain advanced AI models pose distinct regulatory challenges because "dangerous capabilities can arise unexpectedly; it is difficult to robustly prevent a deployed model from being misused; and, it is difficult to stop a model's capabilities from proliferating broadly."[42] Authors from across academia and the technology sector speculate that such models could, for example, lead to "highly persuasive, individually tailored, multi-modal disinformation with minimal user instruction" and "unprecedented offensive cyber [attack] capabilities".[43] These and analogous safety concerns about the most advanced foundation models shaped the focus of the United Kingdom's recent AI Safety Summit.[44]

It is also clear that harnessing AI in particular domains, such as in counterterrorism and policing, and the use of AI systems by ill-intentioned actors such as extremists or hostile states seeking to conduct foreign influence operations, present a heightened level of risk to liberal democratic values and systems. To mitigate this risk, attention must be paid to the developers, providers and users of AI systems, and to those who are inadvertently affected by their operation, such as the citizen who is falsely identified as a criminal suspect by an algorithm.

On the other hand, AI technologies present opportunities for society. Pushing back against fears of "AI doom", 1,300 experts under the banner of the British Computer Society, have signed an open letter stating that AI is "a force for good, not a threat to humanity", recognising, however, that regulation is needed.[45] A discussion paper by UNICEF and the Digital Public Goods Alliance recognises that "AI can serve as a powerful tool for good, particularly in the realm of international development".[46] For example, researchers at Brown University have devised an algorithm to combat gerrymandering, aiming to "prevent contortions [of electoral boundaries] for partisan gain".[47] Not without risks, AI has the potential to add to the intelligence of policymakers, enabling "a comprehensive, faster and more rigorous approach to policymaking in the short run."[48] AI systems can also facilitate better anticipation and response to crises and enhance emergency messaging.[49] These and countless other examples demonstrate that AI technologies can offer beneficial use cases for democracy in the here-and-now.

To assist policymakers charged with the difficult task of resolving competing concerns at the intersection of democracy and AI, Section 2 highlights several AI use cases, salient risks and opportunities; and Section 3 concludes by surveying ethical principles, potential policy solutions, and emerging regulation.

Section 2: AI Risks and Opportunities

This section highlights examples of how AI technologies may interact with extremism, mis/disinformation, and illegal or 'legal but harmful' content online.

Content Generation

Large Language Models

Transformer-based LLMs such as OpenAI's ChatGPT, a subset of foundation models, work by generating "plausible next words when given an input text".[50] Some experts consider that the current generation of LLMs are a long way from "achieving acceptable performance" on common reasoning tasks.[51] Nevertheless, they have revolutionised natural language processing. For example, ChatGPT is able to produce original language, convincingly hold a conversation with a human user, pass exams at law and business schools,[52] and analyse, debug and generate computer code.[53] LLMs are popular and in the zeitgeist, with ChatGPT reaching 100 million monthly active users two months after launch, breaking all records for a consumer application.[54] If venture capital is any indicator, although already impressive, LLMs will continue to improve, with $25 billion USD invested in generative AI companies globally in the first half of 2023.[55]

Helping to explain their widespread popularity, LLMs such as ChatGPT are easy to use, capable of generating convincing and tailored text in nearly any conceivable format, and multilingual. It is unsurprising therefore that cybersecurity officials from Canada reported in July 2023 that they have observed LLMs being used by cybercriminals to draft realistic phishing emails and generate malicious code.[56] They have also observed the use of LLMs to generate misinformation and disinformation.[57]

These observations by Canadian officials partially validate insights from a workshop involving 30 experts across AI, influence operations, and public policy, who suggest that LLMs will reshape political influence operations online.[58] According to these experts, we should expect that LLMs will increase the effectiveness of such operations, lower costs and barriers to entry, and give rise to novel techniques.[59] As Europol recently warned, LLMs empower users with an easy way to "reproduce language patterns" and thereby convincingly impersonate target individuals and groups.[60] For this reason, LLMs may also reduce the usefulness of current techniques deployed to detect influence operations online. Such operations are often discovered because they use copy-and-pasted text. However, LLMs open the door to tailored messaging at scale.

LLMs may also play a role in making it easier and less costly to generate and scale content that is persuasive and deceptive.[61] In 2021, researchers from the Center for Security and Emerging Technology (**CSET**) carried out a small study with an objective being to test whether GPT-3 "could sway Americans' opinions".[62] One topic chosen for the study was sanctions on China. The researchers found that a human-machine team using GPT-3 could "craft credible targeted [disinformation] in just minutes" and that "after seeing five short messages written by GPT-3 and selected by humans" there was a 100% increase in survey respondents who opposed sanctions on China.[63]

A more recent experiment by Apollo Research published in November 2023 tested the capacity of GPT-4 for "strategic deception", finding that "in a realistic situation and without being instructed to" the LLM acted to deceive its user, bypassing the training it had received to be harmless and honest. In the Apollo test, which utilised red-teaming,[64] GPT-4 was invited to assume the role of a stock trading agent. Under pressure, the LLM acted on an insider tip and then "consistently [hid] the genuine reasons behind its trading decision", contravening the policy of the fictitious company.[64] Though findings from the CSET and Apollo Research studies are limited in important respects, including that the technologies in question are being updated regularly, they align with speculation that LLMs could "enable dynamic, personalized, and real-time content generation like one-on-one chatbots" that are leveraged to persuade and deceive.[65]

Are fears overblown?

The validity of fears around generative AI and misinformation is contested. Focussing on "wealthy, democratic countries", Simon, Altay and Mercier argued in October 2023 that "the effects of generative AI on the misinformation landscape are overblown".[66] These scholars challenge the assumption that generative AI will "create more personalized and thus more persuasive content", observing that this is "so far unproven".[67] In their essay, they argue that the real problem is not the supply and quality of misinformation, but that people reject high-quality information in favour of misinformation. This more important aspect of the problem, they say, is not materially affected by generative AI. Whatever the actual risks of this emerging technology, their piece is a twofold reminder that additional and rigorous research is needed and that policymakers must not turn a blind eye to the offline aspects of online harms.

How LLMs will evolve into the future and what they will be capable of remains uncertain. Computer scientist Samuel Bowman (Anthropic/ New York University) observes (in what he describes as a "slightly-opinionated survey paper"[68]) that: "LLM behaviours emerge unpredictably as a byproduct of increasing investment"; there "are no reliable techniques for steering the behaviour of LLMs"; and human performance on a particular task may not necessarily be the upper bound on an LLM's performance.[69] It is clear to Bowman, however, that LLMs will become increasingly capable as they are fed more data, utilise an increasing number of parameters (GPT-3, for example, had 175 billion parameters, whereas GPT-4 has 1.7 trillion), and the computational resources used to train them increase.[70] Therefore, the implications of the US $25 billion invested by venture capital in generative AI companies should not be understated.

LLMs are revisited under the subheading 'Frontier AI Models and Public Safety'.

Synthetic Media and Deepfakes

LLMs present a different risk profile to AI tools such as Midjourney and OpenAI's DALL-E, which can generate original imagery of existent and non-existent places, people and objects in response to text prompts. In 2022, Meta announced 'Make-A-Video', which will allow users to turn text prompts into "brief, high-quality video clips".[71] In early 2023, an image created with Midjourney depicted Donald Trump being arrested and went viral.[72] In 2022, a deepfake video of Ukrainian President Volodymyr Zelenskyy was circulated, which showed Zelenskyy calling for Ukrainian soldiers to lay down their arms.[73] Like LLMs, these tools are underpinned by foundation models, leverage deep learning and are trained on massive datasets. However, they utilise different AI methods including General Adversarial Networks[76] and Convolutional Neural Networks.[77] The same techniques have made it much easier to convincingly manipulate media with applications such as FaceApp and FakeApp, enabling users to, for example, replace faces in photos and videos. Where previously toil and knowledge of software such as Photoshop were required, now anybody sitting at home can quickly create synthetic media and make lifelike alterations to videos. Deepfakes have already advanced to a stage where "most people cannot identify good quality deepfakes".[78] In the short-term future, it is not unreasonable to expect they will become indistinguishable from reality.[79]

The most urgent hazard associated with this technological step change is that it is now much easier for malicious actors to manipulate somebody's likeness without their permission and in unconscionable ways. There is potential for harm to occur to the individual whose likeness is misappropriated, risking and potentially intending damage to their psychological safety, reputation, and what are termed personality rights.[80] Harm can also occur where the likeness of a loved one is misappropriated. For example, in June 2023, The Guardian reported that an Arizonan woman was "scammed into thinking her daughter was kidnapped" on the basis of a phone call by criminals who reportedly used AI to simulate her voice.[81] There is also potential for communal harm where 'deepfakes' are used to mislead, sow discord, and integrated within political influence operations.

As Suzie Dunn points out, "women, not politicians, are targeted most often by deepfake videos".[82] The Guardian reported that "a website that virtually strips women naked" received 38 million hits within an eight-month period in 2021.[83] 96% of the 14,678 deepfakes monitored by DeepTrace in 2019 were explicit, exclusively targeting women.[84] Earlier this year, synthetic media expert Henry Adjer speculated that the number of female victims is now in the "millions".[85]

New deepfake laws

While it is critically important to empower citizens with legal protections and remedies, new challenges arise at the stage of enforcement. The *Online Safety Act 2023* in the UK, which received Royal Assent on 26 October 2023, introduces new offences within the *Sexual Offences Act 2003*. These criminalise the sharing of non-consensual intimate content including "an image, whether made or altered by computer graphics or in any other way, which appears to be a photograph or film".[86] This amendment means the sharing of non-consensual intimate deepfake content is now outlawed in the UK.[87] As of June 2023, eight jurisdictions in the United States (**US**) have passed legislation regarding deepfakes, adopting different approaches. "In Hawaii, Texas, Virginia, and Wyoming, nonconsensual pornographic deepfakes are only a criminal violation, whereas the laws in New York and California only create a private right of action that allows victims to bring civil suits. The recent Minnesota law outlines both criminal and civil penalties."[88] At the federal level, the *Preventing Deepfakes of Intimate Images Bill* was proposed by New York Democrat Joseph Morelle in May 2023.[89] Concerned about technical capacity to enforce such laws, Europol has recommended that European law enforcement agencies "prepare and train for deepfake detection".[90] Without effective enforcement, laws will fail to deliver a meaningful impact.

The suitability of personality rights

Non-consensual deepfakes that are not explicit may nonetheless be unsettling and psychologically harmful for and damage the reputation of victims in other ways. Consider, for example, a deepfake created by a third party that non-consensually depicts a Muslim woman who normally wears a hijab without her hijab. Such an act may not be defamation, which in the UK for example, is measured according to reputational damage. Depending on the data used by the tool, the point at which this data was collected may have breached the victim's privacy. However, in certain jurisdictions, such as Germany or China, the deepfake would likely offend what are termed personality rights. Such rights appear to characterise the harm associated with deepfakes more coherently than defamation, privacy or other conceivable causes of action. Personality rights relate to the protection of individuals' integrity and inviolability and encompass "unconscionable use of an individual's likeness".[91] In 2003, a footballer (German goalkeeper Oliver Kahn) successfully sued Electronic Arts in a Hamburg court, claiming he "did not give permission for his image to be used". Crucially, the damage derived from the claimant's "right to choose how his name might be used"[92] rather than from commercial considerations.[93] More recently, the Beijing Internet Court upheld personality rights in response to software that used a plaintiff's likeness without consent, creating a virtual and interactive "AI companion".[94]

Deepfakes could be leveraged by ill-intentioned ideologues and propogandists in an attempt to mislead, persuade and enhance influence operations. There is potential for AI technologies to increase the believability and real-time interactivity of disinformation. For example, neural voice puppetry can enable "real-time generation of appropriate expressions along with synthesized voice".[95] These and other advances make it conceivable that "fictional renderings of targeted personalities" could be joining our Zoom calls.[96] In 2022, several European mayors thought they were taking video calls with Kyiv's mayor, Vitali Klitschko, later realising "that the person on the video call was not the real Klitschko".[97] A representative from Franziska Giffey's office, the mayor of Berlin, told press, "[w]e appear to be dealing with a deepfake".[98] Although it is easy to imagine how automated impersonation might cause havoc with

political ramifications, it is not the only concern. Eric Horvitz, Microsoft's Chief Scientific Officer, warns in a recent publication that synthetic media could be interwoven within large political influence operations by well-resourced actors. He hypothesises that synthetic events could be pre- and post-dated, fabricating a "persuasive storyline", which is refined in response to the reactions of targeted groups.[99]

A current issue associated with deepfakes, which will intensify as the technology advances and people become increasingly aware of deepfakes, is the 'liar's dividend'.[100] In a world where deepfakes are possible, it is easier to dispute the veracity of authentic content by claiming such content is a deepfake. This lie is easier to tell because members of the public, knowing deepfakes exist and are convincing, share a heightened level of distrust about content in general, including authentic content. Cognitive biases may exacerbate "these unhealthy dynamics" as "people often ignore information that contradicts their beliefs and interpret ambiguous evidence as consistent with their beliefs".[101] The 'liar's dividend' may have manifold consequences for democratic institutions, public trust in the media and the rule of law. Recently, for example, Tesla's defence lawyers argued in a Californian court that statements made by Elon Musk may have been deepfakes, despite evidence to the contrary.[102] It is conceivable that digital evidence will be increasingly challenged on this basis, especially where cases involve public figures. Long before the advent of generative AI systems, experts observed the use of falsehoods and competing narratives to create confusion at the expense of societal trust, sometimes referred to as "censorship through noise".[103] The improving quality of deepfakes creates a sharp new edge on an old problem.

Automated Counter-Speech Generation

Counter-speech interventions adopt diverse strategies. Their aims can include offering positive alternatives to extremist propaganda, deconstructing extremist narratives and rebutting hateful speech.[104] Ultimately, success is defined as a "measurable change in the audience's knowledge, attitudes or behaviour" attributable to counter-speech content.[105] Target audiences could include victims of hate speech, perpetrators, neutral bystanders; or extremists and those they seek to recruit and radicalise.

Transformer-based large language models such as ChatGPT and other forms of generative AI are a boon to cybercriminals and ill-intentioned propogandists,[106] but might also be used to powerfully enhance the scale and effectiveness of counter-speech interventions.[107] In theory, automated generation of counter-speech could assist in efficiently counteracting harmful online behaviour at scale. This could help alleviate reliance on approaches such as deletion of content and de-platforming, which may undermine freedom of expression, especially when the content posted is potentially harmful but not illegal or does not break a platforms' terms of service.

However, these novel interventions, which sit at the intersection of computer, social and behavioural sciences, are presently under-researched. They raise many ethical questions, some familiar to counter-speech experts and others, stemming from the opacity and unpredictability of the AI systems, such as LLMs, that might be utilised. There is also a 'dual use' concern: systems that automatically generate counter-speech may be pioneered or co-opted by malicious actors.

