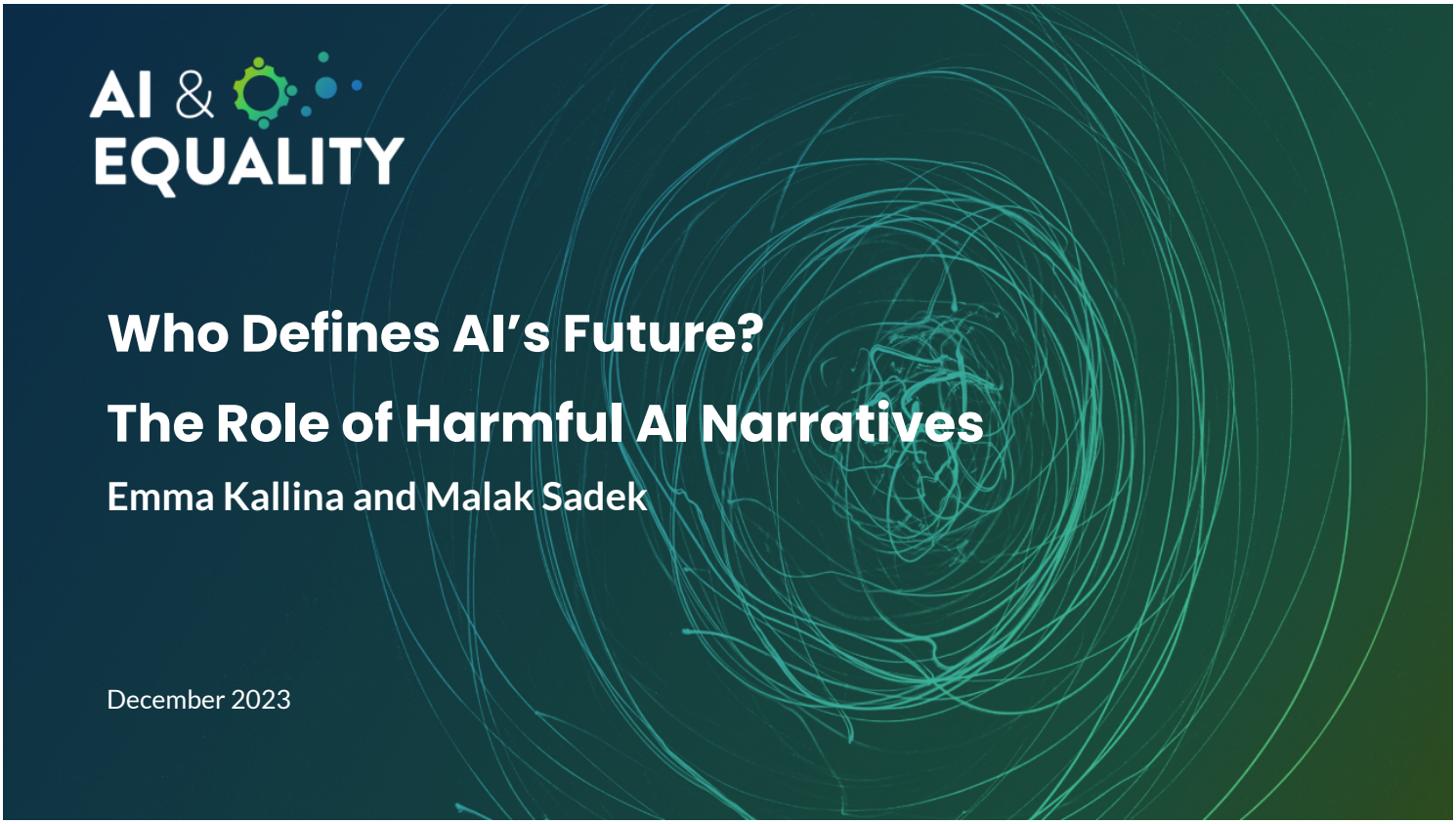

Who Defines AI's Future?
The Role of Harmful AI Narratives

Emma Kallina and Malak Sadek

December 2023

Abstract

As AI-based systems are becoming more mainstream, they are increasingly discussed within public spheres. Thereby, several prevailing narratives around AI systems form almost stereotypical perceptions as to who is in control, who will benefit, and what these systems are capable of. The narratives dominate the development process, resulting in a narrow vision that largely fails to counteract the imprint of historical (and present) power structures in future technologies (Hoffmann, 2019; Lee & Singh, 2021). Since they define expectations of power, they further limit the access of underrepresented groups - both by decreasing the perceived necessity to include them, as well as by lowering their own perception of self-efficacy. The former causes the AI development process to still lack the mandate and process to consolidate stakeholders, especially underrepresented groups. The latter counteracts attempts by these underrepresented groups themselves to claim agency in the development process, further distorting the values included in AI systems. In the following article, we will outline three prominent, yet harmful AI narratives and provide examples of movements, organisations, or tools working against these.

