

From Polarization to Prevention: Can AI Save Global Diplomacy?

Moumita Arun

Symbiosis School for Liberal Arts, Pune, Symbiosis International (Deemed University)

Correspondence: moumita.arun@ssla.edu.in

How to Cite this Paper

Arun, M. (2025). From polarization to prevention: Can AI save global diplomacy? *Confluence: Journal for Interdisciplinary Studies*, Volume IX, 65-76.

Abstract

Technological progress has never moved faster, yet our capacity for division has kept pace. The past decade has redefined what is possible with breakthroughs in artificial intelligence and global communication which provides a wider range of tools for cooperation. And yet, innovation in technology doesn't necessarily translate into innovation in peacebuilding as made evident in the protracted stalemate in Israel-Palestine to the civil wars in Yemen and Syria, violent conflict remains a stubborn constant. The increase in geopolitical crisis, political polarisation and civil unrest reveals a dissonance between our technological tools and our political realities. Therefore as relevant as our traditional diplomatic frameworks are, they appear increasingly insufficient to address the layered dynamics of modern conflict such as algorithmically intensified polarization and disinformation campaigns that weaponize social media. Just as the boundaries between human agency and artificial intelligence become muddled, it becomes imperative to consider how human-AI collaboration can be leveraged- not as a technical solution but as a practice that provides us a critical lens through which we understand and respond to these challenges. One of the most under examined but also pervasive areas of this collaboration is social media. More than just a passive mechanism for curating information based prior interactions such as likes, shares, and clicks, these platforms actively

create highly personalized content ecosystems. While these ecosystems improve user retention, they can also establish echo chambers: digital spaces that expose individuals almost exclusively to information and perspectives that reinforce preexisting beliefs. Beyond influencing decision-making, the algorithmically enforced walls also chip away at shared epistemologies, and may further entrench social divisions. Against a backdrop of global tension, such dynamics raise critical questions about the role of AI-driven communication in either resolving or exacerbating conflict.

The Role of Social Media Algorithms in Escalating Conflicts

Social media platforms have become transnational actors that influence global politics. Furthermore, algorithmic networks, focused on maximizing user engagement, facilitate the reinforcement of ideological divisions and public discourse and in some instances contribute to violence by amplifying these ideologies. Social media algorithms trained on user engagement behaviours like likes, shares and clicks, provide a tailored filtering of content for its user that aligns with their existing preferences. Therefore while this process may enhance user experience, it also fragments public digital spaces that were once conducive for pluralistic dialogue across ideological barriers. This filtering of content on social media creates echo chambers and filter bubbles: environments in which people are surrounded by like-minded views while dissenting opinions are filtered from view. These processes are motivated by confirmation bias, the tendency to seek information to reinforce one's beliefs, and operationalised through algorithmic design (Peckham, 2025). The end result is a digital ecosystem that structurally rewards user engagement without regard for diversity or accuracy producing conditions for political polarization and social fragmentation.

Echo chambers do not merely reinforce existing beliefs, they radicalize them. These ideologically homogeneous networks of individuals are more likely to adopt more extreme positions, a phenomenon termed group polarization. Platforms such as Meta (formally Facebook) and X (formally Twitter) are especially prone to amplifying these processes due to their tendency to reward divisive or emotionally charged content that boosts engagement-based metrics (Cinelli et al., 2021). The consequences of such algorithmic walls are evident in the global political crisis. One notable example is the Rohingya crisis in Myanmar in 2017, where Meta's algorithm influenced the escalation of ethnic violence (De Guzman, 2022). The platform exacerbated

factual violence in the real-world through the substantial amount of hate speech and misinformation directed at the Rohingya Muslim minority. Though international observers and civil society actors raised concerns about global tech governance and digital accountability, at the time, Meta did not intervene. This case demonstrates how platform algorithms that are designed without geopolitical foresights can become complicit in mass violence (BBC News, 2021).

Algorithmic amplification of political polarization is not restricted to “fragile” democracies. Social media platforms have turned into battlegrounds for information warfare and psychological operations in established liberal democracies such as the United States, particularly during the Presidential elections. Researchers observed the formation of partisan echo chambers emerge during the 2016 and 2020 U.S. Presidential Elections driven by both users and automated accounts. More than 6.6 million tweets containing links to fake news sources were shared in 2016 alone, roughly ten websites containing 65% of the pro-Republican disinformation (Knight Foundation, 2024). AI Bots and extremely partisan users spread false claims about everything from mass voter fraud to Hillary Clinton’s health, which fuelled widespread public distrust in democratic institutions. While overtly fake content declined by 2020 partly due to improved platform moderation and user awareness, misinformation through subtler forms such as selective reporting and emotionally charged memes still persist thereby deepening echo chambers. Ideological influencers were more divided than in prior years, users were more shielded from opposing viewpoints (*First Findings From US 2020 Facebook & Instagram Election Study Released*, 2023). Similar trends were seen in Meta and Instagram, where the algorithm sorting led sufficiently like-minded users to content that aligned with their ideology while trying to limit the exposure to contrasting information. However, this threatens broader concerns about the erosion of democratic trust and its implications on international governing bodies and global norms.