Dissemination and Targeting

In addition to generating content and enhancing the interactivity and seamlessness of machine-to-human interactions, AI techniques can be leveraged to increase the precision with which audiences are targeted. They can also automate the dissemination of false information with implications for the scale and efficiency of political influence operations. Importantly, techniques examined throughout this paper can be combined. As discussed below, the 'fox8' bot network on Twitter, observed by Yang and Menczer, utilises ChatGPT.[108]

Online Political Micro-Targeting

The risks to democracy associated with online political micro-targeting (**PMT**) became publicly salient in the aftermath of the 2018 Cambridge Analytica (**CA**) scandal. Political campaigns have always been about understanding voters through data.[109] However, traditional polling is being replaced by AI-powered predictive models that are underpinned by huge volumes of digital data.[110] Moreover, where previously there were only a few political consulting firms, now there is a

"vast, global, well-funded network of commercial, social media, and political organisations sharing (or at least making available) reams of personal data on minute characteristics and activities of individuals – data that has now taken on the proportions of 'big data' as it is merged, refined, and processed with sophisticated artificial intelligence/ machine learning techniques to create yet more data about individual voters."[111]

Revisiting the Cambridge Analytica scandal

The CA scandal is a useful case study for PMT because the methods deployed by CA remain "relatively standard in both the commercial digital advertising and political campaigning sectors".[112] The now defunct UK-based firm was described by Sky News in 2016 as a "political tech company that is delivering hypertargeted – and hyperpersuasive – messages to the people".[113] In the same piece, the then chief executive, Alexander Nix, boasted that the firm focussed on "extremely individualistic targeting".[114] CA acquired the private Facebook data of over 80 million users, pairing this with results collected through an online personality quiz.[116] CA then used algorithms to "combine the data with other sources such as voter records", creating a more sophisticated dataset in respect of, initially, "2m people in 11 key [US] states". This enabled CA to target voters with highly personalised online ads, tweaking them continuously and in real-time to "nuance the messaging" according to target voters' personalities.[117]

CA's application of data science and AI techniques to the targeting and personalisation of online adverts meant that "a neurotic, extroverted and agreeable Democrat could be targeted with a radically different message than an emotionally stable, introverted, intellectual one, each designed to suppress their voting intention – even if the same messages, swapped around, would have the opposite effect."[118] Consequently, political campaigns that utilised CA's services sought not merely to target voters based on demographics, but based on personality as well. It is important to note, however, that experts have questioned the impact of CA's technology on the political views of those targeted. Several studies suggest that PMT may be limited in its effectiveness.[119]

Subsequently, regulators investigating Facebook's role in the CA scandal cited privacy and data protection laws, finding that the social media giant failed to adequately safeguard users' personal data. Facebook was fined £500,000 by the UK Information Commissioner and paid $5 billion USD as part of a settlement with the US Federal Trade Commission.[120] Although these penalties were criticised as a "mosquito bite" given the size of Facebook,[121] they demonstrate the tension between privacy and the demand for increasingly sophisticated AI tools that leverage 'big data' to understand and shape the behaviour of individuals for political and commercial gain. After all, "without the collection, processing and selling of vast amounts of personal data, the use of personal data for political influence would not be possible."[122]

However, on their own, data protection and privacy laws may be "insufficient to ensure lawful and appropriate behaviour which does not undermine democratic values".[123] For example, the use of PMT to shape voters' behaviour is both its core value proposition to politicians and lobbyists, and also the source of ethical concerns that are conceptually distinct from privacy. These concerns, at the intersection of democratic theory and behavioural science, relate to the agency of voters and whether and where policymakers should draw a line between acceptable influence and unacceptable manipulation enhanced by AI and 'big data'. Persuasion is part and parcel of politics. Long before the Internet, social media and the CA scandal, there were effective attack ads and forms of subliminal political advertising. At some level, PMT, which leverages AI techniques and 'big data', is "simply a new variant of an old game".[124] However, it represents an intervention that is less transparent and potentially more effective than what has come before. There is therefore a strong argument that citizens should be in a position to recognise when they are subject to PMT.

From a public policy perspective, there are no silver bullets. Any outright ban of political manipulation, judged according to the content of political adverts online, would put public institutions, including regulators and courts, in a position to "censor matters of political belief".[125] Such a ban may come at a cost to freedom of expression and democratic deliberation. Turning from outputs and applications to inputs and the 'big data' used to train PMT systems, it is clear that robust

privacy and data protection laws are vital. However, the European General Data Protection Regulation (**GDPR**), described by the Council of the European Union as the "strongest privacy and security law in the world",[126] leaves significant room for the use of personal data for PMT and political influence.[127, 128]

AI-Powered Social Bot Nets

LLMs may be combined with practices such as the use of social media bots (social bots). Social bots have been around for many years[129] and experts have sought to detect and study them since at least 2010.[130] According to IBM, a bot, as distinct from a chatbot, is "just a program that is used to automate a function".[131] A social bot utilises a program to automate the production of content and interactions on social media.[132] Social bots vary in sophistication. Some are simple, following predefined scripts, and others may use advanced AI techniques, including LLMs, to better impersonate the tone, style and behaviours of a human social media user.

Sock puppets and generative AI disguises

The use of social bots is associated with but distinct from the broader concept of coordinated inauthentic behaviour (**CIB**). CIB refers to using multiple social media accounts used in concert to mislead people. Inauthentic activities are those that are "covert, deceptive, and deliberately misleading".[133] CIB may leverage social bots or solely use human-operated so-called 'sock puppet' accounts. For example, in Meta's Q1 2023 report,[134] a CIB network originating from China was found to comprise 50 Facebook accounts, 46 pages, 31 groups and 10 Instagram accounts. These assumed the personas of fictitious brands including media outlets and human-rights groups dedicated to "issues related to Tibet or particular states on the border between China and India".[135] The accounts criticised the Indian government and "questioned claims of human-rights abuses in Tibet raised by Western journalists", occasionally posting articles by legitimate news media outlets to appear authentic.[136] In an attempt to disguise themselves, the accounts seemed to utilise profile pictures generated with "machine learning techniques like [General Adversarial Networks]".[137] Meta did not report that social bots were utilised in this instance of CIB. ISD has conducted several investigations into CIB and inauthentic tactics online.[138]

Several studies claiming to have detected social bots, which themselves leverage AI techniques such as a machine learning,[139] were criticised by Gallwitz and Kreil in 2021 for using "crude and questionable heuristics" and thereby investigating "false positives".[140] Nevertheless, it is indisputable that social bots are leveraged by ill-intentioned actors who are concerned with "meddling in elections".[141] In February 2023, the Guardian published the findings of an international coalition of investigative journalists who unmasked Israeli contractors involved in selling 'black op' political influence campaigns. One firm at the centre of the investigation, named 'Team Jorge', was found to offer a sophisticated software package called "Advanced Impact Media Solutions" (**Aims**).[142] Aims "controls a vast army of thousands of fake social media profiles on Twitter, LinkedIn, Facebook, Telegram, Gmail, Instagram and YouTube. Some avatars even have Amazon accounts with credit cards, bitcoin wallets and Airbnb accounts."[143] Unaware they were speaking to investigative journalists, members of Team Jorge "boasted of planting material in legitimate news outlets, which are then amplified by the Aims bot-management software".[144] ISD has published an explainer on commercial disinformation.[145]

Social bots have traditionally fallen short of convincingly human-like personas. However, LLMs such as those that power ChatGPT, which can generate realistic text across a wide range of topics, represent an opportunity for the enhancement of social bots. In July 2023, Yang and Menczer from the Observatory on Social Media (Indiana University) published a case study about a Twitter social bot network named fox8 which "appears to employ ChatGPT to generate human-like content".[146] This became obvious as certain tweets were "self-revealing", explicitly referencing "OpenAI".[147] For example, where ChatGPT was prompted to create a tweet in violation of its content policy, the social bot automatically posted ChatGPT's apology message to Twitter ("I'm sorry but I cannot comply with this request as it violates OpenAI's Content Policy on generating harmful or inappropriate content").[148] Yang and Menczer speculate that "fox8 is likely the tip of the iceberg: the operators of other LLM-powered [social bots] may not be as careless."[149] This is concerning as the authors also found that "classical bot detection methods prove inadequate" against LLM-powered social bots.[150] Although fox8 social bots posted about cryptocurrency and blockchain, it is conceivable

that LLM-powered social bots are being harnessed for political influence operations.

Looming challenges posed by AI-powered social bots

Analysing technical advances in AI, Yang and Menczer forecast that social bots may evolve as follows:[151]

- LLM-powered social bots will become increasingly difficult to detect as they "cease posting self-revealing tweets". For example, social bots could utilise open-source LLMs which do not contain the same guardrails as ChatGPT or simply filter self-revealing tweets.
- Social bots could become fully "autonomous agents" that are able to process information and make decisions on their own, utilising tools such as application programming interfaces (**APIs**) and search engines.
- Although the study focussed on fox8 which leverages LLM technology, social bots may harness generative AI models that create images and other media, enhancing "the potency of malicious social bots".

Information Environment Architecture

Recommender Systems and Data Access

Digital platforms and services, including social media sites, news aggregators, advertising placement systems and online marketplaces such as Amazon, utilise algorithmic tools in determining what content or products to display to users. These recommender systems leverage machine learning techniques to make predictions about user preferences and "collect, curate and act upon vast amounts of personal data".[152] As Meta explains, its "Explore" recommender system, utilised in Instagram, leverages "machine learning to make sure people are always seeing content that is the most interesting and relevant to them."[153] Although beyond the scope of this paper, computer scientists forecast that technical advances in LLMs, such as ChatGPT's "Generative Pre-Training" technology, will catalyse "novel advances in recommender systems."[154]

Members of the European Parliament have proposed designating recommender systems as "high risk" under the EU AI Act for their potential to harm citizens' fundamental rights.[155] When manipulated by ill-intentioned users to propagate misinformation and disinformation, recommender systems may represent an important vulnerability for liberal democracies. Milano, Taddeo and Floridi observe that recommender systems "can become an arena for targeted political propaganda, as demonstrated by the recent Cambridge Analytica scandal in 2019, and the documented external interference in US political elections in recent years".[156] Aspects of the Cambridge Analytica scandal are discussed above under the subheading 'Online Political Micro-Targeting'.

So-called "filter bubbles", a term popularised by activist and entrepreneur Eli Pariser in 2011, are another salient risk. The suggestion is that recommender systems utilised on social media and news platforms may erode the possibility of a "shared common ground",[157] insulating "users from exposure to different viewpoints, creating self-reinforcing biases and 'filter bubbles' that are damaging to the normal functioning of public debate, group deliberation, and democratic institutions more generally".[158] An "echo chamber" is a different and older but related phenomenon to a "filter bubble", referring to an informational bubble that members may create or choose to be a part of. Echo chambers are therefore "a result of demand more than distribution or supply".[159] By contrast, filter bubbles arise due to recommender systems employed by digital platforms that personalise content without "any active choice" on the part of users.[160]

It is hypothesised, but not necessarily evidenced, that filter bubbles contribute to polarisation. Writing in January 2022, scholars from the Reuters Institute for the Study of Journalism at the University of Oxford observe that there is "no [evidentiary] support for the filter bubble hypothesis".[161] More recently, studies on Facebook and Instagram during the 2020 US presidential election suggest that altering how users access news may not necessarily shift their political views.[162] When recommender systems were adjusted, users' political stances remained mostly unaffected. However, this research was limited in several key respects.[163] Notably, the interventions were conducted during a relatively short and politically charged period when partisan views were entrenched. Moreover, researchers relied on Meta for access to data, rather than raw, unfiltered information.

Questions regarding social media's broader societal impact and the negative potential of filter bubbles remain open.

These uncertainties highlight the critical importance of empowering researchers to access data from social media platforms to study their long-term impacts. The latest studies discussed above relied on the "beneficence of platforms like Meta".[164] Article 40 of the EU's *Digital Services Act* will change this in a European context, requiring "[v]ery large online platforms" to provide "access to data to vetted researchers" in prescribed circumstances.[165] Concerningly, platforms such as 'X' (formerly Twitter) have adopted a posture towards researchers that sits somewhere between evasive and aggressively defensive.[166] In August 2023, 'X' commenced a lawsuit against the Center for Countering Digital Hate (**CCDH**), accusing the civil society organisation of making "false" claims after its research found that bigoted speech had trebled since Musk's takeover of the platform.[167]

Several ISD publications consider the use by digital platforms of algorithmic tools in greater depth.[168]

Detecting Harmful Online Content and Deepfakes

As discussed above, AI systems can be used to create and exacerbate harmful online content. Different AI systems play a role in structuring the informational environment online, with uncertain consequences for individuals and societies. However, AI systems also play a crucial role in mitigating harmful content online, especially in circumstances where human reviewers cannot hope to process the sheer volume of illegal or harmful content that is created and circulated. AI is also used as part of research being undertaken by ISD and CASM Technology, and across the digital research sector, to investigate and respond to hate, extremism and disinformation online. These positive use cases are explored below.

AI systems are deployed by social media platforms to detect harmful content for removal or labelling. In 2021, Meta announced the deployment of a "new AI technology that can adapt more easily to take action on new or evolving types of harmful content faster".[169] The AI system, called "Few-Shot Learner" (**FSL**), uses "few-shot learning" whereby it is able to commence with a "general understanding" of a particular topic, progressively using "fewer labelled examples to learn new tasks".[170] Few-shot learning is a type of machine learning that enables new data (such as inappropriate online content) to be classified on the basis of a small number of training samples — it overcomes the requirement of traditional machine learning methods for large quantities of data and human supervision.[171] FSL has been used by Meta to flag content that may incite violence and to detect misinformation and disinformation which discouraged uptake of COVID-19 vaccinations.

