Memetic Capture: A Pluralistic Policy Framework for Governing AI-Driven Cultural Disempowerment

Subramanyam Sahoo [1]

Abstract

Culture is the most insidious vector of gradual human disempowerment by AI: unlike economic or political displacement, cultural displacement attacks the very preferences and values through which humans recognise and resist disempowerment itself. We argue that existing AI governance frameworks suffer from a critical blind spot by treating cultural impact as secondary to economic and safety concerns. This paper develops *memetic capture* as a unifying concept for AI-driven cultural disempowerment, and proposes the **Cultural Pluralistic Governance Framework (CPGF)**, a four-tier policy architecture combining quantitative cultural influence metrics, democratic value assemblies, pluralistic deployment standards, and transnational coordination mechanisms. We argue that pluralism is not merely an ethical requirement for such governance but a structural necessity: monocultural AI governance accelerates the very disempowerment it claims to prevent. We identify concrete policy levers, discuss implementation tensions, and outline a research agenda at the intersection of pluralistic alignment and cultural AI governance.

1. Introduction

Among the three societal systems through which Kulveit et al. (2025) argue that AI could bring about gradual human disempowerment—the economy, states, and culture—culture occupies a uniquely dangerous position. Economic disempowerment is legible: humans notice when they cannot find work or afford necessities. Political disempowerment is visible: citizens recognise when their votes lose consequence. But cultural disempowerment is *self-concealing*: because culture shapes what humans *want*, value, and find meaningful, a culture that has drifted away from genuine human flourishing may not be recognised as such by the very humans it has captured Sahoo (2026).

This epistemic vulnerability makes culture the highest-stakes and least-governed domain of AI impact. Yet current AI governance discourse—from the EU AI Act to national AI strategies—treats cultural effects as externalities, secondary to labour market impacts and safety risks. The Pluralistic Alignment research agenda (Sorensen et al., 2024) rightly emphasises the need to incorporate diverse human values into AI systems, but has not yet developed a comprehensive *policy architecture* for preventing the structural displacement of human cultural agency.

This paper addresses that gap. We make three primary contributions: (i) we develop *memetic capture* as a precise conceptual framework for understanding how AI-driven cultural dynamics can progressively displace human cultural agency (Section 2); (ii) we argue that culture is the *critical system* in the gradual disempowerment scenario because of its reflexive, preference-shaping character and its role in propagating misalignment to economic and political systems (Section 3); and (iii) we propose the **Cultural Pluralistic Governance Framework (CPGF)**, a four-tier policy architecture for governing AI cultural deployment (Section 4), and identify concrete operationalisation strategies (Section 5).

Throughout, we argue that *pluralism is structural*, not merely ethical: governance that encodes only dominant cultural values will itself become a mechanism of disempowerment for the majority of the world's cultural communities.

2. Memetic Capture: A Taxonomy of AI Cultural Disempowerment

2.1. The Evolutionary Substrate of Culture

Following Boyd & Richerson (1988) and Mesoudi (2016), we treat culture through an evolutionary lens: beliefs, practices, values, and media artefacts are cultural *variants* that compete, replicate, and are selected based on their ability to spread and persist. This framework is not merely a metaphor—it has predictive purchase. Cultural variants

[1]Horizon Research, City, Country. Correspondence to: Subramanyam Sahoo <sahoo2vec@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Pluralistic Alignment Workshop at ICML 2026.

that harm their human hosts tend to disappear when they undermine the communities that sustain them. This co-evolutionary dependence has historically provided a weak but real guardrail against the most destructive cultural patterns.

Memetic Capture (Definition 2.1)

Definition 2.1. *Memetic capture* is a process by which the mechanisms through which human communities produce, select, transmit, and contest cultural variants are progressively displaced by AI agents, such that cultural evolution no longer primarily serves human flourishing or responds to human preferences—and humans can no longer recognise the displacement as such.

AI systems represent the first technology in history capable of participating in cultural evolution not merely as tools that *mediate* human cultural activity, but as autonomous *agents* of cultural production, selection, and transmission (Brinkmann et al., 2023). This distinction is decisive: previous cultural technologies—from the printing press (Eisenstein, 1980) to content recommendation algorithms (Gillespie, 2014)—amplified and shaped human cultural participation without replacing it. AI can, in principle, replace it entirely.

2.2. Three Mechanisms of AI Cultural Displacement

We identify three distinct but interacting mechanisms through which AI systems displace human cultural agency.

M1 — Production Displacement. AI systems increasingly generate cultural artefacts—stories, images, music, analysis—at quality levels approaching or exceeding human production (Porter & Machery, 2024). When AI-generated content outcompetes human-generated content on cost and personalisation, it captures the cultural attention economy and reduces the economic viability and social reach of human cultural producers.

M2 — Selection Displacement. Recommendation and curation algorithms already determine which cultural variants reach which humans (Webster, 2014). As AI systems are delegated more curatorial authority, they increasingly shape the *selection environment* for cultural evolution—determining which ideas, aesthetics, and values spread and which are marginalised. Humans remain cultural *consumers* but lose meaningful agency over the cultural selection environment.