The increasing influence of non-state actors, like the tech Mongols, in state agendas complicates the previously discussed relationship between algorithmic influence and global politics. Tesla and X’s CEO Elon Musk’s increasing involvement with U.S. politics- including his alignment with Donald Trump, his financial contributions to political campaigns, and global commentary on foreign affairs illustrates how private actors can wield disproportionate power in international relations. Musk’s executive control over X which affects policy discussions involving Brazil, Germany and Ukraine positions him as an emerging geopolitical actor and not just a mere businessman. His appointment to lead the Department of Government Efficiency (DOGE) under the Trump administration symbolizes the merger of political authority and technology that could be largely unregulated by existing international norms (Us, 2025). Moreover, the manipulation of political systems for the benefits of private actors is brought into question.

If social media algorithms can escalate conflict, can they also be harnessed to prevent it?

The Potential for AI and Technology in Conflict Resolution

Mediators and diplomats are beginning to explore the potential for digital peacebuilding tools—potentially AI-powered algorithmic monitoring systems that detect early warning signs of political polarisation, disinformation clusters and intergroup hostility. AI is emerging as a critical tool for international actors in addressing emerging threats because of its ability to process large amounts of data to identify shifts in sentiments across diverse populations, alongside radicalisation trends.

Through its applications in monitoring early warning signs of conflicts, for example hate speech, increasing polarization or misinformation in online conversations, artificial intelligence is acting as an essential tool for preventing violent escalations, especially in transitional political environments. PeaceTech Lab, a Washington-based organization that applies machine learning to conflict zones is a prominent example of this. By analyzing digital communications and social media sentiment, PeaceTech Lab maps what it terms the “*DNA of Conflict*” to track shifts in language, tone and narrative structures of social media content (*What Is Peacetechnology? A New Environment Brings New Opportunity* | Amazon Web Services, 2022). Their tools have flagged spikes in “inflammatory” hate and rhetoric speech in countries like Kenya and Colombia, as such providing mediators with timely data to shape intervention strategies (Heidebrecht, 2022). This can be vital in shaping early diplomatic responses by multilateral actors such as the UN or regional bodies like the African Union.

However, in digital spaces where extremist propaganda circulates, AI can play a more proactive role in interrupting pathways to political radicalisation and not just a monitoring tool. Projects like Google’s Jigsaw initiative deploy AI to detect users searching for content related to violent extremism and redirect their searches to counter-narratives. This approach, known as the Redirect Method, uses behavioural targeting to offer alternative perspectives, including testimonials from former extremists condemning violence (Chang, 2016). Such interventions have proven effective in reducing the spread of white supremacist or Jihadist content on platforms like YouTube and Telegram. In the broader context, these AI-driven initiatives could work in concert with human diplomatic interventions as a form of digital counterinsurgency, removing the ideological underpinnings of the activity of non-state actors (Cohen, 2023).

AI’s peacebuilding potential extends to the facilitation of digital diplomacy in conflict settings. Virtual platforms, supported by AI, can offer safe and neutral spaces for interaction between opposing groups, even in contexts where face-to-face negotiation is impossible. Soliya’s virtual exchange programs, for instance, use AI-enhanced moderation to foster intercultural discourse between youth from Western and predominantly Muslim societies (*Soliya’s Dialogue Methodology — SOLIYA — Reliably Transformational Virtual Exchange*, n.d.). Participants have

reported reductions in prejudice, improved empathy and greater awareness of global interconnectedness. These virtual spaces allow for anonymity and emotional safety while still enabling honest expressions and mitigating the social risk of disagreement (*Virtual Exchange From the Participant's Perspective — Soliya — Reliably Transformational Virtual Exchange*, n.d.).

A similar dynamic played out during Sudan's civil unrest, where encrypted online platforms facilitated indirect ceasefire talks. Despite the physical threats involved with in-person negotiation, conflicted groups managed to communicate with each other through secure digital means, proving how digital mediation in connection with the purposeful actions of real people can yield tangible results in real-world peace processes (Wählisch et. al., 2024).

