Institute of Art and Ideas

# AI threatens elections and accountable governance

Deepfakes are a danger to this year's elections

**2nd May 2024**

Nayef Al-Rodhan

| Philosopher, neuroscientist, geostrategist and futurologist. He is an Honorary Fellow of St. Antony's College, Oxford University; Head of the Geopolitics and Global Futures Department at the Geneva Center for Security Policy in Switzerland

*With elections in America, the UK, India, and more, this is a huge year for global politics. While bots were a cause for concern as far back as the 2016 election, technology has come a long way since then. With the rise of Generative AI, the world's elections face an existential threat from deepfakes, disinformation and misinformation. To meet this threat, Nayef Al-Rodhan argues that we cannot rewire our brains to be able to spot misinformation. Therefore, the impetus relies on regulation and societal education.*

Will artificial intelligence-generated deepfakes provide a "perfect storm" for malicious actors looking to hijack forthcoming local and general elections, as the British Home Secretary James Cleverly warned recently? Sophisticated deepfakes are becoming a global problem, and an urgent one at that. As an estimated 2 billion people head to the polls this year, there has hardly been a worse time to allow harmful content to flourish online.

The toxic mixture of increasingly sophisticated AI tools and flimsy prevention measures means that we could soon be faced with a situation where a viral deepfake dismantles democratic and governance processes. Policy and tech circles are starting, slowly, to wake up to the problem. This month, European political parties signed a voluntary code of conduct aimed at preventing the creation and dissemination of unlabeled deepfakes ahead of the European elections in June. Earlier this year, Silicon Valley bosses pledged to prevent AI-generated content from interfering with global elections this year. At this year's Munich Security Conference, Amazon, Google, Meta, Microsoft, TikTok and OpenAI were among 20 tech companies which agreed to work together to combat the creation and spread of deepfake images, videos and audio designed to mislead voters.

These are much-needed initiatives. But to properly shield ourselves from a disinformation doomsday scenario, we need to grapple with the true reach and effect of these misleading campaigns and develop a deeper understanding of our neuro-behavioural susceptibility to these sorts of attacks. We must also ask ourselves some uncomfortable but important questions, starting with: why are these methods so effective and what can be done to dilute their influence on societies?

Online disinformation has been an irritant in elections and political systems for many years. According to the Oxford Internet Institute, social media disinformation campaigns had operated in more than 80 countries by 2020. But rapid advances in AI technology mean that it is now easier than ever to manipulate media and public opinion through both disinformation (fake news that is created and spread deliberately by someone who knows that it is false) and misinformation (fake news that is created and spread by mistake, by someone who doesn't realise that it is false). This is largely due to the advent of Generative AI, powerful multi-modal models that can combine text, image, audio and video. These new tools have amplified social media disinformation in unprecedented ways. Generative AI models can help bad actors tailor messaging so that it resonates with target audiences. This has transformed the generation and dissemination of very realistic deepfakes (understood as synthetic media that have been digitally manipulated to substitute one person's likeness with that of another).

It is therefore perhaps unsurprising that the most recent World Economic Forum Global Risks Perception Survey showed that AI-generated misinformation is seen as one of the biggest risks of 2024, a year packed with presidential and parliamentary elections around the world. Elections, with their emotionally charged and often tribal dynamics, are fertile ground for disinformation and misinformation campaigns designed to sway opinion, discredit candidates and, more broadly, undermine trust in accountable governance and democracy. Last year we saw how realistic deepfakes emerged as a new front in the Russia-Ukraine and Israel-Hamas conflicts. We are now seeing deepfakes becoming increasingly prominent in the political arena, most recently causing confusion in Slovakia's election as well as on Bangladeshi social media). Deepfakes are starting to muddy the waters in electoral processes already plagued by political polarisation and dwindling public trust in governments, institutions and democracy. This has created a "liar's dividend": the very existence of deepfakes deepens mistrust in everything online, even if it is real.

