Towards Diverse AI: Can an AI-Human Hybrid Council Prevent Future Apartheids?

Gabriel Diniz Junqueira Barbosa[0000−0003−3929−2736] and
Simone Diniz Junqueira Barbosa[0000−0002−0044−503X]

PUC-Rio, R. Marques de Sao Vicente, 225, Gavea, Rio de Janeiro, RJ, Brazil
gabrieldjb@gmail.com, simone@inf.puc-rio.br

Abstract. Artificial intelligence (AI) is becoming more prevalent in today's society. However, decisions are often made based on single AI models, which we call *single-minded AI*. The use of single-minded AI might bring great harm, while the use of AI-human collectives might help debias the decision-making process and thus promote better decisions. We illustrate through a speculative design fiction some of such potential risks and benefits. Our goal is to help frame the discussion on some of the necessary advances in managing AI to allow for better human-AI collaboration.

Keywords: AI-Human Collective Systems · Collective Intelligence · Decision making.

1 Introduction

During the years of 1948 to 1991 the South African government instituted oppressive policies, aimed at keeping control over a specific segment of their population. If, in that time, the Apartheid-instituting government had access to AI algorithms, perhaps history would have been different. That is the premise of our design fiction. In it, we create a narrative in which a major point in history is negatively impacted by single-minded AI technology, and explain how collective AI might help avoid such situations. By speculating through design fiction, we are able explore ethical and social issues of everyday life, asking what-if questions and opening the space for debate and discussion [5]. This work is meant to serve both as a warning on AI's potential for oppression and as a reframing of that potential for good.

Throughout our narrative, we reference tools that already exist today to illustrate how the narrative could come to pass. Some institutions use such tools to keep control over people [6, pp. 7-13], endangering freedoms that are essential to democracy. The following two-part story seeks to illustrate these potentials for abuse and a path to mitigate them.

2 Single-minded AI: The Menace

As Mandela walked away from prison, he knew he was not the same man he was 27 years earlier. The world had also changed. AI technology had made enormous

strides during his prison sentence. Artificial intelligence tools were now widely available, with many governments using them in various ways.

The Apartheid-enforcing South African government was among the users of this technology. Various AI-based tools decided where you could go, what you could do, and who you could meet [6, pp. 7-13]. All of them were connected by a single artificial consciousness, called Orpheus. Orpheus made most of the governmental day-to-day decisions, with humans creating rules for it to follow but being kept out of the loop from the final decision making.

Mandela knew that Orpheus would be a great opponent in his fight for South African black liberation, as the AI strictly followed the racial separation rules implemented by the government. He decided to fight back by going into politics. However, when he tried to get approval to run for office, he was denied, as Orpheus had decided he was too dangerous, because of his criminal past [2]. He tried to appeal the ruling, but there were no processes in place for the people to challenge Orpheus [6, pp. 36-69].

Having seen the first push-back from this new AI-empowered government, Mandela decided to gather his supporters to protest against Orpheus. Every time his people gathered, they were dispersed by the police. The police had authority to do so because facial recognition software had, in these gatherings, recognized individuals whom Orpheus had labeled as dangerous. This system effectively prevented public gatherings, and protected the inequities supported by the Apartheid-enforcing government [7, 3].

Unable to argue with a machine, Mandela found no sign of hope. How could he and his allies defeat some algorithm that had singular control over many facets of day-to-day life? How could he argue against decisions whose reasoning was opaque? How could he prove that he and his allies had changed and become peaceful when the AI only followed past tendencies?

3 Diverse AI: A New Hope

As these forms of AI-boosted abuses started happening in authoritative regimes, the UN instituted councils where humans and AI could collaborate and discuss these issues. All representatives now have an ensemble of AI assistants who represent their interests and make them explicit, therefore creating a council with diversity of opinion [11]. Since all biases are explicit – and it can be checked that there are various interests being collectively represented in these councils – and humans have the last say, they would not easily succumb to authoritative decisions [1].

These collaborative human-AI councils started thinking about how they could avoid AI technology abuses, like those happening in South Africa. The AI counterparts would provide different rationales, explaining them to humans [10], and allowing for richer discussions [11, 8]. These improved discussions provided arguments for the international community to eventually convince the South African government to abandon Orpheus, allowing for Mandela and his allies to achieve South African liberation.