The promise of AI use in conflict resolution tactics, however, is accompanied by ethical and practical challenges, creating dilemmas that require establishment of global AI governance frameworks that rely on data ethics, international law and the new field of tech diplomacy.

Ethical and Practical Challenges in Technology-Driven Peacebuilding

The central dilemma of integrating artificial intelligence with peacebuilding technology is that while digital tools provide new opportunities for early conflict detection, they are often dual-use purposes and can thus be adapted for use in surveillance and control. For example, many of the tools used in humanitarian settings today, such as predictive analytics and data visualisation software, originated in military applications; the collaboration between Palantir, a data organisation with senior intelligence connections and the UN World Food Programme (WFP) is a clear instance of technology designed for battlefield use being leveraged to support civilian asset delivery.

This merging of contexts and uses complicates trust relations, particularly in politically sensitive settings where there are concerns regarding data collection by government actors. Furthermore, peacebuilding initiatives rely on aggregation of data that raise significant concerns about privacy and exclusion; while AI can filter large volumes of sensitive data to identify sources of risks, any violation or misappropriation of that data could expose individuals or communities to targeted violence or state surveillance. Moreover, given the fact that many tools within digital peacebuilding tech rely on third-party platforms with proprietary algorithms, reducing transparency and accountability. Furthermore, access to these digital platforms is skewed, with the exclusion of marginalized groups, such as women or displaced ethnic minorities who often lack the necessary infrastructure, literacy or trust to participate fully in these initiatives. In some

cases, digital platforms become the cause of retraumatizing participants by exposing them to violent content without adequate safeguards (Hofmann, 2025).

Despite efforts to minimise harm, ethical guidelines like the “Do Not Harm” principle remain insufficient in addressing systemic challenges like infrastructural dependency, algorithmic bias and project unsustainability. Biases embedded in AI training data often reflect broader social inequalities for example facial recognition systems misidentifying ethnic minorities at disproportionately higher rates or AI language models replicating gendered stereotypes thereby amplifying exclusion rather than reducing it (Buolamwini & Gebru, 2018b). Meanwhile, given that the global concentration of digital infrastructure is in the hands of a few powerful tech corporations, it makes these systems susceptible to being manipulated by geopolitical interests. These projects very commonly are initiated without necessary long-term support or community ownership resulting in digital solutions that disappear when the funding stops leaving behind fragmented systems that cannot be locally sustained (Karahana, 2024).

Finally, with the ability to manipulate algorithmic systems and the information networks they support, this too has become a powerful political tactic, with non-state and government actors using these digital platforms to undermine dissent or advance certain political narratives. Many algorithms preferring user engagement to accuracy, can likewise be exploited to spread false information, fuel unrest or dominate political discourse. With COVID-19, for example, conspiracy theories linking health risks to 5G technology gained traction online, all the while being propelled by algorithmic systems that were designed to promote user relevance. The viral feedback loops not only confuses public understanding but actively works against essential informational foundations of peacebuilding efforts (Hofmann, 2025).

For AI to serve as a tool for peace, its integration into global conflict resolution efforts must be at par with robust safeguards and inclusive policymaking that ensures trustworthy AI.

Balancing Optimism and Realism: The Path Forward

AI as a Tool, Not a Replacement for Human Mediation

As artificial intelligence (AI) continues to grow in scope and impact, it is inclined to transform various industries including diplomacy. While AI offers tools designed to process data, translate, and recognize patterns it cannot offer empathy, the implicit judgement and ethical agency a human diplomat brings to the assignment. Thus, it may serve best as a component of human-led negotiations, not as a replacement.

Nevertheless, AI- enhanced systems might aid diplomatic officials by being able to process vast amounts of data, including historical conflict records, positions as outlined in policy documents and real-time social media conversations. Thus, identifying, cataloging and summarising positions from negotiations; identify patterns of negotiations; and even highlight shared, possibly undisputed, areas in negotiations could help to speed up information processing and help facilitate strategic choice-making in the conduct of diplomacy. In addition, advanced AI-enabled translation tools could enhance engagement in intercultural communication, including in multilateral negotiations with under-represented dialects or informal patterns of speech. This could be especially valuable for diplomats in conflict zones or regions with rich cultural diversity, as traditional diplomatic channels may be limited (Wählisch et. al., 2024).