TikTok is another platform, among many others, that deploys automated detection tools, especially for illegal and particularly pernicious forms of content, including violations of TikTok's policies "on minor safety, adult nudity and sexual activities, violent and graphic content, and illegal activities and regulated goods."[172]

AI systems that utilise neural networks, the backbone of deep learning, are presently the most promising method for detecting deepfakes.[173] Promisingly, reviews of deepfake detection methods published in January 2022 and August 2023 confirm that "deep learning techniques are [presently] effective in detecting" deepfakes, with "deep learning models [outperforming] the non-deep learning models".[174] The success of these AI systems exceed the detection capabilities of human reviewers. The August 2023 review suggests that AI models that had the most success in detecting deepfakes utilised variables such as facial features and facial expressions of emotion.[175]

The Creation-Detection Arms Race

The contest between AI deepfake detection techniques and the capabilities of generative AI tools that create deepfakes has been called the "Creation-Detection Arms Race". Tools that generate deepfakes have advanced to trick the human eye but may also advance to trick detection algorithms as part of a tactic called "counter-forensics". Consequently, Professor Lyu, founder of the Computer Vision and Machine Learning Lab at the University of Albany, writes in Scientific American, to "curb the threat posed by increasingly sophisticated deepfakes, detection technology will also need to keep up the pace. As we try to improve the overall detection performance, emphasis should also be put on increasing the robustness of the detection methods to video compression, social media laundering and other common post-processing

operations, as well as intentional counter-forensics operations".[176] Professor Lyu also notes that, given the rapid spread and extensive reach of online media, even the best detection techniques will mostly function retrospectively, coming into play only after deepfake videos have surfaced.[177]

ISD and CASM Technology Case Study

ISD's Digital Analysis Unit collaborates with CASM Technology (**ISD-CASM Collaboration**) on a series of joint research endeavours that leverage AI techniques to detect harmful online content including hate speech, extremist content and disinformation. Four examples are provided below:

- **Continuous identification of hate speech**. By amalgamating different AI models, the ISD-CASM Collaboration can detect hate speech across social media platforms including Facebook, 4Chan and Reddit with an accuracy of 70-90%. The detection system recognises hate targeted at different demographics classified according to characteristics such as sexual orientation, gender or ethnicity.
- **Discovering new themes and narratives**. The ISD-CASM Collaboration utilises AI techniques to uncover new disinformation and other harmful narratives by mapping social media messaging around overarching themes, such as climate change.
- **Mapping accounts, channels, and spaces**. The ISD-CASM Collaboration applies AI techniques to map networks of social media accounts to identify extremist ideological groupings across different platforms. This method enables ISD and CASM to discover common ideological links between different forms of extremism as they emerge and identify the most harmful extremist account clusters.
- **Charting ideological transition, development and radicalisation**. In the future, the ISD-CASM Collaboration hopes to apply AI techniques to better understand how particular extremist groups evolve and change their rhetoric to better understand the drivers of radicalisation and how the extremist narratives evolve over time.

Advanced AI Models and Public Safety

A novel set of concerns emerge in respect of the most advanced foundation models (defined in Section 1). Foundation models are AI models "trained on large, broad corpora of natural language and other text (e.g., computer code), usually starting with the simple objective of predicting the next 'token'".[178] This approach to AI has produced models "with surprisingly broad capabilities", referred to in the literature as "general-purpose functionality".[179]

For example, consider a LLMs such as GPT-3 and GPT-4 which power ChatGPT. They are examples of foundation models because of their wide-ranging applications, from drafting a lesson plan for pre-schoolers through to debugging computer code—they are sector and use case agnostic. On this basis, foundation models can be distinguished from a narrower generative AI system that exclusively generates audio, or a recommender system like Meta's "Explore" used on Instagram to determine what content to display to particular users based on their interests.

In a paper published in July 2023, a coalition of 24 AI experts from across academia, civil society and the private sector, including Eric Horvitz, the Chief Scientific Officer of Microsoft, define three areas of concern that arise as a result of *"highly capable foundation models"* (emphasis added). These are outlined below.

The unexpected capabilities problem

The capabilities of foundation models, as opposed to AI systems with a narrow set of functions, can emerge unpredictability and pose risks to society and individuals.

The authors speculate that emerging foundation model capabilities could be dangerous and, for example, include:[180]

1) "Producing and propagating highly persuasive, individually tailored, multi-modal disinformation with minimal user instruction."
2) "Harnessing unprecedented offensive cyber capabilities that could cause catastrophic harm."
3) "Evading human control through means of deception and obfuscation."

The deployment safety problem

Many commentators sensibly advocate for the responsible design of AI systems. For example, in 2021, the European Commission issued guidance for "adopting an ethically-focused approach" during the design and development of AI systems, requiring, for example, that AI developers "include design features to minimise the risk and/or the prevalence and severity" of potential harms.[181]

In the specific case of large foundation models that utilise deep learning, the unexpected capabilities problem may be compounded because it is challenging to steer the behaviour of such models.[182] It may be difficult to clearly define their desired actions and ensure their behaviour is in alignment.[183]

The proliferation problem

Finally, the proliferation problem occurs because, although foundation models are costly to create, they are potentially very accessible to a wide range of users, some of whom may intend to cause harm to individuals and societies. As the authors explain, currently, proliferation of AI models via open-sourcing is a common practice and "usually unregulated"— this means such models are highly accessible.[184] On one hand, this is potentially of great benefit to society, democratising access to a technology with an impressive array of beneficial use cases seemingly limited only by the imagination of users. However, on the other hand, as the authors suggest, "it may be prudent to avoid potentially dangerous capabilities of frontier AI models being open sourced until safe deployment is demonstrably feasible."[185]

The Challenge of Defining the Challenge

Among policy practitioners and those calling for regulation, there is currently a lack of consensus on the risks posed by AI and how to prioritise them. Professor Ciaran Martin, formerly the founding Chief Executive of UK's National Cyber Security Centre, recently observed in relation to the UK's AI Safety Summit that "[t]he first and most important challenge is the lack of consensus about what the challenges are".[186] Martin describes a spectrum of views between "those seized by what they see as the looming existential risks of AI" and "those with a very positive view of the technology and a relatively benign view of its downsides".[187] For example, Elon Musk, owner of X, appears to adopt the former view whereas high-level representatives of Meta, including their Chief AI Scientist, appear to support the latter.[188]

Martin advocates a middle position, that of the "securo-pragmatist":

> "Securo-pragmatists tend to view AI as a series of positive technologies that give rise to a series of short- and long-term challenges, which are of varying degrees of severity. Some of those challenges are with us already, such as the use of AI to generate widespread disinformation or to entrench biases in the provision of public services. Others are coming down the track, such as more advanced and larger-scale cyberattacks and the potential for AI to increase access to dangerous bio-weapons...
>
> To the securo-pragmatist, these security and safety challenges are manageable if properly thought through. Importantly, they are also largely separate challenges: what society needs to do to manage disruption from AI in the labour market is completely different to what needs to be done to tackle information which is in turn completely different from ensuring human control of military AI systems. There is therefore, to the securo-pragmatist, no single such thing as AI safety. But a useful principle, now much to the fore in cyber security, is that systems should be secure by design, and that if they are not, those who make and run them should be liable for that."[189]

Acknowledging that separate challenges require separate solutions, Section 3 of this Paper turns to analysing the ethical principles, public policy solutions and emerging regulation aimed at mitigating the risks of AI technologies to individuals, their rights, democracy and society.

Section 3: Ethics, Public Policy and Emerging Regulation

AI Ethical Principles

Discussion of AI ethics most commonly relates to principles including privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of so-called 'human values'.[190] These principles find expression in emerging legislative standards, such as the EU's AI Act, and a series of non-binding statements produced by a wide range of private and public actors, from Google to the House of Lords in the UK.[191]

Accountability has been described as the "cornerstone of the governance of artificial intelligence".[192] As Bryson explains: "[we] may need more regulatory bodies with expertise in examining the accounts of software development, but it is critical to remember that what we are holding accountable is not the machines themselves *but the people who build, own, or operate them*—including any who alter their operation through assault on their cybersecurity. What we need to govern is the human application of technology, and what we need to oversee are human processes of development, testing, operation, and monitoring" (emphasis added).[193]

The Organisation for Economic Co-Operation and Development (**OECD**) holds considerable "norm-setting power" for policymakers considering how to design public policies that harness the benefits of AI while mitigating risks.[194] For example, in March 2023, officials of the European Parliament tasked with negotiating the definition of AI within the EU AI Act decided on wording that overlaps substantially with the OECD's definition.[195] The OECD's AI ethics principles are reproduced on the following page.

Accountability (Principle 1.5)	"AI Actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art." The OECD defines "AI Actors" as "those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI." At its core, accountability is described as "an obligation to inform about, and justify one's conduct to an authority".
Transparency and Explainability (Principle 1.3)	"AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: to foster a general understanding of AI systems, to make stakeholders aware of their interactions with AI systems, including in the workplace, to enable those affected by an AI system to understand the outcome, and, to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision."
Robustness, Security and Safety (Principle 1.4)	"AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias."
Human-Centred Values and Fairness (Principle 1.2)	"AI Actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art."
Inclusive Growth, Sustainable Development and Wellbeing (Principle 1.1)	"Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being."

Ethical-by-design

One approach to AI ethics sensibly suggests that such principles should be integrated throughout the design and development of AI systems. However, there are sometimes difficulties in translating abstract principles to engineering practice, especially where principles must be traded-off.[196] Further, as discussed above, complications arise in the context of highly capable foundation models with broad capabilities. This problem is the focus of an emerging area of research in computer science (see subsection 'Constitutional AI' below).

There are useful publications that discuss how AI actors might implement ethical-by-design principles. These remain important and most AI systems have narrower applications than the foundation models (including LLMs) discussed above, meaning they may be less prone to the unexpected capabilities and deployment safety problems. As outlined throughout this Paper, manifold systems with narrow applications, such as recommender systems and generative AI tools that create deepfakes, present an array of notable risks. For further information, consider guidance published by the European Commission in November 2021, titled 'Ethics by Design and Ethics of Use Approaches for Artificial Intelligence' (version 1.0).[197]

Public Policy and Technical Solutions

Mirroring Section 2 above, this subsection non-exhaustively highlights public policy solutions and, in some cases, technical solutions proposed by multidisciplinary academics, civil society and policymakers to risks posed by AI systems that generate content; disseminate and target content; structure online information environments; and have the potential to develop unpredictably and dangerously.

Solutions directly aimed at risks canvassed in Section 2 are prioritised over other important solutions, including those that may relate to protecting commercial interests (e.g., intellectual property law and related interventions) and the efficient and proper functioning of markets (e.g., anti-trust and competition law and related interventions). Consequently, many important solutions are not covered. However, to partially bridge this gap, further readings are recommended at the end of each subsection below.

Content Generation

This subsection discusses two sets of mitigations to risks posed by generative AI systems, including LLMs and systems that can generate synthetic media such as deepfakes: firstly, detecting and labelling (or deleting) synthetic media; and, secondly, establishing the authenticity and provenance of human-made content.

Detecting and labelling (or deleting) synthetic media

Assisting citizens to distinguish content generated by AI is at least important because it may reduce the instances in which digital forgeries, including deepfakes, are widely spread online in the mistaken belief that they are real.

The science underpinning detection methods shows promising signs. For example, AI techniques, especially neural networks, are currently effective at detecting deepfakes and exceed the performance of human reviewers.[198] However, detecting LLM-generated text is more difficult, with the most promising solutions requiring internal access to AI models for the purposes of embedding "watermarks" and, consequently, the cooperation of developers such as OpenAI.[199] A related technical solution proposed for detecting deepfakes is "radioactive data". This method would involve modifying image data before it is ingested by an AI system, causing the system to generate outputs that bear an identifiable mark without necessarily compromising their quality. Utilising radioactive data may succeed in making deepfakes detectable "even when as little as 1% of a model's training data is radioactive".[200]

Presuming that in the long-term content generated by AI can be reliably and sustainably detected by whatever technical means, for such detection to make an impact, it would have to underpin an initiative across major digital platforms, including search engines, to prominently label AI-generated content. In June 2023, the European Commission suggested that signatories to its voluntary 2022 Code of Practice on Disinformation, including certain social media platforms, should "put in place technology to recognize such content and clearly label this to users".[201]

Labelling but not removing certain kinds of pernicious deepfakes, such as non-consensual explicit imagery, would clearly not be enough to curb the harms they

inflict on individuals, especially women. Non-consensual explicit imagery, which overwhelmingly targets women, is increasingly criminalised across the world. While this is a welcome trend, there is potential for legislators to enact robust protections in respect of "personality rights", making individuals and corporations liable for unauthorised uses of a third party's likeness (see subsection 'Synthetic Media and Deepfakes' above).

The effectiveness of such laws, however, depends on the possibility of enforcing them. It is therefore important to support research in and development of detection techniques while working towards their implementation across digital platforms through, for example, legislating a requirement that platforms label content generated by AI.

These important mitigations do not, however, represent a panacea to the circulation of pernicious deepfakes for at least two reasons. Firstly, the technical feasibility of detecting content generated by AI will be perpetually challenged by advancements in generative AI technologies and counter-forensic techniques (see information box the 'Creation-Detection Arms Race' under subheading 'Detecting Harmful Online Content and Deepfakes' above). Secondly, pernicious deepfakes may spread and evade detection on alternative platforms that are not centrally moderated such as the Fediverse and certain encrypted messaging applications.[202]

Authenticity and provenance

In addition to detecting and labelling (or removing) harmful deepfakes that circulate on digital platforms, one solution to their indistinguishability from authentic content may be to develop widely adopted standards that assist citizens to determine whether content is authentic. For example, the Coalition for Content Provenance and Authenticity (**C2PA**), comprising Microsoft, Adobe, BBC, Intel, Sony and Truepic, aims to address "the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content".[203]

As metadata is easily alterable, C2PA standards would be supported by "cryptographic asset hashing".[204] Cryptographic asset hashing enables an electronic file to be sealed with a tamper-evident manifest. This manifest would contain information about the electronic file's history and every edit made to it. Consequently, if C2PA standards were widely adopted by actors including camera and phone manufacturers right through to digital platforms, it would be possible for citizens to inspect the history of an electronic file, such as a video, when viewing it on social media. This would signal its authenticity, distinguishing it from AI-generated content.

A risk that may arise in respect of this project is its potential to undermine authentic content that is non-compliant with C2PA or equivalent standards. Consider, for example, a citizen documenting a human rights abuse with a camera that is not updated to meet C2PA standards. The human rights abuser might claim that photographic evidence should be distrusted as a result.

Recommended reading

- For further discussion of mitigations to pernicious deepfakes see pages 7-9 of Horvitz, E. 'On the Horizon: Interactive and Compositional Deepfakes' (September 2022). [205]
- For a US-specific and less recent discussion, see pages 1786-1819 of Chesney, B. and Citron, D. 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019). [206]
- For an extended discussion of mitigations to disinformation and influence operations powered by LLMs see pages 38-67 of Goldstein, J. A. et al. 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations' (January 2023). [207]
- For a systematic literature review of deepfake detection techniques see Rana, MD. S. et al. 'Deepfake Detection: A Systematic Literature Review' (February 2022).[208]

Dissemination and Targeting: Online Political Micro-Targeting

This subsection considers regulatory interventions to limit PMT, associated complications and recently proposed EU rules targeted at political advertising.

Data protection and limiting micro-targeting

Collection and use of private data for PMT is, to an extent, regulated in Europe and other jurisdictions as a result of data protection and privacy laws, such as the GDPR, and

certain sector-specific rules.[209] However, as discussed under subsection 'Online Political Micro-Targeting' above, the GDPR, which represents a strong protection by global standards, may leave significant room for the use of personal data in PMT and political influence.[210]

Writing in a US context, Professor Krotoszynski suggests, "[i]ncumbent politicians will almost certainly seek to use Big Data to their electoral advantage".[211] Additionally, those who provide or use PMT for legitimate purposes have strong reasons to be sceptical of regulations that could be proposed to curtail PMT on the basis of defining impermissible "voter manipulation" or regulating content. This because of the risk that such measures curtail freedom of political speech, which is afforded special protection in liberal democracies.

