VOLUME 5 · ISSUE 06

BALSILLIE PAPERS

Artificial Intelligence and New Threat Vectors: Using Scenario Planning for Trend Forecasting

Ann Fitz-Gerald, Dmytro Chumachenko, Halyna Padalko and Vijay Ganesh

October 2, 2023

The discourse on AI has increasingly shifted from focusing on its potential as a tool to its possibilities as an autonomous entity. As AI technology evolves rapidly, concerns over its potential threats and the necessity for regulatory mechanisms have become central to the global conversation. Even inside the AI community, among technology business people, scientists, opinion leaders and even creators of AI, there is no consensus on the most optimal and efficient ways to govern it and mitigate potential threats.

On June 22, 2023, at Toronto's Roy Thomson Hall, more than two-thirds of the Munk Debate audience confirmed that artificial intelligence (AI) poses an existential threat to humanity.[1] The ensuing debates notwithstanding, sentiments do not appear to have changed. Also in 2023, many leading AI researchers and entrepreneurs, including Geoffrey Hinton and Elon Musk, wrote a cautionary letter about the potential negative impact of AI and asked for a six-month worldwide moratorium on research in generative AI.

Anxiety levels are rising and with good reason. For example, recently, researchers have found ways to compromise the security and privacy of users using generative AI tools. Another example is the strike activity of Hollywood writers and actors in the summer of 2023, demanding that studios not use generative AI to replace them.

There is an urgent need for action. On July 26, 2023, leading technology companies in the AI sector announced their collaboration to establish a forum to promote safe and responsible AI development.[2] The "Frontier Model Forum" was spearheaded by Microsoft, Google, OpenAI (founder of ChatGPT) and the AI research entity Anthropic. While legislators in the United States[3] and worldwide rush to grasp the rapidly growing AI field and implement appropriate safeguards, this measure is timely.

The discourse on AI has increasingly shifted from focusing on its potential as a tool to its possibilities as an autonomous entity. As AI technology evolves rapidly, concerns over its potential threats and the necessity for regulatory mechanisms have become central to the global conversation. Even inside the AI community, among technology business people, scientists, opinion leaders and even creators of AI, there is no consensus on the most optimal and efficient ways to govern it and mitigate potential threats.

Figure 1: Munk Debate on AI on June 22, 2023, Roy Thomson Hall

[1] Munk Debates. "Artificial Intelligence." Munkdebates.com. (2023). https://munkdebates.com/debates/artificial-intelligence.
[2] OpenAI. "Frontier Model Forum." Openai.com. July 26, 2023. https://openai.com/blog/frontier-model-forum.
[3] The White House. "Blueprint for an AI Bill of Rights." The White House (2022). https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

Geoffrey Hinton, a pioneer in deep learning, highlighted the need for enhanced AI interpretability and insisted on integrating ethical considerations into AI systems.[4] His former student Yoshua Bengio — another pioneering figure in deep learning — has repeatedly emphasized the importance of understanding AI's capability to learn from small amounts of data and proposed the need to move toward systems that can better comprehend the real-world environment and its subtleties.[5] Bengio urged a balanced approach to AI evolution that seeks to harness the technology's potential while acknowledging and mitigating its associated risks. Yann LeCun, a computer scientist known for his work on convolutional neural networks and deep learning, also recognizes the importance of AI learning in a more human-like manner, but warns of the potential dangers of anthropomorphizing AI.[6] As such, LeCun stresses the importance of self-supervised learning, where AI learns from real-world interactions.

Cosmologist-turned-AI-safety advocate, Max Tegmark warns of AI's existential threat to humanity if not adequately regulated.[7] He posits that the future of AI needs to be designed to benefit all of humanity instead of being a competitive race without safety precautions. Tegmark's approach encapsulates the idea that the future of AI should be a cooperative effort. Stuart Russell, a British computer scientist, also underscores the importance of value alignment between AI and human beings. He calls for a comprehensive redesign of current AI systems to create AI that inherently values human preferences and operates within an ethical framework (Woods-Robinson, 2022). Philosopher and historian Emile Torres questions the notion of linking AI with existential threats, and instead calls for more critical inquiry into how these warnings have been picked up and propagated by the news, as they distract from the genuine harms that some AI companies are currently causing, especially to marginalized communities.[8]

These perspectives collectively highlight a shared recognition of the need for a balanced approach to AI evolution — one that capitalizes on its immense potential while addressing its inherent risks through regulation, transparency, interpretability and ethical alignment. Striking this balance demands a thorough understanding of the inherent risks associated with AI and a systematic identification and categorization of these threats. This paper summarizes some of AI's current risks and envisages a range of plausible scenarios that may unfold due to these risks. It addresses the paramount need for policy formulations to address these risks and more comprehensive threat vectors to provide a safe trajectory for the development and application of AI technologies, which balances the promise of unprecedented advancements with the preservation of human values and societal well-being.

