ProSocial AI: What It Is and Why It Matters

Cornelia Walther[1]
Sunway

Essay

Original material authorized for its first publication in the Journal de Ciencias Sociales, the Academic Journal of the Social Sciences School of Universidad de Palermo.

Receipt: 2026-06-27
Acceptance: 2026-10-20

Abstract: Artificial intelligence is being built fast and deployed faster, yet the dominant design logic of most AI systems remains extractive: optimised for engagement, data capture, and shareholder return rather than for human flourishing or ecological integrity. This essay introduces ProSocial AI as a substantive alternative — a framework for designing, deploying, and governing AI systems that are deliberately Tailored, Trained, Tested, and Targeted to bring out the best in and for people and planet. Drawing on the ProSocial AI Index, a 4T×4P matrix that evaluates AI systems across four criteria (Tailored, Trained, Tested, Targeted) and four intent dimensions (Purpose, People, Prosperity, Planet), the essay explains the theoretical rationale for reorienting AI development toward prosocial ends within planetary boundaries, and maps the conditions that make this reorientation urgent. The framework is grounded in a conception of Hybrid Intelligence — the dynamic coevolution of natural human intelligence and artificial intelligence — and governed by the principles of a Hybrid Return on Values (HROV), which counts political empowerment, human development outcomes and ecological integrity alongside financial returns. The essay argues that the Global South, including Latin America, holds particular strategic capacity to lead this transition. Countries that retain epistemic diversity, demographic vitality, and institutional latitude to choose differently are positioned to demonstrate that a Hybrid Future — pro-people, pro-planet, and pro-potential — is achievable. The essay advances the Fourth Path as an orienting philosophy: a deliberate, values-anchored alternative to both techno-utopianism and techno-pessimism, grounded in an expansion of collective human agency.

[1] PhD in Law, is an Associate Professor at Sunway University's Institute for Global Strategy and Competitiveness and Senior Fellow at the Sunway Centre for Planetary Health. She is affiliated with the Wharton Neuroscience Initiative and Harvard's Learning Innovations Laboratory. Her work, grounded in two decades of humanitarian practice with the United Nations, focuses on Hybrid Intelligence, ProSocial AI, human agency, technology governance, and planetary health.
Email: corneliaw@sunway.edu.my

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

Keywords: prosocial artificial intelligence; hybrid intelligence; hybrid return on values; global south.

IA ProSocial: Qué es y por qué importa

Resumen: La inteligencia artificial se construye rápido y se despliega aún más rápido, pero la lógica de diseño dominante de la mayoría de los sistemas de IA sigue siendo extractiva: optimizada para el compromiso del usuario, la captura de datos y el retorno para los accionistas, en lugar de orientarse hacia el florecimiento humano o la integridad ecológica. Este ensayo presenta la IA ProSocial como una alternativa sustantiva — un marco para diseñar, desplegar y gobernar sistemas de IA deliberadamente adaptados, entrenados, probados y focalizados para sacar lo mejor en y para las personas y el planeta. A partir del Índice de IA ProSocial (una matriz 4T×4P que evalúa sistemas de IA según cuatro criterios: Adaptado, Entrenado, Probado, Focalizado; y cuatro dimensiones poblacionales: Propósito, Personas, Prosperidad y Planeta), el ensayo explica el fundamento teórico para reorientar el desarrollo de la IA hacia fines prosociales y mapea las condiciones que hacen urgente esta reorientación. El marco se ancla en una concepción de la Inteligencia Híbrida — la coevolución dinámica de la inteligencia humana natural y la inteligencia artificial — y se gobierna mediante el principio de Retorno Híbrido sobre los Valores (HROV), que incorpora resultados de desarrollo humano e integridad ecológica junto con los retornos financieros. El ensayo sostiene que el Sur Global, incluida América Latina, posee una capacidad estratégica particular para liderar esta transición. Los países que conservan diversidad epistémica, vitalidad demográfica y latitud institucional para elegir de manera diferente están bien posicionados para demostrar que un Futuro Híbrido — pro-personas, pro-planeta y pro-potencial — es alcanzable. El ensayo avanza la Cuarta Vía como filosofía orientadora: una alternativa deliberada y anclada en valores, tanto al tecno-utopismo como al tecno-pesimismo, fundamentada en la expansión de la agencia humana colectiva.

Palabras clave: inteligencia artificial prosocial; inteligencia híbrida; retorno híbrido sobre los valores; sur global.

1. Introduction

The ratio is deceptively simple. On one side: an accelerating buildout of AI systems capable of shaping what people read, how they are hired, what credit they receive, which diseases get diagnosed, and whose voices are amplified in public discourse. On the other: governance frameworks, ethical standards, and distributional safeguards that have not kept pace with the social power of automated systems (Floridi et al., 2018; Jobin et al., 2019; O'Neil,

2016). The result is a values-laden reconfiguration of who gets to shape the future and on whose terms.

This essay argues that the dominant mode of AI development - what can be called extractive AI - is a design choice. And design choices can be revisited. ProSocial AI is defined here with precision: it refers to AI systems that are Tailored, Trained, Tested, and Targeted to bring out the best in and for people and planet. These four criteria are not aspirational adjectives; they constitute a functional standard against which any AI system can be evaluated, extending the psychological lineage of prosocial behaviour toward collective wellbeing into the governance of artificial systems (Eisenberg & Mussen, 1989; Taddeo & Floridi, 2018). A system that is Tailored adjusts to the actual needs and contexts of the communities it serves. A system that is Trained on epistemically diverse, ethically sourced, and culturally representative data builds in breadth from the start. A system that has been Tested across diverse populations and conditions earns its deployment authorisation through evidence, not assumption. A system that is Targeted serves specific human and planetary purposes without concentrating harm on particular groups.

The purpose of this essay is threefold. First, to explain what ProSocial AI is and the conceptual architecture that underpins it - including the ProSocial AI Index (4T×4P), the concept of Hybrid Intelligence and the Hybrid Tipping Zone, the role of Double Literacy in enabling ProSocial AI governance, the Hybrid Return on Values (HROV) as a measurement instrument, and the Fourth Path as an orienting philosophy. Second, to make the case for why this reorientation is urgent now. Third, to advance the argument that the Global South - with particular attention to Latin America - holds strategic capacity to lead the transition toward a Hybrid Future that is pro-people, pro-planet, and pro-potential. This framing draws on capability theory, relational ethics, critical technology studies, and humane technology governance (Birhane, 2021; Nussbaum, 2011; Sen, 1999; Walther, 2024; 2026).

The essay moves through three argumentative registers: the theoretical framework, the analysis of converging urgencies that make the current moment distinctive, and a sustained engagement with Latin America as both a context of concern and a protagonist of change. A note on voice: this essay is written from inside the work it describes. The author has spent the past decade developing the frameworks invoked here across policy, academic, and practitioner registers. That proximity is disclosed as a feature — it is what allows the argument to move between the conceptual and the operational without losing precision at either level.

2. Theoretical Argumentation

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

2.1 The Extractive Default and Its Costs

Extractive AI is not a conspiracy; it is an emergent property of an incentive structure. When the dominant revenue model for digital platforms is advertising, and advertising revenue is proportional to engagement, and engagement is maximised by systems that exploit emotional salience, cognitive shortcuts, and social comparison, the optimisation of AI systems toward those ends follows with a certain mechanical logic (Roberts, 2019; Zuboff, 2019). The outputs include algorithmic amplification of outrage, filter bubbles that narrow epistemic exposure, targeted micro-persuasion at scale, automated hiring systems that encode historical bias, and credit-scoring models that penalise entire postcodes (Eubanks, 2018; O'Neil, 2016).

The costs are unequally distributed. Populations already marginalised by economic, racial, or geographic disadvantage bear disproportionate exposure to extractive AI's harms: discriminatory algorithmic decisions, surveillance infrastructure deployed without consent, and the erasure of local knowledge systems in favour of globally standardised training corpora that reflect predominantly North Atlantic assumptions about what constitutes good data, relevant knowledge, or rational preference (Birhane, 2021; Crawford, 2021; Eubanks, 2018).

What makes this moment distinctive is that the scale and speed of deployment have outrun the social, institutional, and epistemic capacity to absorb, contest, or redirect the technology's effects. The global AI ethics landscape remains fragmented, principle-heavy, and unevenly enforceable; the gap between what AI systems can do and what governance frameworks can evaluate is therefore widening (Coeckelbergh, 2020; Floridi et al., 2018; Jobin et al., 2019). Into that gap flows extractive value: data captured without adequate consent, labour displaced without adequate transition, decisions automated without adequate recourse.

This is the baseline against which ProSocial AI must be understood: a structural alternative, built around a different conception of what intelligence is for and a more explicit account of whose capabilities it should expand (Nussbaum, 2011; Sen, 1999; Toyama, 2015).

2.2. ProSocial AI: Definition and Theoretical Foundation

The term "prosocial" carries a psychological lineage. It refers to voluntary behaviour oriented toward collective wellbeing rather than individual gain (Eisenberg & Mussen, 1989). Applied to artificial intelligence, it designates systems designed and deployed according to a rigorous standard: Tailored, Trained, Tested, and Targeted to bring out the best in and for people and planet. These four criteria map onto a four-dimension evaluation of Purpose, People, Prosperity, and Planet, yielding the 4T×4P ProSocial AI Index. The framework therefore moves beyond abstract principle statements by translating ethical intention into a

diagnostic structure for deployment decisions (Floridi et al., 2018; Jobin et al., 2019; Walther, 2024).

ProSocial AI is not equivalent to "safe AI" or "ethical AI," though it intersects with both. Safety concerns the absence of harm; ethics concerns the application of moral principles (Coeckelbergh, 2020; Floridi et al., 2018). ProSocial AI requires positive purpose, active orientation toward social and ecological benefit - a more demanding standard. A surveillance system that never produces false positives may be technically safe; it fails the ProSocial AI standard if it chills dissent, concentrates state power, or erodes the conditions under which free association is possible. In this sense, ProSocial AI aligns with the proposition that AI can be a force for good only when its governance includes purpose, accountability, and justice as design requirements rather than reputational claims (Taddeo & Floridi, 2018).

The conceptual foundation rests on three interrelated claims. First, intelligence - whether natural or artificial - is always embedded in social and ecological systems, and its effects on those systems are design variables, not side effects (Birhane, 2021; Crawford, 2021). Second, criteria for evaluating AI must be multi-dimensional and transdisciplinary, spanning individual, community, institutional, and planetary levels, consistent with capability-oriented approaches to human development (Nussbaum, 2011; Sen, 1999). Third, AI governance cannot be separated from the governance of the social conditions in which AI operates: power asymmetries, information inequalities, and access gaps are structural features of the deployment environment, not external noise (Eubanks, 2018; Toyama, 2015).

2.3. The ProSocial AI Index: A 4T×4P Diagnostic Framework

To move from philosophical orientation to operational instrument, the ProSocial AI Index offers a structured diagnostic. Its four criteria evaluate whether AI systems are aligned with capability expansion, relational accountability, and transparent evidence about data provenance and model performance (Birhane, 2021; Gebru et al., 2018; Sen, 1999). The criteria are:

Tailored - adjusted to the actual needs, values, cultural contexts, and adaptive capacities of the communities they serve, including linguistic specificity, accessibility, and contextual fit. This requirement is grounded in the insight that technology amplifies existing human and institutional capacity rather than substituting for it (Nussbaum, 2011; Sen, 1999; Toyama, 2015);

Trained - built on epistemically diverse, ethically sourced, culturally representative, and rigorously curated data that reflects the breadth of human experience rather than the sampling biases of dominant technology infrastructure. This criterion draws directly on work showing that documentation, representativeness, and accountability for datasets are

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

prerequisites for responsible model development (Birhane, 2021; Crawford, 2021; Gebru et al., 2018);

Tested - validated across diverse populations, geographies, and conditions before deployment, with evidence-based track records replacing claims of neutrality. This is essential because predictive systems often convert historical inequality into technical authority when their assumptions are not tested under real-world conditions (Mitchell, 2019; O'Neil, 2016); and targeted - directed at specific human and planetary purposes, with clear accountability for outcomes and no systematic concentration of harm on particular groups. The criterion makes explicit the positive-purpose requirement in AI governance: not merely avoiding harm, but demonstrating contribution to human and ecological value (Floridi et al., 2018; Taddeo & Floridi, 2018).

These four criteria are evaluated across four intent dimensions: Purpose (the declared and actual aims of the system), People (effects at individual and community levels), Prosperity (economic and social equity outcomes at institutional and national levels), and Planet (ecological integrity at civilisational and biospheric scales). Each cell in the resulting 4T×4P matrix becomes an evaluative node - sixteen structured accountability questions for any AI deployment. The multi-dimensional structure is consistent with development approaches that treat wellbeing, agency, and institutional opportunity as jointly constitutive rather than reducible to a single efficiency metric (Nussbaum, 2011; Sen, 1999).

A critical governance principle embedded in the Index is the Targeted Veto: any Red score in the Targeted dimension triggers mandatory deployment suspension. This expresses a distributional ethics that refuses to treat justice as a weighted average. A system that is well-trained and well-tested but systematically harms a specific demographic community fails the ProSocial AI standard entirely, regardless of aggregate performance. This position follows from relational ethics and from evidence that algorithmic systems can reproduce structural injustice while appearing efficient at the aggregate level (Birhane, 2021; Eubanks, 2018; O'Neil, 2016).

The Index is designed for multiple uses: developers self-assessing their systems, procurement bodies evaluating AI vendors, regulators calibrating oversight requirements, and civil society organisations conducting independent accountability assessments. In this respect, it responds to a recurring limitation in AI ethics: the distance between high-level principles and operational instruments with consequences (Jobin et al., 2019; Walther, 2024).

2.4. Natural Intelligence, Hybrid Intelligence, and the Hybrid Tipping Zone

ProSocial AI is intelligible only in relation to a theory of the intelligence it is meant to augment. Natural Intelligence - the emotional, aspirational, cognitive, and sensory capacity of

human beings - operates across four dimensions in a specific order: aspirations (what we reach toward), emotions (what we feel), thoughts (what we reason), and sensations (what we perceive through the body). This ordering is consequential. Aspirations orient the whole system; emotions encode value and relational meaning; thoughts process and plan; sensations ground the entire architecture in material reality. The account builds on the author's prior work on humane technology and hybrid intelligence, and aligns with capability theory's emphasis on the conditions that allow people to become and do what they have reason to value (Nussbaum, 2011; Sen, 1999; Walther, 2024, 2026).

When AI systems displace rather than augment these capacities, damage begins at the top of the hierarchy: aspirations become standardised, emotions become managed through algorithmic curation, thoughts become offloaded to automated suggestion, sensations become numbed by mediated experience. The cumulative effect is agency decay - the gradual erosion of the sense that one's choices, judgments, and aspirations are genuinely one's own. This concern resonates with scholarship on surveillance capitalism, content moderation, and the social costs of platform-mediated attention economies (Roberts, 2019; Walther, 2026; Zuboff, 2019).

Hybrid Intelligence describes the condition in which natural human intelligence and artificial intelligence operate in ongoing, dynamic relationship with a holistic understanding of both and their interplays. This opens space for the dynamic co-shaping of cognition, decision-making, emotional life, and collective sense-making. This co-curation is already underway; the question moving forward is how this dynamic is governed and toward what ends (Mitchell, 2019; Walther, 2024, 2026).

The Hybrid Cognition Border demarcates a threshold of concern: the point at which integration becomes so pervasive that qualitative changes in human agency, autonomy, and relational capacity begin to occur. Below it, AI tools augment human capability while leaving the essential structures of human self-determination intact. Beyond it, the relationship inverts - human behaviour begins to be structured primarily around AI outputs rather than AI tools being structured around human purposes. ProSocial AI governance is partly an exercise in Hybrid Cognition Border management: expanding beneficial AI uses while building the institutional, educational, and social resilience that preserves meaningful human agency (Sen, 1999; Toyama, 2015; Walther, 2026).

2.5. Hybrid Return on Values (HROV)

Any governance framework requires measurement instruments capable of capturing what it claims to value. The Hybrid Return on Values (HROV) addresses this need. Where conventional return-on-investment metrics count only financial outputs, HROV integrates four registers of values: political empowerment (agency preservation and expansion), human

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

development outcomes (quality of life, educational attainment, health status, social trust), ecological integrity indicators (carbon footprint, biodiversity effects, resource consumption, regenerative impact), and financial sustainability (revenue, cost structures, market access). Its conceptual basis is close to the human development tradition: value is assessed through expanded freedoms, capabilities, and institutional conditions for flourishing, not only through aggregate output (Nussbaum, 2011; Sen, 1999).

HROV is not philanthropy rebadged as metrics. It is a recognition that extractive AI generates costs - to mental health systems, to democratic functioning, to ecological capacity - that are currently externalised and therefore invisible to standard accounting (Crawford, 2021; Roberts, 2019; Zuboff, 2019). Making those costs visible and attributable to specific design choices is a prerequisite for any market actor to internalise them voluntarily, and for any regulator to require their internalisation structurally.

In practice, HROV functions as a portfolio assessment tool: enabling organisations, governments, and investors to compare AI deployments not only by their financial returns but by their contribution to or detraction from human and planetary flourishing. Applied alongside the ProSocial AI Index, HROV converts governance principles into comparable, auditable performance data - the currency of institutional accountability (Floridi et al., 2018; Taddeo & Floridi, 2018; Walther, 2024).

2.6. Double Literacy as a Governance Prerequisite

No AI governance framework can function without a population capable of engaging with it meaningfully. Double Literacy addresses this: a curriculum and capability framework comprising two strands - Human Literacy (understanding and cultivating natural human intelligence in all its dimensions: aspirational, emotional, cognitive, and somatic) and Algorithmic Literacy (understanding and critically engaging with AI and digital systems, including their data logics, commercial incentives, and governance structures). The pairing reflects the fact that AI is neither magic nor neutral machinery; it is a set of artefacts embedded in human purposes, institutional incentives, and data histories (Gebru et al., 2018; Mitchell, 2019; Toyama, 2015).

The pairing is deliberate. AI literacy programmes that teach people to use tools without teaching them to understand the human capacities those tools interact with produce technically competent users who remain vulnerable to the subtler forms of extractive influence: algorithmic aspiration adjustment, emotional manipulation through content curation, the slow erosion of confidence in unmediated human judgment. Double Literacy addresses both strands because neither is sufficient alone. Human Literacy protects agency from the inside out; Algorithmic Literacy equips people to interrogate systems from the outside in (Walther, 2024, 2026).

In Global South contexts, Double Literacy carries acute strategic weight. Educational systems that integrate it from primary level upward build populations equipped to evaluate AI deployments critically, to advocate for prosocial governance standards, and to generate locally appropriate AI applications rather than simply consuming externally designed ones. This is consistent with evidence that technology produces equitable outcomes only when human capability, institutional capacity, and local context are strengthened alongside technical access (Sen, 1999; Toyama, 2015).

2.7. The Fourth Path

The political and philosophical framework within which ProSocial AI sits is the Fourth Path. It is distinguished by contrast with three other orientations toward AI: techno-utopianism, which holds that AI will solve human problems if allowed to develop freely; techno-pessimism, which holds that AI represents an unmitigable threat; and techno-pragmatism, which accepts existing AI paradigms while advocating for incremental reform within them. The typology responds to the recurrent tension in technology ethics between principled optimism, risk containment, and institutional reform (Coeckelbergh, 2020; Floridi et al., 2018; Taddeo & Floridi, 2018).

The Fourth Path refuses all three. It holds that the direction of AI development is genuinely open - that neither optimism nor pessimism is warranted by the technical facts alone, and that incremental reform of an extractive paradigm is insufficient. What is required is a deliberate, values-anchored reorientation: a choice, made collectively and institutionally, to build AI systems and governance structures around a different conception of what intelligence is for. This insistence on social choice echoes the development-as-freedom tradition and critical accounts of technological solutionism (Sen, 1999; Toyama, 2015).

The Fourth Path is pro-agency. The measure of a good AI system is its effect on the aspiration-forming, emotion-experiencing, thought-generating, sensation-grounded human beings who interact with it - and on the planetary systems on which all human life depends. This framework gives ProSocial AI its normative grounding and its political distinctiveness (Crawford, 2021; Walther, 2024, 2026).

2.8. Why Now: The Convergence of Urgencies

Arguments for AI reorientation have circulated since the earliest days of commercial AI deployment. What makes the current moment qualitatively different is a convergence of urgencies that creates both pressure and opportunity.

The first urgency is ecological. The computational demands of large AI models carry a material footprint: energy consumption, water use for cooling, rare earth mineral extraction for hardware. Planetary systems are under maximum stress from carbon emissions,

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

biodiversity collapse, and freshwater depletion. A ProSocial AI agenda that ignores those ecological costs is not prosocial; it is selectively attentive. The Planet dimension of the 4T×4P Index must be built in, not bolted on (Crawford, 2021; ECLAC, 2024).

The second urgency is epistemic. The mass deployment of generative AI at the content layer of public discourse is reshaping what people believe, how they form opinions, and what knowledge they can access. The Verbal Value Void - the homogenisation of language and argument produced by AI-generated content trained on a narrow slice of global textual output - represents a structural threat to epistemic diversity. Marginalised languages, oral knowledge traditions, and non-Western analytical frameworks are systematically underrepresented in the training corpora of dominant AI systems. The result is a form of cognitive colonialism in which the epistemological assumptions of a handful of developers are quietly embedded in the world's knowledge infrastructure (Birhane, 2021; Crawford, 2021; Gebru et al., 2018).

The third urgency is political. The concentration of AI capability in a small number of private actors - primarily headquartered in the United States and China - creates geopolitical asymmetries that dwarf anything produced by earlier technology waves. Access to frontier AI models, ownership of the semiconductor supply chain, control over cloud infrastructure, and dominance of the data-labelling economy are structural determinants of whose values get encoded in the world's cognitive infrastructure. Countries and regions that do not develop indigenous AI capacity and governance frameworks will find themselves subject to AI systems designed for someone else's purposes (Crawford, 2021; Zuboff, 2019).

The fourth urgency is social. The psychological and relational effects of AI-mediated social life are accumulating at population scale. Agency decay - the erosion of the sense that one's choices and aspirations are genuinely one's own - is emerging as a public health concern. The emotional commons, the shared reservoir of attunement and relational trust that makes community life possible, is under pressure from algorithmic curation that rewards polarisation. These are not abstract philosophical concerns; they have measurable effects on mental health, civic participation, and democratic functioning (Roberts, 2019; Walther, 2026; Zuboff, 2019).

Together, these four urgencies constitute an argument for urgency rather than gradualism. The window for shaping AI development toward prosocial ends is not indefinite. Path dependencies harden, institutional arrangements consolidate, and populations adapt to extractive defaults in ways that make alternatives progressively harder to imagine (Jobin et al., 2019; O'Neil, 2016; Toyama, 2015).

2.9. The Global South as Protagonist: Strategic Capacity for a Hybrid Future

The standard framing of AI and the Global South is one of gap and risk: the AI capability gap between wealthy and poor countries, the risk of AI-amplified inequality, the danger of digital colonialism. These concerns are documented and real. Yet the framing of gap and risk alone produces a passive politics - one in which Global South countries are perpetual recipients of AI ethics frameworks designed elsewhere and adaptation targets for AI systems built for different contexts. Relational accounts of algorithmic justice suggest that such passivity is itself part of the problem, because power over classification, data, and deployment shapes whose interests become technically legible (Birhane, 2021; Crawford, 2021).

A different framing is available, and this essay advances it: the Global South holds particular strategic capacity to lead the transition toward ProSocial AI, for reasons that are structural.

The first structural reason is leapfrog latitude. Countries that did not build their infrastructure around extractive industrial templates of the twentieth century are not locked into their AI extensions. Mobile banking emerged faster in East Africa than in Western Europe precisely because there was no incumbent branch-banking infrastructure to defend. Community health workers in South Asia and Sub-Saharan Africa have developed relational, holistic health delivery models that AI augmentation could scale without first dismantling a medicalised bureaucracy. The absence of legacy infrastructure is a design freedom, provided technology is treated as an amplifier of locally held capability rather than a substitute for it (Sen, 1999; Toyama, 2015).

The second structural reason is demographic and epistemic diversity. The Global South contains the majority of the world's population, the majority of its linguistic diversity, and most of the knowledge traditions historically excluded from dominant AI training corpora. These are resources - for building AI systems that actually work for the full range of human communities - not deficits to be corrected (Birhane, 2021; Gebru et al., 2018; Nussbaum, 2011).

The third structural reason is political will. Several Global South governments have demonstrated a capacity for governance innovation that exceeds what is achievable in contexts where AI industry lobbying power is most concentrated. The African Union's Continental AI Strategy (endorsed July 2024) and the ASEAN Guide on AI Governance and Ethics (February 2024) reflect a genuine appetite for shaping AI development rather than merely receiving it (African Union, 2024; ASEAN, 2024).

