KHAZANAH
RESEARCH
INSTITUTE

VIEWS 22/23 | 9 NOVEMBER 2023

Visualising Societal Harms of Artificial Intelligence

Jun-E Tan

Introduction

Artificial intelligence (AI) has been a subject of much discussion lately, as industry leaders and experts within the tech space come out one after another, and in groups, to warn about its existential risks[1]. Some AI experts are calling to pause development and to regulate the space, especially within the context of runaway artificial general intelligence in the future[2]. Others have dismissed these warnings as hype, insisting that policymakers should not be distracted from the harms of narrow AI systems currently in use[3].

Essentially, these debates conceptualise AI and its problems differently, leading to different priorities in governing the technology. In this article, I contribute to the discussion by

[1] Centre for AI Safety (n.d.)
[2] Future of Life Institute (2023)
[3] Gebru et al. (2023)

Views are short opinion pieces by the author(s) to encourage the exchange of ideas on current issues. They may not necessarily represent the official views of KRI. All errors remain the authors' own.

This view was prepared by Jun-E Tan, a researcher from the Khazanah Research Institute (KRI). The author is grateful for the valuable comments from Rachel Gong, Gregory Ho Wai Son and Nik Syafiah Anis Nik Sharifulden.

Author's email address:
june.tan@krinstitute.org

An earlier version of this article was published in the LSE Southeast Asia Blog, under the same title.

Information on Khazanah Research Institute publications and digital products can be found at www.KRInstitute.org.

considering different levels of AI harms, focusing on existing technologies and impacts. I introduce relationality as a conceptual lens to untangle societal harm from individual and collective harm, and provide a visualisation to make the distinction sharper. At this crucial point when the world is moving towards AI regulation, clearer conceptualisation of AI's societal impacts will expand our options for better policy and governance mechanisms.

AI and AI governance

There is no agreed definition of AI[4], but to set the scene I point to the OECD Framework for the Classification of AI Systems[5] to emphasise the wide array of tasks common for AI technologies, such as recognition (of patterns in image, voice, video, etc), event detection, forecasting, personalisation, interaction support, goal-driven optimisation and reasoning with knowledge structures. These functions and outputs are usually provided through machines learning from large swathes of data.

To bring it closer to home, the types of AI that are in use in everyday life include recommender systems on social media and e-commerce sites, traffic navigation systems, facial recognition systems, and more recently generative AI like ChatGPT and Dall-E which have taken the world by storm. These are some technologies that have transformed, within a short span of time, how we interact with the world and with each other.

The nascent field of AI governance has had to keep up with the pace of technology evolution, as it seeks to shape the development, use, and infrastructures of AI[6]. Inadvertently, this necessitates discussions about potential risks and harms brought about by the technologies, so that appropriate rules and monitoring systems can be put into place to ensure safe and responsible use.

Different levels of AI harms

Defining "harm" as "a wrongful setback to or thwarting of an *interest*" (emphasis in original), Nathalie Smuha classifies AI harms into individual, collective and societal harms[7].

Individual harms affect an identifiable individual, such as a person whose privacy has been violated or who is wrongfully prosecuted because of a biased facial recognition system[8]. Collective harms affect a group of individuals with shared characteristics (such as skin colour or similar behaviour in browsing the Internet) and can be viewed as a sum of individual harms sustained. Societal harms however go beyond and above individual and collective concerns, such as in cases where the use of AI risks harming the democratic process, eroding the rule of law, or exacerbating inequality.

---

[4] Evans (2019)
[5] OECD (2022)
[6] Veale, Matus, and Gorwa (2023)
[7] Smuha (2021)
[8] Hao (2019)

Smuha argues that most of the work done within the area of legal frameworks for AI governance, such as data protection law and also procedural law more generally has focused disproportionately on setbacks to individual interests and relies on the individual to seek damage and take legal action. In her paper, she takes a pragmatic approach of focusing on governance mechanisms for societal harms of AI, drawing from the area of environmental law to propose measures that have been used in the context of environmental pollution. Ideas include public oversight mechanisms such as impact assessments and public monitoring mechanisms, to ensure that companies are transparent and accountable.

These interventions are useful to lead policy thinking away from the current paradigm of relying on individuals to shoulder the responsibility of protection against AI risks. However, Smuha stops short at offering a conceptual understanding of societal harms of AI. In particular, the difference between collective harms and societal harms in her classification begs for further clarification.

Viewing societal harms with a relational perspective

I argue that a framework of relationality might be able to provide us with the conceptual clarity needed to make this distinction, for clearer communication of AI's potential risks and to open up new pathways towards imagining solutions.

According to Jennifer Nedelsky, the relationality framework refers to the idea that human beings are fundamentally interconnected, starting from familial and romantic relationships, widening to more distant relationships for example with teachers and employers, and then to social structural relationships, such as gender and class relations[9]. Looking through a relational lens, a core value such as individual autonomy needs to be understood as a capacity made possible by constructive relationships (for instance, with the family and with the state), instead of as independence from others.

The relational perspective is not incidental but fundamental to our understanding of societal harms of AI and the associated downstream effects on individuals. In a paper establishing a relational theory of data governance, Salome Viljoen makes the point that data production in a digital economy generates much of its revenue (and harms) from putting people into "population-based relations with one another", in identifying patterns and providing predictions through creating profiles of people with shared characteristics[10].

Scholarship on AI ethics grappling with systemic harms have also touched on relationality as the missing piece in countering dehumanising effects of AI systems that further marginalise the marginalised. For instance, Sabelo Mhlambi has argued that automated decision-making systems (ADMS) built without considering the interconnectivity among individuals perpetuate social and

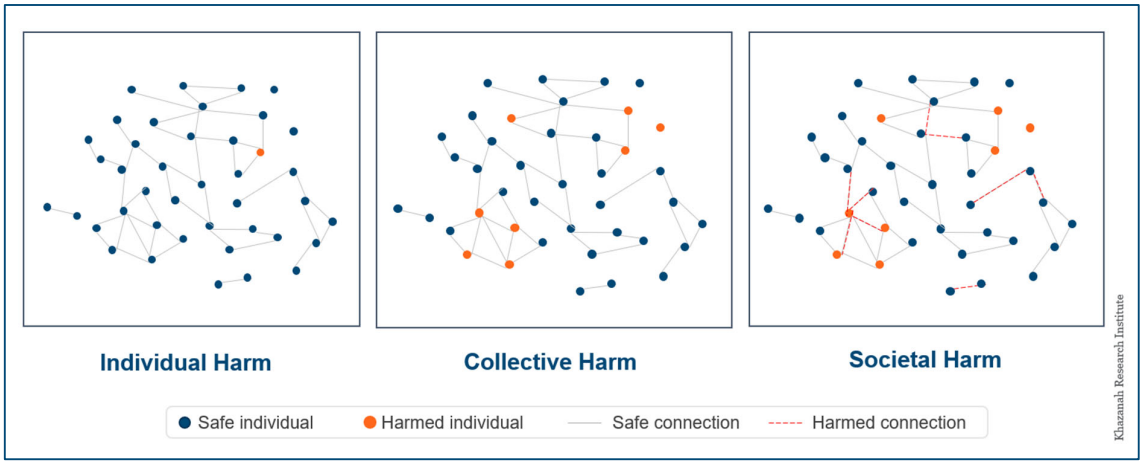
[9] Nedelsky (2011)
[10] Viljoen (2020)

economic inequalities by design[11]; Abeba Birhane has also recommended a relational ethics approach in thinking about personhood, data and justice to address algorithmic injustice[12].

Visualising societal harms of AI

To simplify the notion of relationality, I provide the following visualisation for consideration, to illustrate the different types of AI harms.

Figure 1 Visualisation of different types of AI harms



Source: Khazanah Research Institute

The boxes within Figure 1 contain the same social network graph, representing individuals and their connections within a given society. In the first box illustrating individual harm, the orange dot represents an identifiable person whose interests have been affected by AI, while the blue dots are unaffected. In the second box on collective harm, orange dots indicate groups of individuals who have suffered some negative consequences as a result of algorithms targeting their population based on some shared characteristics.

In the third box on societal harm, we shift the emphasis from the dots to the lines, moving our focus to the connections between the individuals. Red dashed lines represent harmed connections between individuals. As indicated in the figure, even connections between persons who have not sustained direct harms from AI can be affected.

For example, differences in beliefs on deeply polarising and emotionally charged issues such as climate change can foment hostility within and between communities[13], a phenomenon known as affective polarisation[14]. Researchers from Persuasion Lab have demonstrated that political

[11] Mhlambi (2020)
[12] Birhane (2021)
[13] Dresden University of Technology (2023)
[14] Iyengar et al. (2019)

advertising on climate issues in Europe targets younger users disproportionately, with a "marked decrease" in the reach of such content to older people[15].

This provides an indication of how different segments of society are exposed to different perceived realities within their media environments, due to the use of automated recommender systems to push content to groups most susceptible to given messaging. Linking back to affective polarisation, an unintended consequence of such algorithmic optimisation presumably affects social ties, visualised in Figure 1 as red dashed lines between individual dots.

The resulting impact on social cohesion and trust because of differing baseline beliefs are not captured in the individual or collective harm scenarios. Direct harms on individuals and groups may be invasion of privacy or behavioural manipulation, but impacts on relationships and ties (and indeed, resulting climate politics that would have downstream impacts on public and planetary health) are invisible.

Policy implications

The conceptualisation of societal harms expands the scope and visibility of harms brought about by AI technologies, addressing a crucial challenge in the understanding and communication of the issue. Widespread erosion or disruption of social connections without appropriate safeguards risks societal fragmentation or in worse cases, collapse, which has been observed in the case of the genocide in Myanmar linked to the amplification of hate speech on Facebook[16].

As we move into a phase of AI regulation, making sense of potential risks becomes increasingly urgent. Policymakers and researchers globally are considering different governance mechanisms for AI, and these include legal and ethical frameworks, technical standards, AI impact assessments, auditing and reporting tools, databases of AI harms, and so on.

With a relational framework considering the nature and quality of relationships or social relations, a range of societal harms of AI becomes visible. Ongoing research in the Khazanah Research Institute (KRI) aims to unpack societal impacts of AI using this lens, looking at social inequality and justice, as well as social roles and institutions that have been altered as a result of rapid advancements in the field.

Conclusion

Within the space of this article, I have distinguished societal harms from individual and collective harms by moving the focus of analysis from individual units within society to the connections and relationships between them. At this juncture when debates on AI governance are being translated into actual mechanisms and regulatory frameworks, it is crucial to ensure that all forms of AI harms are identified and addressed.

---

[15] Persuasion Lab (n.d.)
[16] Milmo (2021)

AI technologies have the potential to bring great benefits to society if they are designed well and optimised for pro-social goals[17]. Flipping the logic, if such powerful technologies are allowed to disrupt and break social connections without appropriate safeguards, societal development and progress may well be reversed. Future research and policy direction should therefore devote more attention and resources into the areas of relational impacts of AI, such as issues of social cohesion, inequality, and institutions.

References

Birhane, Abeba. 2021. "Algorithmic Injustice: A Relational Ethics Approach." *Patterns (New York, N.Y.)* 2 (2):100205. https://doi.org/10.1016/j.patter.2021.100205.

Centre for AI Safety. n.d. "Statement on AI Risk." Accessed October 13, 2023. https://www.safe.ai/statement-on-ai-risk.

Dresden University of Technology. 2023. "Immigration Polarizes the Right, and Climate Change Polarizes the Left: Study Reveals Europe's Fault Lines." *Phys.Org* (blog). July 18, 2023. https://phys.org/news/2023-07-immigration-polarizes-climate-left-reveals.html.

Evans, Ian. 2019. "'Nobody Agrees on What AI Is' – How Elsevier's Report Used AI to Define the Undefinable." *Elsevier Connect* (blog). January 18, 2019. https://www.elsevier.com/connect/nobody-agrees-on-what-ai-is-how-elseviers-report-used-ai-to-define-the-undefinable.

Future of Life Institute. 2023. "Pause Giant AI Experiments: An Open Letter." *Future of Life Institute* (blog). March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gebru, Timnit, Emily Bender, Angelina McMillan-Major, and Margaret Mitchell. 2023. "Statement from the Listed Authors of Stochastic Parrots on the 'AI Pause' Letter." DAIR Institute. March 31, 2023. https://www.dair-institute.org/blog/letter-statement-March2023/.

Hao, Karen. 2019. "AI Is Sending People to Jail—and Getting It Wrong." MIT Technology Review. January 21, 2019. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22 (1):129–46. https://doi.org/10.1146/annurev-polisci-051117-073034.

Mhlambi, Sabelo. 2020. "From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance." Carr Center Discussion Paper Series. Cambridge, MA: Harvard Kennedy School, Harvard University. https://carrcenter.hks.harvard.edu/publications/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial.

[17] Overgaard and Woolley (2022)

Milmo, Dan. 2021. "Rohingya Sue Facebook for £150bn over Myanmar Genocide." *The Guardian*, December 6, 2021, sec. Technology. https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence.

Nedelsky, Jennifer. 2011. *Law's Relations: A Relational Theory of Self, Autonomy, and Law*. Oxford University Press, USA.

OECD. 2022. "OECD Framework for the Classification of AI Systems." 323. OECD Digital Economy Papers. OECD. https://www.oecd-ilibrary.org/docserver/cb6d9eca-en.pdf?expires=1658971383&id=id&accname=guest&checksum=189AA5CC6AD125404EA182E173EB33AD.

Overgaard, Christian Staal Bruun, and Samuel Woolley. 2022. "How Social Media Platforms Can Reduce Polarization." *Brookings* (blog). December 21, 2022. https://www.brookings.edu/articles/how-social-media-platforms-can-reduce-polarization/.

Persuasion Lab. n.d. "Are Climate Crisis Debates in the European Union Reaching Older People?" Accessed October 13, 2023. https://ad.watch/story/story0.html.

Smuha, Nathalie A. 2021. "Beyond the Individual: Governing AI's Societal Harm." SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3941956.

Veale, Michael, Kira Matus, and Robert Gorwa. 2023. "AI and Global Governance: Modalities, Rationales, Tensions." *Annual Review of Law and Social Science* 19. https://discovery.ucl.ac.uk/id/eprint/10171121/1/Veale%20Matus%20Gorwa%202023.pdf.

Viljoen, Salome. 2020. "A Relational Theory of Data Governance." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3727562.