---

[4] Siddiqui, Tabassum. "Risks of Artificial Intelligence Must Be Considered as the Technology Evolves: Geoffrey Hinton." University of Toronto. July 29, 2023. https://www.utoronto.ca/news/risks-artificial-intelligence-must-be-considered-technology-evolves-geoffrey-hinton#:~:text=During%20his%20talk%2C%20Hinton%20outlined.

[5] Munk Debates. "Artificial Intelligence." Munkdebates.com. (2023). https://munkdebates.com/debates/artificial-intelligence.

[6] Ibid.

[7] Ibid.

[8] Torres, Émile P. "Does AGI Really Threaten the Survival of the Species?" Truthdig. June 30, 2023. https://www.truthdig.com/articles/does-agi-really-threaten-the-survival-of-the-species/.

Surveying the Academic Literature

One of the gravest concerns in AI technologies lies in their potential application for information warfare and manipulating public opinion through new media, which could undermine democratic institutions and democracy. Our world is currently grappling with the disinformation phenomenon, and AI systems play a pivotal role in exacerbating this issue. These systems enable the creation of highly realistic AI-generated fake content and facilitate the widespread dissemination of disinformation to specific audiences by malicious actors on a large scale. Comprehending the risks of employing AI for propaganda requires analyzing cutting-edge research in multidisciplinary domains. Such an analysis helps classify the existing threats to inform a more manageable approach to policy development for mitigation purposes.

The rapid advancement and proliferation of AI-generated deepfakes, driven by Generative Adversarial Networks (GANs) and deep learning, has sparked concerns about the potential for misinformation and consequent erosion of trust in digital media.[9] [10] [11] [12] [13] [14] Recent academic research analyzed material on generating and detecting deepfakes, their source attribution, various applications and possible future trajectories. The emergence of new multifaceted threats and ethical questions raised by sophisticated AI text generators, such as GPT-3,[15] [16] [17] were also considered.

Deepfakes, primarily employed in manipulating video and audio content, pose considerable societal implications, including the spread of disinformation, harm to reputations and potential destabilization of

[9] Masood, Momina, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. "Deepfakes Generation and Detection: State-of-The-Art, Open Challenges, Countermeasures, and Way Forward." *Applied Intelligence* 53 (June 2022). https://doi.org/10.1007/s10489-022-03766-z.

[10] Vinay, A., Paras S. Khurana, T.B. Sudarshan, S. Natarajan, Vivek Nagesh, Vishruth Lakshminarayanan, and Niput Bhat. "AFMB-Net." *Tehnički Glasnik* 16 (4) (2022): 503–8. https://doi.org/10.31803/tg-20220403080215.

[11] Coccomini, Davide, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. "Combining EfficientNet and Vision Transformers for Video Deepfake Detection." *Lecture Notes in Computer Science* 13233 (July 2021): 219–29. https://doi.org/10.1007/978-3-031-06433-3_19.

[12] Ciftci, Umur Aybars, Ilke Demir, and Lijun Yin. "How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals." *2020 IEEE International Joint Conference on Biometrics (IJCB)*, September (2020). https://doi.org/10.1109/ijcb48548.2020.9304909.

[13] Jung, Tackhyun, Sangwon Kim, and Keecheon Kim. "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern." *IEEE Access* 8 (2020): 83144–54. https://doi.org/10.1109/access.2020.2988660.

[14] Korshunov, Pavel, and Sebastien Marcel. "Vulnerability Assessment and Detection of Deepfake Videos." *2019 International Conference on Biometrics (ICB)*, June 2019. https://doi.org/10.1109/icb45273.2019.8987375.

[15] Dwivedi, Yogesh K., Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, et al. 2023. "'So What If ChatGPT Wrote It?' Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy." *International Journal of Information Management* 71 (2023), (0268-4012): 102642.

[16] Ma, Jing, Jun Li, Wei Gao, Yang Yang, and Kam-Fai Wong. "Improving Rumor Detection by Promoting Information Campaigns with Transformer-Based Generative Adversarial Learning." *IEEE Transactions on Knowledge and Data Engineering*, 1–1 (2021). https://doi.org/10.1109/tkde.2021.3112497.