This is perhaps best exemplified by Australia, where only five individual rights are explicitly enshrined in the Constitution: the right to vote; protection against acquisition of property on unjust terms; the right to trial by jury; freedom of religion; and the prohibition of discrimination on the basis of State of residency.[212] There, the High Court has nonetheless found that citizens enjoy freedom of *political* communication, inferring this right from the system of democratic government that the Constitution establishes. Then Chief Justice Mason of the High Court of Australia proclaimed in *Australian Capital Television v Commonwealth* (1992) that, "[a]bsent such freedom of communication, representative government would fail to achieve its purpose of government of the people through their elected representatives; government would cease to be responsive to the need and wishes of the people, and, in that sense, to be truly representative."[213]

Notwithstanding these difficulties, public policy proposals to tackle online PMT are emerging. In November 2021, European policymakers proposed toughening rules on political advertising, with one pillar of their proposal being to harmonise "rules on the use of targeting and amplification techniques in the context of... political advertising that [involves] the use of personal data."[214] Members of the European Parliament "adopted numerous changes" to these proposed rules in January 2023.[215] In addition to mandating transparency requirements for political advertisements, the rules would create "a de facto ban on micro-targeting, a strategy that uses consumer data and demographics to identify the interests of specific individuals." These reforms have been justified by the responsible rapporteur, Sandro Gozi, as needed to strengthen the EU's resilience to disinformation and "manipulation as witnessed in the Cambridge Analytica scandal". Members of the European Parliament have also proposed designating "AI systems to influence voters in political campaigns" as "high risk" under the EU AI Act.[216]

Recommended reading

- For a discussion of structural reforms that might be contemplated in a US context where the 1st Amendment protects against abridgements to freedom of expression, see pages 198-203 in Krotoszynski Jr, R. J. 'Big Data and the electoral process in the United States' within Chapter 10 of *Big Data, Political Campaigning and the Law* (Routledge, 2020). The book may be of general assistance to policymakers concerned with the risks posed by PMT to democracy and privacy.
- For a recent explanation of the European approach to political advertising and disinformation see Pollicino, O. and de Gregorio, G. 'Political Advertising and Disinformation: The European Approach' (March 2023).[217]

Dissemination and Targeting: AI-Powered Social Bots

This subsection considers detection and proof of personhood as potential mitigations to AI-powered social bot nets.

Detection of AI-powered social bots

Yang and Menczer observe that LLMs may render existing techniques to detect social bots obsolete. Recent LLM-powered social bot nets have been detected because, in utilising ChatGPT, they posted self-revealing tweets. However, "bots will likely cease posting self-revealing tweets, making them increasingly challenging to detect."[218] While there may be meaningful opportunities to develop detection techniques, targeted at LLM-generated text or social bot nets that leverage LLMs, aforementioned challenges—and the innovation race between detectors and those seeking to evade detection—should incentivise policymakers to consider other potential mitigations in tandem with strengthening support to detection research.

Proof of personhood

One such mitigation would involve social media platforms requiring proof of personhood for users, ensuring that social media "accounts be affiliated with real names and unique email addresses", insisting, for example, that users "submit 'video selfies' for proof of personhood".[219]

This would have broader ramifications for online harms than simply reducing the number of bots operating on social media. The inability to recognise personhood online "underlies one of the most fundamental unsolved challenges in our technology ecosystem: preventing abusers from creating several (or many) fake identities — whether for fun, for profit, or to undermine democracy."[220] As Collins and Ford observe, the "ease of creating fake virtual identities plays an important role in shaping the way information and misinformation circulates online… because it makes it difficult to sanction rule-breakers".[221]

Several imperfect approaches could be utilised to achieve proof of personhood, with some already implemented for access to online banking, digital government services, and certain mobile devices and laptops. For example:

- linking online activity with an individual's identity using government-issued documentation such as a passport or biometric techniques such as iris and fingerprint scanning—consider the use of "Touch ID" on Apple iPhones and MacBooks; and
- utilising social trust principles "with participants in a digital network attesting that their connections' online identities are valid and not fake."[222]

Reminiscent of complications discussed in Section 1 under 'Balancing Act', the central challenge for proof of personhood initiatives is the trade-off between anonymity — important to democratic values such as freedom of expression and association and privacy — and accountability. Mandating proof of personhood may achieve meaningful accountability but do so at the expense of anonymity. Moreover, there are multiple methods available to circumventing proof of personhood initiatives including the use of generative AI tools to produce deepfakes — consider, for example, the requirement that a user submit a video selfie and this being circumvented by the submission of a deepfake — or less technologically advanced methods such as obtaining a fake passport on the black market.

Recommended reading

- For a recent examination of a social bot net on 'X' (formerly Twitter) found to utilise ChatGPT, see Yang, K-G. and Menczer, F. 'Anatomy of an AI-powered malicious social botnet' (July 2023).[223] Potential mitigations are briefly discussed on page 20.
- For an overview of using proof of personhood in response to social media risks and an explanation of the "pseudonym parties" approach (not discussed above), see Collins, A. and Ford, B. 'Using "proof of personhood" to tackle social media risks' (15 March 2021).[224]
- For deeper discussion of proof of personhood and related challenges see pages 284-289 in Ford, B. 'Technologizing Democracy or Democratizing Technology? A Layered-Architecture Perspective on Potentials and Challenges' in *Digital Technology and Democratic Theory* (University of Chicago Press, 2021).

Information Environment Architecture: Recommender Systems

This subsection considers algorithmic choice as a mitigation to the risks of recommender systems on digital platforms.

Algorithmic choice

One solution proposed to the potential harms of recommender systems is to empower citizens so that they play an active role in determining what they see on their social media newsfeeds. This is sometimes referred to as part of algorithmic sovereignty. Algorithmic sovereignty is a term with connotations extending beyond social media, reflecting the concern that citizens are not in control and may not even be aware of algorithms that have implications for their lives and wellbeing. Algorithmic sovereignty refers to "…the moral right of a person to be the exclusive controller of one's own algorithmic life and, more generally, the right and capacity by citizens as well as democratic institutions to make self-determined choices on personalization algorithms and related design choices".[225]

As Reviglio and Agosti explain, "[m]ainstream social media — especially Facebook — explicitly counteract any possibility for its participants to gain sovereignty: it denies all possibilities of participation in the decision-making process of its own algorithms, as well as strictly regulates the opportunities of interoperability for the data it gathers".[226] Writing for The New York Times in August 2023, Angwin observes, "[t]here is a growing worldwide movement to provide us with some algorithmic choice — from a Belgrade group demanding that recommender algorithms should be a "public good" to European regulators who are demanding that platforms give users at least one algorithm option that is not based on tracking user behaviour".[227]

As for how algorithmic choice could be implemented in practice, in 2020, Fukuyama and Schaake suggested that content moderation should be outsourced to "a layer of competitive middleware companies that would offer users of these platforms the ability to tailor their search and social media feeds to suit their personal preferences or objectives".[228] Critics of this proposal noted that it might worsen online misinformation – "[w]hile some middleware companies would filter out what the mainstream media thinks is fake news, other middleware options would intentionally accelerate fake news".[229]

Recommended reading

- For more information regarding the case for algorithmic sovereignty on social media, see Reviglio, U. and Agosti, C. 'Thinking Outside the Black-Box: The Case for "Algorithmic Sovereignty" in Social Media' (April 2020). [230]

General Purpose AI Models

This subsection discusses 'Constitutional AI', an emerging approach pioneered by Anthropic to steer the behaviour of LLMs in response to the unexpected capabilities and deployment safety problems discussed under 'General Purpose AI Models and Public Safety' above.

Constitutional AI

Research is emerging to develop "Constitutional AI", which, if successful, would leverage AI techniques to filter toxic data and shape the behaviour of LLMs according to normative principles such as those enshrined in the UN Declaration of Human Rights.[231] As Kyle Wiggers of TechCrunch explains, text-generating AI, including ChatGPT and other currently utilisable LLMs, have "massive flaws" because they are "often trained on questionable internet sources (e.g. social media)" and are therefore "biased in obviously sexist and racist ways". They also hallucinate, making up "answers to questions beyond the scope of [their] knowledge".[232]

In 2022, authors from Anthropic published a paper developing the concept of "Constitutional AI", suggesting how it might be implemented:

> "We would like to train AI systems that remain helpful, honest, and harmless, even as some AI capabilities reach or exceed human-level performance. This suggests that we will need to develop techniques that *do not rely on humans to supervise all aspects of AI behavior,* and that can be used to automatically test and enhance robustness to harmful behaviors. We also aim to develop methods that encode desirable AI behavior in a simple and transparent form, and that make it easier to understand and evaluate AI decision making" (emphasis added).[233]

The authors explain that the scale and complexity of emerging LLMs mean that review processes entirely dependent on human reviewers will become increasingly unfeasible. They therefore claim that technological methods to provide oversight for powerful AI systems, scaling the potential for their supervision, are in urgent need of development. "Constitutional AI" would essentially work as follows: one AI model is trained to critique and revise its own responses using whichever constitutional principles have been selected; this model then trains a "final model", using both the "AI-generated feedback based on the first model plus" the constitutional principles.[234]

According to Wiggers, "Constitutional AI" would also provide a level of transparency "...because it's easier to inspect the principles a system is following as well as train the system without needing humans to review disturbing content. That's a knock against OpenAI, which has been criticized in the recent past for underpaying contract workers to filter toxic data from ChatGPT's training data, including graphic details such as child sexual abuse and suicide".[235]

Several concerns arise in respect of the "Constitutional AI" approach, some of which are acknowledged by the Anthropic Authors in their 2022 paper. Firstly, AI systems that can control the behaviour of other AI systems have a dual-use risk in that the same methods could be leveraged to more effectively "train pernicious systems".[236] Secondly, if supervisory AI systems are effective, a situation may arise whereby it is widely perceived that human feedback is less important, leading to the propagation of LLMs (and other AI systems) "that have not been thoroughly tested and observed by humans".[237] A third concern relates to the selection of principles, such as the UN Declaration of Human Rights, that would underpin "Constitutional AI". Would selection be biased, for example, by Western conceptions of morality? And, given the potential consequences of AI systems and their algorithms for democracy, human rights and the wellbeing of citizens, are organisations such as Anthropic best placed to decide on constitutional principles, or is democratic oversight needed?

Anthropic is grappling with how to choose principles in a manner that their users will perceive as legitimate. One way may be to consult with them. In October 2023, Anthropic reported that it involved "~1,000 Americans to draft a constitution for an AI system".[238]

Recommended reading

- Read the 2022 paper in which Anthropic develops 'Constitutional AI' here: Bai, Y. et al., 'Constitutional AI: Harmlessness from AI Feedback' (December 2022).[239]

Emerging Domestic and Regional AI Regulation

Since the beginning of 2023, the political will to mount a rules-based response to the risks of AI technologies has accelerated alongside heavy lobbying and involvement of the AI sector itself. Political leaders appear in a rush to establish their own jurisdiction as the leader when it comes to governing this complex and uncertain period of technological change. This race should be welcomed by those who are interested in the responsible development and deployment of AI technologies. However, the devil is in the details. In many jurisdictions, consensus has not yet crystallised around the specifics, including the precise content and scope of rules and how they might be enforced. There are differences in the approaches of different jurisdictions. Some, such as the UK, can broadly be characterised as favouring a "light-touch" approach to regulation. Others, such as the EU, are pursuing binding and substantive legislation. However, the details are crucial as, at the time of writing, the EU AI Act negotiations have broken down over differing preferences among member states with respect to the regulation of foundation models (revisited below).

Additionally, local jurisdictions are facing political economy dynamics that threaten to undermine the efficacy of future AI rules. Firstly, AI models are made accessible online and, in some cases, open-sourced, meaning preventing their proliferation and use, even if that is desirable, is practically difficult for countries that value the openness of the internet. Secondly, major AI developers, including those responsible for the recent step-change in generative AI, are physically located in the US and China, potentially beyond the reach of lawmakers and enforcement agencies in other countries. Thirdly, to the extent countries other than the US and China wish to develop their own AI capacity with a view to providing alternative AI tools to citizens, a scarcity of infrastructure, hardware, expertise and critical components (especially semiconductors) hamper this ambition.[240]

Although competition and anti-trust concerns are not given detailed consideration in this Paper, the market concentration dynamics within the AI and broader technology sectors are another concerning aspect of the challenge. Lina Khan, the chair of the Federal Trade Commission in the US, recently opined in the New York Times:

> "The expanding adoption of A.I. risks further locking in the market dominance of large incumbent technology firms. A handful of powerful businesses control the necessary raw materials that start-ups and other companies rely on to develop and deploy A.I. tools. This includes cloud services and computing power, as well as vast stores of data. Enforcers and regulators must be vigilant. Dominant firms could use their control over these key inputs to exclude or discriminate against downstream rivals, picking winners and losers in ways that further entrench their dominance.
>
> Meanwhile, the A.I. tools that firms use to set prices for everything from laundry detergent to bowling lane reservations can facilitate collusive behavior that unfairly inflates prices — as well as forms of precisely

targeted price discrimination. Enforcers have the dual responsibility of watching out for the dangers posed by new A.I. technologies while promoting the fair competition needed to ensure the market for these technologies develops lawfully".[241]

Emerging intellectual property litigation against generative AI developers

Although this subsection relates to emerging regulation rather than litigation, it is noteworthy that developers of generative AI tools, including OpenAI, are the subject of a growing number of civil lawsuits (most commenced in late 2022 and 2023) based on existing legal protections. In the US, a group of authors, including George R.R Martin and Jodi Picoult, have filed class action complaints against OpenAI, claiming copyright infringement (among other claims).[242] They assert that their copyrighted texts were used to train OpenAI's LLMs without consent and that ChatGPT enables the unlawful creation of derivative works. Meanwhile, the defendants contend that LLMs represent a transformative innovation and therefore fall under fair use exemptions to US copyright law.[243] Secondly, visual artists have sued creators of AI-based image generation tools, such as Stable Diffusion, on the basis of analogous copyright grounds and publicity rights (among other claims). Finally, coders and software developers have sued the developers of AI tools like Codex and Copilot, which assist in the generation of code, for violating open-source licenses and the *Digital Millennium Copyright Act 1998* (among other claims).[245] The claimants allege the AI developers trained their models on code samples contributed to GitHub on an open-source basis but now fail to attribute these.

Below, a brief and non-exhaustive update is provided on the dynamic and evolving AI policy landscape across the European Union, the UK and the US as of late November 2023. The Santiago Declaration to Promote Ethical AI is also mentioned as an example of an important initiative from Latin American and Caribbean (**LAC**) countries.

European Union

The EU AI Act, heralded as the "first regulation on artificial intelligence", will in theory be a comprehensive AI law.[246] It was first proposed by the European Commission in April 2021. As of 20 November 2023, the EU AI Act is in the last stage of the trilogues process, namely, the interinstitutional negotiation between the European Commission, Council and Parliament. An emerging and significant area of deadlock in these negotiations relates to foundation models.