The fourth structural reason is problem-fit. Climate adaptation, pandemic preparedness, informal economy inclusion, multilingual public services, community-based resource management - these problems are disproportionately concentrated in Global South

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

contexts. Building AI systems that work well for these domains requires design teams, training data, and governance frameworks drawn from those contexts. This is also why capability expansion, rather than simple tool diffusion, is the appropriate development lens for ProSocial AI (Nussbaum, 2011; Sen, 1999).

2.10. Latin America: From Policy Declaration to ProSocial Architecture

Latin America deserves sustained attention within the Global South analysis - both because its AI governance landscape is already among the world's most institutionally developed and because the structural conditions for ProSocial AI leadership are particularly visible there. The region's significance is not only regulatory; it is epistemic, linguistic, environmental, and developmental (ECLAC & CENIA, 2025; UNESCO & CAF, 2024).

The region's governance record is striking. In October 2023, Chile convened the first Ministerial and High-Level Authorities Summit on the Ethics of AI in Latin America and the Caribbean, organised by UNESCO, CAF (Development Bank of Latin America and the Caribbean), and the Chilean government. Twenty nations signed the Santiago Declaration, committing to ethical AI development principles grounded in UNESCO's Recommendation on AI Ethics and establishing the Regional Artificial Intelligence Council for Latin America and the Caribbean (UNESCO, 2022). The 2024 Montevideo Declaration deepened this architecture, adopting a regional roadmap with five priority lines: governance and regulation, talent and the future of work, protection of vulnerable groups, environment and sustainability, and AI infrastructure (UNESCO & CAF, 2024). Summits now occur annually, a mechanism for sustained accountability rather than one-off declarations.

These regional commitments are not merely symbolic. Brazil's Senate approved a risk-based AI governance bill in late 2024; Chile updated its National AI Policy and introduced aligned legislation; Colombia adopted CONPES 4144, a national AI governance policy in 2025; Mexico, Argentina, Uruguay, and Peru are all advancing legislative frameworks (IADB, 2025). The Economic Commission for Latin America and the Caribbean (ECLAC) and Chile's National Center for Artificial Intelligence (CENIA) developed the Latin American Artificial Intelligence Index (ILIA), a tool tracking AI preparedness, adoption, and governance across nineteen countries in the region — a remarkable act of regional self-knowledge (ECLAC & CENIA, 2025).

Technically, the most instructive development is Latam-GPT. Coordinated by CENIA, backed by CAF, the Chilean government, and AWS, and developed by more than thirty institutions across the region, Latam-GPT is an open-source large language model trained specifically on Latin American data — court decisions from Buenos Aires, library records from Peru, school textbooks from Colombia — and including Indigenous languages such as Nahuatl, Quechua, and Mapudungun (CENIA & CAF, 2025; Brookings Institution, 2025b; Rest

of World, 2025). Its release marks a concrete move from digital consumption to creation: a region building AI that reflects its own languages, histories, and epistemologies rather than importing Silicon Valley's assumptions.

Latam-GPT embodies three of the four ProSocial AI criteria in its design intent. It is Tailored to regional linguistic and cultural specificity. It is being Trained on locally sourced, ethically obtained, contextually rich data - the inverse of a generic English-dominant corpus. And its open-source architecture creates conditions for ongoing Testing and accountability by communities across the region. The remaining criterion, Targeted deployment toward clear human and planetary purposes, is where governance frameworks like the Santiago and Montevideo Declarations must do their work - translating the technical artifact into accountable social infrastructure. In this sense, Latam-GPT is not only a model; it is a test case for whether data documentation, contextual design, and regional governance can be aligned (CENIA & CAF, 2025; Gebru et al., 2018; UNESCO & CAF, 2024).

Latin America also offers a specific lesson in structural fit between the region's development challenges and ProSocial AI's potential applications. Informality is a defining feature of Latin American economies: between 50 and 60 percent of employment in most countries operates outside formal labour protections (ILO, 2024). AI systems Tailored to informal economy contexts — tools for financial inclusion, health navigation, legal aid, skills recognition — could generate substantial human benefit precisely because dominant AI systems, built for formal, documented, institutionalised life, serve those contexts poorly. The region's linguistic diversity — Spanish, Portuguese, and more than 500 Indigenous languages — makes the Trained criterion particularly consequential: training data choices have direct effects on who benefits and who is excluded.

The environmental dimension is equally specific. Latin America hosts 40 percent of the world's biodiversity and contains the Amazon basin, whose regulatory function for global carbon and water cycles is irreplaceable. The Planet dimension of the ProSocial AI Index maps directly onto some of the most consequential land-use, agricultural, and water-management decisions on Earth. AI deployed in service of extractive industries without adequate governance is an ecological risk; AI Targeted toward Amazon monitoring, deforestation detection, smallholder agricultural support, and water system management is a planetary asset (ECLAC, 2024).