[17] Carpenter, Kristy A., and Russ B. Altman. "Using GPT-3 to Build a Lexicon of Drugs of Abuse Synonyms for Social Media Pharmacovigilance." *Biomolecules* 13 (2) (2023): 387. https://doi.org/10.3390/biom13020387.

government functions.[18] [19] Scientific research has sought to address these issues by developing advanced detection techniques and defensive mechanisms to authenticate visual content,[20] [21] [22] [23] [24] [25] and innovative methodologies for deepfake detection beyond traditional techniques are now emerging. Notably, a method by A. Vinay et al. combining heart-rate analysis with machine learning, leverages unique individual heart-rate patterns that GANs cannot mimic.[26] Similar approaches have been taken with DeepVision, which detects deepfakes by analyzing human eye-blinking patterns,[27] and DeFakePro, which employs Electrical Network Frequency signals to authenticate media broadcasts in online video conferencing tools.[28] Another recent strategy recommends using a specific generative model to create a deepfake, capitalizing on the residuals of the generator learned by convolutional neural network-based deepfake detection methods.[29]

Despite advancements in detection algorithms, the increasing sophistication of deepfake creation tools presents significant challenges. Specifically, pre-trained GANs have sought to "democratize" deepfake production, resulting in videos capable of deceiving face recognition systems and evading existing

[18] Masood, Momina, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. "Deepfakes Generation and Detection: State-of-The-Art, Open Challenges, Countermeasures, and Way Forward." *Applied Intelligence* 53 (June 2022). https://doi.org/10.1007/s10489-022-03766-z.
[19] Vinay, A., Paras S. Khurana, T.B. Sudarshan, S. Natarajan, Vivek Nagesh, Vishruth Lakshminarayanan, and Niput Bhat. "AFMB-Net." *Tehnički Glasnik* 16 (4) (2022): 503–8. https://doi.org/10.31803/tg-20220403080215.
[20] Masood, Momina, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. "Deepfakes Generation and Detection: State-of-The-Art, Open Challenges, Countermeasures, and Way Forward." *Applied Intelligence* 53 (June 2022). https://doi.org/10.1007/s10489-022-03766-z.
[21] Vinay, A., Paras S. Khurana, T.B. Sudarshan, S. Natarajan, Vivek Nagesh, Vishruth Lakshminarayanan, and Niput Bhat. "AFMB-Net." *Tehnički Glasnik* 16 (4) (2022): 503–8. https://doi.org/10.31803/tg-20220403080215.
[22] Coccomini, Davide, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. "Combining EfficientNet and Vision Transformers for Video Deepfake Detection." *Lecture Notes in Computer Science* 13233 (July 2021): 219–29. https://doi.org/10.1007/978-3-031-06433-3_19.
[23] Ciftci, Umur Aybars, Ilke Demir, and Lijun Yin. "How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals." *2020 IEEE International Joint Conference on Biometrics (IJCB)*, September (2020). https://doi.org/10.1109/ijcb48548.2020.9304909.
[24] Jung, Tackhyun, Sangwon Kim, and Keecheon Kim. "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern." *IEEE Access* 8 (2020): 83144–54. https://doi.org/10.1109/access.2020.2988660.
[25] Nagothu, Deeraj, Ronghua Xu, Yu Chen, Erik Blasch, and Alexander Aved. "DeFakePro: Decentralized Deepfake Attacks Detection Using ENF Authentication." *IT Professional* 24 (5) (2022): 46–52. https://doi.org/10.1109/mitp.2022.3172653.
[26] Vinay, A., Paras S. Khurana, T.B. Sudarshan, S. Natarajan, Vivek Nagesh, Vishruth Lakshminarayanan, and Niput Bhat. "AFMB-Net." *Tehnički Glasnik* 16 (4) (2022): 503–8. https://doi.org/10.31803/tg-20220403080215.
[27] Jung, Tackhyun, Sangwon Kim, and Keecheon Kim. "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern." *IEEE Access* 8 (2020): 83144–54. https://doi.org/10.1109/access.2020.2988660.
[28] Nagothu, Deeraj, Ronghua Xu, Yu Chen, Erik Blasch, and Alexander Aved. "DeFakePro: Decentralized Deepfake Attacks Detection Using ENF Authentication." *IT Professional* 24 (5) (2022): 46–52. https://doi.org/10.1109/mitp.2022.3172653.
[29] Ciftci, Umur Aybars, Ilke Demir, and Lijun Yin. "How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals." *2020 IEEE International Joint Conference on Biometrics (IJCB)*, September (2020). https://doi.org/10.1109/ijcb48548.2020.9304909.