Single-minded AI, such as Orpheus, threatened freedom, but collective AI allowed humans to arrive at better decisions and prevent many humanitarian disasters. Humans work better together, and so might AI. By making biases and human interests explicit, they allowed for more transparent human-AI collaboration. By explaining the AI's rationale, humans were allowed to question them, and decide how much weight they should assign to their suggestions. By making AI work in a collective structure, people can ensure that no one AI can skew the collective, thus leading to more measured decisions [1].

AI offers humans great power. It is up to humans to decide how to use it, for good or for bad.

4 Discussion

AI will probably have increasing impact in the coming years. As illustrated in our Mandela story, oppressive regimes might seek to use these tools to perpetuate inequities and control their people [6, pp. 140-160]. AI is agnostic, and can serve for either good or bad. What will determine the outcome is the way in which humans will use these tools. There is much work to be done to ensure that humans use AI wisely, and that opportunities for abuse are mitigated. Explainability is still a major challenge for AI. For there to be proper human-AI collaboration, it ought to be possible for humans to understand how AI models arrive at their decisions [10, 4]. Moreover, humans also need to have some oversight or even control over them, to ensure that these AI models behave in the way that humans expect them to [1].

Human-AI collaboration offers great potential, but this form of interaction should be structured properly. By having AI models work together and having their biases explicit, it might be possible to avoid potential skews in collective decision-making [8]. We can organize these collectives in different ways, each with their own advantages and disadvantages. In Figure 1, we show a couple of such organizational structures. In addition to different structures, there can also

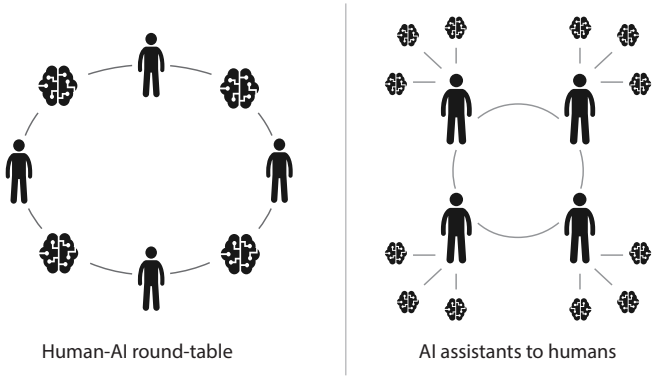

Fig. 1. A couple of structures for decision-making teams

be different ways to weigh each member's opinion, be they humans or AI. Selecting adequate structures and weights will then be essential for better collective decision-making.

Single minded AI, however, may be a source of concern. Individual models may not be transparent, and may generate extremely biased outcomes. If humans put these individual models in positions of power, and do not allow for transparency and appeal, negative consequences become quite likely [9].

References

1. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E.: Guidelines for Human-AI Interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 3:1–3:13. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300233
2. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research p. 0049124118782533 (Jul 2018). https://doi.org/10.1177/0049124118782533
3. Bowyer, K.W.: Face recognition technology: security versus privacy. IEEE Technology and Society Magazine **23**(1), 9–19 (2004). https://doi.org/10.1109/MTAS.2004.1273467
4. Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H.: Explaining Decision-Making Algorithms Through UI: Strategies to Help Non-Expert Stakeholders. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 559:1–559:12. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300789
5. Dunne, A., Raby, F.: Speculative Everything: Design, Fiction, and Social Dreaming. The MIT Press (Dec 2013)
6. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press (Jan 2018)
7. Introna, L., Wood, D.: Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems. Surveillance & Society **2**(2/3) (Sep 2002). https://doi.org/10.24908/ss.v2i2/3.3373
8. Mohammed, S., Ringseis, E.: Cognitive Diversity and Consensus in Group Decision Making: The Role of Inputs, Processes, and Outcomes. Organizational Behavior and Human Decision Processes **85**(2), 310–335 (Jul 2001). https://doi.org/10.1006/obhd.2000.2943
9. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, New York, 1 edition edn. (Sep 2016)
10. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing Theory-Driven User-Centric Explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 601:1–601:15. CHI '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300831
11. Wang, X.H.F., Kim, T.Y., Lee, D.R.: Cognitive diversity and team creativity: Effects of team intrinsic motivation and transformational leadership. Journal of Business Research **69**(9), 3231–3239 (Sep 2016). https://doi.org/10.1016/j.jbusres.2016.02.026