What would it mean for Latin America to be a ProSocial AI champion in practice? It would mean operationalising the ProSocial AI Index as a regional procurement standard - requiring AI systems deployed in public services, health, education, and financial inclusion to meet the 4T×4P criteria before deployment approval. It would mean integrating Double Literacy into national curricula across the region, building a generation of citizens equipped to

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

evaluate AI deployments critically and generate locally appropriate solutions. It would mean using HROV as the measurement framework for all AI investments receiving public finance through development banks - CAF, IADB, and national development institutions - shifting the standard of accountability from cost efficiency alone to human and planetary value generation (IADB, 2025; Sen, 1999; Walther, 2024).

It would also mean asserting, in international AI governance forums, that the Regional Artificial Intelligence Council for Latin America and the Caribbean is not a consultative decoration but a standard-setting institution with genuine authority to develop binding AI governance criteria for the region. The Santiago and Montevideo processes have created the institutional seed; the work ahead is building it into something with teeth.

2.11. What ProSocial AI Governance Requires: Architecture for a New Paradigm

Articulating ProSocial AI as a framework is necessary but insufficient without governance architecture. Four requirements stand out. Each requirement responds to a documented weakness in existing AI ethics regimes: the tendency to remain voluntary, abstract, and detached from affected communities' power to contest deployment (Coeckelbergh, 2020; Jobin et al., 2019).

The first is multi-stakeholder accountability with structural substance. Not the corporate-led, NGO-decorated consultative processes that pass for stakeholder engagement in most AI governance forums, but processes in which affected communities hold formal decision-making power over AI systems that affect them. Community AI councils with formal advisory roles in procurement decisions, participatory impact assessments that activate before deployment rather than after harm, and redress mechanisms with genuine consequences are the institutional forms this requires. Relational ethics and work on automated inequality both point to the same lesson: those who bear algorithmic risk must have voice, remedy, and institutional power before systems are normalised (Birhane, 2021; Eubanks, 2018).

The second is the Hybrid Return on Values (HROV) as the standard measurement instrument. Applied consistently across public sector AI deployments, development bank financing criteria, and national procurement standards, HROV converts the aspiration of prosocial governance into auditable performance data. It makes costs visible, enables comparison across deployments, and creates market pressure on developers to compete on human and planetary value, not only on capability benchmarks (Crawford, 2021; Floridi et al., 2018; Taddeo & Floridi, 2018).

The third is educational infrastructure built around Double Literacy. Scaling Double Literacy through formal education systems, professional development programmes, and community learning networks creates the human capital base without which ProSocial AI governance remains an elite discourse disconnected from the communities it nominally serves. In Latin America, where educational systems are already engaged in significant curriculum reform conversations, Double Literacy integration represents a tractable near-term priority and a practical expression of capability development in a hybrid era (Nussbaum, 2011; Sen, 1999; Walther, 2026).

The fourth is international cooperation through reciprocal rather than hierarchical relationships. The current landscape of AI ethics initiatives - dominated by declarations from wealthy-country governments and technology companies - will not produce globally legitimate governance. A genuine international ProSocial AI architecture would require regional bodies with democratic mandates, funding mechanisms that do not route Global South participation through donor dependency, and standards designed to generate binding obligations rather than voluntary guidelines that extractive actors can selectively adopt or ignore. Regional initiatives such as those advanced by UNESCO, CAF, ASEAN, and the African Union indicate that standard-setting capacity already exists beyond the conventional centres of AI power (African Union, 2024; ASEAN, 2024; Jobin et al., 2019; UNESCO & CAF, 2024).

3. Conclusions

The argument assembled here is that the dominant direction of AI development encodes values of extraction, concentration, and short-term return, and that a deliberate, structurally grounded alternative exists. This claim is supported by scholarship on surveillance capitalism, algorithmic injustice, automated inequality, and the planetary costs of AI, all of which show that technological systems organise power as well as performance (Birhane, 2021; Crawford, 2021; Eubanks, 2018; Zuboff, 2019).

ProSocial AI names that alternative with precision: AI systems that are Tailored, Trained, Tested, and Targeted to bring out the best in and for people and planet. The 4T×4P ProSocial AI Index provides the diagnostic framework. Hybrid Intelligence and the Hybrid Tipping Zone locate the stakes for human agency. The Hybrid Return on Values (HROV) converts prosocial principles into auditable performance data. Double Literacy equips populations to participate meaningfully in governance. The Fourth Path provides the philosophical orientation. Together, these constitute a toolkit - instruments that can be applied in procurement decisions, curriculum design, regulatory frameworks, and international negotiations today (Floridi et al., 2018; Taddeo & Floridi, 2018; Walther, 2024, 2026).

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

The Latin America dimension of the argument deserves emphasis. The region has moved further and faster on AI governance institution-building than most of the world - from the Santiago Declaration and Montevideo Declaration to the Regional AI Council, from ILIA to Latam-GPT. This is not coincidental. Latin America's combination of institutional creativity, epistemic richness, demographic vitality, and clear-eyed development priorities creates conditions for ProSocial AI leadership that are structural rather than incidental (ECLAC & CENIA, 2025; UNESCO & CAF, 2024).

