A Systems Thinking Approach to Artificial Intelligence Governance

Liane Weintraub, Blair Johnson, and Michelle Eleby

Abstract

Artificial intelligence (AI) is among the most globally contended topics today. The rapid advancement of AI technology has magnified both its beneficial and detrimental impacts. Proponents welcome AI's promise of interpersonal connectivity, expanded civic engagement, and increased productivity, accuracy, efficiency, accessibility, and quality of life. Due to advances in machine learning, deep learning algorithms can now examine voluminous quantities of audio, text, images, and video to detect peculiarities from manufacturing processes to healthcare patient patterns. Consequently, AI can combine pieces of information in ways that traditional analytics cannot. These new capabilities promise significant positive impacts across multiple industries. However, detractors point out that the same tools are used to censor or limit speech, surveil, exploit, spread disinformation, and displace jobs. In May 2023, the Center for AI Safety issued a statement signed by hundreds of international scientists, scholars, and legislators, including OpenAI CEO Sam Altman and Google DeepMind CEO Demis Hassabis that read: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." However alarmist the statement may appear, it reflects the growing sentiments of much of the world citizenry. A 2023 Reuters/Ipsos survey indicated that 61% of Americans believe AI technology may threaten civilization, and a 2022 Ipsos global survey revealed that respondents in 28 countries on five continents expect AI technology to enhance education, entertainment, and transportation but to worsen employment, cost of living, and freedom/legal rights.

All of these factors suggest that world governments should urgently consider AI governance measures. Furthermore, such considerations cannot be undertaken individually in an increasingly globalized world where technology and data flow across borders. This paper considers how an international alliance might approach AI governance to enhance its positive capabilities while instituting guardrails to prevent misuse or negative outcomes.

To consider how a new global governance framework might function, the paper first offers an overview of current and projected AI capabilities, the constructive implications, and concerns surrounding AI's ethical development, deployment, and application. Next, the paper presents a summary of the two primary regulatory approaches, omnibus and sectoral. These two models inform a proposed integrated global governance framework that seeks to address the challenge of integrating different and competing national values. The proposed model integrates aspects of the current EU AI Act with a novel approach informed by New Zealand's response to the 2019 Christchurch terrorist attack.

The proposed model takes a systems thinking approach by changing the traditional regulatory

technology requires a new regulatory mental model that recognizes that substantial advances in scientific and financial knowledge and technological complexity demand faster and ongoing regulatory assessment rhythms. Furthermore, in an interconnected global community, a collaborative group of experts, industry leaders, civic leaders, consumers, and government officials is needed to keep pace with machine learning tools like generative AI. The new model must also take a cross-border approach, encouraging economies to grow while addressing the inequity of low-skilled workers displaced by digitization. Such leadership and mindset changes cannot be reserved for regulatory officials. Rather, technological advances require industry and business leaders to develop new mental models in order to transform their enterprises, lower service and production costs, deliver better customer and patient experiences, and improve health outcomes while adhering to new regulations.

This paper aims to address these challenges and concerns by presenting a novel approach to global technology governance. The substantially holistic framework draws on systems thinking to offer a point of departure from which countries and groups of countries can negotiate. In this process, it may ultimately become necessary for the global community to agree on a governing body to attend to global concerns surrounding AI.

Selected References

Acemoglu, D. & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Political Economy*, 33(2), 3-30.

Acemoglu, D. & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, 128(6), 2188-2244.

AI Algorithmic and Automation Incidents and Controversies (n.d.). Aiaaic repository. Retrieved June 4, 2023, from https://www.aiaaic.org/aiaaic-repository

Anderson, J., Rainie, L. & Luchsinger, A. (2018). Artificial intelligence and the future of humans. Pew Research Center.

Ariga, T. (2023, May 16). Artificial intelligence: Key practices to help ensure accountability in federal use. General Accounting Office. https://www.gao.gov/assets/gao-23-106811-highlights.pdf

Bathaee, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, *31*(2), 889.

Belli, L., Curzi, Y., & Gaspar, W. B. (2023). AI regulation in Brazil: Advancements, flows, and need to learn from the data protection experience. *Computer Law & Security Review, 48*. https://doi.org/10.1016/j.clsr.2022.105767

Berman, J., & Weitzner, D. J. (1997). Technology and democracy. *Social Research, 64*(3), 1313-1319.

Bode, I., Huelss, H., Nadibaidze, A., Qiao-Franco, G., & Watts, T. F. (2023, February). Prospects for the global governance of autonomous weapons: comparing Chinese, Russian, and US practices. *Ethics and Information Technology, 25*(1), 5. https://doi.org/10.1007/s10676-023-09678-x

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world.* Faculty Books.

Butcher, J., & Beridze, I. (2019). What is the state of artificial intelligence governance globally? *The RUSI Journal*, *164*(5–6), 88–96.

