Theories of AI and the Necessity of Moving Between Disciplines

Joe Burton

Do we need a revolution in theory for AI, or do we already have the theoretical tools to understand this transformative technology? The answer to that question is an important one for all scholars studying this new technology.

On the one hand, we already have a range of theoretical frameworks that can help us understand AI technologies. Realist International Relations (IR) theory is a useful starting point. Are we to discount core concepts like the 'security dilemma' when analysing AI?[1] This concept suggests that development of offensive capabilities by one state causes fear and mistrust in others, which leads to arms races. Are arms race dynamics in evidence in the sphere of military AI? Many seem to think so.[2] Other realist theory posits that AI will affect the global balance of power, which seems a compelling argument in the context of Vladimir Putin's comment, the state that leads in AI will be the 'ruler of the world'.[3] Realism has been an influential theory within IR for over a century and there's a reason why – its assumptions and propositions can be applied across time and space to new and emerging issue areas such as AI.

The same could be said of liberal internationalist and institutionalist IR theory. Whether it's the emergence of a new AI 'regime'[4] - a system of rules, protocols, and governance mechanisms to mitigate threats to international security, the 'absolute gains' that international institutions can provide, including through providing greater transparency around technology and opportunities for consultation, mediation and dialogues, or indeed the emergence of AI as a form of 'soft power'.[5] These tenets of liberal IR theory would seem to have a lot to offer for our understanding of AI and to provide a pre-existing conceptual platform that we shouldn't ignore.

The more 'critical' body of theory also has a lot to say about the emergence of AI, including in military and defence contexts. My own work on the securitisation of AI[6] suggests that the discourse of existential threat that is often associated with AI (General AI and Superintelligence in particular) serves political and commercial purposes, including generating funding for new military AI projects.

---

[1] Johnson, James, 'AI-security dilemma: Insecurity, mistrust, and misperception under the nuclear shadow', *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (Oxford, 2023; online edn, Oxford Academic, 23 Mar. 2023), https://doi.org/10.1093/oso/9780192858184.003.0005, accessed 10 Apr. 2024.

[2] https://hir.harvard.edu/a-race-to-extinction-how-great-power-competition-is-making-artificial-intelligence-existentially-dangerous/

[3] https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world

[4] https://carnegieendowment.org/2024/03/21/envisioning-global-regime-complex-to-govern-artificial-intelligence-pub-92022

[5] https://www.brookings.edu/articles/malevolent-soft-power-ai-and-the-threat-to-democracy/

[6] Burton, J. (2023) "Algorithmic extremism? The securitization of artificial intelligence (AI) and its impact on radicalism, polarization and political violence," *Technology in society*, 75, pp. 102262-. Available at: https://doi.org/10.1016/j.techsoc.2023.102262.

The literature that has already emerged on AI and data colonialism,[7] AI as an extractive technology,[8] and the racial bias that often exists in the data that AI relies on,[9] is illustrative of how Marxist, Feminist, Postcolonial and Poststructuralist approaches to IR can enhance our understanding of AI's social and political effects and implications.

All these theories are applicable and revealing for our understanding of AI – their 'explanatory power' is something predicated on the assumption that technologies are embedded in human, social and political systems, and we already know quite a lot about how these work.

The challenge that must occupy us as Defence, Security and International Relations scholars is to push these theories forward, to refine them and develop them for a new technological context. In doing do, we need to think beyond our disciplinary boundaries and silos. As scholars of AI, we should spend as much time reading computer science journals as IR ones. Take for example the practice of 'data poisoning' – one of the most prevalent ways that AI models and algorithms can be corrupted and attacked. Computer science academics have written extensively on the technical means and mechanisms available to poison data.[10] But there has been almost nothing written on the social and political causes and consequences of this increasingly important threat to AI security.

You'd think we might want to work with and engage with Linguistics as a discipline as we tackle the risks posed by Large Language Models, and, if AI systems are 'neural networks', talking to psychologists about the cognitive effects of the deployment of AI technologies in military strategic contexts will be as important as understanding how AI will influence the way we think and behave.

If we want to understand why AI tools are being 'rushed to release', often with adverse consequences (the cyber and social risks associated with the release of Chat GPT, for example),[11] then the field of economics may provide crucial insights, including in understanding commercial AI processes in the private sector, such as how security is too often an afterthought in software development.

Our problem is we are not doing enough of this type of cross disciplinary engagement and collaboration. If that doesn't change, IR and Security Studies will stagnate and be less and less able to provide the analytical insight and organising concepts that are relevant to policy makers.

---

[7] Arora, A. *et al.* (2023) "Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization," *Information and organization*, 33(3), pp. 100478-. Available at: https://doi.org/10.1016/j.infoandorg.2023.100478.

[8] Crawford, K. (2021) *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. 1st edn. New Haven: Yale University Press. Available at: https://doi.org/10.2307/j.ctv1ghv45t.

[9] Noble, S.U. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. 1st edn. New York: NYU Press. Available at: https://doi.org/10.18574/9781479833641.

[10] Sun, G. *et al.* (2022) "Data Poisoning Attacks on Federated Machine Learning," *IEEE internet of things journal*, 9(13), pp. 11365–11375. Available at: https://doi.org/10.1109/JIOT.2021.3128646.

[11] NCSC (2023). **ChatGPT and large language models: what's the risk?, available:** https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk