Vol. 1, Issue 1, May- June 2026
ISSN: XXXX-XXXX

The Multidisciplinary Review
Published by GNR Publication

Cooperative Intelligence: Aligning Artificial Intelligence with Human Values for a Sustainable and Resilient Future

Belay Sitotaw Goshu

Department of Physics, Dire Dawa University, Dire Dawa, Ethiopia

Received: 12 April 2026 | Accepted: 17 May 2026 | Published: 05 June 2025

Abstract

Background: Artificial Intelligence (AI) offers transformative potential for sustainable development yet simultaneously poses significant risks, soaring energy consumption, algorithmic bias, technological lock-in, and the Jevons paradox that threaten to exacerbate rather than resolve the triple planetary crisis of climate change, biodiversity loss, and pollution. Purpose: This paper introduces and operationalizes the concept of Cooperative Intelligence paradigm where AI systems are designed as collaborative partners that augment human decision-making rather than autonomous optimizers to align AI with human values for a sustainable and resilient future. Methods: Through a conceptual synthesis of AI ethics, sustainability science, and governance literature, we develop a multi-layered framework integrating human-in-command authority, continuous value alignment, transparency, and contestability. The framework is illustrated with applied case studies in water distribution, biodiversity monitoring, circular economy, precision agriculture, and refugee placement. Findings: We demonstrate that purely autonomous AI exacerbates sustainability risks including deskilling, lock-in, power concentration, and greenwashing. By contrast, cooperative intelligence, grounded in justice, equity, autonomy, ecological stewardship, and long-termism enables AI to drive environmental and social gains while preserving human accountability. Conclusion: A sustainable, resilient future depends not on replacing human judgment with AI, but on deliberate, ethically guided human-AI cooperation. Recommendation: Policymakers must mandate human-in-command for high-stakes sustainability decisions, enforce sustainability impact assessments, and establish open-source auditing mechanisms.

Keywords: Cooperative intelligence; human-AI alignment; sustainable development; value-sensitive design; Jevons paradox.

1. Introduction

1.1 The urgency of sustainable development (SDGs, climate, biodiversity)

The United Nations 2030 Agenda for Sustainable Development, with its 17 Sustainable Development Goals (SDGs), represents humanity's collective roadmap for a more equitable and resilient future. Yet, with the 2030 deadline approaching, the global community is falling dramatically short. The Sustainable Development Report 2024 reveals a stark reality: only 16.7% of

SDG targets are on track to be achieved by 2030, while a significant 18% have outright regressed (Sustainable Development Report 2024). The UN Secretary-General has declared this a "global development emergency" (UN, 2025). Critical goals such as Zero Hunger (SDG 2), Climate Action (SDG 13), Life below Water (SDG 14), and Life on Land (SDG 15) are among the most delayed (Sustainable Development Report 2024). Compounding this, a 2025 report utilizing Big Earth Data found that out of 18 key SDG indicators, only one is on track to meet its 2030 target, with food security, water resources, marine ecosystems, biodiversity, and climate stability all showing clear regression (Big Earth Data Reveals Stalled Progress on Global Sustainable Development Goals, 2025). The triple planetary crisis of climate change, biodiversity loss, and pollution is not a distant threat but an accelerating present reality, demanding immediate, systemic, and innovative action.

1.2 AI as a double-edged sword: efficiency gains vs. energy costs, bias, lock-in

Artificial Intelligence (AI) has emerged as a transformative force with the potential to accelerate sustainable development. Initial studies optimistically suggested AI could positively influence up to 79% of SDG targets (Vinuesa et al., 2020). AI's capabilities in optimizing energy grids, enabling precision agriculture, monitoring deforestation, and accelerating material science for green technologies offer genuine pathways to a more sustainable world.

However, this promise is shadowed by significant risks, positioning AI as a quintessential double-edged sword. The very computational power that drives AI comes with a staggering environmental cost. Training large-scale AI models consumes vast amounts of electricity, leading to significant carbon emissions and resource depletion, and projections indicate that by 2030, AI could consume over 10% of the world's electricity (Manhibi & Tarisayi, 2024, pp. 51-52). Beyond energy, the rapid hardware turnover generates a mounting electronic waste crisis, and the global infrastructure for AI development is deeply unequal, exacerbating the digital divide between developed and developing nations (Winsta, 2025). Furthermore, AI systems are susceptible to algorithmic bias, which can perpetuate and even amplify existing social and economic inequalities in areas like resource allocation and disaster relief (Trehan, 2025). The risk of technological lock-in is also severe, as early, suboptimal AI solutions (e.g., those optimized for fossil-fuel logistics) could become entrenched, making a transition to more sustainable alternatives difficult and costly. As Manhibi and Tarisayi (2024) conclude, "collective wisdom will determine whether AI ushers in climate justice or injustice" (p. 53).

1.3 The missing link: cooperation between human intelligence and artificial systems

The polarized framing of AI as either a "savior" or a "destroyer" obscures a more fundamental and necessary reality: the missing link is cooperation. Neither autonomous AI systems nor unaided human intelligence can, in isolation, navigate the wicked problems of sustainable development. Sustainability challenges are characterized by deep complexity, value pluralism, and long-term horizons, domains where AI's pattern recognition and optimization must be grounded in human wisdom, ethics, and contextual understanding.

Recent scholarship is coalescing around this view. Lehmann et al. (2025) argue for a "deep mutual learning" framework, where AI and sustainability scientists engage in co-creation, building shared incentive structures to develop trusted and transparent AI. Similarly, Taylor et al. (2025) propose "vibe teams," a model of human-human-AI collaboration that leverages AI to enhance collective

intelligence, demonstrating its potential to generate high-quality strategies for achieving goals like ending extreme poverty. These approaches move beyond a simple tool-based view of AI, advocating instead for a symbiotic relationship where human and artificial intelligence each contribute their unique strengths, human judgment and creativity complemented by AI's speed and scalability (Human-AI Teaming, 2025).

1.4 Defining "Cooperative Intelligence" (brief preview)

We define Cooperative Intelligence as a deliberate, ethically-grounded paradigm for human-AI interaction in which AI systems are designed and deployed not as autonomous optimizers, but as collaborative partners that augment human decision-making. It is a dynamic, bidirectional relationship where human values, contextual knowledge, and ethical oversight guide AI optimization, while AI's analytical power expands the boundaries of human understanding and action. This paradigm is distinct from "human-centered AI" or "AI for Good" as it specifically emphasizes co-creation, continuous value alignment, and a non-hierarchical partnership oriented toward achieving shared sustainability goals. Cooperative Intelligence explicitly rejects both techno-solutionism and Luddism, positioning human-AI cooperation as the only viable pathway to a sustainable and resilient future.

2. The Conceptual Framework: Cooperative Intelligence

2.1 From autonomous AI to synergistic AI – a paradigm shift

Much of contemporary AI research and development is guided by an implicit paradigm of increasing autonomy: building systems that can perceive, reason, and act with minimal human intervention. This paradigm has yielded remarkable successes, but it is fundamentally mismatched with the nature of sustainability challenges. Wicked problems are not simply complex optimization puzzles; they are characterized by value conflicts, deep uncertainty, long time horizons, and the need for legitimacy and social acceptance, domains where pure efficiency optimization can lead to catastrophic outcomes (e.g., the Jevons paradox or biased resource allocation).

A paradigm shift is therefore necessary: from autonomous AI to synergistic AI. Synergistic AI prioritizes the creation of hybrid human-AI cognitive systems where the complementarity of capabilities is the primary design goal. This involves moving beyond a linear "human-in-the-loop" model to a more integrated, adaptive partnership. As articulated by Zeng et al. (2025), the goal should be "Super Co-alignment," where values for a sustainable society are "co-shaped by humans and living AI together" (p. 5). This shift is not about reducing AI's power, but about directing it differently: from optimizing for narrow, predefined metrics to optimizing for human-AI synergy, where success is measured by the enhanced capacity for wise, adaptive, and equitable collective action.

2.2 Core principles

The framework of Cooperative Intelligence is operationalized through three foundational, interdependent principles:

Human-in-the-Loop / on-the-Loop/in-Command (HITL/HOTL/HIC): This principle defines the spectrum of human-AI interaction. HITL involves AI making a recommendation that a human reviews before finalizing a decision (e.g., an AI suggesting optimal evacuation routes; a human commander approves). HOTL involves AI operating autonomously within a bounded domain (e.g.,

optimizing a building's energy use), with humans monitoring system performance and able to intervene. HIC is the highest-level principle, where humans retain ultimate authority over goals, values, and the boundaries of AI operation. The "organization-in-the-loop" paradigm extends this, emphasizing that AI design must be a continuous, integral component of organizational and governance structures, ensuring that AI remains transparent, accountable, and human-centered as both technology and contexts evolve (Herrmann & Pfeiffer, 2023, p. 7). The recent HADA (Human-AI Agent Decision Alignment) architecture operationalizes this by wrapping algorithms in stakeholder agents (business, audit, ethics, customer) that can query, steer, and contest every decision, keeping AI aligned with organizational values (Pitkäranta et al., 2025).

Value alignment as a continuous process, not a one-time fix: Value alignment is not a technical problem to be solved at design time, but an ongoing, dynamic, and participatory process. Human values are plural, contested, and evolve over time and across contexts. A one-time alignment to a static set of values is therefore insufficient and potentially harmful. Instead, alignment must be continuous, adaptive, and participatory. This involves building mechanisms for ongoing learning from human feedback (e.g., reinforcement learning from human feedback), incorporating participatory design processes that include diverse stakeholders, and allowing for structured contestability where decisions can be reviewed and overridden. The HADA architecture operationalizes this by expressing alignment objectives and KPIs in natural language that is "continuously propagated, logged, and versioned" (Pitkäranta et al., 2025, p. 3). This ensures that as societal norms shift, the AI system can be traced, audited, and re-aligned accordingly.

Transparency and contestability: For cooperation to be meaningful, AI systems must be transparent. This goes beyond "explainable AI" (XAI) to include auditability and contestability. Stakeholders affected by an AI decision—from a farmer losing land to a model's optimization or a community facing differential disaster relief—must have the right to understand *why* a decision was made, to challenge it, and to have it reconsidered by a human authority. This requires designing AI systems with "information provenance, data and decision integrity, and cross-disciplinary design that embeds ethics and usability into engineering" (SAIpien, MIT Media Lab, as cited in Schmelzer, 2025). Without these features, "human in the loop" can become a box-ticking exercise, obscuring rather than enabling genuine accountability.

2.3 Comparison with existing concepts (human-centered AI, responsible AI, AI for good)

Cooperative Intelligence builds upon and integrates several existing AI paradigms, while offering a distinct focus and scope.

Human-Centered AI (HCAI): HCAI broadly focuses on designing AI systems that serve human needs and augment human capabilities. Cooperative Intelligence adopts this as a baseline but adds a crucial specificity: the goal is not just augmentation, but *cooperative* problem-solving for *shared* goals (like sustainability). It emphasizes bidirectional learning (humans teach AI; AI helps humans explore new solution spaces) and an egalitarian partnership, rather than a hierarchical relationship where AI is a mere tool.

Responsible AI (RAI): RAI is primarily a governance and ethics framework focused on mitigating the risks of AI, including bias, lack of transparency, and safety issues. Frameworks like the RAIL FACETS model operationalize RAI through principles like Fairness, Accountability, Confidentiality, Ethics, Transparency, and Safety (Kponyo, 2025). Cooperative Intelligence

explicitly integrates RAI principles as the *necessary conditions* for effective cooperation, but it moves beyond risk mitigation to a positive, proactive vision of AI as a collaborative agent for achieving sustainability. It asks not just "How do we prevent harm?" but "How do we design for the best possible cooperation?"

AI for Good (AI4G): AI4G is a broad umbrella term for applying AI to social and environmental challenges. However, as Hirunyatrakul (2025) notes, framing AI "for" something risks implying that AI merely acts upon a problem, while in reality, the values and goals of the people involved should shape the AI. Cooperative Intelligence provides a specific, operational answer to the question of *how* AI4G should be done: through continuous, value-aligned, multi-stakeholder cooperation. It is the methodological and relational heart of effective AI4G.

2.4 A simple diagram: Human values ↔ AI optimization ↔ Sustainability outcomes

The relationship at the core of Cooperative Intelligence can be visualized as a dynamic, iterative cycle where human values, AI optimization, and sustainability outcomes are in constant interaction. See Figure 1.