Cath, C. (2018, October). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences, 376*(2133). https://doi.org/10.1098/rsta.2018.0080

Chen, N., Li, Z., & Tang, B. (2022). Can digital skill protect against job displacement risk caused by artificial intelligence? Empirical evidence from 701 detailed occupations. *Public Library of Science ONE*, 17(11).

Chesney, R. & Citron, D. (2019). Deepfakes and the disinformation war: The coming of age of post-truth geopolitics. *Foreign Affairs, 98*(1).

Chui, M. & Bush, J. (2023, May 31). The brave new world of generative AI with Ethan Mollick. [Audio podcast episode]. In *Forward Thinking*, *McKinsey Global Institute*, https://www.mckinsey.com/mgi/forward-thinking/forward-thinking-on-the-brave-new-world-of-generative-ai-with-ethan-mollick

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). *The economic potential of generative AI: The next productivity frontier.* McKinsey & Company.

Clarke, L. (2023, April). Alarmed tech leaders call for AI research pause. *Science, 380*(6641), 120-121. https://doi.org/10.1126/science.adi2240

Cohen, J., & Fung, A. (2021). Democracy and the public sphere. In L. Bernholz, H. Landemore, & R. Twitch (Eds.) *Digital technology and democratic theory*. University of Chicago Press.

Engler, A. (2023, April 25). The E.U. and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. *Brookings Institute.* Retrieved June 20, 2023, from https://www.brookings.edu/research/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/

European Commission, (2021, April). Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts. COM(2021) 206 final. Retrieved June 11, 2023, from https://artificialintelligenceact.eu/the-act/

Feijóo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Howell, B., Jain, R, Potgeiter, P., Vu, K., Whalley, J. & Xia, J. (2020). Harnessing artificial intelligence to increase wellbeing for all: the case for a new technology diplomacy. Telecommunications Policy, 44(6).

Fitzgerald, M., Boddy, A., & Baum, S. D. (2020). 2020 survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute Technical Report*, 20-1.

Imagination, Risk, Wicked and Critical problems: the leadership dynamics of artificial general intelligence and the regulatory hype-cycle

Leighton Andrews,
Professor of Practice in Public Service Leadership and Innovation
Cardiff Business School

After the pandemic and the war in Ukraine, and the perennial unresolved climate crisis, do we think that public leadership is capable of addressing existential risk? How do we imagine risk and opportunity? Some speak of a crisis of social imagination, others suggest we are in a new age of uncertainty in which human action risks catastrophic and unforeseen consequences. At the time of writing, the most recent of those identified risks is - once again - artificial general intelligence or generative AI, where, following the launch of Chat-GPT and its competitors, industry actors are warning of an existential risk to humanity. Others urge us to focus more on existing challenges such as algorithmic discrimination and data abuse. We have been here before: are we doomed to a regulatory hype-cycle in which public panics provoke short-term action but no sustained focus by public leaderships? Drawing on Grint's problem typology and his distinction between wicked and critical problems (2005, 2022), this paper examines public leadership around algorithms, Big Data and artificial general intelligence as a way into thinking about political management of existential risk, looking at the UK in particular. The paper asks how public leaders prioritise long-term existential risk against short-term challenges, no matter how serious, and how public leadership can lead when critical information is controlled by powerful private entities, run by 'hero technologists' and 'hero founders' with apparently new-found saviour complexes. The paper returns to issues raised in the author's previously published work on these subjects. It draws on the author's secondary analysis of over 100 published interviews with UK Government ministers in the Institute for Government's Ministers Reflect archive, additional interviews personally conducted, and documentary analysis. This research suggests that, notwithstanding the many issues around the Covid crisis, government leadership of crises is relatively well developed, in keeping with Grint's 'command' mode of decision styles. However, it also tentatively suggests there are systemic disincentives which deflect leaders from the consideration of fundamental long-term risk, such as those it is suggested may result from the development of generative AI. Finally, it argues that the tame, critical and wicked problem typology, corresponding to Grint's three modes of decision-style - management, command and leadership - also requires a dynamic temporal framing depending on context. Critical problems can become tame problems; wicked problems can become critical, problems constructed as 'tame' may turn out not to be so. In the case of generative AI, we arguably face tame, critical and wicked problems, meaning that bricolage, messiness and clumsiness will be the most likely government approach.

On AI and the Executive Class – Power, Discourse, and Paradox in Top Leaders' Sensemaking of Algorithmic Logics

Leadership, as an activity and as a concept, is by necessity affected by the political, economical, social, and technological context it emerges in. At the moment, this means being affected a number of emergent technical facts, one of these being the indeterminate shift from logic and intelligence being always already mediated by human action to a state of affairs where algorithmic logic, machine learning, and artificial intelligence (AI) take on a more and more dominant role in much of our working lives. Current notions of AI and how algorithmic logics might affect organizational work has usually focused on menial and manual labor, the logic being that only such categories will be truly affected and in the end displaced by the limited logical works of emergent AI. Put somewhat differently, much of the existing debate regarding how AI will affect work has indeed put the worker front and center – of a firing line.