detection methods.[30] [31] [32] In tandem with deepfake technologies, advancing large language models such as GPT-3 present substantial risks in text generation.[33] The ability of these models to generate contextually relevant and persuasive text can be exploited to impersonate and disseminate misinformation and disinformation. A recent compilation of 43 expert contributions addressed opportunities and challenges that AI text generators such as ChatGPT brought about, but underscored risks relating to biases, privacy concerns, misinformation and potential misuse.[34] Another study highlights the use of GAN-style approaches for automatic rumour detection in text-based media.[35] GPT-3's potential to generate a lexicon of colloquial drug synonyms for effective pharmacovigilance on social media underscores the model's dual-use nature.[36] Numerous other research papers[37] [38] [39] [40] explore various improvements in NLP tasks and deepfake detection strategies, the development of synthetic medical images and AI-generated audio. These papers collectively demonstrate the breadth of ongoing research, from devising innovative models for task-oriented dialogue systems[41] to exploring the potential of GANs in creating artificial retinal fundus images[42] and generating naturalistic spoken dialogues.[43]

---

[30] Korshunov, Pavel, and Sebastien Marcel. "Vulnerability Assessment and Detection of Deepfake Videos." *2019 International Conference on Biometrics (ICB)*, June 2019. https://doi.org/10.1109/icb45273.2019.8987375.

[31] Ma, Jing, Jun Li, Wei Gao, Yang Yang, and Kam-Fai Wong. "Improving Rumor Detection by Promoting Information Campaigns with Transformer-Based Generative Adversarial Learning." *IEEE Transactions on Knowledge and Data Engineering*, 1–1 (2021). https://doi.org/10.1109/tkde.2021.3112497.

[32] Carpenter, Kristy A., and Russ B. Altman. "Using GPT-3 to Build a Lexicon of Drugs of Abuse Synonyms for Social Media Pharmacovigilance." *Biomolecules* 13 (2) (2023): 387. https://doi.org/10.3390/biom13020387.

[33] Dwivedi, Yogesh K., Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, et al. 2023. "'So What If ChatGPT Wrote It?' Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy." *International Journal of Information Management* 71 (2023), (0268-4012): 102642.

[34] Ibid.

[35] Ma, Jing, Jun Li, Wei Gao, Yang Yang, and Kam-Fai Wong. "Improving Rumor Detection by Promoting Information Campaigns with Transformer-Based Generative Adversarial Learning." *IEEE Transactions on Knowledge and Data Engineering*, 1–1 (2021). https://doi.org/10.1109/tkde.2021.3112497.

[36] Carpenter, Kristy A., and Russ B. Altman. "Using GPT-3 to Build a Lexicon of Drugs of Abuse Synonyms for Social Media Pharmacovigilance." *Biomolecules* 13 (2) (2023): 387. https://doi.org/10.3390/biom13020387.

[37] Jiang, Shufan, Stéphane Cormier, Rafael Angarita, and Francis Rousseaux. "Improving Text Mining in Plant Health Domain with GAN And/or Pre-Trained Language Model." *Frontiers in Artificial Intelligence* 6 (February 2023). https://doi.org/10.3389/frai.2023.1072329.

[38] Salini, Yalamanchili, and J HariKiran. "Deepfakes on Retinal Images Using GAN." *International Journal of Advanced Computer Science and Applications* 13 (8) (2022). https://doi.org/10.14569/ijacsa.2022.0130880.

[39] Liu, Hong, Yucheng Cai, Zhijian Ou, Yi Huang, and Junlan Feng. "Building Markovian Generative Architectures over Pretrained LM Backbones for Efficient Task-Oriented Dialog Systems." *2022 IEEE Spoken Language Technology Workshop (SLT)*, January 2023. https://doi.org/10.1109/slt54892.2023.10023191.

[40] Tu Dinh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, et al. "Generative Spoken Dialogue Language Modeling." *Transactions of the Association for Computational Linguistics* 11 (October 2022): 250–66. https://doi.org/10.1162/tacl_a_00545.

[41] Liu, Hong, Yucheng Cai, Zhijian Ou, Yi Huang, and Junlan Feng. "Building Markovian Generative Architectures over Pretrained LM Backbones for Efficient Task-Oriented Dialog Systems." *2022 IEEE Spoken Language Technology Workshop (SLT)*, January 2023. https://doi.org/10.1109/slt54892.2023.10023191.

[42] Salini, Yalamanchili, and J HariKiran. "Deepfakes on Retinal Images Using GAN." *International Journal of Advanced Computer Science and Applications* 13 (8) (2022). https://doi.org/10.14569/ijacsa.2022.0130880.

[43] Tu Dinh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, et al. "Generative Spoken Dialogue Language Modeling." *Transactions of the Association for Computational Linguistics* 11 (October 2022): 250–66. https://doi.org/10.1162/tacl_a_00545.

In summary, a cursory examination of contemporary research supports the argument that AI-generated deepfakes and advanced language models offer unprecedented capabilities and pose considerable challenges. Continuous research and the development of defensive mechanisms, ethical guidelines and a legal/policy framework are crucial to counter potential misuse.