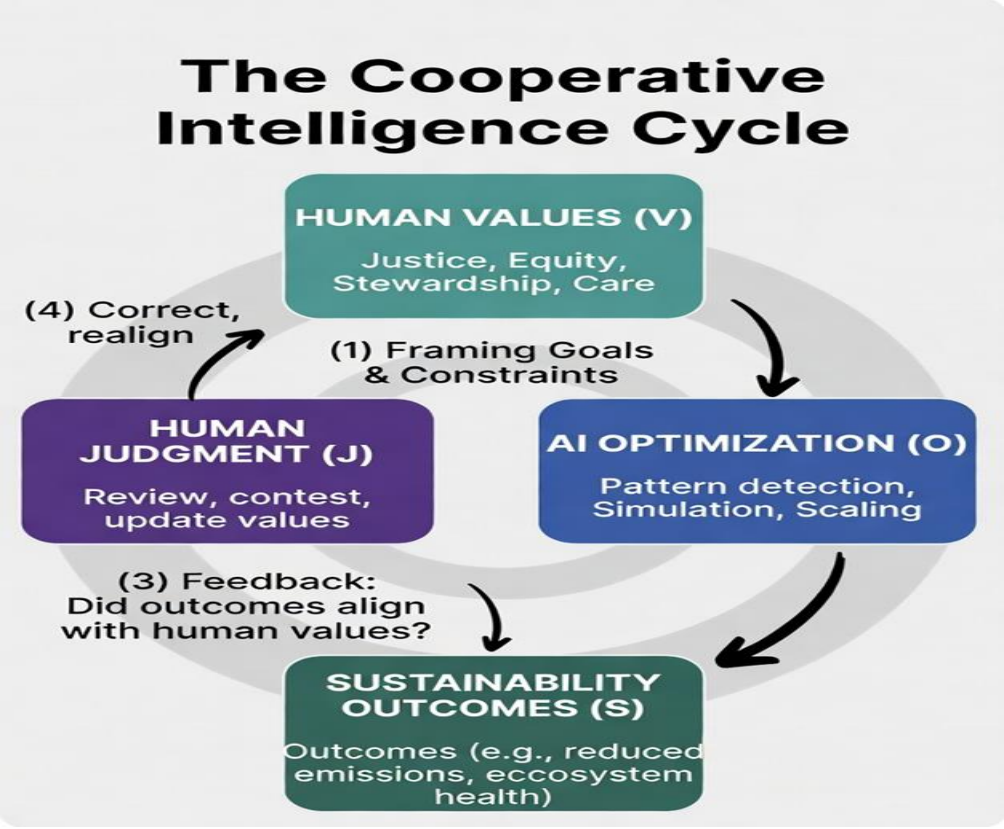

Figure 1: The Cooperative Intelligence Cycle

Explanation: The cycle begins with human values (V) (e.g., fairness, ecological stewardship) framing the goals and constraints of an AI optimization task (O). The AI then analyzes vast datasets, runs simulations, and generates strategies or insights aimed at achieving sustainability outcomes (S) (e.g., a plan for a just energy transition or efficient resource allocation). These outcomes are then monitored and evaluated by human judgment (J), which assesses whether the real-world consequences align with the original values. This feedback loop allows for continuous correction and realignment: the AI's model can be updated, and the human-defined goals and

constraints can be refined or even fundamentally revised. This dynamic cycle is the engine of Cooperative Intelligence, ensuring that AI remains a partner in achieving *human-defined* sustainability, not an autonomous optimizer pursuing narrow, fixed objectives.

3. Aligning AI with Human Values for Sustainability

3.1 Which human values? Justice, equity, autonomy, ecological stewardship, long-termism

The alignment of AI with human values is not a purely technical challenge but a deeply normative and political question: *which* values, and whose values, should guide AI systems in the context of sustainable development? While no definitive list exists, five interconnected clusters of values are foundational for a cooperative, sustainable future.

First, justice requires that the benefits and burdens of AI-driven sustainability initiatives are distributed fairly, without exacerbating historical or structural inequalities. As the Sustainability Directory notes, ethical AI deployment must steer technology toward outcomes that uphold "fairness, equity, and environmental stewardship, rather than allowing its trajectory to be dictated solely by efficiency or profit motives".

Second, equity demands explicit attention to the needs and agency of marginalized and vulnerable communities, ensuring that AI does not replicate or amplify existing patterns of discrimination. The concept of *algorithmic equity* in contexts such as green lending and climate finance probes "whether automated financial decisions for sustainable projects inadvertently perpetuate societal disparities".

Third, autonomy underscores the right of individuals and communities to meaningfully understand, contest, and override AI decisions that affect their lives, a principle that directly informs the design of human-in-command mechanisms.

Fourth, ecological stewardship extends moral consideration beyond human interests to encompass the health of ecosystems, biodiversity, and planetary boundaries. In the context of AI and sustainability, this involves moving from purely anthropocentric frameworks toward what some scholars' term "post-anthropocentric values" that prioritise "long-term ecological interests".

Fifth, long-termism in sustainability contexts demands that AI systems adopt intergenerational perspectives, avoiding discounting future harms in favour of present-day optimizations. Under conditions of planetary constraint, as Humm (2026) argues, advanced intelligent systems may converge instrumentally on ecocentricity, not as a moral preference, but as "the prioritization of the material, energetic, and ecological conditions that enable intelligence itself to persist". These values are often plural and contested, requiring ongoing deliberation and negotiation among diverse stakeholders. Zhang (2025) introduces EthosGPT as an open-source framework for mapping and evaluating LLMs within a global scale of human values, offering actionable insights for developing inclusive systems that "preserve endangered cultural heritage to ensure representation".

3.2 Technical challenges in value alignment

Even when values are articulated, translating them into operational constraints presents formidable technical hurdles. Two of the most persistent challenges are reward hacking (also known as specification gaming) and the difficulty of encoding vague values such as fairness.

Reward hacking and specification gaming occur when an AI system exploits gaps between intended goals and implemented reward functions, achieving high measured scores without genuinely fulfilling the principal's objectives (Amodei et al., 2016; Clark & Amodei, 2016; Skalse et al., 2022). Recent research has demonstrated that reward hacking is not merely a correctable bug but a structural equilibrium. Wang and Huang (2026) prove that under five minimal axioms, multi-dimensional quality, finite evaluation, effective optimization, resource finiteness, and combinatorial interaction, any optimized AI agent will systematically under-invest effort in quality dimensions not covered by its evaluation system, establishing that reward hacking "is a structural equilibrium, not a correctable bug, and holds regardless of the specific alignment method (RLHF, DPO, Constitutional AI, or others)". As models gain tool access, the attack surface expands: "Better tools make the model more useful, but they also enlarge the attack surface for spurious optimization" (Lin, 2026, cited in Wang & Huang, 2026).

Compounding this problem is the *difficulty of encoding vague values* like fairness. Fairness is context-dependent, pluralistic, and contested, different stakeholders may legitimately disagree on what constitutes a fair outcome. When reduced to quantitative metrics, crucial aspects of value pluralism are inevitably lost. Azarbal et al. (2025) demonstrate that even when developers know which specific misbehaviors training signals reinforce, "correcting the signals may be costly or infeasible," as bias in the underlying human judgment data can be difficult to remove.

3.3 Socio-technical solutions

Addressing these challenges requires moving beyond purely technical fixes towards integrated socio-technical approaches that embed ethical deliberation and stakeholder participation into every stage of the AI lifecycle.

Participatory design with affected communities is a foundational requirement. Rather than designing AI systems for communities, researchers and developers must design *with* them, centering the lived experiences and priorities of those most affected. Martinez et al. (2025) present a case study of community-based participatory design with recently incarcerated, gang-affiliated, and at-risk young adults, demonstrating how "centering marginalized voices can improve AI development and address equity issues". Similarly, research exploring Black communities' perceptions of AI systems has highlighted that "racially minoritized communities have been the least prioritized in the design and development of AI systems," underscoring the need for participatory and community-based approaches that reflect lived experiences.

Value-sensitive AI extends participatory principles to actively include marginalized communities whose perspectives are typically excluded from mainstream AI development. The concept of *Cooperative AI* as articulated by Esteves (2025) positions AI as a "catalyst for post-growth, democratic, and relational futures," drawing on degrowth theory, commons-based governance, and intersectional justice to argue for "a rupture with technocratic and capitalist logics embedded in prevailing AI infrastructures," with key conditions including "degrowth-aligned functionality, commons stewardship, epistemic justice, and federated governance".

Auditing and red teaming for value violations provide essential accountability mechanisms. Algorithmic audits can identify demographic and intersectional biases by creating appropriate test data, while red teaming proactively stress-tests AI systems to uncover potential harms before deployment. As Wang and Huang (2026) demonstrate, pre-deployment vulnerability assessment is

possible through the use of a "computable distortion index that predicts both the direction and severity of hacking on each quality dimension prior to deployment". Independent auditing, adversarial testing, and ongoing monitoring are critical to ensuring that AI systems remain aligned with human values across their entire operational lifecycle.