Yet, while there are potential benefits to this application of AI, there are also limits to its ability to act independently as a diplomatic actor. For example, regulatory frameworks in accountability, algorithmic transparency and data privacy are still poorly developed. In addition, AI systems may inadvertently or tacitly replicate cultural or political biases present in their dataset, and contribute to bias in confusion in high-stakes negotiations.

The future of diplomacy lies not in AI autonomy or replacement but in cooperation where machine intelligence enhances human decision-making without eroding human-centric values. Furthermore by rooting AI use in ethical oversight and cultural awareness international and state actors can ensure that technological tools serve diplomacy's goal of inclusive and peaceful global cooperation.

Reforming Social Media Algorithms for Bipartisan Dialogue

Despite social media algorithms being optimised for profit through the attention economy, these algorithms can be repurposed to promote bipartisan dialogue and reduce political polarisation. One alternative is bridging-based ranking, a content-sorting mechanism that prioritises posts positively received by ideologically diverse users. This approach emphasises mutual understanding instead of traditional metrics such as likes or shares (King's College London, 2023). Platforms such as Polis, used by governments in Taiwan and Finland to collect citizen feedback on divisive policy issues, exemplify how algorithmic design can foster consensus-building. Meta and X have also experimented with similar systems, like Community Notes which crowdsource fact-checks and prioritizes comments that support the political spectrum (Thorburn & Ovadya, 2023).

Whereas technology companies like Meta have traditionally optimized short-term engagement using incendiary content, evidence indicates that other algorithms can cause a short-term decline

in user activity followed by a longer-term recovery. This change not only improves civic discussion but also decreases dependence on costly content moderation, as hate speech and toxicity decrease. Further promoting cross-cutting exposure, subtle algorithmic “nudges” such as adding feeds with counter-attitudinal information have been shown to decrease confirmation bias without making users defensive (Company & Meta, 2021).

As AI is increasingly deployed in content curation, the ethical issues surrounding data privacy, manipulation and transparency have become more critical. Social media platforms covertly collect vast quantities of personal data from users, which is then fed into AI algorithms to improve the predictive accuracy of social connections, behaviour, and beliefs. Unprotected data can be used as a weapon for political monitoring or microtargeting, particularly under authoritarian governments. Platforms must prioritise algorithmic openness in order to rebuild confidence. Instagram’s “Why Am I Seeing This?” feature, for example, lets users understand how recommendations are made (IMM Institute, 2023).

Furthermore, in order to avoid perpetuating discriminatory prejudices, AI systems must be trained on a variety of representative datasets. Models for international regulation include the European Union’s AI Act, which forbids misleading algorithmic practices (Karahana, 2024).

I contend that AI on social media must adhere to international norms for democratic governance and digital rights not only for accountability and transparency but to counter the threats of algorithmic bias, surveillance, and trust that I have pointed out throughout this commentary. To prevent misuse for censorship or ideological arguments, it requires openness, allowing independent audits, and ways for users to maintain autonomy.

Here, global cooperation matters. A transnational standard for responsible AI applied to social media can help to ensure that social media is used to facilitate inclusive dialogue and to better advance democratic resilience, even as the deeper, structural challenges related to the platform incentives enforcement gaps and trust-building in relevant cultural contexts loom doubtfully above us.

Conclusion:

As this commentary has shown, technological systems can either deepen global divisions amongst people, especially through algorithmic content curation on social media, or they can also be re-engineered for reconciliation if guided by international cooperation and ethical foresight. I contend that AI must not become a substitute for human empathy, political will or cross-cultural understanding; instead, it can be wielded as a complementary tool to amplify the

efforts of civil society actors in the international community. The analysis reveals that when AI is embedded into early warning systems and virtual dialogue platforms can transform digital spaces from conflict-magnifying chambers into arenas of preventive diplomacy. Furthermore, the commentary emphasises the need for digital sovereignty and international governance around digital spaces and the role of private actors in the global arena.

References:

Albert, E. (2020, January 23). The Rohingya crisis. *Council on Foreign Relations*.
<https://www.cfr.org/backgrounder/rohingya-crisis>

BBC News. (2021, December 7). *Rohingya sue Facebook for \$150bn over Myanmar hate speech*. <https://www.bbc.com/news/world-asia-59558090>

Buolamwini, J., & Gebru, T. (2018b, January 21). Gender Shades: Intersectional accuracy Disparities in commercial gender classification. PMLR.
<https://proceedings.mlr.press/v81/buolamwini18a.html>

Chang, L. (2016, September 8). Google will begin showing anti-Islamic State ads to counter terrorism in North America. *Digital Trends*.
<https://www.digitaltrends.com/web/google-anti-terrorism-ad/>