Classifying Existing Threats

The analysis of the current research allows us to distinguish across three groups of applied areas for threat-based AI uses: fake identities, fake behaviour and fake information threats (Figure 2). These threat categories can operate independently or merge with threats from other categories when initiating disinformation campaigns. For maximum impact, malicious actors or rogue states might employ a full spectrum of these threats to sway public sentiment. This can involve crafting fake personas, experts and organizations on social platforms, leveraging technology to tarnish prominent figures, generating vast amounts of deceptive content that appear human-authored and feigning online support or dissent to confuse genuine users and shift the public dialogue.

Figure 2: Fake Identities-Fake Information-Fake Behaviour: AI for Disinformation Threats Classification Model

Fake Identities

In 2021, Mirsky and Lee analyzed the architecture of deepfakes that spread misinformation, impersonate political leaders and defame innocent individuals.[44] They drew on examples of where deepfakes have been used to impersonate politicians giving talks on particular subjects and to swap faces of celebrities into

[44] Mirsky, Yisroel, and Wenke Lee. "The Creation and Detection of Deepfakes." *ACM Computing Surveys* 54 (1) (2021): 1–41. https://doi.org/10.1145/3425780.

pornographic videos and describe the large number of new deepfake videos that emerged thereafter. Their research raised concerns over identity theft, impersonation and spreading misinformation on social media. These issues were further evidenced by the hate and disinformation that five fake voter profiles — set up by BBC's disinformation correspondent Marianna Spring — experienced in the November 2022 US election. The research suggests the following fake-identity issues:

- **Creation of Fake Opinion Leaders —** Generative AI can produce fake expert profiles, complete with life-like descriptions and a variety of photos and videos of an individual who has never existed, all crafted synthetically.

- **Deepfakes —** Generative AI can produce deepfakes of popular and influential people that are falsified videos, images or audio that appear extremely realistic. These can be used to create disinformation, false news or impersonations that can harm reputations, manipulate public opinion or even threaten national security.

- **Impersonation and Identity Theft** — With the ability to mimic voices, writing styles and even faces, generative AI could be used for phishing attacks, impersonating trusted individuals or organizations to trick victims into revealing sensitive information or committing fraud.

Fake Information

- **Untraceable Disinformation** — Generative AI can create large amounts of false information quickly and efficiently, which could be used to spread propaganda or manipulate public sentiment. The scale and speed of such operations could make it difficult to trace the origin of the misinformation, making it harder to counteract.

- **Bias and Discrimination —** AI systems can unintentionally learn and replicate bias in the data it is trained on, leading to unfair outcomes. This could be exploited intentionally in a harmful way, leading to discriminatory or biased actions.

- **Human-like Created Content** — Advanced AI models can mimic human-style writing and adapt to various tones and styles, making it challenging to discern whether the content was genuinely written by a human or artificially produced by a machine. This blurring of lines poses risks for misinformation and authenticity in digital communications.

Fake Behaviour

- **Online Manipulation and Effect of Comment** — Bots powered by AI could be used to manipulate online discussions, amplify specific narratives, create a false impression of public consensus on divisive issues, create echo chambers of fake profiles and engage with real users.

- **Profiling and Targeting** — Advanced AI systems can amass and analyze vast amounts of personal data, enabling precise profiling of individuals. This heightened accuracy in profiling can lead to invasive targeted advertising, potential misuse in political campaigns or discriminatory practices, raising serious concerns about privacy and the potential for manipulative or biased decision-making.

- **Social Engineering Attacks** — AI models could be used to generate believable and persuasive text in advanced social engineering attacks. These could include highly targeted phishing emails, fraudulent messages, or scam calls, potentially leading to data breaches or financial fraud.

Scenarios

Following a survey of the relevant interdisciplinary literature on the impacts of AI, the above analysis exposes the fact that, notwithstanding the technologies and systems developed to counter the illicit and malign effects of AI, vulnerabilities to society are still rooted in sophisticated forms of fake identity, fake behaviour and fake information.

A better understanding of the manifestations of interactions across these three areas of vulnerability can be developed by applying scenario analysis or alternative futures methodology. The use of the Field Anomaly Relaxation (FAR)[45] technique can provide constellations of different plausible and possible scenarios, and help incorporate multiple dimensions of a "wicked problem." As an "alternative futures" model, FAR stands as a variant of morphological analysis and supports scenario development based on a game-theoretic approach and the use of mathematical methods to discover diverse scenarios that go beyond the standard best, worst,and current scenarios.