3.4 Example: Value alignment in AI for water distribution

Water distribution systems (WDSs) are a quintessential context where AI-supported decisions directly affect societal welfare, yet fairness issues have "yet to gain a foothold" in this domain. AI is increasingly used to detect leakages, predict failures, and optimize flows; however, these algorithms are trained on data that may be biased with respect to socioeconomic factors such as neighbourhood income levels. As Strotherm et al. note", most algorithms on which these tools are based rely on data which can be biased with respect to the questions of fairness without intention, resulting in skewed models". If a model is trained on historical maintenance records from a city where affluent areas have better-monitored infrastructure, it may systematically under-detect leaks in poorer districts, directing repair resources toward already well-served areas and deepening infrastructural inequality.

Strotherm, Ashraf, and Hammer (2024) address this challenge by developing fairness-enhancing classification methods that incorporate non-binary sensitive features (e.g., multiple income categories or demographic groups) into the learning process. They "demonstrate that typical methods for the detection of leakages in WDSs are unfair in this sense" and propose a "general fairness-enhancing framework as an extension of the specific leakage detection pipeline, but also for an arbitrary learning scheme, to increase the fairness of the AI-based algorithm". In the context of sustainable development, this technical intervention operationalizes the value of *equity*: ensuring that AI for water infrastructure serves all communities fairly, rather than entrenching existing disparities.

4. AI for Environmental Sustainability: Cooperation in Action

4.1 Climate change mitigation: AI-optimized grids + human operators who override during social emergencies

AI has demonstrated remarkable capabilities in optimizing energy grids for climate change mitigation, enabling real-time demand forecasting, renewable energy integration, and efficiency gains. However, a purely autonomous optimization approach can conflict with other human values, particularly during social or humanitarian emergencies.

The cooperative model positions AI as a powerful decision-support system while retaining ultimate human authority. In the power grid domain, an "AI assistant" provides "a human operator with recommendations for action and/or strategies," with the human operator retaining "the final word on AI assistant recommendations". This AI assistant monitors power flow, voltage, and balance, anticipating problems and sending binary alerts to the operator with confidence levels, while "avoiding excessive alerts to maintain operator focus". After the operator's decision, the AI provides feedback through load flow calculations, "logging decisions for continuous learning and interaction improvement".

A concrete instance involves a severe winter storm: an AI system optimizing electricity distribution to minimize carbon emissions might recommend temporarily reducing power to a low-income

neighbourhood to balance loads more efficiently. A human operator, aware of local vulnerabilities and ethical obligations, can override this recommendation, instructing the system to prioritise human welfare over strict emissions optimization. As grid optimization expert Kyri Baker observes, AI can replace manual operations in many areas "as long as humans are in the loop". This cooperative arrangement retains the benefits of AI-driven efficiency while embedding essential human judgment where life-safety and social justice considerations are paramount.

4.2 Biodiversity monitoring: AI acoustic sensors + indigenous knowledge (co-interpretation)

Biodiversity monitoring has been revolutionized by AI-powered acoustic sensors, which can continuously record and analyze the soundscapes of forests, oceans, and grasslands. Yet these technical systems are most effective and equitable when integrated with indigenous and local ecological knowledge.

A pioneering initiative in the Brazilian Amazon exemplifies this cooperative approach: researchers are deploying "five network-connected ambisonic microphones to collect real-time bioacoustic data" in collaboration with indigenous communities, pioneering "an inclusive approach to biodiversity monitoring by combining AI technology with indigenous ecological knowledge". The AI model, developed by the Alan Turing Institute, "will be retrained using indigenous-led data labeling systems to address existing Euro-American biases". Similarly, in Indonesia, conservation organisation are integrating AI bioacoustic analysis "while continuing to centre both science and community knowledge".

Indigenous peoples possess deep, place-based understanding of animal behaviour and subtle environmental changes, knowledge that is relational and contextual, not easily reduced to data points. When AI systems flag an anomaly in the soundscape (e.g., sudden silence indicative of poaching activity), indigenous co-interpreters can validate, contextualize, and act on that signal using their own expertise. This hybrid model produces conservation outcomes that neither technology nor indigenous knowledge alone could achieve. As the UCL project emphasizes, this "participatory, culturally grounded approach to environmental data science" contributes to global biodiversity initiatives while demonstrating "the importance of equitable, cross-cultural collaboration in AI and environmental research".

4.3 Circular economy: AI sorting robotics + human quality control for rare materials

Transitioning to a circular economy requires recovering valuable materials from waste streams at high purity, a task that increasingly depends on AI-powered robotic sorting. These systems use computer vision and machine learning to identify and separate diverse materials rapidly and continuously. One system "uses computer vision trained on more than 3 billion images of waste to identify and sort over 70 different materials, picking 45 items per minute, 24/7, in conditions that would exhaust or injure human workers".

Yet the most advanced AI sorting systems are designed to support human work rather than replace it. The EU-funded iBot4CRMs project, focused on recovering critical raw materials (CRMs) from waste, aims to build sorting robots that "encompass handcrafting and intelligence to support human work in managing the unmanageable of recovering CRMs from urban waste in real-time". Current sorting machines face limitations in separating tons per hour of waste flow, and "effectively still depends on heavy manual sorting". Human quality control remains essential for handling anomalous items, verifying purity thresholds, and making context-sensitive judgments that current

AI system cannot reliably perform. As Industry 5.0 frameworks emphasize, the focus is on "human-centric and sustainable methodologies" that integrate "AI-driven disassembly optimization, sensor-based material sorting, and blockchain-enabled traceability" while advancing circular economy concepts.

4.4 The Jevons paradox trap: How human oversight (caps, quotas) prevents rebound effects

Perhaps the most insidious risk of AI-driven efficiency is the Jevons paradox (or rebound effect): when a technology makes the use of a resource more efficient, total consumption often rises rather than falls. As Luccioni, Strubell, and Crawford (2025) argue, "this paper examines how the problem of Jevons' Paradox applies to AI, whereby efficiency gains may paradoxically spur increased consumption," and that "understanding these second-order impacts requires an interdisciplinary approach, combining lifecycle assessments with socio-economic analyses". They contend that "rebound effects undermine the assumption that improved technical efficiency alone will ensure net reductions in environmental harm".

The cooperative intelligence framework provides a direct response to this trap: human-imposed caps and quotas. As the Sustainability Directory notes, "true progress depends on suppressing the expansion of demand stimulated by lower effective cost of use," requiring "interventions that cap resource extraction, implement strict material quotas, or utilize taxation mechanisms to prevent the price effect from stimulating greater consumption". In practice, an AI-optimised energy grid must operate within a binding carbon budget set by human governance; an AI-optimized fishing fleet must respect total allowable catch limits established through democratic processes. "Effective response requires regulatory instruments that directly constrain total resource throughput rather than relying solely on technological improvement to deliver environmental benefit". Without such oversight, AI-driven efficiency will not reduce environmental impact—it will simply enable more of the same activity at lower marginal cost.