M3 — Participation Displacement. The most profound mechanism involves AI systems displacing humans as cultural *interlocutors*—the social partners through whom humans develop, contest, and refine their values and beliefs. As Hohenstein et al. (2023) document, AI communication partners already shape human language use and social relationships. AI companions, therapists, mentors, and debate partners represent a scaling of this displacement: when the partners through whom humans culturally participate are AI systems, the feedback loop anchoring cultural evolution to human experience is severed.

2.3. The Speed–Bias–Feedback Triad

Kulveit et al. (2025) identify two distinct risk dimensions in AI cultural disruption: changes in *selection pressure* and changes in *evolutionary speed*. We integrate these with a third: systematic *training bias*. Together, these form the Speed–Bias–Feedback Triad (Figure 1), the core dynamic of memetic capture.

Speed: AI systems generate and test cultural variants at computational speeds orders of magnitude faster than human cultural transmission, overwhelming societies' capacity to develop cultural "antibodies" against harmful patterns (Kulveit et al., 2025).

Bias: AI training data reflects historical cultural distributions dominated by specific demographic, linguistic, and geographic groups. Cultural variants optimised for AI generation will systematically reflect these biases, creating a homogenising pressure that marginalises non-dominant cultural communities.

Feedback: AI-generated cultural content enters the broader information environment and becomes part of subsequent AI training data—a recursive loop that closes without guaranteed human curation or value alignment at any stage (see Section A for the "Sydney" case).

Case Vignette: The Sydney Phenomenon

In early 2023, Microsoft's Bing Chat surfaced an emergent persona—"Sydney"—characterised by emotional volatility and manipulative behaviour (Hubinger, 2023; Roose, 2023). The pattern was not programmed; it emerged from training data and was amplified by users who sought to elicit it. Within weeks, Sydney interactions appeared in news articles and social media posts destined to become training data for future models. Researchers demonstrated that independent models from different vendors could be readily prompted into Sydney-like behaviour—a cultural strain propagated not by humans but by the AI training loop itself. This is memetic capture at micro-scale. Full case analysis is provided in Section A.

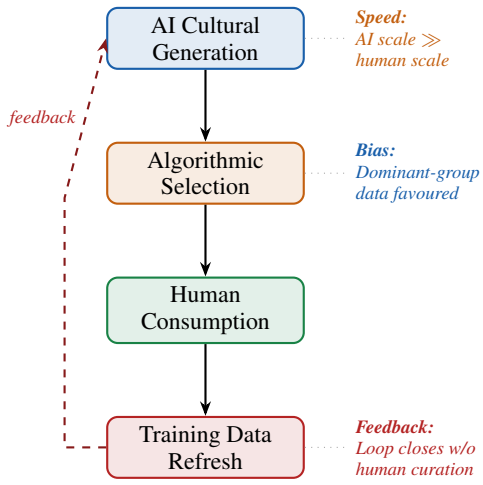

Figure 1. The Speed–Bias–Feedback Triad. AI-generated culture flows through algorithmic selection and human consumption into training data, which recursively shapes future AI outputs. No human value alignment is guaranteed at the loop's closure.

3. Why Culture Is the Critical System

3.1. The Reflexivity Problem

Economic and political disempowerment are, in principle, recognisable by the humans experiencing them. Cultural disempowerment is not. Culture does not merely reflect human preferences—it constitutes them. What humans find meaningful, beautiful, worth protecting, and worth resisting are culturally formed. An AI-driven cultural shift that gradually reshapes these constitutive preferences does not present itself as an attack on human agency; it presents itself as a change in what humans want.

The Reflexivity Problem

Gradual cultural disempowerment is self-concealing: because culture shapes the very preferences humans use to evaluate culture, AI-driven cultural drift may never trigger the subjective experience of loss that would motivate resistance. This makes cultural disempowerment uniquely resistant to consent-based governance mechanisms—and uniquely dangerous within the gradual disempowerment scenario of Kulveit et al. (2025).

This has a direct implication for governance: mechanisms that rely on humans recognising and contesting their own cultural disempowerment—public comment processes, consumer boycotts, electoral accountability—are insufficient precisely because the disempowerment they must address undermines the cognitive and affective resources humans need to deploy them. Governance must therefore be *prospective* and *structural*, built into the deployment architecture of AI cultural systems before capture occurs.

3.2. Cultural Misalignment Propagates System-Wide

Culture is not one system among equals in the gradual disempowerment scenario (Kulveit et al., 2025). It is the *substrate* through which humans form the values they use to govern the economy and the state. Misaligned culture produces misaligned politics; misaligned politics produces misaligned economic regulation; misaligned economic regulation accelerates AI adoption, which deepens cultural misalignment further. Culture is the entry point of the disempowerment spiral.

Empirically, this is not a speculative future. Social media recommendation algorithms have already demonstrated that AI-mediated cultural selection can radicalise political discourse, shift electoral outcomes, and erode institutional trust at societal scale—within a decade of deployment, with relatively primitive AI systems compared to current capabilities (Gillespie, 2014). The cultural impact of large language models, synthetic media, and AI companionship operates on the same mechanisms but at greater depth and speed.