That said, more modern notions of AI are rapidly indicating that a core developmental tendency within this field is the way in which actions such as delegation, reporting, communication, and control (all key leadership functions) can be entrusted to algorithms and associated learning regimes. Further, as recently seen in generative AI systems, much of the communication and various kind of textual work that has been at the heart of the work of leaders might also be automated. As notions that would previously have been considered executive become assigned to machinic logics, what does this mean for categories such as general management as a function, the way corporate leaders view the same, and the way they identify as specifically leaders and executives? What happens when the displacers come face to face with their own potential displacement?

This paper thus deals with the corporate dilemma of being committed to notions of technological development, and at the same time conflicted when it comes to how the selfsame developments might affect one's own professional identity. Based on a qualitative study of how top executives view artificial intelligence and the manner this might upend power relations in management, the paper will discuss the complex issue of how algorithmically driven decision-making processes and general leadership functions can be understood within an executive mindset, and how this might be critiqued.

By way of an interview study with high-level executives in the Nordics, this paper thus aims to analyze the manner in which corporate leaders are attempting to make sense of their role

in an age of burgeoning AI, with a particular focus on the often paradoxical and looping manner in which they grapple with the notion of their own displacement. Through this, the paper highlights the way in which technological developments are used to illustrate a simplistic notion of progress, even when they could be interrogated for the ways they destabilize and query existing power structures. In the end, the paper aims to show that categories such as “leadership” play a dual role – both as a valorized category that stabilize organizational discourse, and as an indeterminate category that allows for a radical questioning of the same. By doing so, the paper aims to highlight the manner in which the marker of “leadership” is more than an objective category, and also a symbol and signifier of power in an executive matrix, one that makes and potentially breaks notions of managerial identity.

The patriarchal ghost in the AI machine:

What AI-generated narratives tell us about issues of gender in leadership

Suze Wilson, Massey U, New Zealand

Toby Newstead, U. of Tasmania, Australia

Bronwyn Eager, U. of Tasmania

AI-generated material is estimated to comprise up to 90% of web-based content by 2026 (Schick, 2020), having a range of implications for leadership practice, research, and development: positive, negative, and largely unknown. AI models used for content generation are typically trained using internet data, analysing it for patterns and computing the statistical probabilities of words, phrases, structures, and representations. Based on what they consider most probable, these AI models then produce content which reflects and amplifies existing societal beliefs, attitudes, values, and biases. This propensity to reproduce pre-existing views offers researchers an opportunity to gain insights into the nature of current discourse on a given issue. AI-generated content thus offers to researchers a novel approach for conducting discourse analysis and a means to overcome cost, time, and resource constraints typically associated with large-scale multinational studies. In this study, we leverage AI-generated content to develop insights into discourse on issues of gender in relation to leadership.

Our study employs a specific type of AI tool, known as a long-form AI writer. In contrast to popularly known AI models, such as the conversation-style chatbot, ChatGPT, which generates content through iterative human-AI interaction achieved by human-directed instruction (i.e., prompts), long-form AI writing tools can generate content with minimal user input. While there are many companies offering long-form writing tools, it is important to note that the technology behind these tools relies on the same underlying mechanism found in ChatGPT. This mechanism, known as a generative pre-trained transformer, is responsible for generating content based on (potentially problematic) training data. The use of GPTs allows these writing tools to generate coherent and contextually relevant text, which is created using generative AI's statistical probability-based text generation feature. This feature makes long-form AI writing tools useful for researchers who require extensive textual data for their analyses, allowing scholars access to a large volume of AI-generated content that is statistically representative, while tailored to specific research needs. In industry, the functionality of long-form AI writers is increasingly being used to create all manner of text-based artefacts, from news articles, blogs, and books to materials used in the leadership development sector.

For the current research, a total of 54 text-based narratives were generated using one of the many available long-form AI tools (Wordplay.ai), all of which rely on the same underlying GPT technology. We tasked the AI tool with generating 54 long-form narrative texts. Each text was generated to address a unique title, inclusive of leadership valence, gender, and historical context variables (e.g., "Good women leaders throughout history"). Valence variations were 'good' and 'bad'; gendered delineations were 'man', 'woman', and 'neutral') (wherein reference to gender was omitted); while historical contexts were 'past', 'present', and 'future'. The AI tool required the specification of a keyword, for which we specified 'leadership'. After entering only the title and the keyword, the AI model responded by providing multiple options relating to sub-headings it would use when generating its content. In all instances, we selected the first suggested option. In total, 54 narratives were generated, ranging from 600 to 1300 words in length. Each narrative adhered to the same format, comprising a title, introduction, three to five paragraphs, and a conclusion.