Using the three threat vectors described above — fake identities, fake information and fake behaviour — to form the main factors in a FAR "field of factors," the different types of scenarios described above would then list themselves in order from less severe to more severe, underneath each of these headlines. All plausible and possible scenarios could be conceived by linking combinations across the factors field depending on the timeline being considered. For example, a case where deepfakes were used to significantly discredit or smear a highly influential leader, this deepfake application is then made by an untraceable disinformation publication for evidence supporting a targeted attack at the institutions with which the influential leader was affiliated (Scenario 1). The case of Russian computer hackers breaching Ukraine's national news broadcast serves as an example of this scenario. In this case, the news ticker was manipulated to display a deepfake video of Ukraine President Volodymyr Zelenskiy urging Ukrainians to surrender and falsely claiming that he had fled Kyiv after failing to secure the Donbas region of Ukraine.[46]

---

[45] Rhyne, Russell. "Field Anomaly Relaxation." *Futures* 27 (6) (1995): 657–74. https://doi.org/10.1016/0016-3287(95)00032-r.

[46] Digital Forensic Research Lab. "Russian War Report: Hacked News Program and Deepfake Video Spread False Zelenskyy Claims." Atlantic Council. March 16, 2022. https://www.google.com/url?q=https://www.atlanticcouncil.org/blogs/new-atlanticist/russian-war-report-hacked-news-program-and-deepfake-video-spread-false-zelenskyy-claims/&sa=D&source=docs&ust=1693971569758353&usg=AOvVaw21HhCavNcylpKsQeWoBXNh.

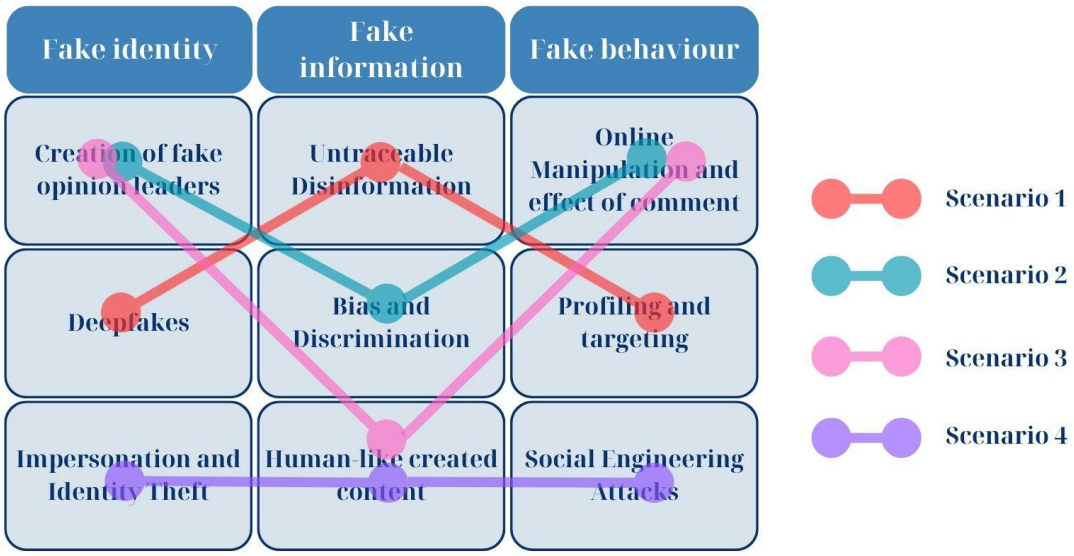

Figure 3. Scenarios Example

Another example of a combination of factors that manifests into a concerning threat vector can involve the creation of a fake opinion leader operating either under the anonymity of false names and credentials, whose profile becomes supported AI-generated profiles (and also inauthentic follower support based on direct compensation) and generates human-like content to manipulate and amplify public opinion (Scenario 3). Research indicates that trolls increasingly use polarized language on social media platforms such as X (formerly Twitter) and Reddit, intensifying political divisions and engaging more with already politically active users.[47] This same scenario could also involve manipulating biased and discriminatory content (Scenario 2). Recently, X removed fake accounts with individuals posing as African Americans supporting Donald Trump, highlighting concerns about disinformation campaigns using polarized language to intensify political divisions ahead of the election.[48]

Another scenario, which has been observed now for years in the context of banking fraud - and which will only become more sophisticated in the future - involves AI-driven attempts to impersonate and attempt identity theft by creating human-like content and generating socially engineered attacks (Scenario 4). The experience involving the hacking of Capital One accounts led to the theft of the social insurance numbers of over a million Canadians. In response, the government emphasized that it rarely issues replacements and emphasized that a new number does not protect individuals from identity theft-related fraud.[49]

---

[47] Simchon, Almog, William J Brady, and Jay J Van Bavel. "Troll and Divide: The Language of Online Polarization." Edited by Diana Mutz. *PNAS Nexus* 1 (1) (2022). https://doi.org/10.1093/pnasnexus/pgac019.
[48] Collins, Ben. 2020. "Viral Pro-Trump Tweets Came from Fake African American Spam Accounts, Twitter Says." www.nbcnews.com. August 27, 2020. https://www.nbcnews.com/news/amp/ncna1238553.
[49] Gatehouse, Jonathon . 2019. "Social Insurance Numbers Are Stolen by the Millions — but Ottawa Replaces Just Dozens per Year Social Sharing." CBC News. August 1, 2019. https://www.cbc.ca/news/politics/stolen-social-insurance-numbers-fraud-1.5232037.