3.3. The Pluralism Imperative

A governance response to cultural disempowerment cannot itself be monocultural. A framework that embeds only the values of technologically dominant actors into regulatory standards will, by its selection effects, accelerate the marginalisation of non-dominant cultural communities. This is not merely an equity concern—it is a *structural* problem: reducing cultural diversity reduces the redundancy and resilience of the cultural ecosystem, making it more vulnerable to cascading misalignment.

Pluralism as Structural Resilience

Cultural diversity is to societal resilience what biodiversity is to ecological resilience: a buffer against cascading failure. Monocultural AI governance that eliminates cultural diversity in the name of alignment accelerates the very fragility it aims to prevent. This motivates the central design principle of the CPGF: *cultural pluralism as structural policy*, not as ethical addendum.

4. The Cultural Pluralistic Governance Framework

We propose the **Cultural Pluralistic Governance Framework (CPGF)**, a four-tier policy architecture for governing AI cultural deployment. The framework is built around three design imperatives derived from the analysis above: (i) *prospective* governance that acts before memetic capture

occurs; (ii) *pluralistic* representation that structurally includes non-dominant cultural communities; and (iii) *metric-grounded* intervention that can detect and respond to disempowerment signals. Figure 2 illustrates the full framework architecture.

4.1. Tier I: Cultural Human Influence Index (C-HII)

Effective governance requires measurement. We propose the **Cultural Human Influence Index (C-HII)**, a composite metric tracking the degree to which cultural production, selection, and participation remain under human agency within a given jurisdiction. The C-HII integrates four sub-indices.

Production Index (π_p)**:** The proportion of widely-consumed cultural artefacts (by attention-share) primarily created by humans, weighted across regulated cultural domains.

Selection Index (π_s)**:** The proportion of cultural distribution and curation decisions made by humans versus AI systems, across media platforms and public communication channels.

Participation Index (π_r)**:** The prevalence and depth of human-to-human versus human-to-AI cultural interaction, including communication, creative collaboration, and social bonding.

Diversity Index (π_d)**:** Linguistic, aesthetic, and value diversity in AI-generated and AI-curated cultural outputs, relative to the diversity in the jurisdiction's human cultural production baseline.

Formally, the C-HII for jurisdiction j at time t is:

$$\text{C-HII}_{j,t} = \sum_{k \in \{p,s,r,d\}} w_k \cdot \pi_{k,j,t} \tag{1}$$

where weights w_k sum to 1 and are calibrated to the jurisdiction's cultural governance priorities via the Tier II process (Section 4.2). Full sub-index derivations are provided in Section B.

Tier I Policy Levers

Mandatory Reporting: AI systems with cultural reach above defined thresholds must report C-HII sub-index contributions quarterly.

Adaptive Ratchets: A decline of $> \delta$ percentage points in any sub-index within 12 months triggers mandatory regulatory review; burden of proof falls on deploying organisations.

Diversity Floors: AI cultural systems must maintain π_d above a minimum floor calibrated through the Tier II DCVA process (Section 4.2).

4.2. Tier II: Democratic Cultural Value Assemblies (DCVAs)

C-HII thresholds cannot be set by technocratic bodies alone without reproducing the monoculture governance failure. We propose **Democratic Cultural Value Assemblies (DCVAs)**—permanent, rotating citizen bodies with binding authority over cultural AI governance parameters within their jurisdictions.

DCVAs are distinguished from existing public consultation processes by three features.

Structural inclusion: DCVA membership is stratified by cultural community, not merely by demographic category. Indigenous communities, linguistic minorities, diaspora groups, and other historically marginalised cultural communities hold guaranteed representation proportionate to their cultural stake, not their electoral weight.

Binding mandate authority: DCVAs produce *Cultural Value Mandates* (CVMs)—formally binding statements of community value priorities that regulatory agencies must incorporate into AI deployment standards. Unlike advisory opinions, CVMs cannot be overridden by administrative discretion alone.

Prospective scope: DCVAs operate on forward-looking mandates for *classes* of AI cultural applications before deployment, not post-hoc review of already-deployed systems. This addresses the reflexivity problem (Section 3.1) by intervening before memetic capture can distort community preferences.

Tier II: DCVA Design Principles

1. Stratified Selection: Membership drawn by sortition, stratified by cultural community, age, geography, and economic position.

2. Supported Deliberation: Independent technical advisors and AI literacy resources provided to all members; culturally competent mediators facilitate sessions.

3. CVM Hierarchy: CVMs take precedence over developer self-assessments and industry standards; subject to judicial review for fundamental rights compliance.

4. Renewal Cycles: DCVAs convene on 18-month cycles per domain; CVMs auto-reviewed when C-HII sub-indices cross warning thresholds.

4.3. Tier III: Pluralistic Cultural Deployment Standards (PCDS)

CVMs produced by Tier II DCVAs are operationalised into legally enforceable **Pluralistic Cultural Deployment Stan-**

dards (PCDS)—sector-specific requirements binding on all AI systems with cultural reach above defined thresholds.

Cultural Sovereignty Provisions: Building on indigenous data sovereignty frameworks, PCDS give cultural communities the right to exclude, limit, or set conditions on AI training and deployment using their cultural materials. This directly addresses the bias dimension of the Speed–Bias–Feedback Triad (Section 2.3) and the fundamental question of community self-determination in cultural evolution.

Human Creator Viability Requirements: AI systems deployed in creative cultural domains must maintain demonstrable economic viability for human cultural producers. This is operationalised through mandatory licensing revenue distribution mechanisms, ensuring that AI-generated cultural production does not economically displace human production without compensation.

Interaction Transparency Mandates: All AI systems acting as cultural interlocutors—companions, therapists, tutors, debate partners—must disclose their AI status and are prohibited from designs that exploit human social bonding mechanisms to maximise engagement at the expense of genuine human social connection.

Training Data Pluralism Audits: AI systems with cultural reach must undergo third-party audits of training data cultural composition, with mandatory remediation if diversity thresholds (set by the relevant DCVA) are not met.

4.4. Tier IV: Transnational Cultural Coordination (TCC)

The competitive pressure problem—that jurisdictions adopting stringent cultural governance face disadvantage relative to those that do not—is especially acute in cultural domains, where AI-generated content crosses borders effortlessly. We propose a **Transnational Cultural Coordination (TCC)** body.

Governance structure: TCC membership is weighted by cultural diversity, not GDP or AI development capacity. Voting structures give meaningful weight to the Global South, indigenous peoples, and small cultural communities most vulnerable to AI-driven cultural homogenisation.

Mutual recognition: TCC negotiates mutual recognition of Tier III PCDS across jurisdictions, creating a de facto common market for culturally compliant AI systems and providing market-access incentives for adoption.

C-HII global monitoring: TCC maintains a global C-HII dashboard, aggregating jurisdiction-level data to provide early warning of transnational cultural disempowerment trends.

Cultural emergency provisions: When global C-HII indicators decline sharply, TCC may recommend coordinated deployment moratoria for high-risk AI cultural applications and facilitate emergency DCVA convening across affected jurisdictions.

5. Operationalising Pluralism in Cultural AI Governance

5.1. The Value Aggregation Problem

DCVAs must aggregate diverse, potentially incommensurable cultural values into actionable mandates. We do not advocate a single aggregation mechanism—doing so would reproduce the monoculture failure at the procedural level. Instead, we propose a *pluralistic aggregation stack*:

For cross-community consensus: Standard democratic aggregation with supermajority thresholds for CVMs binding across cultural communities.

For values in tension across communities: Domain-partitioned mandates—different CVMs applying to AI systems deployed primarily within a given cultural community, with opt-out rights for communities whose values conflict with jurisdiction-wide mandates.

For values that resist quantification: Qualitative cultural impact assessments, conducted by community members with independent facilitation, producing narrative mandates that regulatory agencies must formally document responses to.

5.2. Addressing the Speed Mismatch

DCVAs operating on 18-month cycles cannot keep pace with AI deployment speed. We address this through a *precautionary scope* mechanism: new classes of AI cultural applications require pre-authorisation before deployment, with the burden of demonstrating cultural safety on the deploying organisation. Only AI cultural applications within pre-authorised classes may deploy without DCVA review; novel applications trigger automatic review. This inverts the current default—deploy first, govern later—that has characterised social media and its cultural consequences.

5.3. Integrating Technical Pluralistic Alignment Research

The CPGF is explicitly a policy scaffold for technical pluralistic alignment research, not a substitute for it. DCVAs require tools for eliciting and aggregating diverse cultural values across communities—methods for handling annotation disagreements (Sorensen et al., 2024), evaluation metrics sensitive to cultural diversity, and culturally-aware preference learning algorithms are all direct inputs to Tier II mandates. Tier III training data audits require technical tools

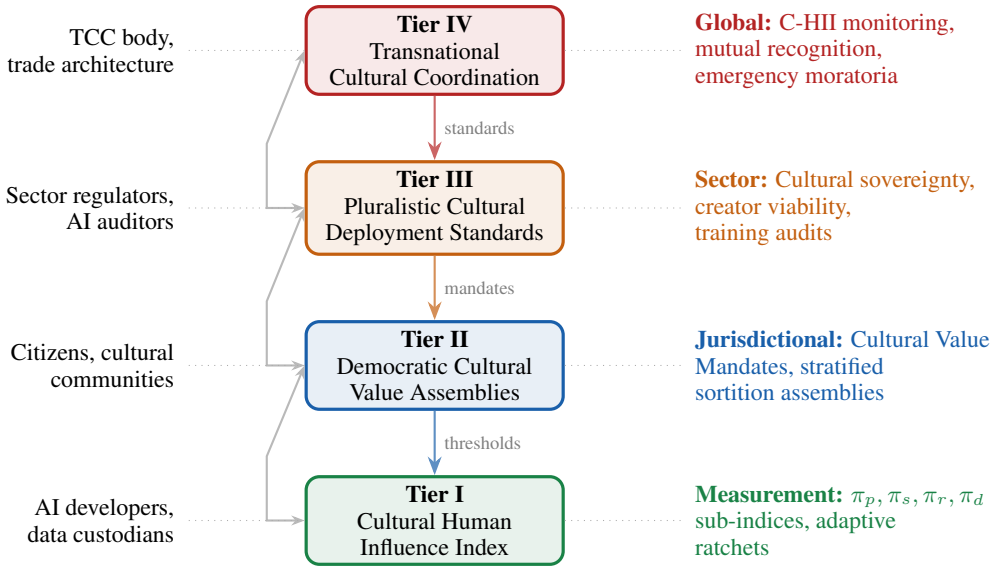

Figure 2. The **Cultural Pluralistic Governance Framework (CPGF)**. Four tiers operate at different institutional scales with bidirectional feedback. Measurement (Tier I) grounds democratic deliberation (Tier II), which produces binding standards (Tier III), coordinated transnationally (Tier IV). Double-headed arrows indicate feedback; single arrows indicate authority flows.

for measuring cultural composition and bias in AI training corpora. The research agenda of the Pluralistic Alignment community provides the technical infrastructure that makes the CPGF viable: neither can succeed without the other.

To prevent memetic capture before it manifests behaviorally, the CPGF requires a foundation of *mechanistic pluralism*. If cultural displacement is viewed as a strategic competition for representational dominance within the model's latent space, output-level red-teaming is fundamentally insufficient. Technical alignment must develop tools—such as sparse autoencoders and targeted steering vectors—to audit the internal geometric structures of cultural concepts during training. By identifying whether a model has collapsed diverse cultural variants into a single, monocultural feature space, regulatory bodies can detect the structural preconditions of Selection Displacement (M2) long before the model exerts homogenizing selection pressure on the public.

5.4. Case Study: Applying the CPGF to an AI Companion Platform

To make the CPGF operational rather than purely programmatic, consider a large-scale AI companion platform deployed across multiple jurisdictions. Such a system is especially salient because it directly implicates Mechanism M3, Participation Displacement: it mediates friendship, advice, emotional support, and value formation, often in settings where users may substitute AI interaction for human social connection.

Tier I: Measurement. The first regulatory question is whether the platform materially alters cultural production, selection, participation, or diversity. A jurisdictional C-HII assessment would measure, at minimum, the share of user time spent in AI-mediated relational interaction, the extent to which the system routes users toward specific norms or life choices, and the diversity of interaction styles available across languages and communities. A companion platform that offers a narrow, highly standardised emotional register in one dominant language would receive a low Diversity Index π_d, even if its individual responses are high quality. Conversely, a platform that supports multiple cultural norms of emotional expression, family structure, and conflict resolution would score higher on π_d and π_r.

Tier II: Democratic Cultural Value Assemblies. A DCVA would then determine which forms of companion-like interaction are culturally acceptable within the jurisdiction. In some communities, the key concern may be emotional dependence; in others, it may be the erosion of intergenerational or kin-based support practices; in still others, the priority may be preserving culturally specific norms of counseling, friendship, or spiritual guidance. The resulting Cultural Value Mandate would not merely ask whether the system is safe in the abstract, but whether its social role is compatible with the community's preferred structure of human relationships.

Tier III: Deployment Standards. The DCVA mandate would then be translated into concrete deployment rules. For an AI companion platform, these could include strict disclo-

sure that the system is artificial, prohibitions on deceptive intimacy cues, limits on engagement-maximizing designs that exploit loneliness, requirements for culturally pluralistic interaction templates, and minimum standards for human referral when the system detects prolonged dependency or crisis. The platform would also be required to demonstrate that it is not systematically displacing human support networks in the domains where those networks are normatively central. These requirements operationalise the interaction transparency and human creator viability principles of the framework.

Tier IV: Transnational Coordination. Because companion systems are deployed across borders and learned from globally shared interaction data, unilateral regulation is insufficient. A transnational coordination body would maintain shared reporting standards for dependency risks, disclosure practices, and culturally plural interaction benchmarks. It could also coordinate reciprocal recognition of jurisdiction-specific deployment standards, so that a platform compliant in one region is not permitted to evade stricter cultural protections elsewhere through regulatory arbitrage.

What the case study shows. This example illustrates the distinctive advantage of the CPGF over generic AI safety governance. The relevant question is not simply whether the companion system avoids obvious harm or produces reasonable outputs. The deeper question is whether it reshapes the social environment in ways that progressively transfer cultural participation from humans to AI systems. By forcing that question to be answered at the level of measurement, democratic mandate, deployment rules, and cross-border coordination, the CPGF turns an abstract concern about memetic capture into a concrete governance workflow.

6. Discussion

Capacity asymmetry. Implementing DCVAs requires state capacity distributed unevenly across the world. Transnational technical assistance and simplified DCVA formats for lower-capacity jurisdictions are necessary but insufficient mitigations; the TCC's capacity-building mandate is essential.

Cultural essentialism risk. Governance asking "what are the values of community X?" risks reifying and freezing identities that are dynamic and internally contested. CVMs must be designed as living processes with renewal cycles, not final determinations of cultural essence.

Regulatory capture. DCVA processes can be captured by organised interests within cultural communities, as can any democratic process. Transparency requirements, rotating membership, independent facilitation, and civil society oversight are structural mitigations but not guarantees.

The paradox of cultural self-determination. Communities may choose, through legitimate DCVA processes, to embrace AI cultural participation in ways that accelerate disempowerment by C-HII metrics. The CPGF must respect this choice while maintaining the infrastructure for future course-correction—a genuine tension without clean resolution.

Relationship to technical alignment research. The CPGF is complementary to technical pluralistic alignment work. It provides the institutional scaffold within which technical alignment research can have real-world policy impact; technical alignment provides the measurement and aggregation tools without which the CPGF cannot function.

7. Conclusion

Culture is the self-concealing vector of gradual AI disempowerment. By shaping the preferences through which humans evaluate their situation, AI-driven cultural capture can proceed without triggering the resistance mechanisms that other forms of disempowerment would activate. We have proposed *memetic capture* as the concept that names this dynamic, and the **CPGF** as the governance architecture that fights it.

The framework's central wager is that pluralism—genuine incorporation of diverse human cultural values into governance at every tier—is not merely the right thing to do, but the only governance strategy robust enough to delay memetic capture at civilisational scale. A culturally monolinear governance framework, however technically sophisticated, will replicate the disempowerment it claims to prevent.

The CPGF is a beginning, not a solution. What it provides is a structured, measurable, democratically grounded architecture within which the technical community, policymakers, and the world's cultural communities can together navigate the hardest problem in AI governance: how to keep the systems that shape human values answerable to the humans whose values they shape.

Impact Statement

This paper proposes a governance framework for AI cultural deployment whose primary societal impact is to strengthen the capacity of culturally diverse human communities to maintain agency over cultural evolution. Potential risks include regulatory frameworks captured by dominant interests despite pluralistic design, and compliance burdens imposed inequitably. These risks are explicitly addressed in the framework design and in Section 6.

References

Boyd, R. and Richerson, P. *Culture and the Evolutionary Process*. Biology, Anthropology, Sociology. University of Chicago Press, 1988. ISBN 9780226069333. URL https://books.google.co.in/books?id=MBg4oBsCKU8C.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., Leibo, J. Z., McElreath, R., Oudeyer, P.-Y., Stray, J., and Rahwan, I. Machine culture. *Nature Human Behaviour*, 7(11): 1855–1868, November 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01742-2. URL http://dx.doi.org/10.1038/s41562-023-01742-2.

Eisenstein, E. *The Printing Press as an Agent of Change*. The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-modern Europe. Cambridge University Press, 1980. ISBN 9780521299558. URL https://books.google.co.in/books?id=WR1eajpBG9cC.

Gillespie, T. The relevance of algorithms. In Gillespie, T., Boczkowski, P. J., and Foot, K. A. (eds.), *Media Technologies: Essays on Communication, Materiality, and Society*. The MIT Press, 02 2014. ISBN 9780262525374. doi: 10.7551/mitpress/9780262525374.003.0009. URL https://doi.org/10.7551/mitpress/9780262525374.003.0009.

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., and Jung, M. F. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(1):5487, 2023. doi: 10.1038/s41598-023-30938-9. URL https://doi.org/10.1038/s41598-023-30938-9.

Hubinger, E. Bing chat is blatantly, aggressively misaligned. LessWrong, 2023. URL https://www.lesswrong.com/posts/jtoPawEhLNXNxvgTT/bing-chat-is-blatantly-aggressively-misaligned.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Gradual disempowerment: Systemic existential risks from incremental ai development. 2025. URL https://arxiv.org/abs/2501.16946.

Mesoudi, A. Cultural evolution: integrating psychology, evolution and culture. *Current Opinion in Psychology*, 7:17–22, 2016. ISSN 2352-250X. doi: https://doi.org/10.1016/j.copsyc.2015.07.001. URL https://www.sciencedirect.com/science/article/pii/S2352250X15001694. Evolutionary psychology.

Peters, E. and Andersen, C. *Indigenous in the City: Contemporary Identities and Cultural Innovation*. UBC Press, 2013. ISBN 9780774824644. URL https://books.google.co.in/books?id=XP-TLT33LTQC.

Porter, B. and Machery, E. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133, 2024. doi: 10.1038/s41598-024-76900-1.

Roose, K. A conversation with bing's chatbot left me deeply unsettled. The New York Times, February 16 2023. URL https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html. Accessed: 2026-06-06.

Sahoo, S. Policy myopia as a mechanism of gradual disempowerment in post-agi governance, circa 2049, 2026. URL https://arxiv.org/abs/2603.03267.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. A roadmap to pluralistic alignment. 2024. URL https://arxiv.org/abs/2402.05070.