The 'Whiteness' of AI and Other Structural Injustices

Most humanoid instances of AI, both in media and in real life are portrayed as 'white' (Cave & Dihal, 2020 [1]). This subtle racialization of anthropomorphised machines is both a representation of the race of the people creating them, where whiteness creates whiteness, and a dissuading barrier for people of colour to break the status-quo of machines not made by them nor for them, perpetuating a "feedback loop" (West, Whittaker, & Crawford., 2019 [2]) or a "cycle of social injustice" (Cave & Dihal, 2020). These injustices do not only stop at race but also affect minorities across a number of diversity axes. Since a homogeneous group of people tends to generalise broadly based on the characteristics represented among them, people and elements that do not fit into the creators' categories are often disadvantaged by the systems created by them (Hollaneck, 2021 [3]; West, Whittaker, & Crawford, 2019). This need to generalise and simplify complex problem spaces erases people's experiences and histories because of experts' reluctance to tackle them, or their lack of awareness of their existence in the first place (Majlesein, 202 [4]) (referred to as the "privilege hazard" by D'ignazio and Klein (2020)[5]).

The extremely complex and interrelated nature of existing biases in AI systems motivated the Leverhulme Centre for the Future of Intelligence [6] to call biases in AI systems a "vicious cycle": The lack of diversity among researchers and practitioners leads to data-based and algorithmic biases, as well as a narrow vision of the interests and values prioritised during system development. These, in turn, create an image and a narrative around AI systems and their design that limit the types of people attempting to become AI researchers and practitioners, further perpetuating existing power imbalances.

Best Practice Examples: AymurAI

AymurAI [7] is an open-source tool, designed to support criminal justice courts in Latin America in their collection and publication of data about gender-based violence. AymurAI was planned by a female team, illustrating how a non-traditional group of AI creators can envision and build AI tools that serve highly important, but often overlooked use cases.

Technological Determinism

Technological determinism describes a reductionist understanding of technological innovation. It asserts that society's technology progresses by itself and the rules of efficiency, influencing the societal structures and values around it (Chandler, 1995 [8]). In other words, technological determinism describes "the belief in technology as a key governing force in society" (Smith & Marx, 1994, p. 2 [9]). This theory neglects that technology is shaped by society and that societal factors enable - or prevent - the development of a specific technology (e.g. D'ignazio & Klein, 2020). Instead, technology is regarded as neutral, free of any power structures or bias, merely a tool. This characterises AI systems as unable to disadvantage someone - harm is solely created by its users. Additionally, technological determinism renders efficient high-tech systems unavoidable, depicting the communities around them as passive spectators whose interventions will be unable to change technological advancements.

BigTech companies often harness technological determinism to justify their products. For example, in Mark Zuckerberg's view, Facebook is optimised for "showing people what they think is meaningful" (Metz, 2017 [10]). By implying that the algorithm neutrally reacts to human desires, the company pushes away any responsibility for the detrimental consequences of their platform (ironic, if we consider that their RAI guidelines include 'Accountability & Governance', Pesenti (2022) [11]), leading to an "environment blind[ness]" with regards to relevant socio-cultural facets of technology and its implications (Whitworth and Ahmed, 2014 [12]). Additionally, it neglects the values that the company instils in the algorithm to serve its management goals thus making it inherently biased (Martinho et al., 2021 [13]). Such a perspective is detrimental.

Furthermore, endeavours related to AI are generally framed as a solely technical and abstract pursuit. This leverages technical practitioners as superior 'wizards' whilst diminishing the participation from communities and non-technical experts (D'ignazio & Klein, 2020). However, a solely technical approach creates a socio-technical gap between social requirements and technical designs (Ehsan et al., 2023 [14]). Aside from the obvious disconnect that this gap creates between the technologies created and those using them or being affected by them, it can also lead to serious ethical problems and misuses (Ehsan et al., 2023; Schniderman, 2020 [15]; Selbst, 2019 [16]). This is especially true for AI-based

systems which are considered highly socio-technical (van de Poel, 2020 [17]). The result is a number of gaps between the creators of the AI system, the roles responsible for maintaining its wider infrastructure, those making business decisions, and the people using it and being affected by it.