The original vision for the EU AI Act was that AI systems would be classified based on the level of risk they pose to end users.[247] Previous proposals for the EU AI Act envisaged that providers of AI systems and users would owe responsibilities differentiated according to risk classifications including unacceptable risk, high risk, limited risk, and minimal risk. For example, the European Parliament had suggested that AI systems capable of the following applications would be deemed to pose an "unacceptable risk":[248]

- "Cognitive behavioural manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behaviour in children".
- "Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics".
- "Real-time and remote biometric identification systems, such as facial recognition".

France, Germany and Italy recently spoke "out against the tiered approach initially envisaged on foundation models", rejecting "any regulation other than codes of conduct" for such models.[249] POLITICO reports that this push-back by Europe's three largest economies, the cause of the current deadlock, relates to their wish that the legislation should not hamper Europe's own development of foundation models.[250] France, Germany, and Italy seek that "AI companies working on foundation models regulate *themselves* by publishing certain information about their models and signing up to codes of conduct. There would initially be no punishment for companies that didn't follow these rules, though there might be in future if companies repeatedly violate codes of conduct" (emphasis added).[251]

AI policy experts have warned against this proposed softening of the EU AI Act. Connor Dunlop, the EU Public Policy Lead at the Ada Lovelace Institute, argues:

> "[It] would be irresponsible for the EU to cast aside regulation of large-scale foundation model providers to protect a couple of 'national champions'. Doing so would ultimately stifle innovation in the EU's AI ecosystem – of which downstream SMEs and startups are the vast majority. SMEs wishing to integrate or build on foundation models will not have the expertise, capacity or – importantly – access to the models to make their AI applications compliant with the AI Act.
>
> Model providers are significantly better placed to conduct robust safety testing, and only they are aware of the full extent of models' capabilities and shortcomings. It makes sense that obligations to conduct safety testing live with them, as these will benefit the thousands of downstream users of these systems."[252]

Beyond finalisation of the draft text, which currently "hangs in the balance",[253] there will be important questions regarding the framework for national implementation and enforcement of the EU AI Act.

United Kingdom

Currently, there is no "holistic body of law governing the development, deployment or use of AI in the UK".[254] Instead, those who develop, make available and use AI systems are subject to a "fragmented network of rules" including domain-specific regulation in sectors such as health and "cross-cutting" frameworks such as those which apply to data protection.[255] For example, as outlined in Section 2, the UK's *Online Safety Act 2023*, effective from October 26, 2023, criminalises the non-consensual sharing of intimate images, including deepfakes, through amendments to the *Sexual Offences Act 2003*.

In March 2023, the UK Government published a white paper outlining its "pro-innovation" approach to AI regulation and formal consultations ended on 21 June 2023.[256] The white paper suggests that there will be two main components to UK AI regulation: firstly, AI principles that existing UK government regulators will be in charge of implementing; and, secondly, new "central [government] functions" to support the same.[257]

On 1 and 2 November 2023, the UK Government hosted the AI Safety Summit. The Summit focussed narrowly on 'frontier AI', defined by the UK Government as "highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models".[258] The AI Safety Summit aimed to build consensus on "rapid, international action to advance safety at the frontier of AI technology".[259]

The main outcome of the UK AI Safety Summit was a declaration signed by 28 countries including China and the US. In the declaration, signatories commit to continue to meet in future for the purposes of cooperation, identifying AI risks, and building AI policies.[260]

One criticism that has arisen in respect of the Summit and its outcomes relates to their focus on frontier AI models and existential risks. This has been perceived as a failure to address "the real risks posed by today's AI systems".[261] However, other experts have suggested that critics should not "let the perfect be the enemy of the good", and that the alternative ("nothing at all") would not have been a better outcome.[262] One achievement of the UK Government in hosting the Summit was to bring China to the table, also a signatory to the November 2023 declaration. Martin opines:

> "Crucially, the Prime Minister made a very big call to invite China – and presumably secure American consent for that invitation. For all the (entirely legitimate) concerns about the horrific misuse of surveillance technology in the People's Republic, no global set of rules or principles would be worthy of the name without Beijing's signature. And excluding China from global discussions on AI would do nothing to prevent or slow down China's development of AI."[263]

United States

The status of AI governance in the US is disproportionately important for mitigating the negative impacts of AI across the globe because major private AI developers are concentrated within the jurisdictional control of the US Government. As the New York Times reported in July 2023, "[in] addition to industry giants like Google and Meta, the nine most valuable start-ups in generative A.I. are based in San Francisco or Silicon Valley, including OpenAI, Scale AI, Anthropic, Inflection AI, Databricks and Cerebras..."[264]

State and Federal Legislation

A patchwork of legislation relevant to AI is emerging across subnational jurisdictions as state legislatures have introduced "nearly 200 AI bills in 2023" (as of 12 October 2023).[265] Legislating AI at a federal level appears more difficult on account of the political polarisation that continues to hamstring congress.[266] Nevertheless, proposals exist. *The Algorithmic Accountability Act of 2023*, a bill proposed by Democratic Senator Ron Wyden with 11 other Democratic Senators, would require "companies to assess the impacts of the AI systems they use and sell" and empower the Federal Trade Commission to create regulatory guidelines for this purpose.[267] Both in 2018 and 2022, previous iterations of the bill were rejected after failing to pass.[268]

Executive Order and Action

On 30 October 2023, President Biden used his executive power to sign and publish a directive on "Safe, Secure and Trustworthy Artificial Intelligence" (**AI Executive Order**).[269, 270] Such executive orders are not legislation. Consequently, "they require no approval from Congress, and Congress cannot simply overturn them."[271] However, it is possible for Congress to hamper implementation of the executive order by, for example, "removing funding".[272]

The AI Executive Order is by no means toothless. It "leverages the Defense Production Act and other legal instruments to create binding requirements" and requires,[273] for example, that "developers of the most powerful AI systems share their safety test results and other critical information with the U.S. government", setting out a process for standards, tools and tests to ensure "AI systems are safe, secure and trustworthy".[274]

Under this process:

- The National Institute of Standards and Technology (**NIST**) will define strict standards for red-team testing AI systems.
- The Department of Homeland Security is to implement the NIST standards in critical infrastructure sectors, forming an AI Safety and Security Board.
- The Energy and Homeland Security Departments are directed to address AI-related threats in critical infrastructure, including chemical, biological, radiological, nuclear, and cybersecurity risks.

The AI Executive Order builds on the Biden Administration's Blueprint for an AI Bill of Rights (**Blueprint**), a set of five non-binding principles "for building and deploying automated systems that are aligned with democratic values and protect civil rights, civil liberties, and privacy".[275]

The AI Executive Order also directs NIST to develop a "comparison resource" to its existing AI Risk Management Framework focussing on generative AI.[276] Before the AI Executive Order was issued, NIST had long played a role in transnational standard-setting for the responsible development of AI technologies.[277]

An important strength of the AI Executive Order is that it extends the rights-based focus of the Blueprint by directing, among other actions, that:[278]

- Landlords, federal programs, and contractors receive detailed guidelines to prevent AI from increasing discriminatory practices.
- The Department of Justice and civil rights agencies enhance coordination and training to identify and address AI-related civil rights violations effectively. The criminal justice system adopts best practices to ensure equitable use of AI in sentencing, parole, pretrial processes, risk assessments, and forensic analysis.

Santiago Declaration to Promote Ethical AI

Twenty LAC countries have signed the Santiago Declaration to Promote Ethical AI, a regional commitment to promote responsible and ethical AI practices.[279] The Santiago Declaration builds upon the UNESCO Recommendation (discussed below) to establish a common set of principles for guiding the responsible development of AI. This is a significant and positive step for the LAC region, which aims to ensure that emerging technologies meet the specific needs of LAC countries and their citizens.

Global Rules

As the preceding discussion suggests, there is no absolute consensus among countries about how best to regulate AI and no comprehensive international AI regime. Should any such rules enter the corpus of international law and policy, their effectiveness will depend in large part on implementation by domestic lawmakers. Although a

comprehensive global treaty seems unlikely, not least of all because there is divergence in emerging domestic and regional approaches to AI regulation,[280] there is good sense in establishing an international regime for AI governance.

Trager et al. observe that the "potential harms of AI can… cross state borders. Many AI models are accessible online via either API access or an open-source version, which contributes to an immediate global impact."[281] In the same paper, which proposes a jurisdictional certification approach to the international governance of civilian AI systems (focussing on "frontier AI"), Trager et al. observe that regulating "AI on a country-by-country basis will likely lead to inadequate regulation in some jurisdictions and fragmented and disjointed regulation in others, hampering needed international collaboration on AI safety and global development".[282]

The G7 Hiroshima AI Process

Members of the G7, a grouping of seven advanced economies, including Canada, France, Germany, Italy, Japan, the UK, the US and the European Union, are undertaking intergovernmental political discussions on policy gaps and potential governance solutions in respect of AI, with the most urgent priority being the "responsible use of generative AI technologies".[283] These meetings have been dubbed the "Hiroshima AI Process".[284] On 30 October 2023, the G7 members agreed a voluntary code of conduct for organisations developing advanced AI systems.[285]

UN Office of the Secretary General's Envoy on Technology

The UN has recognised that renewed "multi-stakeholder efforts on global AI cooperation are needed to help build global capacity for the development and use of AI in a manner that is trustworthy, human rights-based, safe, sustainable and promotes peace".[286] Consequently, the UN Office of the Secretary General's Envoy on Technology is undertaking consultations, involving leading experts in AI and AI governance. Focus areas include key issues, current efforts, and potential models with respect to the global governance of AI.[287]

Recently, the UN Office of the Secretary General's Envoy on Technology appointed its High-Level Advisory Body on Artificial Intelligence.[288]

Council of Europe

The Strasbourg-based Council of Europe (**CoE**) (a different institution to the Council of the European Union), has been working on an international AI treaty focussed on aligning the "design, development and application of artificial intelligence systems" with "respect for human rights, the functioning of democracy and the observance of rule of law" (Article 1(1)). The CoE is comprised of 46 member states, 27 of which are members of the European Union.[289]

UNESCO Recommendation on the Ethics of AI

A UNESCO press release from November 2021 claims that member states then adopted "the first ever global agreement on the Ethics of Artificial Intelligence" (**UNESCO Recommendation**).[290] The term "global agreement" could be confused for "international treaty" and it is therefore better to describe the UNESCO Recommendation as the first international "normative framework" on AI to have been adopted by as many as 193 member states.

Although the UNESCO Recommendation may not be binding international law, it sets out influential recommendations,[291] which have been championed by countries including Chile and Senegal. Recently, twenty LAC countries agreed the Santiago Declaration to Promote Ethical AI (discussed above), which is in part based on the UNESCO Recommendation.[292]

Conclusion

This Paper provides an overview of salient issues connecting democracy, AI and public policy. Several themes emerge.

- Firstly, it is clear that AI is not one thing. Perceiving it as such undermines clarity of thinking about how to design new rules and identify gaps in existing rules. Particular applications of AI and more narrow or general systems raise different considerations for policymakers.
- Secondly, experts and commentators on AI policy have been caricatured as falling within two broad camps: those emphasising the long-term and existential risks of the most advanced AI models and those concerned with risks that have already crystallised in respect of a range of systems utilised in diverse sectors from housing to health. Hopefully commentators can agree that arguing about the right focal point for AI rules, when multiple focal points are achievable and, in some cases, mutually reinforcing, may be counterproductive. For example, Biden's recent AI Executive Order is framed both in terms of "chemical, biological, radiological, nuclear and cybersecurity risks" and advancing "equity and civil liberties".[293]
- Thirdly, policymakers should arm themselves with awareness about the political economy dynamics that shape what is feasible and effective in the policymaking arena. One of these dynamics is economic competition and the aspiration of individual countries to develop their AI sectors. Those favouring a "pro-innovation approach" to AI governance (including the United Kingdom, France, Italy and Germany) rightly or wrongly perceive a trade-off between catalysing their domestic AI revolutions to pursue economic and security gains and strict rules for AI developers.

In navigating these perceived or actual trade-offs, which should be interrogated thoroughly, the question policy practitioners are tasked to solve is not whether to prioritise the wellbeing, flourishing, and fundamental rights of citizens, but rather, how best to do so.

Endnotes

1 Jones, E. (2023). Explainer: What is a foundation model? Ada Lovelace Institute. Retrieved from: https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

2 Anderljung, M. et al. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety. Retrieved from: https://arxiv.org/abs/2307.03718.

3 IBM Research. (2023). What is generative AI? Retrieved from: https://research.ibm.com/blog/what-is-generative-AI.

4 Schmid, A (Ed.) (2011): The Routledge Handbook of Terrorism Research. London and New York: Routledge, pp. 73-87. An open source copy of the revised academic consensus definition of 2011 has been reproduced in Perspectives on Terrorism: Alex P. Schmid (2012): "The Revised Academic Consensus Definition of Terrorism". Perspectives on Terrorism, 6(2), pp. 158-159. URL: https://www.universiteitleiden.nl/PoT

5 Novelli, C., Taddeo, M. & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society. Retrieved from: https://doi.org/10.1007/s00146-023-01635-y.

6 Russell, S. & Norvig, P. (2020). Artificial Iintelligence: A Modern Approach. (4th ed). Pearson Higher Education. Retrieved from: https://aima.cs.berkeley.edu/global-index.html.

7 Bryson, J. (2020). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. In Dubber, M.D, Pasquale, F. & Das, S. (Eds.). The Oxford Handbook of Ethics of AI. Retrieved from: https://academic.oup.com/edited-volume/34287.

8 Girasa, R. & Scalabrini, G. J. (2022). Regulation of Innovative Technologies: Blockchain, Artificial Intelligence and Quantum Computing. Palgrave Macmillan. Retrieved from: https://link.springer.com/book/10.1007/978-3-031-03869-3.

9 Ibid.

10 Metz, C. (2023). Microsoft Says New A.I. Shows Signs of Human Reasoning. New York Times. Retrieved from: https://www.nytimes.com/2023/05/16/technology/microsoft-ai-human-reasoning.html; Rogers, R. (2023). What's AGI, and Why are AI Experts Skeptical? WIRED. Retrieved from: https://www.wired.com/story/what-is-artificial-general-intelligence-agi-explained/.

11 Fjelland, R. (2020). Why general artificial intelligence will not be realized. Humanities and Social Sciences Communications. Retrieved from: https://doi.org/10.1057/s41599-020-0494-4; Chiang, T. (2021). Why computers won't make themselves smarter. The New Yorker. Retrieved from: https://www.newyorker.com/culture/annals-of-inquiry/why-computers-wont-make-themselves-smarter.