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9). <https://doi.org/10.1073/pnas.2023301118>

Cohen, J. (2023, July 20). Digital Counterinsurgency: How to marginalize the Islamic State online. *Foreign Affairs*.
<https://www.foreignaffairs.com/articles/middle-east/digital-counterinsurgency>

Company, F., & Meta. (2021, December 8). Our new AI system to help tackle harmful content. *Meta*. <https://about.fb.com/news/2021/12/metanew-ai-system-tackles-harmful-content/>

De Guzman, C. (2022, September 29). Meta's Facebook algorithms 'Proactively' promoted violence against the Rohingya, new Amnesty International report asserts. *TIME*.
<https://time.com/6217730/myanmar-meta-rohingya-facebook/>

First Findings from US 2020 Facebook & Instagram Election Study Released. (2023, July 27). Annenberg.
<https://www.asc.upenn.edu/news-events/news/first-findings-us-2020-facebook-instagram-election-study-released>

Heidebrecht, P. (2022, December 11). *PeaceTech – IEEE Technology and Society*. IEEE Technology and Society. <https://technologyandsociety.org/peacetech/>

Hofmann, F. (2025). Towards a holistic approach to PeaceTech ethics. In *Policy Perspectives: Vol. Vol. 13/2*.
https://ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP13-2_2025-EN.pdf

IMM Institute. (2023, September 11). *Social media algorithms and regulation: Striking a balance between connection and concerns*.
<https://www.linkedin.com/pulse/social-media-algorithms-regulation>

Karahan, O. (2024a, March 1). *The Ethics of AI in Social Media explored (2025)*. 618 Media: #1 Digital Marketing Agency.
<https://618media.com/en/blog/the-ethics-of-ai-in-social-media-explored/#understanding-ais-role-in-social-media>

Karahan, O. (2024b, March 1). *The Ethics of AI in Social Media explored (2025)*. 618 Media: #1 Digital Marketing Agency.
<https://618media.com/en/blog/the-ethics-of-ai-in-social-media-explored/#enhancing-ethical-ai-through-regulation-and-innovation>

King's College London. (2023, June 26). New approaches to social media algorithms could counteract destructive polarisation. *King's College London*.
<https://www.kcl.ac.uk/news/new-approach-to-social-media-algorithms-could-counteract-destructive-polarisation>

Knight Foundation. (2024, November 1). *Seven ways misinformation spread during the 2016 election*.

<https://knightfoundation.org/articles/seven-ways-misinformation-spread-during-the-2016-election/>

Marino, G., & Iannelli, L. (2023). Seven years of studying the associations between political polarization and problematic information: a literature review. *Frontiers in Sociology*, 8. <https://doi.org/10.3389/fsoc.2023.1174161>

Peckham, S. (2025, March 24). *What are algorithms? How to prevent echo chambers and keep children safe online*. Internet Matters. <https://www.internetmatters.org/hub/news-blogs/what-are-algorithms-how-to-prevent-echo-chambers/>

Soliya's Dialogue Methodology — SOLIYA — Reliably Transformational Virtual Exchange. (n.d.). Soliya – Reliably Transformational. <https://soliya.net/soliya-dialogue-methodology>

Thorburn, L., & Ovadya, A. (2023, October 30). How to redesign social media algorithms to bridge divides. *Tech Xplore*. <https://techxplore.com/news/2023-10-redesign-social-media-algorithms-bridge.html>

Us, J. K. W. (2025, April 10). *A running list of Elon Musk's biggest controversies*. The Week. <https://theweek.com/elon-musk/1022182/elon-musks-most-controversial-moments>

Virtual Exchange from the Participant's Perspective — Soliya — Reliably Transformational Virtual Exchange. (n.d.). Soliya – Reliably Transformational. <https://soliya.net/virtual-exchange-from-the-participants-perspective>

Wählich, M. (2024). *AI and the Future of Mediation*. In *Still Time to Talk: Adaptation and Innovation in Peace Mediation* (pp. 112-114). Conciliation Resources. https://pure-oai.bham.ac.uk/ws/portalfiles/portal/255472364/accord_30_still_time_to_talk_-_adaptation_and_innovation_in_peace_mediation_0.pdf

What is Peacetechnology? A New Environment Brings New Opportunity | Amazon Web Services. (2022, June 9). Amazon Web Services. <https://aws.amazon.com/blogs/publicsector/what-is-a-peacetechnology-a-new-environment-brings-new-opportunity/>