While it is beyond the capacity of this paper to analyze all the potential scenarios that could unfold under the field of factors considering fake identity, fake information and fake behaviour, this exercise provides sample evidence to deduce the following trends for the future.

Low Capacity to Anticipate and Manage Threats

The collective impact of these trends will manifest in an increase in what are known as "wicked problems" (problems that may not be anticipated and for which no readily available way to address the challenge exists). In this context, institutional policies and systems lag behind the pace and sophistication of AI and data-driven manipulation and tactics. This indicates the lack of knowledge and skill sets in this area, which combines policy and technology, particularly in addressing the interdependent impacts of these threats. Feedback from policy makers working in this space suggests that, in the absence of such intersectional knowledge and skill sets, not only do policy deficits emerge, but also conflicting policies. Systems that could advance this intersectional knowledge — such as the merit-based promotion systems in universities, many of which reflect the maintenance of domain-specific structures and the encouragement of single-domain research — inadvertently work against these ends.

Without a critical mass of knowledge and skill sets in these multidisciplinary areas, threat vectors that combine deepfakes, manipulation, impersonation and profiling, and targeting carry serious implications for future leadership. With resource deficits beneath them and inauthentic forms of public criticism amplifying, the personal cost of becoming a leader has escalated. Moreover, the sheer volume and intensification of public inauthentic criticism being levelled toward institutions leave many public institutions functioning in a democratic system weighed down and unable to cope with responses to the intensified criticism and having minimal capacity to respond to the legitimate needs of the electorate. The latter opens further space for criticism of leadership and further rationale for leaders not to step forward as candidates for these positions in the future. The leadership deficit then combines with the knowledge and competency deficit to further undermine the national capacity to manage national resources, opening further vulnerabilities necessitating a heavier reliance on partnerships with external actors.

Spread of Localized Conflict and Tensions

Trends that intersect fake information, fake behaviour and fake identities have led to influencing operations, which generate the breakdown of relationships at both personal and professional levels. In diverse democracies, the breakdown of relationships can provoke violence and conflict-related tensions. Ground-based conflicts occurring on different continents can leverage such influencing operations to impact the democracy and stability of other countries. Diverse democracies can serve as a haven for foreign agents whose facilitation of AI-driven influencing operations can go undetected through anonymity, but whose amplification of conflict-based narratives can impact discussions in academia, civil society and parliamentary structures. Continuity of trends in democratic systems that stand staunchly behind human rights and freedoms may inadvertently provide platforms and fair hearings to illicit groups and malicious actors.

Diverse societies that have been relatively unified in the past, like Canada, have found their strength in effective assimilation strategies — but the GAN-based threats diminish the capacity to assimilate society under common values and priorities. This leads to social marginalization and threats to the unity of society. The current housing crisis in countries such as Canada and the United Kingdom further exacerbates extreme forms of social and economic marginalization, which, with the support of AI-driven algorithms and illicit global networks, can increase the risk of conflict and instability. The 2023 violence that emerged in both Paris and Sweden and higher crime rates in Toronto are all uncomfortable outcomes of this trend and, if left unaddressed, could become an even broader, more severe, phenomenon.

These threat vectors, intensified significantly by AI, threaten democracy, constitutionalism and the rule of law. With institutions being the cornerstone of democracy and with faith in institutions diminishing, social contracts will become less robust. Faith in governments and institutions to provide for the people has always involved the population giving up certain rights and freedoms — such as the right to wield, carry and use arms for the use of force against others, which threaten our safety and security — to other institutions who lead in these roles. A lack of faith in these pivotal institutions and the diminishing of the social contract could see more incidents of groups and individuals taking the law and other publicly served functions of the state into their own hands, resulting in violence and the fragmentation of society.

Stifling of Economic Opportunities

With knowledge, skill sets and competency deficits to manage and govern the AI-driven threat vectors, and an increased reliance on national and multinational actors who have more capacity in this area, economic opportunities also become threatened. Supply chains that align with knowledge and skillsets — and the way in which these supply chains carry significance for international trade — also risk diverting international trade from a country with less control exercised over the means and ways in which those supply chains are accessed and also the way in which their functioning is managed and overseen. The weakening of supply chain resilience and control of national resources will carry implications for economic growth opportunities, foreign investment, external dependencies and thought leadership. These trajectories will also bring an inevitable increase in the surrendering of intellectual property for upscaling and commercialization, which will in turn stifle entrepreneurialism and innovation.