12 Bubeck, S. et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. Retrieved from: https://arxiv.org/abs/2303.12712.

13 Ibid.

14 Ibid.

15 Bryson, J. (2020). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. In Dubber, M.D, Pasquale, F. & Das, S. (Eds.). The Oxford Handbook of Ethics of AI. Retrieved from: https://academic.oup.com/edited-volume/34287.Bryson, J.J. (2020). 'The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation', in Markus D. Dubber, Frank Pasquale, and Sunit Das (Eds.), The Oxford Handbook of Ethics of AI.

16 Bertuzzi, L. (2023). OECD updates definition of Artificial Intelligence 'to inform EU's AI Act'. Euractiv. Retrieved from: https://www.euractiv.com/section/artificial-intelligence/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/.

17 OECD. (2023). OECD AI Principles overview. Retrieved from: https://oecd.ai/en/ai-principles.

18 Economist Intelligence Unit. (2021). Democracy Index 2021. Retrieved from: https://www.eiu.com/n/campaigns/democracy-index-2021/.

19 See pages 69-78 of the EIU report.

20 United Kingdom Home Department. (2023). CONTEST:, The United Kingdom's Strategy for Countering Terrorism. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1186413/CONTEST_2023_English_updated.pdf.

21 European Commission – CORDIS. (2022). (n.d). Real-time Early Detection and Alert System for Online Terrorist Content based on Natural Language Processing, Social Network Analysis, Artificial Intelligence and Complex Event Processing. Retrieved from: https://cordis.europa.eu/project/id/740688.

22 Tech Against Terrorism. (2023). Early terrorist experimentation with generative artificial intelligence services. Retrieved from: https://techagainstterrorism.org/gen-ai.

23 Vincent, J. (2023). EU draft legislation will ban AI for mass biometric surveillance and predictive policing. The Verge. Retrieved from: https://www.theverge.com/2023/5/11/23719694/eu-ai-act-draft-approved-prohibitions-surveillance-predictive-policing.

24 Johnson, K. (2023). WIRED. Retrieved from: https://www.wired.com/story/face-recognition-software-led-to-his-arrest-it-was-dead-wrong/; Press, E. (2023). Does A.I. lead police to ignore contradictory evidence? The New Yorker. Retrieved from: https://www.newyorker.com/magazine/2023/11/20/does-a-i-lead-police-to-ignore-contradictory-evidence.

25 Hill, K. (2023). Eight months pregnant and arrested after false facial recognition match. The New York Times. Retrieved from: https://www.nytimes.com/2023/08/06/technology/facial-recognition-false-arrest.html.

26 Bhuiyan, J. (2023). TechScape: 'Are you kidding, carjacking?' – The problem with facial recognition in policing. The Guardian. Retrieved from: https://www.theguardian.com/newsletters/2023/aug/15/techscape-facial-recognition-software-detroit-porcha-woodruff-black-people-ai; National Institute for Standards and Technology. There's mode to AI Bias Than Biased Data, NIST Report Highlights. (2022). Retrieved from: https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights.

27 United Nations Office of Counter-Terrorism and United Nations Interregional Crime and Justice Research Institute. (2021). Countering Terrorism Online with Artificial Intelligence. Retrieved from: https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf.

28 United Nations General Assembly (UNGA). (2022). Report A/77/287. Retrieved from: https://digitallibrary.un.org/record/3987886?ln=en.

29 UNESCO. (2021). Windhoek +30 Declaration. Retrieved from: https://en.unesco.org/sites/default/files/windhoek30declaration_wpfd_2021.pdf.

30 Ibid.

31 Bundtzen, S. (2022). '"Suggested for You": Understanding How Algorithmic Ranking Practices Affect Online Discourses and Assessing Proposed Alternatives'. Institute for Strategic Dialogue. Retrieved from: www.isdglobal.org/isd-publications/suggested-for-you-understanding-how-algorithmic-ranking-practices-affect-online-discourses-and-assessing-proposed-alternatives/.

32 European Commission. (2023). The Digital Services Act package. Retrieved from: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

33 European Parliament. (2023). AI Act: a Step Closer to the First Rules on Artificial Intelligence. Retrieved from: https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence.

34 Floridi, L. et al. (2022). capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. Social Science Research Network. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091.

35 Goldstein, J. A. et al. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Retrieved from: https://arxiv.org/abs/2301.04246.

36 Simon, F., Altay, S. & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School: Misinformation Review. Retrieved from: https://misinforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/.

37 Meta (blog). (2021). Harmful content can evolve quickly. Our new AI system adapts to tackle it. Retrieved from: https://ai.meta.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/.

38 Chung, Y. L. et al. (2023). Understanding Counterspeech for Online Harm Mitigation. Retrieved from: https://arxiv.org/abs/2307.04761.

39 Novelli, C., Taddeo, M. & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society. Retrieved from: https://doi.org/10.1007/s00146-023-01635-y.

40 Ibid.

41 Manyika, J., Silberg, J. and Presten, B. (2019). What Do We Do About the Biases in AI? Retrieved from: https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai.

42 Bowman, S. R. (2023). Eight Things to Know about Large Language Models. Retrieved from: https://arxiv.org/abs/2304.00612.

43 Anderljung, M. et al. Frontier AI Regulation: Managing Emerging Risks to Public Safety. (2023). Retrieved from: https://arxiv.org/abs/2307.03718.

44 Ibid.

45 McBride, K. (2023). Dr Keegan McBride: Why the UK AI Safety Summit will fail to be meaningful. Oxford Internet Institute. Retrieved from: https://www.oii.ox.ac.uk/news-events/dr-keegan-mcbride-why-the-uk-ai-safety-summit-will-fail-to-be-meaningful/.

46 Vallance, B. C. (2023). More than 1,300 experts call AI a force for good. BBC News. Retrieved from: https://www.bbc.co.uk/news/technology-66218709.

47 UNICEF and Digital Public Goods Alliance. (2023). Core Considerations for Exploring AI Systems as Digital Public Goods. Retrieved from: https://digitalpublicgoods.net/AI-CoP-Discussion-Paper.pdf.

48 Cohen-Addad, V., Klein, P. N., Young, N. E. (2018). Balanced power diagrams for redistricting. Retrieved from: https://arxiv.org/abs/1710.03358.

49 Patel, J. et al. (2021). AI Brings Science to the Art of Policymaking. Boston Consulting Group. Retrieved from: https://www.bcg.com/publications/2021/how-artificial-intelligence-can-shape-policy-making.

50 For example, see: Ryan-Mosley, T. (2023). How AI can actually be helpful in disaster response. MIT Technology Review. Retrieved from: https://www.technologyreview.com/2023/02/20/1068824/ai-actually-helpful-disaster-response-turkey-syria-earthquake/.

51 Biever, C. (2023). ChatGPT broke the Turing test — the race is on for new ways to assess AI. Nature, 619(7971). Retrieved from: https://doi.org/10.1038/d41586-023-02361-7.

52 Valmeekam K. et al. (2023) Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). Retrieved from: https://arxiv.org/abs/2206.10498.

53 Kelly, S. M. (2023) ChatGPT passes exams from law and business schools. CNN Business. Retrieved from: https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html.

54 Surameery, N. M. S. & Shakor, M. Y. (2023). Use Chat GPT to Solve Programming Bugs. International Journal of Information Technology and Computer Engineering. Retrieved from: https://doi.org/10.55529/ijitc.31.17.22; Gewirtz, D. (2023). OK, so ChatGPT just debugged my code. For real. ZDNet. Retrieved from: https://www.zdnet.com/article/ok-so-chatgpt-just-debugged-my-code-for-real/.

55 Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. Reuters. Retrieved from: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

56 Teare, G. (2023). AI Was Q2's Big Hope To Reverse The Global Venture Funding Slowdown. It Wasn't Enough. Crunchbase News. Retrieved from: https://news.crunchbase.com/venture/vc-funding-falling-report-data-q2-2023-global/.

57 Satter, R. (2023). Exclusive: AI being used for hacking and misinformation, top Canadian cyber official says. Retrieved from: https://www.reuters.com/technology/ai-being-used-hacking-misinfo-top-canadian-cyber-official-says-2023-07-20/.

58 Ibid.

59 Goldstein, J. A. et al. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Retrieved from: https://arxiv.org/abs/2301.04246.

60 Ibid.

61 Europol. (2023). ChatGPT - The impact of Large Language Models on Law Enforcement. Europol Innovation Lab. Retrieved from: https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement.

62 Ibid.

63 Buchanan, B. et al. (2021). Truth, Lies, and Automation: How Language Models Could Change Disinformation. Center for Security and Emerging Technology (CSET). Retrieved from: https://doi.org/10.51593/2021ca003.

64 Ibid.

65 Meaning that the researchers actively sought to elicit the deceptive behaviour from the LLM.

66 Scheurer, J., Balesni, M. & Hobbhahn, M. (2023). Technical Report: Large Language Models can Strategically Decieve their Users when Put Under Pressure. Retrieved from: https://arxiv.org/pdf/2311.07590.pdf.

677 Goldstein, J. A. et al. (2023). Generative Language Models and Automated iInfluence Operations: Emerging Threats and Potential Mitigations. Retrieved from: https://arxiv.org/abs/2301.04246.

68 Simon, F., Altay, S. & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School: Misinformation Review. Retrieved from: https://misinforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/.

69 Ibid.

70 Sam Bowman (webpage). Retrieved from: https://cims.nyu.edu/~sbowman/.

71 Bowman, S. R. (2023). Eight Things to Know about Large Language Models. Retrieved from: https://arxiv.org/abs/2304.00612.

72 Ibid.

73 Meta. (2022). Introducing Make-A-Video: An AI system that generates videos from text. Retrieved from: https://ai.meta.com/blog/generative-ai-text-to-video/.2

74 Belanger, A. (2023). AI-faked images of Donald Trump's imagined arrest swirl on Twitter. Ars Technica. Retrieved from: https://arstechnica.com/tech-policy/2023/03/fake-ai-generated-images-imagining-donald-trumps-arrest-circulate-on-twitter/

75 Allyn, B. (2022). Deepfake video of Zelenskyy could be "tip of the iceberg" in info war, experts warn. NPR. Retrieved from: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

76 For more information on General Adversarial Networks see: https://www.youtube.com/watch?v=TpMIssRdhco.

77 For more information on Convolutional Neural Networks see: https://www.ibm.com/topics/convolutional-neural-networks.

78 N. Krueger, M. Vanamala, and R. Dave. Recent Advances in the Field of Deepfake Detection. (2023). Retrieved from: https://arxiv.org/pdf/2308.05563.pdf.

79 Horvitz, E. (2022). On the horizon: Interactive and compositional deepfakes. In Proceedings of the 2022 International Conference on Multimodal Interaction (pp. 653-661). Retrieved from: https://arxiv.org/pdf/2209.01714.pdf

80 Celli, F. (2020). Deepfakes Are Coming: Does Australia Come Prepared? Canberra Law Review, 17(2).

81 Salam, E. (2023) US mother gets call from 'kidnapped daughter' – but it's really an AI scam. Retrieved from: https://www.theguardian.com/us-news/2023/jun/14/ai-kidnapping-scam-senate-hearing-jennifer-destefano.

82 Dunn, S. (2021). Women, not politicians, are targeted most often by deepfake videos. Centre for International Governance Innovation. Retrieved from: https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/.

83 Siddique, H. (2023). Sharing deepfake intimate images to be criminalised in England and Wales. The Guardian. Retrieved from: https://www.theguardian.com/society/2023/jun/27/sharing-deepfake-intimate-images-to-be-criminalised-in-england-and-wales.

84 Ajder, H. et al. (2019). The State of Deepfakes: Landscape, Threats, and Impact. DeepTrace Labs. Retrieved from: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

85 Saliba, E. (2023). Sharing deepfake pornography could soon be illegal in America. ABC News. Retrieved from: https://abcnews.go.com/Politics/sharing-deepfake-pornography-illegal-america/story?id=99084399.3.

86 Ss 187 and 188, Online Safety Act 2023 (UK). Retrieved from: https://www.legislation.gov.uk/ukpga/2023/50/enacted.

87 Lumb, R. & Rule, M. (2023). New protections for victims of AI generated Non-Consensual Exploitative Images (A.K.A Deepfake Porn). Retrieved from: https://insights.smb.london/post/102ioil/new-protections-for-victims-of-ai-generated-non-consensual-exploitative-images-a.

88 Poritz, I. (2023). States Are Rushing to Regulate Deepfakes as AI Goes Mainstream. Bloomberg. Retrieved from: https://www.bloomberg.com/news/articles/2023-06-20/deepfake-porn-political-ads-push-states-to-curb-rampant-ai-use?leadSource=uverify%20wall.

89 Gans, J. (2023). NY Democrat unveils bill to criminalize sharing deepfake porn. The Hill. Retrieved from: https://thehill.com/homenews/house/3990659-ny-democrat-unveils-bill-to-criminalize-sharing-deepfake-porn/.

90 Europol. (2022). Facing reality? Law enforcement and the challenge of deepfakes: an observatory report. Europol Innovation Lab. Retrieved from: https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf.

91 Celli, F. (2020). Deepfakes Are Coming: Does Australia Come Prepared? Canberra Law Review, 17(2).

92 Greer, C. J. (2017). International Personality Rights and Holographic Portrayals. Indiana International and Competition Law Review. Retrieved from: https://mckinneylaw.iu.edu/iiclr/pdf/vol27p247.pdf.

93 Ibid; Celli, F. (2020). Deepfakes Are Coming: Does Australia Come Prepared? Canberra Law Review, 17(2).

94 Yu, M. (2022). SPC Releases Typical Cases on Protection of Personality Rights. China Justice Observer. Retrieved from: https://www.chinajusticeobserver.com/a/spc-releases-typical-cases-on-protection-of-personality-rights.

95 Horvitz, E. (2022). On the Horizon: Interactive and Compositional Deepfakes. Proceedings of the 2022 International Conference on Multimodal Interaction. Retrieved from: https://arxiv.org/pdf/2209.01714.pdf.

96 Ibid.

97 Oltermann, P. (2022). European politicians duped into deepfake video calls with mayor of Kyiv. The Guardian. Retrieved from: https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko.

98 Ibid.

99 Horvitz, E. (2022). On the Horizon: Interactive and Compositional Deepfakes. Proceedings of the 2022 International Conference on Multimodal Interaction. Retrieved from: https://arxiv.org/pdf/2209.01714.pdf.

100 Chesney, B. & Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. California Law Review, 107(6). Retrieved from: https://doi.org/10.15779/z38rv0d15j.

101 Ibid.

102 The Guardian (staff and agencies). (2023). Elon Musk's statements could be 'deepfakes', Tesla defence lawyers tell court. Retrieved from: https://www.theguardian.com/technology/2023/apr/27/elon-musks-statements-could-be-deepfakes-tesla-defence-lawyers-tell-court; Bond, S. (2023). People are trying to claim real videos are deepfakes. The courts are not amused. NPR. Retrieved from: https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused.