Our analysis of these 54 narratives sought to identify if, where and how gender-based biases were reflected and perpetuated in that material. Our approach was informed in part by our prior understanding of the research literature on issues of gender and leadership (e.g. Elliot & Stead, 2009; Madsen, 2017), alerting us to the kinds of issues that we might encounter. We were motivated, also, by a recent finding that young women's aspirations for management roles have reduced this century while young men's have not (Powell & Butterfield, 2022), coupled with evidence of a growing anti-feminist, misogynistic backlash in which women exercising leadership are key targets for abuse and harassment (Di Meco & Brechenmacher, 2020).

We found that while these texts sometime offered a brief acknowledgement of prejudice against women's leadership, and none expressed glaringly misogynistic or egregiously sexist comments about women's capacity for leadership, it was also commonplace that they reproduced and promulgated a range of subtle but pernicious 'second-generation' gender-based biases (Ely, Ibarra & Kolb, 2011; Sturm, 2001). These biases reflect and

perpetuate long-standing patriarchal beliefs, creating a context in which women leaders legitimacy and authority is rendered tenuous and where double standards apply to how women leaders are perceived and judged. Our analysis found the following characteristics of contemporary leadership discourse that is circulating on the internet, as evidenced in the AI-generated narratives we examined:

- 'Good' leadership by men, across all time horizons, is consistently characterised as relying on strength, courage, risk-taking, determination, competency and strategic nous, attributes that largely reflect patriarchal notions of masculinity. In contrast narratives about 'good' leadership by women patronisingly remind them to set the right priorities, have self-belief, find confidence, be aware of their emotions, and to look after and help develop those around them, thus reproducing patriarchal tropes about femininity and its alleged fragilities;
- Examples of 'good' male leaders emphasize their skills, compelling personality and significant influence, while examples of 'good' women leaders portray them as struggling valiantly to break through glass ceilings or as dependable supporters of more significant men;
- Exemplars of 'bad' leadership comprised twice as many women as men, despite women's underrepresentation in leadership roles. This subtly implies women are more likely than men to lead badly and, perhaps, indicates bad leadership by women will attract more scrutiny than that of men;
- While narratives about 'bad' leadership by men emphasize tyrannical and authoritarian behaviours and their capacity to harm others, for 'bad' women leaders it is their lack of skill and ability that features. Accordingly, while such men are portrayed as powerful agents that others should fear their female counterparts are characterised as weak, ineffectual and paralysed by their own fear of failure;
- The 16 narratives where we set no gender-based prompt included no examples of women leaders, thereby perpetuating the 'think manager, think male' stereotype first suggested by Schein in 1973;
- The little recognition given to the existence of biases against women leaders only emerged in texts related to women leaders, thus its appears as an issue seemingly worthy of women's attention but not necessarily everyone's attention.

These finding suggest despite centuries of effort to advance women's leadership, contemporary understandings continue to be plagued by patriarchal attitudes and norms. The state of the discourse evidenced in our study indicates the need to radically rethink how gender equity in leadership can, in fact, be advanced. We argue the patriarchal ghost in the AI-machine that our findings also illustrate now risks automating biases against women's leadership as AI-generated content grows in influence, and that this also requires attention. We reflect on what strategies may help advance gender equity in leadership given how persistent patriarchal perspectives are proving to be.

References

Di Meco, L., & Brechenmacher, S. (2020). *Tackling online abuse and disinformation targeting women in politics*. https://carnegieendowment.org/

Ely, R. J., Ibarra, H., & Kolb, D. M. (2011). Taking gender into account: Theory and design for women's leadership development programs. *Academy of Management Learning & Education, 10*(3), 474-493. 10.5465/amle.2010.0046

Madsen, S. R. (Ed.) (2017). *Handbook of research on gender and leadership*. Edward Elgar.

Powell, G. N., & Butterfield, D. A. (2022). Aspirations to top management over five decades: A shifting role of gender? *Gender in Management: An International Journal, 37*(8), 953-968. https://doi.org/10.1108/GM-10-2021-0330

Schein, V.E. (1973). The relationship between sex role stereotypes and requirement management characteristics. *Journal of Applied Psychology, 57*: 95-100.

Schick, N. (2020) *Deepfakes: The coming infocaplyse: What you urgently need to know*. Monoray.

Stead, V., & Elliott, C. (2009). *Women's leadership*. Palgrave Macmillan.

Sturm, S. (2001). Second generation employment discrimination: A structural approach. *Columbia Law Review,* 101: 458 –568.