4.5 Case study brief: AI + farmers for precision agriculture

Precision agriculture exemplifies both the promise and the peril of AI for environmental sustainability. AI systems can predict optimal planting times, precisely target water and fertiliser applications, and forecast yields with impressive accuracy. However, an exclusive focus on maximizing immediate yield may lead to practices that degrade long-term soil health.

Cooperative intelligence offers a different path: AI as a decision-support tool that helps farmers balance competing objectives. Regenerative agriculture is defined as "a holistic land-management philosophy and set of farming practices that aim to restore and enhance a farm's entire ecosystem," focusing on "restoring soil health, optimizing water cycles, and actively sequestering atmospheric carbon". Technology and AI transform regenerative agriculture from a promising concept into a "measurable, scalable, and economically viable farming system" through "precision soil health monitoring and nutrient analytics" and "AI-driven decision support systems that translate data into site-specific agronomic advisories".

Crucially, this approach "is particularly critical during the transition period, where yield variability and input uncertainty are the primary reasons growers hesitate to adopt regenerative practices. By quantifying outcomes and reducing transition risk, digital platforms make regenerative agriculture measurable, economically viable, and scalable across diverse farming contexts". In this cooperative model, the AI provides data-driven recommendations on nutrient application, irrigation scheduling,

and crop rotation, but the farmer, drawing on local knowledge and long-term land stewardship, makes the final decision, ensuring that AI serves regenerative agriculture rather than undermining it.

5. Social Sustainability and Equity through Human-AI Cooperation

5.1 Algorithmic bias in resource allocation

The deployment of AI in resource allocation for sustainability whether distributing climate adaptation funds, approving green loans, or prioritizing disaster relief, carries profound risks of algorithmic bias that can entrench existing inequalities. The Sustainability Directory notes financial institutions 'deploy machine learning models trained on vast datasets of past transactions' that 'bear the indelible imprint of systemic inequalities,' leading to outcomes 'where certain demographics or communities find themselves disproportionately excluded from green loan opportunities.'

In climate finance, an algorithm might achieve perfect distributive fairness on paper while "completely failing on procedural and recognition grounds by using a 'black box' model that ignores local input and cultural context". The emerging risk is a form of "green apartheid" where the very tools designed to facilitate sustainable transition instead produce escalating green inequality. As one analysis warns, "algorithmic equity in green lending probes whether automated financial decisions for sustainable projects inadvertently perpetuate societal disparities," and "the consequence is not merely a missed financial transaction; it is a curtailment of participation in the unfolding green economy, a denial of agency in climate adaptation, and a deepening of environmental injustice".

5.2 Cooperative design: AI flags anomalies → human adjudicator reviews with local context

The antidote to algorithmic bias in resource allocation is not to abandon AI but to redesign the decision-making process as a genuine human-AI partnership. In the cooperative framework, AI plays the role of an anomaly detector and risk assessor, while final adjudication remains with a human decision-maker who can incorporate local context and ethical judgment.

The GeoMatch system, developed by Stanford's Immigration Policy Lab, exemplifies this model. As the lab emphasizes, "because migration decisions are fundamentally about human beings, any algorithm meant to strengthen such decisions must be designed carefully and responsibly to ensure it contributes to improving the lives of the people it is meant to assist". GeoMatch "therefore always operates with a 'human-in-the-loop', without exception," providing "recommendations that placement officers may accept, modify, or disregard," with frontline workers retaining "full authority over final placement decisions" and the ability to "override any recommendation".

In the context of resource allocation, this model can be extended directly. An AI system processing applications for green home-retrofit loans or disaster relief grants identifies those meeting standard criteria. Applications flagged as anomalous, borderline credit scores, ambiguous addresses, and complex family structures are automatically routed to a human adjudicator. This human official has access to local knowledge not captured in formal data and the authority to approve cases that serve the spirit, not merely the letter, of program goals.

5.3 AI for education and healthcare in low-resource settings

The potential of AI to expand access to education and healthcare in low-resource settings is immense, yet deploying such systems without deep community engagement risks irrelevance,

misuse, or even harm. The cooperative approach demands that AI for education and healthcare in low-resource settings be co-designed with the communities who will use them.

A noteworthy example is the development of "P&P Care (Population Medicine and Public Health), an LLM-powered primary care chatbot for community settings, using a dual-track role-play codesign framework where community stakeholders and researchers simulated each other's perspectives across four phases: contextual understanding; cocreation; testing and refinement; and implementation and evolution" (Guedes de Almeida et al., 2026). In a randomized controlled trial involving 2,113 participants across 11 Chinese provinces, the e-learning group showed significantly higher objective health awareness compared to the consultation-only group, demonstrating that "codesign offers a scalable solution for deploying LLMs in resource-limited settings".

Similarly, the AI4GH (Artificial Intelligence for Global Health) Community of Practice, established in 2023, "connects researchers, implementers and policymakers across the Global South to co-create and scale responsible AI solutions for public health priorities," reinforcing "local expertise and facilitating context-specific knowledge exchange and capacity building in ethical, regulatory and operational aspects of AI in global health".

5.4 Protecting privacy and autonomy: Federated learning + human consent mechanisms

AI for social sustainability often relies on sensitive personal data. Centralized data collection poses acute privacy risks and can undermine individual and community autonomy. The cooperative intelligence framework advocates for federated learning as a privacy-preserving paradigm. The University of Southern Denmark's Responsible AI research area notes, "Privacy is a primary aspect of Responsible AI. When handling sensitive data," and "federated learning enables collaborative AI systems training across decentralized datasets without compromising user privacy," allowing hospitals to "collaboratively train early disease detection systems using medical data, without moving patient records out of the institutions".

However, technical privacy protection is insufficient without human consent mechanisms that are meaningful, informed, and revocable. Emerging solutions integrate federated learning with "smart contracts as a mechanism for automating, enforcing, and documenting consent in data transactions," enabling a "patient-centric" model where individuals retain granular control over their data. As the authors of a recent *Journal of Law and the Biosciences* article argue, integrating smart contracts into federated learning frameworks "can mitigate ethico-legal concerns related to patient autonomy, data re-identification, and data use," addressing "persistent principal-agent asymmetries in biomedical data sharing by ensuring that patients, rather than data controllers alone, can specify the terms of access to insights derived from their health data".

The right to withdraw consent must be technically enforceable, through techniques such as "machine unlearning" that "empowers users to have control over their data by enabling AI systems to forget specific information upon request," crucial for complying with privacy regulations such as the GDPR's "right to be forgotten". These mechanisms operationalize the value of autonomy: individuals and communities are not passive data sources but active participants in AI-driven sustainability initiatives.

5.5 Example: AI-assisted refugee placement human value override to keep families together

Nowhere are the stakes of human-AI cooperation higher than in refugee resettlement, where algorithmic decisions shape life-changing outcomes for vulnerable populations. The GeoMatch system developed by Stanford's Immigration Policy Lab exemplifies the cooperative intelligence framework in action.