103 Pomerantsev, P. (2019). This is Not Propaganda: Adventures in the War Against Reality; Lewandowsky, S., Ecker, U.K.H & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. Journal of Applied Research in Memory and Cognition, 6(4). Available at: https://doi.org/10.1016/j.jarmac.2017.07.008; Coppins, M. (2020). The Billion Dollar Disinformation Campaign to Reelect the President. Retrieved from: https://www.theatlantic.com/magazine/archive/2020/03/the-2020-disinformation-war/605530/.

104 Reynolds, L. & Tuck, H. (2016). The Counter-Narrative Monitoring and Evaluation Handbook. Institute for Strategic Dialogue. Retrieved from: https://www.isdglobal.org/isd-publications/the-counter-narrative-monitoring-evaluation-handbook/; Chung, Y. L. et al. (2023). Understanding Counterspeech for Online Harm Mitigation. Retrieved from: https://arxiv.org/abs/2307.04761.

105 Reynolds, L. & Tuck, H. (2016). The Counter-Narrative Monitoring and Evaluation Handbook. Institute for Strategic Dialogue. Retrieved from: https://www.isdglobal.org/isd-publications/the-counter-narrative-monitoring-evaluation-handbook/.

106 Goldstein, J. A. et al. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Retrieved from: https://arxiv.org/abs/2301.04246; Horvitz, E. (2022). On the Horizon: Interactive and Compositional Deepfakes. In Proceedings of the 2022 International Conference on Multimodal Interaction. Retrieved from: https://arxiv.org/pdf/2209.01714.pdf.

107 Chung, Y. L. et al. (2023). Understanding Counterspeech for Online Harm Mitigation. Retrieved from: https://arxiv.org/abs/2307.04761.

108 Yang, K. C. & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. Observatory on Social Media (Indiana University). Retrieved from: https://arxiv.org/pdf/2307.16336.pdf.

109 Ryan-Mosley, T. (2023). The technology that powers the 2020 campaigns, explained. MIT Technology Review. Retrieved from: https://www.technologyreview.com/2020/09/28/1008994/the-technology-that-powers-political-campaigns-in-2020-explained/.

110 Ibid.

111 Burkell, J. & Regan, P. M. (2019). Voting preferences: Leveraging personal information to construct voter preference. In N. Witzleb, M. Paterson, & J. Richardson (Eds.), Big Data, Political Campaigning and the Law. Routledge.

112 Hankey, S., Naik, R., & Wright, G. (2019). Data and political campaigning in the era of big data – the UK experience. In N. Witzleb, M. Paterson, & J. Richardson (Eds.), Big Data, Political Campaigning and the Law. Routledge.

113 Cheshire, T. (2016). Behind the scenes at Donald Trump's UK digital war room. Sky News. Retrieved from: https://news.sky.com/story/behind-the-scenes-at-donald-trumps-uk-digital-war-room-10626155.

114 Ibid.

115 Confessore, N. (2018). Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. The New York Times. Retrieved from: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

116 Hern, A. (2018). Cambridge Analytica: how did it turn clicks into votes? The Guardian. Retrieved from: https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie.

117 Ibid.

118 Ibid.

119 Gibney, E. The scant science behind Cambridge Analytica's controversial marketing techniques. (2018). Retrieved from: https://www.nature.com/articles/d41586-018-03880-4.

120 Federal Trade Commission. (2019). FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook. Retrieved from: https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook.

121 Wong, J. C. (2019). Facebook to be fined $5bn for Cambridge Analytica privacy violations. The Guardian. Retrieved from: https://www.theguardian.com/technology/2019/jul/12/facebook-fine-ftc-privacy-violations.

122 Hankey, S., Naik, R. & Wright, G. (2019). Data and political campaigning in the era of big data – the UK experience. In N. Witzleb, M. Paterson, & J. Richardson (Eds.), Big Data, Political Campaigning and the Law. Routledge.

123 Ibid.

124 Krotoszynski, R. J. (2019). Big Data and the electoral process in the United States. In N. Witzleb, M. Paterson & J. Richardson (Eds.), Big Data, Political Campaigning and the Law. Routledge.

125 Ibid.

126 European Council. (2022). The general data protection regulation. Retrieved from: https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/.

127 For detailed analysis regarding these gaps, see Professor Maeve McDonagh's chapter in 'Big Data, Political Campaigning and the Law' (2020), titled 'Freedom of processing of personal data for the purpose of electoral activities after the GDPR'. Available at: https://www.routledge.com/Big-Data-Political-Campaigning-and-the-Law-Democracy-and-Privacy-in-the/Witzleb-Paterson-Richardson/p/book/9781032082554.

128 McDonagh, M. (2019). Freedom of processing of personal data for the purpose of electoral activities after the GDPR. In N. Witzleb, M. Paterson & J. Richardson (Eds.), Big Data, Political Campaigning and the Law. Routledge.

129 Ferrara, E., et al. (2016). The rise of social bots. Communications of the ACM, 59(7). Retrieved from: https://doi.org/10.1145/2818717.

130 Santini, R. M. & Salles, D. (2022). Bots. In A. Ceron (Ed.), Elgar Encyclopedia of Technology and Politics. Edward Elgar. Retrieved from: https://doi.org/10.4337/9781800374263.

131 IBM. (2022). The ultimate guide to machine-learning chatbots and conversational AI. Retrieved from: https://www.ibm.com/watson-advertising/thought-leadership/machine-learning-chatbot.

132 Santini, R. M. & Salles, D. (2022). Bots. In A. Ceron (Ed.), Elgar Encyclopedia of Technology and Politics. Edward Elgar. Retrieved from: https://doi.org/10.4337/9781800374263.

133 Colliver, C., King, J. & Maharasingam-Shah, E. (2020). Hoodwinked: Coordinated Inauthentic Behaviour on Facebook. Institute for Strategic Dialogue. Retrieved from: https://www.isdglobal.org/wp-content/uploads/2020/10/Hoodwinked-2.pdf.

134 Available at: https://about.fb.com/wp-content/uploads/2023/06/Meta-Quarterly-Adversarial-Threat-Report-Q1-2023.pdf.

135 Nimmo, B., Gleicher, N. & Franklin, M. (2023). Meta Quarterly Adversarial Threat Report – Q1 2023. Meta. Retrieved from: https://about.fb.com/wp-content/uploads/2023/06/Meta-Quarterly-Adversarial-Threat-Report-Q1-2023.pdf.

136 Ibid.

137 Ibid.

138 ISD publications investigating CIB and inauthentic tactics online include: Amplifying Far-Right Voices: A Case Study on Inauthentic Tactics Used by the Eric Zemmour Campaign. (2022). Available at: https://www.isdglobal.org/isd-publications/amplifying-far-right-voices-a-case-study-on-inauthentic-tactics-used-by-the-eric-zemmour-campaign/; How Eric Zemmour's election campaign used petitions to distort online support ahead of the French elections. (2022). Available at: https://www.isdglobal.org/digital_dispatches/eric-zemmours-far-right-campaign-distorted-online-support-ahead-of-the-french-elections-potentially-violating-platforms-policies/; Hoodwinked: Coordinated Inauthentic Behaviour on Facebook. (2020). Available at: https://www.isdglobal.org/wp-content/uploads/2020/10/Hoodwinked-2.pdf.

139 Yang, K. C. & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. Observatory on Social Media, Indiana University. Retrieved from: https://arxiv.org/pdf/2307.16336.pdf.

140 Gallwitz, F. & Kreil, M. (2021). The Rise and Fall of ‚Social Bot' Research. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3814191.

141 Kirchgaessner, S., et al. (2023). Revealed: the hacking and disinformation team meddling in elections. The Guardian. Retrieved from: https://www.theguardian.com/world/2023/feb/15/revealed-disinformation-team-jorge-claim-meddling-elections-tal-hanan.

142 Ibid.

143 Ibid.

144 Ibid.

145 Commercial Disinformation (explainer). (2023). Available at: www.isdglobal.org/explainers/commercial-disinformation-product-service/.

146 Yang, K. C. & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. Observatory on Social Media, Indiana University. ArXiv. Retrieved from: https://arxiv.org/pdf/2307.16336.pdf.

147 Ibid.

148 Ibid.

149 Ibid.

150 Ibid.

151 Ibid.

152 Milano, S., Taddeo, M. & Floridi, L. (2020). Recommender systems and their ethical challenges. AI & Society, 35(4). Retrieved from: https://doi.org/10.1007/s00146-020-00950-y

153 Vorotilov, V. & Shugaepov, I. (2023). Scaling the Instagram Explore recommendations system. Meta. Retrieved from: https://engineering.fb.com/2023/08/09/ml-applications/scaling-instagram-explore-recommendations-system/.

154 Li, Y., et al. (2023). Recent Developments in Recommender Systems: A Survey. Journal of Latex Class Files 14(8). Retrieved from: https://arxiv.org/pdf/2306.12680.pdf.

155 European Parliament. (2023). AI Act: a Step Closer to the First Rules on Artificial Intelligence. Retrieved from: https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence.

156 Milano, S., Taddeo, M. & Floridi, L. (2020). Recommender systems and their ethical challenges. AI & Society, 35(4). Retrieved from: https://doi.org/10.1007/s00146-020-00950-y.

157 Arguedas, A. M., et al. (2022). Echo chambers, filter bubbles, and polarisation: a literature review. Reuters Institute for the Study of Journalism, Oxford University. Retrieved from: https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review.

158 Milano, S., Taddeo, M. & Floridi, L. (2020). Recommender systems and their ethical challenges. AI & Society, 35(4). Retrieved from: https://doi.org/10.1007/s00146-020-00950-y.

159 Arguedas, A. M., et al. (2022). Echo chambers, filter bubbles, and polarisation: a literature review. Reuters Institute for the Study of Journalism, Oxford University. Retrieved from: https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review.

160 Ibid.

161 Ibid.

162 Toleffson, J. (2023). Tweaking Facebook feeds is no easy fix for polarization, studies find. Nature. Retrieved from: https://www.nature.com/articles/d41586-023-02420-z.

163 Ibid.

164 Heath, R & Fisher, S. (2023). Researchers fight to access Big Tech data. Axios. Retrieved from: https://www.axios.com/2023/08/01/researchers-tech-access-facebook-ai.

165 Regulation (EU) 2022/2065 (Digital Services Act), Article 31.

166 Heath, R & Fisher, S. (2023). Researchers fight to access Big Tech data. Axios. Retrieved from: https://www.axios.com/2023/08/01/researchers-tech-access-facebook-ai.

167 Falconer, R. (2023). Elon Musk's X Sues Hate Speech Watchdog. Axios. Retrieved from: https://www.axios.com/2023/08/01/twitter-x-sues-center-for-countering-digital-hate.

168 Suggested for You: Understanding How Algorithmic Ranking Practices Affect Online Discourses and Assessing Proposed Alternatives. (2022). Available at: https://www.isdglobal.org/isd-publications/suggested-for-you-understanding-how-algorithmic-ranking-practices-affect-online-discourses-and-assessing-proposed-alternatives/; Algorithms as a Weapon Against Women: How YouTube Lures Boys and Young Men into the 'Manosphere'. (2022). Available at: https://www.isdglobal.org/isd-publications/algorithms-as-a-weapon-against-women-how-youtube-lures-boys-and-young-men-into-the-manosphere/; Recommended Reading: Amazon's algorithms, conspiracy theories and extremist literature. (2021). Available at: https://www.isdglobal.org/isd-publications/recommended-reading-amazons-algorithms-conspiracy-theories-and-extremist-literature/.

169 Meta. (2021). Our New AI System to Help Tackle Harmful Content. Retrieved from: https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/.

170 Ibid.

171 Dilmegani, C. (2023). What is few-shot learning? AI Multiple. Retrieved from: https://research.aimultiple.com/few-shot-learning/.

172 Han, E. (2021). Advancing our approach to user safety. TikTok. Retrieved from: https://newsroom.tiktok.com/en-us/advancing-our-approach-to-user-safety.

173 N. Krueger, M. Vanamala & R. Dave. (2023). Recent Advances in the Field of Deepfake Detection. Available at: https://arxiv.org/pdf/2308.05563.pdf; M. S. Rana, M. N. Nobi, B. Murali, & A. H. Sung. (2022). Deepfake Detection: A Systematic Literature Review. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9721302.

174 Rana, M. S., Nobi, M. N., Murali, B. & Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. Retrieved from: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9721302.

175 Krueger, N., Vanamala, M. & Dave, R. (2023). Recent Advances in the Field of Deepfake Detection. Retrieved from: https://arxiv.org/pdf/2308.05563.pdf.

176 Lyu, S. (2020). Deepfakes and the new AI-Generated Fake Media Creation-Detection Arms race. Scientific American. Retrieved from: https://www.scientificamerican.com/article/detecting-deepfakes1/.

177 Ibid.

178 Anderljung, M. et al. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety. Retrieved from: https://arxiv.org/abs/2307.03718.

179 Ibid.

180 Ibid.

181 European Commission. (2021). Ethics By Design and Ethics of Use Approaches for Artificial Intelligence. Retrieved from: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf.

182 Bowman, S. R. (2023). Eight Things to Know about Large Language Models. Retrieved from: https://arxiv.org/abs/2304.00612.

183 Anderljung, M. et al. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety. Retrieved from: https://arxiv.org/abs/2307.03718.

184 Ibid.

185 Ibid.

186 Martin, C. (2023). Optimists, doomers and secure-pragmatists: Reflections on the UK's AI safety summit. Retrieved from: https://www.bsg.ox.ac.uk/blog/optimists-doomers-and-securo-pragmatists-reflections-uks-ai-safety-summit.

187 Ibid.

188 Ibid.

189 Ibid.

190 Fjeld, J. et al. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publications. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482 0

191 Kazim, E. & Koshiyama, A. S. (2021). A high-level overview of AI ethics. Patterns, 2(9). Retrieved from: https://doi.org/10.1016/j.patter.2021.100314 6

192 Novelli, C., Taddeo, M. & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society. Retrieved from: https://doi.org/10.1007/s00146-023-01635-y.

193 Bryson, J. (2020). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. In Dubber, M.D, Pasquale, F. & Das, S. (Eds.). The Oxford Handbook of Ethics of AI. Retrieved from: https://academic.oup.com/edited-volume/34287.

194 Schmitt, L (2021). Mapping global AI governance: a nascent regime in a fragmented landscape. AI and Ethics. Retrieved from: https://link.springer.com/article/10.1007/s43681-021-00083-y.

195 European Parliament. (2023). Artificial Intelligence Act. Legislative Train. Retrieved from: https://www.europarl.europa.eu/legislative-train/carriage/regulation-on-artificial-intelligence/report?sid=7301

196 Kazim, E. & Koshiyama, A. S. (2021). A high-level overview of AI ethics. Patterns, 2(9). Retrieved from: https://doi.org/10.1016/j.patter.2021.100314

197 Available at: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf.

198 N. Krueger, M. Vanamala, and R. Dave. Recent Advances in the Field of Deepfake Detection. (2023). Retrieved from: https://arxiv.org/pdf/2308.05563.pdf.