Best Practice Example: EU AI Act Stakeholder Consultation

The EU AI Act is a proposed law on AI regulation by the European Union, the first of its kind (Future of Life Institute, 2022 [18]). During the process of drafting the act, citizens and stakeholders were able to provide feedback on the draft from February to June 2020 via a custom-made platform (European Commission, 2020 [19]). This counteracts the narrative of AI determinism in two ways: Firstly, the EU AI Act is a form of external regulation, designed to guide the further development of AI systems. This is an active stance, contradicting the passive notion of AI determinism. Secondly, it empowers the people that will be users and stakeholders of these systems by giving them a voice in shaping the regulation that in turn will shape the technology of the future.

AI Race

In recent policy and strategic documents from governments, a strong narrative around the "race for technological superiority" in AI emerged (Cave & Oh´ Eigeartaigh, 2018, p. 36 [20]). It is fueled by the perception that the country establishing leadership will profit from scientific, infrastructural, and economic advantages, i.e. "the winner takes all" (Cave and Oh´ Eigeartaigh 2018). This is enhanced by the conviction that AI advantages can be applied on a vast scale across sectors, initiating a ubiquitous transformation towards increased efficiency (for more details see Cave and Oh´ Eigeartaigh (2018) who systematically analyse potential states of the AI race). Such a narrative comes with several problems: It prioritises speed over thoroughness and reflection. However, most steps to "not only [make] AI more capable, but also [maximise] the societal benefit of AI" (Open Letter of the Future of Life Institute of Life Institute (2015) [21]) - e.g. the iterative involvement of multiple stakeholders, participatory design and careful roll-outs with continuous evaluation - require time and care.

If confronted with ethical issues, the AI Race narrative pushes at best towards the creation of reactionary, passive measures such as guidelines and checklists to enable a "comply and deploy" mindset among practitioners (Crawford & Calo, 2016 [22]). As a result, little active

reflection is practised while creating these systems, leading to harmful practitioner mindsets such as “rejecting practices or downplaying the importance of values or the possible threats of ignoring them" (Manders-Huits & Zimmer, 2009 [23]) or shifting responsibility onto other stakeholders alone (Mitchell, 2020 [24]).

Best Practice Example: Microsoft’s Judgement Call Cards & MIT’s AI Blindspot Cards

The card deck ‘Judgement Call Cards’ [25] was developed to support AI practitioners in considering ethical aspects and stakeholder values early in the creation of an AI-based system. The tool tackles the typical ‘move fast and break things’ mentality that focuses solely on technical perspectives. The practitioners are prompted to consider who their system’s stakeholders are and then write reviews for the planned system from their perspectives, representing their various, potentially contradicting values. This counteracts the AI race narrative by allowing practitioners to reflect on the experiences that their AI-based system might provide for different stakeholders, triggering an awareness of the kinds of harms they might cause. The resulting empathy (hopefully) results in increased motivation to identify mitigation strategies for these negative experiences, e.g. to explore potential countermeasures in collaboration with the communities affected.

MIT’s AI Blindspot Cards [26] are another deck of cards for AI practitioners. The cards highlight a number of ‘blindspots’ that AI practitioners are likely to suffer from along the various design phases for AI-based systems. Covering planning, building, deploying and monitoring, each card presents a potential blindspot that might cause harm if overlooked. The cards include potential questions to ask, stakeholders to engage with, and a real-world example illustrating the problems that might occur if the corresponding blindspot is not considered. By highlighting specific stakeholders and illustrating concrete and real harms that specified blindspots might cause, these cards work towards combating the AI race narrative by encouraging practitioners to slow down and reflect on the direct and specific problems they might enable if they do not engage in proactive measures to address potential blindspots before they occur. Similar to Microsoft’s Judgement Call cards, the hope is that solidifying these harms as real and present through the use of named stakeholders or specific examples will motivate practitioners to work actively towards ensuring these harms do not occur.

Conclusion

Narratives around AI systems shape perceptions about who is in control, who benefits, and what these systems can or should achieve. This article has shed light on three harmful AI narratives that form inaccurate misconceptions among practitioners and the public which, in some cases, lead them to becoming self-fulfilling prophecies: Firstly, the "whiteness of AI", perpetuating the cycle of social injustice and data-based biases that affect minority groups. Secondly, technological determinism that portrays AI systems as neutral tools, absolving companies of responsibility for the detrimental consequences of their platforms. Lastly, the idea of an AI race creates a gap between the required time-intensive socio-technical considerations and the AI development practice.

This work hopes to raise awareness regarding these narratives and how they might harm valuable and meaningful progress towards the creation of human-centred, responsible AI. The motivation is to investigate deeper, systemic problems regarding the creation and perception of AI-based instead of focusing on the technology alone. Through providing examples of how these narratives can be reversed, we hope to inspire more projects that work towards a more diverse, feminist, and democratic future of AI.