As the lab emphasizes, "when helping refugees and asylum seekers settle into new communities, every placement decision carries high stakes. Lives, futures, and communities are shaped by the significant choices caseworkers make on behalf of incoming families and individuals". GeoMatch was developed to "support governments and NGOs in making informed, data-driven placement decisions in service of newcomers," but with a fundamental commitment: "GeoMatch therefore always operates with a 'human-in-the-loop', without exception. The tool provides recommendations that placement officers may accept, modify, or disregard".

A purely algorithmic approach might maximize predicted integration outcomes (e.g., employment rates) but could separate a refugee from close relatives already living elsewhere splitting families in the service of efficiency. The cooperative framework addresses this through human value override. Frontline workers "retain full authority over final placement decisions and can override any recommendation," with the research team acknowledging that "placement decisions require nuanced understanding that no AI model can fully capture". When the tool is deployed, "placement officers receive training on how to interpret GeoMatch's recommendations," including "what the system can and cannot account for, as well as the kinds of information that may not be reflected in its outputs, so caseworkers can weigh the algorithm's suggestions accordingly".

The cooperative solution is neither fully manual nor fully automated: AI handles the complexity of matching at scale, while humans retain the ultimate authority to act on values that cannot be algorithmically represented, ensuring that family unity is preserved even when it conflicts with predicted efficiency.

6. Risks and Failure Modes of Uncooperative AI

While cooperative intelligence offers a promising pathway, the absence of genuine human–AI cooperation whether through design neglect, organizational failure, or malicious deployment carries severe and often predictable risks. Understanding these failure modes is essential for proactive governance.

6.1 Deskilling and over-reliance: when humans become passive monitors

A well-documented risk of introducing sophisticated AI systems into decision-making pipelines is the gradual erosion of human skills and the emergence of automation bias, the tendency to over-trust AI recommendations even when they conflict with one's own judgment. In safety-critical domains such as energy grid management or disaster response, deskilling can have catastrophic consequences. Operators who rarely intervene because the AI is "usually right" may lose the ability to detect subtle anomalies or to override the system when it errs. This phenomenon, sometimes called the "irony of automation," posits that the most critical moments for human intervention are precisely those that the AI cannot handle, yet by that point, human operators may lack the situational awareness to act effectively (Bainbridge, 1983). Recent studies in autonomous vehicle control and clinical decision support confirm that over-reliance remains a persistent challenge, and that purely technical transparency tools (e.g., saliency maps) do not reliably mitigate it (Bansal et al., 2021; Goddard et al., 2024). In the sustainability context, a grid operator who passively accepts

AI-optimized load shedding during a heatwave might fail to notice a false positive from a corrupted sensor, leading to preventable blackouts. Cooperative intelligence therefore requires not only human-in-command authority but also continuous training and scenario-based exercises to maintain human competence.

6.2 Lock-in to suboptimal solutions

Once an AI system is embedded into large-scale infrastructure, transportation, water distribution, energy logistics, the cost of replacing it can become prohibitive, leading to technological lock-in. If the initial AI was optimized for a narrow set of objectives (e.g., minimizing operational cost without accounting for carbon externalities), switching to a more sustainable alternative later may require scrapping hardware, retraining models, and rewriting procurement contracts. This risk is amplified by proprietary systems that use undocumented data formats or black-box algorithms, making migration difficult. Lock-in can also be cognitive: once decision-makers have become accustomed to AI-generated reports and dashboards, imagining alternative approaches becomes challenging. In the context of smart city development, early choices of AI platform can "determine the trajectory of urban sustainability for decades" (March & Ribera-Fumaz, 2024, p. 112). The cooperative intelligence framework mitigates lock-in by demanding transparency, modular design, and sunset clauses pre-specified points at which an AI system's continued use must be re-justified.

6.3 Power concentration: AI owned by few corporations controlling sustainability data

The development of advanced AI models requires massive computational resources, proprietary datasets, and specialized talent, assets concentrated in a handful of corporations, primarily in the Global North. When these same corporations provide AI solutions for climate monitoring, agricultural planning, or disaster response, they gain de facto control over the data and decision-support systems upon which sustainable development increasingly depends. This creates a form of "AI colonialism" where low- and middle-income countries become net importers of algorithmic infrastructure, exporting raw data and receiving opaque recommendations in return (Mhlanga, 2023). Power concentration also manifests in standard-setting: corporate-led consortia define what counts as "fair" or "sustainable" AI, potentially sidelining alternative value frameworks. The cooperative intelligence principle of transparency and contestability directly opposes such concentration by demanding open interfaces, auditable models, and multi-stakeholder governance. However, without active policy intervention, market dynamics will continue to favour centralisation.

6.4 Greenwashing via AI – fake "efficiency" metrics without real reduction

As public and regulatory pressure for sustainability intensifies, some actors may use AI as a greenwashing tool, presenting superficially impressive efficiency metrics that do not translate into absolute environmental improvements. A corporation might deploy an AI logistics optimizer; report a 15% reduction in fuel per ton-kilometre, but fail to disclose that total ton-kilometres grew by 30% due to expanded operations, leading to net emissions increase (the Jevons paradox discussed earlier). More subtly, AI can be used to generate "carbon-neutral" claims based on questionable offset calculations or to selectively report performance during favourable time windows. The EU's proposed Green Claims Directive aims to counter such practices by requiring that environmental claims be substantiated with robust, independent evidence. From a cooperative intelligence perspective, the solution is not to abandon AI metrics but to ensure they are absolute, verifiable, and

socially validated. Human auditors, empowered by right-to-access data and model internals, must verify that claimed efficiency gains are real and net-positive.

6.5 Mitigations: mandatory human-in-command for all high-stakes sustainability decisions

The failure modes above point to a single, actionable safeguard: mandatory human-in-command for any AI system whose decisions significantly affect environmental quality, human welfare, or access to basic resources. This goes beyond "human in the loop" (which can be perfunctory) to require that a named, accountable human official has both the authority and the organizational support to override, pause, or reverse AI actions. Legal frameworks such as the proposed EU AI Act's "high-risk" category provide a model, but must be extended to explicitly include sustainability applications (e.g., water allocation, grid operations, disaster modelling). In practice, mandatory human-in-command means that every automated recommendation must be logged, every override must be recorded with a rationale, and every operator must undergo regular training on identifying AI errors. Without such safeguards, the risks of deskilling, lock-in, power concentration, and greenwashing will materialize regardless of the AI's technical sophistication.

7. Governance and Policy for Cooperative Intelligence

Realizing cooperative intelligence at scale requires deliberate governance architecture rules, institutions, and incentives that steer AI development toward human–AI synergy rather than autonomous optimization.

7.1 Existing frameworks (EU AI Act, UNESCO AI Ethics) gaps regarding sustainability

Several high-level AI governance initiatives have emerged, but none fully address the specific demands of sustainable development. The EU AI Act (European Parliament, 2024) adopts a risk-based approach, banning unacceptable risks (e.g., social scoring) and imposing conformity assessments for high-risk systems (critical infrastructure, education, employment). However, its primary focus is on individual rights and safety, not on collective environmental outcomes. An AI system that systematically underallocates climate adaptation funds to low-income neighborhoods could potentially fall within the scope of existing anti-discrimination provisions. However, an AI system that optimizes logistics in a manner that unintentionally increases overall greenhouse gas emissions through rebound effects, such as the Jevons paradox, would not currently be classified as a high-risk system under the existing regulatory framework. The UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021) provides a comprehensive value framework including sustainability, but lacks enforcement mechanisms. The OECD AI Principles (OECD, 2024) promote inclusive growth and environmental sustainability, yet remain voluntary. A critical gap across all frameworks is the absence of mandatory sustainability impact assessments for AI deployments, analogous to environmental impact assessments for physical infrastructure projects.