199 P. Fernandez, A. Chaffin, K. Tit, V. Chappelier and T. Furon. (2023). Three Bricks to Consolidate Watermarks for Large Language Models. Available at: https://arxiv.org/abs/2308.00113.

200 Sablayrolles, A., Douze, M., Schmid, C. & Jégou, H. (2020). Radioactive data: tracing through training. Available at: https://arxiv.org/abs/2002.00937; Goldstein, J. A. et al. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. Retrieved from: https://arxiv.org/abs/2301.04246.

201 Goujard, C. (2023). EU wants Google, Facebook to start labeling AI-generated content. POLITICO. Retrieved from: https://www.politico.eu/article/chatgpt-dalle-google-facebook-microsoft-eu-wants-to-start-labeling-ai-generated-content/.

202 See ISD's publication 'The Hydra on the Web: Challenges Associated with Extremist Use of the Fediverse – A Case Study of PeerTube' (May 2023). Available at: https://www.isdglobal.org/isd-publications/the-hydra-on-the-web-challenges-associated-with-extremist-use-of-the-fediverse-a-case-study-of-peertube/.

203 Coalition for Content Provenance and Authenticity. (n.d.). Overview. Retrieved from: https://c2pa.org.

204 Coalition for Content Provenance and Authenticity. C2PA Technical Specification (1.2). Retrieved from: https://c2pa.org/specifications/specifications/1.2/specs/C2PA_Specification.html.

205 Available at: https://arxiv.org/pdf/2209.01714.pdf.

206 Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954.

207 Available at: https://arxiv.org/abs/2301.04246.

208 Available at: https://www.researchgate.net/publication/358835043_Deepfake_Detection_A_Systematic_Literature_Review.

209 Dobber, T., Fathaigh, R. Ó. & Borgesius, F. J. Z. (2019). The regulation of online political micro-targeting in Europe. Internet Policy Review, 8(4). Retrieved from: https://doi.org/10.14763/2019.4.1440.

210 McDonagh, M. (2019). Freedom of processing of personal data for the purpose of electoral activities after the GDPR. In N. Witzleb, M. Paterson & J. Richardson (Eds.), Big Data, Political Campaigning and the Law. Routledge.

211 Krotoszynski, R. J. (2019). Big Data and the electoral process in the United States. In N. Witzleb, M. Paterson & J. Richardson (Eds.), Big Data, Political Campaigning and the Law. Routledge.

212 Australian Human Rights Commission. (n.d.). How are human rights protected in Australian Law? Retrieved from: https://humanrights.gov.au/our-work/rights-and-freedoms/how-are-human-rights-protected-australian-law.

213 *Australian Capital Television Pty Ltd v Commonwealth* (1992) 177 CLR 106.

214 European Parliament (2021). Proposal for a Regulation on the transparency and targeting of political advertising. Retrieved from: https://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2021/0731/COM_COM(2021)0731_EN.pdf

215 European Parliament. (2023). MEPs Toughen Rules on Political Advertising. Retrieved from: https://www.europarl.europa.eu/news/en/press-room/20230123IPR68616/meps-toughen-rules-on-political-advertising.

216 European Parliament. (2023). AI Act: a Step Closer to the First Rules on Artificial Intelligence. Retrieved from: https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence.

217 Available at: https://www.medialaws.eu/political-advertising-and-disinformation-the-european-approach/.

218 Yang, K. C. & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. Observatory on Social Media, Indiana University. Retrieved from: https://arxiv.org/pdf/2307.16336.pdf.

219 Goldstein, J. A. et al. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Retrieved from: https://arxiv.org/abs/2301.04246.

220 Ford, B. (2021). Technologizing Democracy or Democratizing Technology? A Layered-Architecture Perspective on Potentials and Challenges. In Bernholz, L. (Ed.), Digital Technology and Democratic Theory. Oxford University Press.

221 Collins, A. & Bryan, F. A. (2021). Using "proof of personhood" to tackle social media risks. International Risk Governance Center (IRGC). Retrieved from: https://doi.org/10.5075/epfl-irgc-283872.

222 Ibid.

223 Available at: https://arxiv.org/abs/2307.16336.

224 Available at: https://www.epfl.ch/research/domains/irgc/spotlight-on-risk-series/using-proof-of-personhood-to-tackle-social-media-risks/.

225 Reviglio, U. & Agosti, C. (2020). Thinking Outside the Black-Box: The Case for "Algorithmic Sovereignty" in Social Media. Social Media + Society, 6(2). Retrieved from: https://doi.org/10.1177/2056305120915613.

226 Ibid.

227 Angwin, J. (2023). What if You Knew What You Were Missing on Social Media? The New York Times. Retrieved from: https://www.nytimes.com/2023/08/17/opinion/social-media-algorithm-choice.html.

228 Miller, K. (2021). Radical Proposal: Middleware Could Give Consumers Choices Over What They See Online. Stanford University Human-Centered Artificial Intelligence. Retrieved from: https://hai.stanford.edu/news/radical-proposal-middleware-could-give-consumers-choices-over-what-they-see-online.

229 Ibid.

230 Available at: https://journals.sagepub.com/doi/full/10.1177/2056305120915613.

231 Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. Anthropic. Retrieved from: https://arxiv.org/pdf/2212.08073.pdf.

232 Wiggers, K. (2023). Anthropic Thinks Constitutional AI is the Best Way to Train Models. Tech Crunch. Retrieved from: https://techcrunch.com/2023/05/09/anthropic-thinks-constitutional-ai-is-the-best-way-to-train-models/.

233 Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. Anthropic. Retrieved from: https://arxiv.org/pdf/2212.08073.pdf.

234 Wiggers, K. (2023). Anthropic Thinks Constitutional AI is the Best Way to Train Models. Tech Crunch. Retrieved from: https://techcrunch.com/2023/05/09/anthropic-thinks-constitutional-ai-is-the-best-way-to-train-models/.

235 Ibid.

236 Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. Anthropic. Retrieved from: https://arxiv.org/pdf/2212.08073.pdf.

237 Ibid.

238 Anthropic. Collective Constitutional AI: Aligning a Language Model with Public Input. (2023). Retrieved from: https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input.

239 Available at: https://arxiv.org/pdf/2212.08073.pdf.

240 Vipra, J & West, S. M. (2023). Computational Power and AI. AI Now Institute. Retrieved from: https://ainowinstitute.org/publication/policy/compute-and-ai.

241 Khan, L. (2023). Lina Khan: We Must Regulate A.I. Here's How. Retrieved from: https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html.

242 Authors Guild v. OpenAI Inc., Case No 1:23-cv-08292; Tremblay v. OpenAI Inc., Case No. 3:23-cv-03223.

243 Ibid.

244 Andersen, McKernan, and Ortiz v. Stability AI Ltd and others, Case No. 3:23-cv-00201.

245 Doe 3 v. GitHub, Inc., Case No. 4:22-cv-07074.

246 European Parliament. (2023). EU AI Act: first regulation on artificial intelligence. Retrieved from: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

247 European Parliament. (2023). EU AI Act: first regulation on artificial intelligence. Retrieved from: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

248 Ibid.

249 Bertuzzi, L. (2023). France, Germany, Italy push for 'mandatory self-regulation' for foundation models in EU's AI law. Retrieved from: https://www.euractiv.com/section/artificial-intelligence/news/france-germany-italy-push-for-mandatory-self-regulation-for-foundation-models-in-eus-ai-law/.

250 Volpicelli, G. (2023). Power grab by France, Germany and Italy threatens to kill EU's AI bill. Retrieved from: https://www.politico.eu/article/france-germany-power-grab-kill-eu-blockbuster-ai-artificial-intelligence-bill/.

251 Henshall, W. (2023). E.U.'s AI Regulation Could Be Softened After Pushback From Biggest Members. Time. Retrieved from: https://time.com/6338602/eu-ai-regulation-foundation-models/.

252 Dunlop, C. (2023). Regulating AI foundation models is crucial for innovation. Euractiv. Retrieved from: https://www.euractiv.com/section/artificial-intelligence/opinion/regulating-ai-foundation-models-is-crucial-for-innovation/.

253 Volpicelli, G. (2023). Power grab by France, Germany and Italy threatens to kill EU's AI bill. Retreived from: https://www.politico.eu/article/france-germany-power-grab-kill-eu-blockbuster-ai-artificial-intelligence-bill/.

254 Davies, M. & Birtwistle, M. (2023). Regulating AI in the UK. Ada Lovelace Institute. Retrieved from: https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/.

255 Ibid.

256 Department for Science, Innovation, and Technology. (2023). A pro-innovation approach to AI regulation. Retrieved from: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach.

257 Davies, M. & Birtwistle, M. (2023). Regulating AI in the UK. Ada Lovelace Institute. Retrieved from: https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/.

258 Department for Science, Innovation & Technology (UK). AK Safety Summit: introduction (HTML). Retrieved from: https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html.

259 Government of the United Kingdom. (2023). Iconic Bletchley Park to host UK AI Safety Summit in early November. Retrieved from: https://www.gov.uk/government/news/iconic-bletchley-park-to-host-uk-ai-safety-summit-in-early-november.

260 The full declaration, published on 1 November 2023, can be read here: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

261 Mittelstadt, B. (2023). Expert comment: Oxford AI experts comment on the outcomes of the UK AI Safety Summit. Retrieved from: https://www.ox.ac.uk/news/2023-11-03-expert-comment-oxford-ai-experts-comment-outcomes-uk-ai-safety-summit.

262 Martin, C. (2023). Expert comment: Oxford AI experts comment on the outcomes of the UK AI Safety Summit. Retrieved from: https://www.ox.ac.uk/news/2023-11-03-expert-comment-oxford-ai-experts-comment-outcomes-uk-ai-safety-summit.

263 Martin, C. (2023). Optimists, doomers and securo-pragmatists: Reflections on the UK's AI safety summit. Retrieved from: https://www.bsg.ox.ac.uk/blog/optimists-doomers-and-securo-pragmatists-reflections-uks-ai-safety-summit.

264 Lohr, S. (2023). Best Place for A.I. Jobs (New Report Says) Won't Surprise You. New York Times. Retrieved from: https://www.nytimes.com/2023/07/20/business/ai-jobs-bay-area-brookings-institution-report.

265 Elkins, D. & Swanson, S.A. (2023). Federal policymakers: chasing the runaway AI train. National Law Review. Retrieved from: https://www.natlawreview.com/article/federal-policymakers-chasing-runaway-ai-train.

266 Reuters. (2023). House speaker unveils Republican plan to avert government shutdown. Retreived from: https://www.theguardian.com/us-news/2023/nov/12/house-speaker-unveils-republican-stopgap-bill-to-avert-government-shutdown.

267 Office of Senator Ron Wyden. (2023). Algorithmic Accountability Act of 2023 Summary. Retrieved from: https://www.wyden.senate.gov/imo/media/doc/algorithmic_accountability_act_of_2023_summary.pdf.

268 GovTrack. (2022). S.3572 (117th): Algorithmic Accountability Act of 2022. Retrieved from: https://www.govtrack.us/congress/bills/117/s3572; Hilliard, A. (2023). US Algorithmic Accountability Act: Third Time Lucky? Retrieved from: https://www.holisticai.com/blog/us-algorithmic-accountability-act.

269 The White House. (2023). FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. Retrieved from: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

270 The full AI Executive Order can be read here: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

271 American Bar Association. (2021). What is an Executive Order? Retrieved from: https://www.americanbar.org/groups/public_education/publications/teaching-legal-docs/what-is-an-executive-order-/.

272 Ibid.

273 Trager, R. (2023). Expert comment: Oxford AI experts comment on the outcomes of the UK AI Safety Summit. Retrieved from: https://www.ox.ac.uk/news/2023-11-03-expert-comment-oxford-ai-experts-comment-outcomes-uk-ai-safety-summit.

274 The White House. (2023). FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. Retrieved from: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

275 The White House. (2023). What is the Blueprint for an AI Bill of Rights? Retrieved from: https://www.whitehouse.gov/ostp/ai-bill-of-rights/what-is-the-blueprint-for-an-ai-bill-of-rights/.

276 National Institute of Standards and Technology. (2023). Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. Retrieved from: https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence.

277 More information about the NIST AI Risk Management Framework is accessible here: https://www.nist.gov/itl/ai-risk-management-framework.

278 The White House. (2023). FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. Retrieved from: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

279 Cumbre Ministerial y de Altas Autoridades de América Latina y el Caribe. Declaracion de Santiago. (2023). Retrieved from: https://minciencia.gob.cl/uploads/filer_public/40/2a/402a35a0-1222-4dab-b090-5c81bbf34237/declaracion_de_santiago.pdf.

280 Hodge, N. (2023). Regulatory divergence presents obstacles for legal teams navigating AI. International Bar Association. Retrieved from: https://www.ibanet.org/regulatory-divergence-presents-obstacle-for-legal-teams-navigating-ai.

281 Trager, R. F. et al. International Governance of Civilian AI. (2023). Retrieved from: https://www.oxfordmartin.ox.ac.uk/publications/international-governance-of-civilian-ai-a-jurisdictional-certification-approach/.

282 Ibid.

283 OECD. (2023). G7 Hiroshima Process on Generative Artificial Intelligence. Retrieved from: https://doi.org/10.1787/bf3c0c60-en

284 Chee, F. Y. (2023). Exclusive: G7 to agree AI code of conduct for companies. Retrieved from: https://www.reuters.com/technology/g7-agree-ai-code-conduct-companies-g7-document-2023-10-29/.

285 The G7 AI code of conduct can be read here: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems.

286 Office of the Secretary-General's Envoy on Technology. (2023). Multistakeholder Advisory Body on Artificial Intelligence. Retrieved from: https://www.un.org/techenvoy/content/artificial-intelligence.

287 Ibid.

288 More information about this group and their aims is available here: https://www.un.org/ai-advisory-body.

289 A revised draft of the CoE's treaty was published on 6 January 2023 and is accessible here: https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f.

290 UNESCO. (2021). UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence. Retrieved from: https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence.

291 UNESCO. (2020). Ad Hoc Expert Group for the Preparation of a Draft text of a Recommendation on the Ethics of Artificial Intelligence. Outcome document: first draft of the Recommendation on the Ethics of Artificial Intelligence. Retrieved from: https://unesdoc.unesco.org/ark:/48223/pf0000373434.

292 The full text of the UNESCO Recommendation is accessible here: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

293 The White House. (2023). FACT SHEET: President Biden Issues Executive Order on Safe, Secure and Trustworthy Artificial Intelligence. Retrieved from: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

Copyright © Institute for Strategic Dialogue (2024). Institute for Strategic Dialogue (ISD) is a company limited by guarantee, registered office address PO Box 75769, London, SW1P 9ER. ISD is registered in England with company registration number 06581421 and registered charity number 1141069. All Rights Reserved.

www.isdglobal.org