References

[1] Cave, S., Dihal, K. The Whiteness of AI. Philos. Technol. 33, 685–703 (2020). https://doi.org/10.1007/s13347-020-00415-6

[2] West, S., Whittaker, M., Crawford, K. (2019). Discriminating systems: Gender, race, and power (Tech. Rep.). AI Now Institute.

[3] Hollaneck, T. (2021). Design choices in ML systems – cambridge digital humanities' social data school. Retrieved from https://www.youtube.com/watch?v=OvcsaRj6dDo&feature=emb_logo

[4] Majlesein, S. (2021), Building solid experiences on unsolid ground, Presentation at World Interaction Design Day (IxDD) and can be accessed at: https://vimeo.com/618306283.

[5] D'ignazio, C., & Klein, L.F. (2020). Data feminism. MIT press. https://data-feminism.mitpress.mit.edu/

[6] The centre's work can be viewed on their website at: http://lcfi.ac.uk/

[7] Their work can be seen at: https://sites.google.com/view/aymurai-en

[8] Chandler, D. (1995). Technological or media determinism.

[9] Smith, M.R., & Marx, L. (1994). Does technology drive history?: The dilemma of technological determinism. Mit Press.

[10] Metz, C. (2017). Mark Zuckerberg's Answer to a World Divided by Facebook Is More Facebook, https://www.wired.com/2017/02/mark-zuckerbergs-answer-world-divided-facebook-facebook/

[11] Pesenti, J. (2022). Facebook's five pillars of Responsible AI. Retrieved from: https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/

[12] Whitworth, B. and Ahmad, A. (2014), Socio-Technical System Design, Interaction Design Foundation.

[13] Martinho, A., Poulsen, A., Kroesen, M., Chorus, C. (2021). Perspectives about artificial moral agents. AI and Ethics, 1 (4), 477–490. https://doi.org/10.1007/s43681-021-00055-2

[14] Ehsan, U., Saha, K., Choudhury, M. and Riedl, M. (2023), 'Charting the sociotechnical gap in explainable ai: A framework to address the gap in XAI, Proceedings of the ACM on Human-Computer Interaction (34). https://doi.org/10.48550/arXiv.2302.00799

[15] Schneiderman, B. (2020), 'Human-centered artificial intelligence: Three fresh ideas', AIS Transactions on Human-Computer Interaction (THCI) 12(3). https://doi.org/10.17705/1thci.00131

[16] Selbst, A. (2019), 'Accountable algorithmic futures'. Retrieved at: https://points.datasociety.net/building-empirical-research-into-the-future-of-algorithmic-accountability-act-d230183bb826

[17] van de Poel, I. (2020), 'Embedding values in artificial intelligence (ai) systems', Mind and Machines 30, 385–409.

[18] Future of Life Institute (2022), The Artificial Intelligence Act. https://artificialintelligenceact.eu/

[19] European Commission (2020), White Paper on Artificial Intelligence: Public consultation towards a European approach for excellence and trust. https://digital-strategy.ec.europa.eu/en/library/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence-and

[20] Cave, S., & ´Oh´Eigeartaigh, S.S. (2018). An AI race for strategic advantage: Rhetoric and risks. Proceedings of the 2018 aaai/acm conference on ai, ethics, and society (p. 36–40). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3278721.3278780

[21] The letter is readable at: https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/

[22] Crawford, K. and Calo, R. (2016), 'There is a blind spot in AI research', Nature 538, 311–313. https://doi.org/10.1038/538311a

[23] Manders-Huits, N. and Zimmer, M. (2009), 'Values and pragmatic action: The challenges of introducing ethical intelligence in technical design

communities', The International Review of Information Ethics 10, 37–44. https://doi.org/10.29173/irie87

[24]Mitchel, M. (2020). Intentional Ignorance is a Value-Laden Choice. Presentation at the PAIR Symposium. Accessible at: https://www.youtube.com/watch?v=TcE6_NPjvuo

[25] Accessible at: https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgment call

[26] Accessible at: https://aiblindspot.media.mit.edu/

Biographies

Emma Kallina is a PhD student at the University of Cambridge. Her research focuses on democratising the AI development process through improved stakeholder involvement along the entire lifecycle of AI systems. She believes that everyone who will be impacted by a system should have a say in how it is created.

Malak Sadek is a Design Engineering PhD student at Imperial College London. Her work lies at the intersection of AI and design. Specifically, she's interested in creating tools that help collaboratively designing conversational AI that respects stakeholders' values. The aim is to work towards ensuring AI-based systems are better aligned with the values of those involved in making and using them.