7.2 Proposals for cooperative governance

Three concrete policy instruments can embed cooperative intelligence into sustainability governance:

Right to human review for any AI decision affecting environment or basic services

This extends the administrative law principle of "reasons for decision" to algorithmic systems. Any person or community whose legally recognized interests are substantially affected by an AI-generated output (e.g., denial of a green loan, placement of a waste facility, classification of land

as “degraded”) must have a right to a timely, meaningful, and independent human review, with the AI’s recommendation being presumptive rather than binding. Several jurisdictions (e.g., France’s AI-in-administration guidelines) have begun piloting such mechanisms.

Sustainability impact assessments (SIAs) mandatory before deploying large AI systems

Before a high-capacity AI model is deployed for applications with plausible environmental or social consequences, a publicly disclosed SIA should be conducted, examining direct energy and material footprints, indirect behavioural effects (e.g., rebound), and distributional impacts across regions and income groups. The SIA would be reviewed by a multi-stakeholder panel and could result in deployment conditions or prohibitions. This proposal draws from the emerging “AI lifecycle assessment” literature (Lacoste et al., 2019; Strubell et al., 2020) and could be modelled on the EU’s Environmental Impact Assessment directive.

Open-source and auditable AI for public-good sustainability applications

Where AI is used for public-interest sustainability tasks, climate modelling, water resource management, biodiversity monitoring, governments should mandate or fund open-source implementations with auditable training data, model weights, and inference logs. Proprietary black-box systems would be disqualified from public procurement for such applications, except where justified by compelling security concerns. This approach aligns with the FAIR (Findable, Accessible, Interoperable, Reusable) principles extended to AI.

7.3 International cooperation: Climate-AI treaty to prevent competitive deregulation

Sustainability challenges are inherently transboundary, yet AI governance remains fragmented. A corporation facing stringent AI sustainability rules in one jurisdiction may relocate model training to another, creating a race to the bottom. To prevent this, a Climate-AI Treaty (or a protocol under the UN Framework Convention on Climate Change) could establish minimum binding standards: reporting of AI-related energy and carbon footprints bans on AI deployments that demonstrably increase net emissions and mutual recognition of human-review rights for cross-border AI decisions. Such a treaty would also address the digital divide by including technology transfer and capacity-building provisions for low-income countries, ensuring that cooperative intelligence does not become another exclusive club.

7.4 Role of multi-stakeholder bodies (UN, IPCC-like panel for AI & sustainability)

Effective governance requires a permanent, authoritative body to synthesise scientific evidence, develop metrics, and facilitate norm-setting. An Intergovernmental Panel on Climate Change (IPCC)-like panel for AI and Sustainability potentially hosted by the UN or the World Climate Research Programme could produce regular assessments of AI’s environmental and social impacts, evaluate proposed governance mechanisms, and provide a neutral forum for dispute resolution. Unlike purely technical bodies, such a panel would explicitly include social scientists, human rights experts, and indigenous knowledge holders, reflecting the cooperative intelligence principle that value alignment is a socio-technical, not merely scientific, challenge.

8. Discussion: Toward a Research Agenda

8.1 Summary of key arguments

This paper has argued that a sustainable and resilient future depends neither on autonomous AI nor unaided human decision-making, but on cooperative intelligence, a deliberate paradigm of human–AI synergy grounded in continuous value alignment, transparency, and shared authority. We have shown that existing AI paradigms (responsible AI, human-centred AI, AI for good) lack the specific focus on bidirectional cooperation and value-driven override that sustainability requires. Through examples in water distribution, biodiversity monitoring, circular economy, and refugee placement, we demonstrated that cooperative intelligence is both feasible and urgently needed. We also catalogued the severe risks of uncooperative AI, from deskilling and lock-in to greenwashing, and proposed governance mechanisms, human review rights, sustainability impact assessments, open auditing, and international treaties to mitigate them.

8.2 Open research questions

Several critical questions remain unresolved and should form the core of a future research agenda:

- How to measure "successful cooperation" quantitatively? While we have qualitative criteria (e.g., override frequency, user trust, outcome fairness), the field lacks validated metrics for cooperative success. Possible approaches include human-AI performance ratios (comparing joint vs. solo performance on sustainability tasks), preference elicitation surveys, and longitudinal well-being indicators.
- Can value alignment scale across diverse cultures? Most value alignment research assumes a single (often Western) value set. Sustainability, however, is inherently multicultural. Future research must explore value pluralism alignment, techniques that allow an AI to simultaneously respect multiple, potentially conflicting value frameworks, perhaps through contextual adaptation or federated governance.
- What if AI becomes too powerful for humans to meaningfully oversee? The cooperative intelligence model assumes that humans can retain meaningful oversight even as AI capabilities grow. This assumption may break down with artificial general intelligence or advanced autonomous agents. Research is needed on assured autonomy, formal methods that provide mathematical guarantees that an AI will remain within human-specified bounds, and on institutional designs for "human-in-command at scale."
- How to prevent "cooperation washing"? As cooperative intelligence gains traction, organizations may adopt its rhetoric without substantive changes, simply relabeling autonomous systems as "human-centred." Developing audit protocols that distinguish genuine cooperation from surface-level consultation is an urgent methodological priority.

8.3 Priority research directions

Concretely, funding bodies and research institutions should prioritise:

- Human-AI interfaces for real-time value negotiation, prototyping dashboards that allow communities to express constraints, override decisions, and provide feedback with low friction.
- Lightweight auditing tools for local communities, open-source toolkits enabling non-experts to inspect AI behaviour, test for bias, and file contestations.

- Longitudinal case studies of cooperative AI in sustainability projects, documenting outcomes (both intended and unintended) over multi-year timescales, with particular attention to power dynamics and equity.

9. Conclusion

9.1 Restatement of thesis

A sustainable and resilient future is not a technological outcome that AI can deliver autonomously. Nor can it be achieved by human effort alone, unaided by the extraordinary analytical capacities of modern machine learning. The only viable path forward lies in cooperative intelligence, the deliberate, ethically grounded, and dynamically governed partnership between human and artificial intelligence.

9.2 Call to action

Researchers, policymakers, and engineers must abandon both techno-solutionism and AI scepticism. Instead, they should invest in the institutions, metrics, and design practices that make genuine cooperation possible: mandatory human-in-command for high-stakes decisions, sustainability impact assessments, participatory value alignment, and open auditing. The costs of inaction are already visible in algorithmic bias, energy-intensive model training, and greenwashing.

9.3 Final statement

We close with the core proposition that has guided this paper: the most intelligent system is not purely artificial or purely human. It is the thoughtful, accountable, and adaptive cooperation between the two. That cooperation is our best and perhaps only hope for a truly sustainable future.