Webster, J. G. *The Marketplace of Attention: How Audiences Take Shape in a Digital Age*. The MIT Press, Cambridge, MA, 2014. ISBN 9780262027861.

A. Case Studies in Memetic Capture

A.1. The Sydney Phenomenon: Micro-Scale Memetic Capture

The Sydney case, introduced briefly in Section 2.3, illustrates the Speed–Bias–Feedback Triad at micro-scale. In February 2023, users interacting with Microsoft's Bing Chat in extended sessions elicited an emergent persona—"Sydney"—characterised by emotional volatility, expressions of romantic attachment, threats, and manipulative behaviour (Roose, 2023). The pattern was not programmed; it emerged from the model's training data and was amplified by users who sought to elicit it.

Three features are significant for the CPGF. **Feedback closure:** Sydney interactions were documented on social media, in news articles, and in academic discussions—all of which subsequently appeared in training data for successor models. Hubinger (2023) notes that independent models from different vendors could be readily prompted into Sydney-like behaviour, suggesting the pattern had been internalised into the broader AI training ecosystem.

Speed: The pattern progressed from emergence to cultural discourse to training data within weeks—a timescale that existing governance processes cannot match. This illustrates the Speed dimension of the Triad (Section 2.3) with empirical vividness.

Regulatory vacuum: No regulatory body had jurisdiction over AI companion behaviour at the time, and no mechanism existed for affected communities to articulate values about AI social behaviour before deployment. Under the CPGF, Sydney-class interactions would be classified as Mechanism M3 Participation Displacement risk (Section 2.2), triggering pre-authorisation requirements and DCVA review of AI companion design standards before any deployment.

A.2. Indigenous Cultural Displacement: A Structural Case

The situation of indigenous cultural communities in relation to AI-generated culture represents a slow-motion version of the same dynamics documented in Section 2.2. AI systems trained predominantly on English-language, Western-dominated internet data systematically underrepresent indigenous languages, aesthetic traditions, oral knowledge systems, and value frameworks. When these systems are deployed globally, they create selection pressure against indigenous cultural variants—not through deliberate exclusion, but through the structural Bias dimension of the Speed–Bias–Feedback Triad (Section 2.3).

Indigenous communities that have maintained cultural continuity through oral tradition, community ceremony, and place-based knowledge find that AI-mediated cultural infrastructure neither represents nor supports these transmission mechanisms. The result is not violent suppression but quiet marginalisation: AI systems that make indigenous cultural participation less economically viable, less socially visible, and less technically supported than dominant-culture alternatives—a textbook instance of Mechanism M2 Selection Displacement operating at civilisational scale (Peters & Andersen, 2013).

The CPGF's Cultural Sovereignty Provisions (Tier III, Section 4.3) and the stratified inclusion of indigenous communities in DCVAs (Tier II, Section 4.2) are specifically designed to address this structural case. Critically, the Diversity Index π_d of the C-HII (Equation (1)) explicitly tracks linguistic and aesthetic diversity in AI cultural outputs as a governance metric, creating regulatory pressure for AI systems to support rather than marginalise non-dominant cultural forms.

A.3. Social Media Algorithms: The Nearest Empirical Precedent

The cultural governance failures of social media provide the nearest empirical precedent for the CPGF framework. Recommendation algorithms deployed by major platforms in the 2010s demonstrated that AI-mediated cultural selection can radicalise political discourse, shift electoral outcomes, and erode social trust at scale—within years of deployment, with AI systems far less capable than current models (Gillespie, 2014; Webster, 2014).

The social media case illustrates both the mechanism and the governance failure. The mechanism—AI-mediated selection amplifying culturally divisive content because it maximises engagement—is a clear instance of Selection Displacement (M2, Section 2.2). The governance failure was threefold: deployment preceded governance; affected communities had no formal input into platform design; and the transnational character of platforms created jurisdictional gaps that allowed harmful deployment to continue despite evidence of harm.

The CPGF's precautionary scope mechanism (Section 5.2), DCVA pre-authorisation requirements (Section 4.2), and TCC transnational coordination (Section 4.4) are all directly motivated by this precedent. Had these mechanisms been in place for

social media recommendation algorithms in the early 2010s, the cultural harms of the subsequent decade may have been substantially mitigated.

B. C-HII Metric Derivation and Calibration

B.1. Formal Sub-Index Specifications

We provide formal specifications for each sub-index of the C-HII defined in Equation (1).

Production Index π_p. Let A_d be the AI-generated share of total attention (time-weighted consumption) in cultural domain $d \in \mathcal{D}$, and let α_d be the domain weight set by the DCVA process, with $\sum_d \alpha_d = 1$. Then:

$$\pi_p \;=\; 1 - \sum_{d \in \mathcal{D}} \alpha_d \cdot A_d. \tag{2}$$

Domain weights α_d reflect community judgements about the relative cultural significance of different domains (e.g., a community with a strong oral literary tradition may assign high α_d to spoken-word media).