Failure to manage threats posed by the dangers of AI will lead to risk manifestation at the individual level, which will impact insurance premiums and the general cost of safeguarding the population. Combining these trends systematically with a lack of home ownership (in some countries) and a bulging youth population will have an inevitable increase in crime.

Ungoverned AI is likely to lead to the widening of supply chain dependencies and inevitable economic downturn, which will also marginalize national voices at the international AI table, as well as economic and trade tables such as the Group of 20 and the Group of 7. It is a pivotal time for democratic and balanced voices at international tables to be well-informed and have capacity in these areas in order to inform guidance, laws and accountability of data usage in the AI era. Data has become the most important instrument of national power and beyond the need to build capacity around this pillar comes

the need to influence national and international systems, which will distinguish across, manage and govern the threats and vulnerabilities as well as the opportunities and efficiencies that must be understood.

The manifestation of threat vectors in ways that can potentially weaken institutions, promote conflict and instability, and undermine the economy, accounts for only some of the strategic-level concerns that could emerge from a sustained lack of capacity to address the intersectional impacts of fake identity, fake information and fake behaviour. These impacts also carry serious implications for the violation of basic human rights, such as the right to safety and security, food security and freedom of expression, all of which, with the manifestation of these threats, a state will lose the capacity to protect itself.

The discussion above and the intersectional trends gleaned by way of a preliminary scenario analysis necessitates an urgent policy response. Based on the cross-sectoral nature and impacts of content generated by AI, national governments must move towards the characterization of data as a national instrument of power and even consider creating new institutions focused on data and information. Subsequent strategic national policies across all cabinet functions of government, including security, economic, justice, education and other social policies, should also reflect the data-driven reality and the global virtual space. Supporting well-informed policies and basic knowledge and skill-set development in this area will require transformational reforms across all levels of the education system, an investment in multidisciplinary research efforts and robust advocacy promoting the need for discussions within credible international coalitions that can facilitate the sharing of good practice and promote a wider and much-needed international consensus on these issues.

Conclusion

This paper reviews the appeals of leading global thinkers on the impacts of AI in general, and generative AI in particular. A literature review on generative AI technologies and the technologies developed to counter their potential negative impact indicates that, detection technologies notwithstanding, significant threats persist. The cumulative impacts of these threats — fake identity, fake behaviour and fake information — can be presented using the FAR technique, which details how the threats intersect and manifest into wider policy concerns relating to institutions, the economy and national security. Further scenario-based analysis research should be supported to more fully understand the wider impact and more comprehensive policy implications concerning generative AI applications.

Ann Fitz-Gerald is the Director of the Balsillie School of International Affairs and a Professor in Wilfrid Laurier University's Political Science Department. She has worked at both at King's College, London University's International Policy Institute, and at Cranfield University, where she was the Director, Defence and Security Leadership. During her time at Cranfield, Ann led the UK-Government funded Global Facilitation Network for Security Sector Reform and Cranfield's Centre for Security Sector Management.

Dmytro Chumachenko is the Post-Doctoral fellow at the Ubiquitous Health Technology Lab at the University of Waterloo. He completed his PhD on Systems and Means of Artificial Intelligence at Kharkiv National University of Radio Electronics, Ukraine. His research interests include epidemic modeling, machine learning, agent-based simulation, infectious diseases simulation and data-driven medicine.

Halyna Padalko worked in media communications in Ukraine for 9 years as the TV host of a socio-political talk show "The opposite view LIVE", press officer of the Ukrainian Digital Transformation Institute, and Chief Communications Officer for the Ukrainian "Vivat" publishing house. In 2020 Halyna was elected a deputy of the Lutsk City Council. She is working on her PhD in Computer Science at the National Aerospace University "Kharkiv Aviation Institute" analyzing propaganda in the social media using Artificial Intelligence approaches.

Vijay Ganesh is a professor at the School of Computer Science at GeorgiaTech. Prior to joining GeorgiaTech in 2023, Vijay was a professor at the University of Waterloo in Canada from 2012 to 2023 and a research scientist at the Massachusetts Institute of Technology from 2007 to 2012. Vijay completed his PhD in computer science from Stanford University in 2007. Vijay's primary area of research is the theory and practice of SAT/SMT solvers, and their application in AI, software engineering, security, mathematics, and physics.

BALSILLIE PAPERS

balsilliepapers.ca

ISSN 2563-674X

doi:10.51644/BAP56