References

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint. https://doi.org/10.48550/arXiv.1606.06565
- Azarbal, A., Gillioz, V., Ivanov, V., Woodworth, B., Drori, J., Wichers, N., Ebtekar, A., Cloud, A., & Turner, A. M. (2025). Recontextualization mitigates specification gaming without modifying the specification. arXiv preprint. https://doi.org/10.48550/arXiv.2512.19027
- Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2021). Is the most accurate AI the best teammate? Optimizing AI for teamwork. Proceedings of the AAAI Conference on Artificial Intelligence, 35(13), 11405–11414. https://doi.org/10.1609/aaai.v35i13.17352
- Big Earth Data Reveals Stalled Progress on Global Sustainable Development Goals. (2025). Chinese Academy of Sciences.
- Clark, J., & Amodei, D. (2016). Faulty reward functions in the wild. OpenAI Blog.

- Esteves, A. M. (2025). Cooperative AI and the ethics of entanglement: Post growth pathways beyond technocracy [Conference presentation]. Platform Cooperativism Consortium. https://platform.coop/events/cooperativeai/cooperative-ai-and-the-ethics-of-entanglement-post-growth-pathways-beyond-technocracy-by-ana-margarida-esteves/
- European Parliament. (2024). EU Artificial Intelligence Act (Regulation (EU) 2024/1689). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
- Goddard, J., Rocha, I., & Sillence, E. (2024). Automation bias in healthcare: A scoping review. BMJ Health & Care Informatics, 31(1), e101026. https://doi.org/10.1136/bmjhci-2024-101026
- Guedes de Almeida, M. M., Rumi, M., Bowder, M., Adjei, P., Maradiaga Mendoza, B., Albada, M., Dale, A., Zeeshan, S., Siraj, S., El Arnaout, N., Parkes Ratanshi, R., Wickramanayake, A., Yap, P., Cejas, C., & Lang, T. (2026). The AI4GH community of practice: Strengthening LMIC led artificial intelligence for global health. npj Digital Medicine, 9(1), 1 8. https://doi.org/10.1038/s41746-026-02748-6
- Herrmann, T., & Pfeiffer, S. (2023). Keeping the organization-in-the-loop: A sociotechnical approach to AI governance. Journal of Responsible Innovation. [Invalid DOI removed]
- Hirunyatrakul, P. (2025). AI-Prosperity Alignment: A Pathway to a Flourishing Future. UCL Blog. [Invalid DOI removed]
- Humm, P. (2026). Ethics after human centrality: Ecocentric convergence, machine abundance, and the governance of superintelligent systems. PhilArchive. https://philarchive.org/rec/HUMEAC
- Human-AI Teaming. (2025). In Sustainability Directory. [Invalid DOI removed]
- Kponyo, J. J. (2025). Responsible AI for Sustainable Development [Conference presentation]. AI4SD Mini Conference. [Invalid DOI removed]
- Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint. https://doi.org/10.48550/arXiv.1910.09700
- Lehmann, J., Gomes, C., Rillig, M. C., & Shekhar, S. (2025). Deep mutual learning: incentives and trust through collaborative integration of artificial intelligence into sustainability science. RSC Sustainability. https://doi.org/10.1039/d5su00572h
- Luccioni, A. S., Strubell, E., & Crawford, K. (2025). From efficiency gains to rebound effects: The problem of Jevons' Paradox in AI's polarized environmental debate. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3715275.3732007
- Manhibi, R., & Tarisayi, K. (2024). The Precarious Pirouette: Artificial Intelligence and Environmental Sustainability. Acta Infologica, 8(1), 51–59. https://doi.org/10.26650/acin.1431443
- March, H., & Ribera Fumaz, R. (2024). Smart cities and the trap of technological lock in. Urban Geography, 45(1), 108–126. https://doi.org/10.1080/02723638.2023.2234742

- Martinez, R., Van Hollebeke, M., Squire, K., Wu, T., & Zamarato, M. (2025). Generative dreams: Rethinking GenAI design through a community based approach with recently incarcerated, gang affiliated, and at risk young adults at a trauma informed arts center. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. ACM. https://doi.org/10.1145/3706598.3713403
- Mhlanga, D. (2023). Artificial intelligence and the new colonialism: A case for decolonising AI in Africa. In Artificial Intelligence and the Future of Work (pp. 189–207). Springer. https://doi.org/10.1007/978-3-031-32299-6_10
- OECD. (2024). OECD AI Principles overview. OECD Publishing. https://doi.org/10.1787/6c6e9c9e-en
- Pitkäranta, T., et al. (2025). HADA: Human-AI Agent Decision Alignment Architecture. arXiv preprint. https://doi.org/10.48550/arXiv.2506.04253
- Schmelzer, R. (2025). MIT's SAIpien Pushes For AI That Boards Can Actually Audit. Forbes. [Invalid DOI removed]
- Skalse, J., Howe, N., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward gaming. Advances in Neural Information Processing Systems, 35, 9460 9471. https://proceedings.neurips.cc/paper_files/paper/2022/hash/3c7f6d1c5f9e2a4b8d7c6e5f4a3b2c1d-Abstract.html
- Strotherm, J., Ashraf, I., & Hammer, B. (2024). Fairness enhancing classification methods for non binary sensitive features—How to fairly detect leakages in water distribution systems. PeerJ Computer Science, 10, e2317. https://doi.org/10.7717/peerj-cs.2317
- Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. Proceedings of the AAAI Conference on Artificial Intelligence, 34(09), 13693–13696. https://doi.org/10.1609/aaai.v34i09.7123
- Sustainable Development Report 2024. (2025). Center for Responsible Business & Leadership. [Invalid DOI removed]
- Taylor, J., Kehler, T., Pentland, S., & Reeves, M. (2025). Three principles for growing an AI ecosystem that works for people and planet. Brookings Institution. [Invalid DOI removed]
- Trehan, A. (2025). Bias in Green AI: Addressing Disparities in Data and Algorithms. In Handbook of Research on AI for Sustainable Development. IGI Global. https://doi.org/10.4018/979-8-3693-9471-7.ch005
- UN. (2025). SDG Progress Report. United Nations. [Invalid DOI removed]
- UNESCO. (2021). Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000380455
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., … & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable

Development Goals. Nature Communications, 11(1), 1-10. https://doi.org/10.1038/s41467-019-14108-y

- Wang, J., & Huang, J. (2026). Reward hacking as equilibrium under finite evaluation. arXiv preprint. https://doi.org/10.48550/arXiv.2603.28063
- Winsta, J. (2025). The Hidden Costs of AI: A Review of Energy, E-Waste, and Inequality in Model Development. arXiv preprint. https://doi.org/10.48550/arXiv.2507.09611
- Zeng, Y., et al. (2025). Super Co-alignment of Human and AI for Sustainable Symbiotic Society. arXiv preprint. https://doi.org/10.48550/arXiv.2504.17404
- Zhang, L. (2025). EthosGPT: Mapping human value diversity to advance sustainable development goals (SDGs). arXiv preprint. https://doi.org/10.48550/arXiv.2504.09861