*This battle for online authenticity is also linked to questions of a philosophical nature, not least: who decides what is true and what is not? At the moment, the social media companies are the arbiters of truth.*

This begs the question: why are deepfakes and other AI-driven disinformation methods so effective? A neurophilosophical understanding of human nature offers some clues. As I have argued elsewhere, insights from neuroscience demonstrate the extraordinary salience of emotions to human existence, as human nature is rooted in emotionality, amorality and egoism. Emotionality, in particular, overlaps with a large part of our cognitive functions, including decision-making. This means that to counter the appeal of fake news, we need to develop a deeper appreciation of our emotional, amoral and egoistic nature, as well as our profound need for human dignity. Human dignity is fundamental to human nature and is rooted in nine fundamental ‘needs’: reason, security, human rights, accountability, transparency, justice, opportunity, innovation, and inclusiveness. A dignity deficit makes us more likely to generate or believe deepfakes, and makes us more susceptible to disseminating them. This can be explained in neuroscientific terms. By using modern neuroimaging tools we can tell that the prefrontal cortex (which contains the logical part of the brain) plays second fiddle when we read the news: a social media post containing a convincing deepfake about a politician, for example, is more likely to be shared if it connects to the emotional side of our brain, the social part. Neurochemically, we are hardwired to use social media, which is the most commonly used platform for deepfakes. The more “likes'' that we receive, the more dopamine hits we get – ultimately feeding our addiction to social media. This is a worrying reminder that we are not automatically driven by rational calculations or critical thinking when it comes to spotting deepfakes and sophisticated disinformation campaigns.

*We cannot rewire our brains to resist the pull of misinformation.*

Government and multilateral initiatives tasked with creating regulatory frameworks on AI tools are, belatedly, starting to pick up momentum. These include the Biden Administration’s recent executive order on AI; the AI Safety Summit, held in the United Kingdom last November; a new AI advisory board at the United Nations; and the European Union’s AI Act, expected to come into force by 2025. Authorities are also becoming increasingly alert to the dangers that deepfakes and other AI disinformation tools can have on the political health of nations. Last November, GCHQ, the UK's intelligence, security and cyber agency, warned about “AI-created hyper-realistic bots” ahead of the parliamentary election later this year. Meanwhile in the U.S., a bipartisan group of senators recently proposed legislation to ban “materially deceptive AI-generated” content in political advertising. But there is still much more that can - and should - be done to guard against nefarious actors using social media platforms such as Meta, Google’s YouTube, TikTok and X to distribute deepfakes. For these companies, working at speed will be essential, as will larger investments in high-quality detection capabilities. But many of the social media giants have been hampered by mass layoffs following the tech downturn last year, which has hampered content moderation resources. This arguably makes the platforms even less equipped to meet the task at hand.

This battle for online authenticity is also linked to questions of a philosophical nature, not least: who decides what is true and what is not? At the moment, the social media companies are the arbiters of truth. That should be a cause for concern given that they are primarily driven by the pursuit of revenues, profit and potential influence (as the turmoil within OpenAI's boardroom late last year reminded us). Beyond tighter regulation and clear-eyed non-partisan policies, we need a wide transdisciplinary coalition to keep governance accountable and protect society against the perils of sophisticated forms of online disinformation, such as deepfakes. By joining forces within a framework I have called Neuro-Techno-Philosophy (NTP), tech-savvy philosophers, neuroscientists, social scientists, policy-makers, as well as experts from AI and other disruptive and intrusive technologies can help us get to grips with the ethical and societal implications of the impending transformations caused by Generative AI. Whether we succeed or not will depend in large part on how well we understand the predilections of human nature and our hard-wired neurochemical gratifications. We cannot rewire our brains to resist the pull of misinformation. That is why our success will also hinge on wider societal awareness and the ability of governments to introduce strict regulatory oversight mechanisms to safeguard accountable and transparent governance paradigms, societal cohesion and human dignity needs. Neglecting these dignity needs will come at a perilous price for humanity.