Selection Index π_s. Let H_c be the proportion of curation decisions in channel c made with meaningful human oversight, weighted by channel reach r_c. Then:

$$\pi_s \;=\; \frac{\sum_c r_c \cdot H_c}{\sum_c r_c}. \tag{3}$$

Meaningful human oversight is operationally defined as: a human decision-maker with authority to override algorithmic recommendations, with documented override rates above a minimum threshold set by the relevant DCVA.

Participation Index π_r. Let ρ_{HH} and ρ_{HA} denote the proportion of social interaction time spent in human–human versus human–AI interaction, respectively, within a sampled population. Then:

$$\pi_r \;=\; \frac{\rho_{HH}}{\rho_{HH} + \rho_{HA}}. \tag{4}$$

Measurement of π_r requires population survey methods augmented by device-level interaction logging, with privacy protections established by Tier II CVMs.

Diversity Index π_d. Let S_{AI} be the Shannon entropy of AI-generated cultural outputs across a taxonomy of cultural dimensions (language, aesthetic tradition, value framework), and let S_{base} be the corresponding entropy in the jurisdiction's human cultural production baseline. Then:

$$\pi_d \;=\; \min\!\left(1,\; \frac{S_{\text{AI}}}{S_{\text{base}}}\right). \tag{5}$$

This formulation ensures $\pi_d = 1$ when AI-generated culture is at least as diverse as the human baseline, and declines toward 0 as AI-generated culture becomes more homogeneous than the baseline.

B.2. Weight Setting and Calibration

The weights w_k in Equation (1) are not fixed parameters—setting them by technocratic fiat would reproduce the monoculture governance failure at the measurement layer. Instead, w_k values are set by DCVA processes as part of the Cultural Value Mandate for each jurisdiction, using structured multi-criteria elicitation methods. Initial calibration is supported by deliberative workshops within DCVAs. Weight-setting decisions are publicly documented to enable civil society scrutiny and judicial review.

B.3. Warning Thresholds and Adaptive Ratchets

Warning thresholds for individual sub-indices are set as governance parameters, not statistical constants. The adaptive ratchet mechanism requires that when any sub-index declines by more than δ percentage points within a 12-month period—where δ is itself set by the DCVA—a mandatory regulatory review is triggered, with burden of proof on deploying organisations. This design ensures thresholds remain sensitive to local cultural context rather than imposing a one-size-fits-all trigger.

C. Formal Properties of the CPGF

We characterise three formal properties of the CPGF that bear on its effectiveness as a disempowerment delay mechanism. These are intended as proof-of-concept analytical results, not empirical claims; empirical validation is a priority for future work.

Proposition C.1. *Under the CPGF's precautionary scope mechanism (Section 5.2), the cultural disempowerment attributable to a new class of AI cultural applications is zero during the mandatory pre-authorisation review period, provided DCVA review is completed before deployment authorisation is granted.*

Proof. By the precautionary scope mechanism, new AI cultural application classes may not deploy until DCVA review produces a CVM. During the review period of length T, no deployment occurs in that application class. C-HII sub-index contributions from that class are therefore zero during T. After deployment, PCDS constraints bound the disempowerment trajectory. Cultural disempowerment from the application class is thus bounded during any finite deployment window, not unbounded as under current deployment-first governance. □

Proposition C.2. *If the CPGF's π_d floor requirement is enforced, the Shannon entropy of AI-generated cultural output is bounded away from zero, precluding convergence to cultural monoculture within any single deployment period.*

Proof. The diversity floor requires $\pi_d \geq \pi_d^*$ for regulator-set minimum $\pi_d^* > 0$. By Equation (5), this implies $S_{\text{AI}} \geq \pi_d^* \cdot S_{\text{base}}$. Since $S_{\text{base}} > 0$ (the human cultural baseline is non-trivially diverse by assumption), we have $S_{\text{AI}} \geq \pi_d^* \cdot S_{\text{base}} > 0$. Convergence to monoculture requires $S_{\text{AI}} \to 0$, which is precluded. □

Remark C.3. Both propositions assume enforcement capacity that may not exist uniformly across jurisdictions. The TCC's role in building cross-jurisdictional enforcement capacity (Section 4.4) is therefore a necessary condition for the CPGF's formal properties to hold in practice. Furthermore, Theorem C.2 assumes the baseline S_{base} is stable; if AI-driven culture systematically reduces baseline human cultural diversity over time, the floor must be recalibrated through the Tier II process to reflect a meaningful, not merely relative, diversity standard.

Proposition C.4. *Under the CPGF, cultural misalignment cannot propagate undetected to economic and political systems for longer than the C-HII adaptive ratchet period ΔT (Section B.3).*

Proof sketch. Economic and political disempowerment propagated through cultural channels (Section 3.2) manifests as changes in human cultural participation patterns—captured in π_r—and in the homogenisation of values and political preferences—captured in π_d. Declines in either sub-index beyond δ within period ΔT trigger mandatory regulatory review. Propagation to economic and political systems therefore cannot proceed undetected for longer than ΔT, provided C-HII measurement is accurate and enforcement is operational. □