Three specific calls emerge. The first: operationalise the ProSocial AI Index as a regional procurement and governance standard in Latin America and the Caribbean, using the existing ministerial summit architecture as the legitimating framework. The second: integrate Double Literacy into national curriculum frameworks across the region - a complement to the digital skills agendas already underway, but one that takes the human half of the human-AI relationship as seriously as the algorithmic half. The third: establish HROV as the measurement standard for all AI investments receiving public development finance in the region, beginning with CAF and IADB-backed projects, and extending the standard through South-South cooperation channels (IADB, 2025; Sen, 1999; Walther, 2026).

The Hybrid Future will be built wherever communities insist on AI systems that serve their aspirations, honour their knowledge, respect their emotional realities, and sustain the planetary systems on which their survival depends. Latin America is generating that insistence already - in governance rooms and in research labs, in policy declarations and in open-source model architectures. What it has, and what the world needs to recognise, is that ProSocial AI is already being imagined here. The work ahead is making it the standard (Birhane, 2021; Crawford, 2021; ECLAC & CENIA, 2025; Walther, 2024).

Bibliographic References

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns, 2*(2), Article 100205. https://doi.org/10.1016/j.patter.2021.100205

Coeckelbergh, M. (2020). *AI ethics.* MIT Press.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Eisenberg, N., & Mussen, P. H. (1989). *The roots of prosocial behavior in children*. Cambridge University Press.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2018). *Datasheets for datasets*. arXiv. https://doi.org/10.48550/arXiv.1803.09010

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Farrar, Straus and Giroux.

Nussbaum, M. C. (2011). *Creating capabilities: The human development approach.* Harvard University Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Sen, A. (1999). *Development as freedom*. Oxford University Press.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.

Toyama, K. (2015). *Geek heresy: Rescuing social change from the cult of technology.* PublicAffairs.

Walther, C. C. (2024). *Human leadership for humane technology*. Palgrave Macmillan.

Walther, C. C. (2026). *From natural to hybrid intelligence*. Palgrave Macmillan.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* PublicAffairs.

Additional References and Policy Documents

African Union. (2024). *Continental Artificial Intelligence Strategy*. African Union Commission. https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy

Association of Southeast Asian Nations (ASEAN). (2024). *ASEAN Guide on AI Governance and Ethics*. ASEAN Secretariat. https://asean.org/book/asean-guide-on-ai-governance-and-ethics/

Brookings Institution. (2025a, February). *Regional cooperation crucial for AI safety and governance in Latin America*. https://www.brookings.edu/articles/regional-cooperation-crucial-for-ai-safety-and-governance-in-latin-america/

Brookings Institution. (2025b, November). *Latam-GPT and the search for AI sovereignty*. https://www.brookings.edu/articles/latam-gpt-and-the-search-for-ai-sovereignty/

Centro Nacional de Inteligencia Artificial de Chile (CENIA) & Development Bank of Latin America and the Caribbean (CAF). (2025). *Latam-GPT: Open large language model for Latin America and the Caribbean.* https://accesspartnership.com/opinion/launch-of-the-first-open-large-language-model-for-latin-america-the-caribbean-latam-gpt/

Development Bank of Latin America and the Caribbean (CAF), & United Nations Educational, Scientific and Cultural Organization (UNESCO). (2023*). Santiago Declaration to promote ethical artificial intelligence in Latin America and the Caribbean*. UNESCO. https://www.unesco.org/en/articles/authorities-24-countries-participate-1st-forum-high-authorities-ethics-artificial-intelligence-latin

Economic Commission for Latin America and the Caribbean (ECLAC). (2024). *Digital agenda for Latin America and the Caribbean*. United Nations.

Economic Commission for Latin America and the Caribbean (ECLAC), & Centro Nacional de Inteligencia Artificial de Chile (CENIA). (2025). *Latin American Artificial Intelligence Index (ILIA 2025)*. United Nations.

https://www.cepal.org/en/pressreleases/latin-america-and-caribbean-accelerate-adoption-artificial-intelligence-though

Inter-American Development Bank (IADB). (2025). *An enabling regulatory framework for artificial intelligence in Latin America and the Caribbean.* https://publications.iadb.org/en/enabling-regulatory-framework-artificial-intelligence-latin-america-and-caribbean

International Labour Organization (ILO). (2024). *World employment and social outlook: Trends 2024.* International Labour Office. https://www.ilo.org/global/research/global-reports/weso/WCMS_900148/lang--en/index.htm

Rest of World. (2025, August). *Latin America is building LatamGPT to rival ChatGPT.* https://restofworld.org/2025/chatgpt-latin-america-alternative-latamgpt/

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2022). *Recommendation on the ethics of artificial intelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000381137

United Nations Educational, Scientific and Cultural Organization (UNESCO), & Development Bank of Latin America and the Caribbean (CAF). (2024). *Montevideo Declaration and regional AI roadmap for Latin America and the Caribbean.* https://www.telefonica.com/en/communication-room/blog/2nd-unesco-ministerial-summit-ethics-ai-latin-america-caribbean/

Esta obra está bajo una Licencia Creative Commons Atribución 4.0 Internacional.