Position: AI Must Become Planet-Centered, Not Just Human-Centered

Anonymous Authors[1]

Abstract

This position paper argues that contemporary AI paradigms are insufficient for supporting complex global goals and introduces Planet-Centered AI (PCAI) as a design philosophy and research agenda that reorients AI toward planetary-scale socio-ecological systems and their long-term trajectories. A planet-centered approach is grounded in systems thinking, treating Earth as an interconnected whole of which humans are part. We diagnose recurring limitations across AI frameworks—many of which remain human-centered—and show why these become especially consequential under current planetary conditions characterized by systemic risk, non-stationarity, and deep uncertainty. We then articulate how PCAI reshapes the AI lifecycle, from problem formulation and model design to evaluation and deployment, by emphasizing alignment with global agendas, developing system-aware AI foundations, trajectory-oriented evaluation, and monitorability. Finally, we advance a falsifiable claim: AI systems optimized without explicit consideration of systemic consequences are more likely to exacerbate systemic instability than to mitigate it.

1. Introduction

Over the past decade, the AI community has developed a range of paradigms to address the ethical, social, and technical risks of AI. Frameworks such as Human-Centered AI, Responsible AI, AI for Social Good/Sustainability, and AI safety have been essential in establishing that AI has far-reaching consequences, and that potential harms and broader impacts should inform algorithmic design and deployment.

Despite this progress, **this position paper argues that these paradigms are insufficient for enabling AI to meaningfully support societies in confronting complex challenges: ML must be reoriented toward a planet-centered paradigm that treats systemic risk, long-term impact, and global goals as first-order design objectives.**

As the world enters what is described as a polycrisis (Lawrence et al., 2024), risks arise not from isolated failures but from coupled social, technological, and environmental systems whose interactions generate self-reinforcing and systemic dynamics. In such settings, feedback loops, nonlinear interactions, and path dependence—where early interventions shape and constrain future outcomes—can amplify risks and lock societies into trajectories that are difficult to reverse (Delannoy et al., 2025; Steffen et al., 2018). Recent work shows that AI is increasingly entangled with polycrisis dynamics through material pathways (e.g., energy use, resource extraction, infrastructure lock-in) and informational pathways (e.g., shaping behavior, accelerating decision cycles, synchronizing systems), with the potential to intensify systemic instability (Creutzig et al., 2022).

It is precisely this systems perspective (Meadows, 2008)—focused on feedbacks, interactions, and trajectories—that remains largely absent from AI frameworks (Kondor et al., 2024). Specifically, we argue that AI methods are poorly suited to supporting planetary challenges, which exhibit properties of "wicked problems" (Rittel & Webber, 1973): e.g. non-stationary nature and feedback-driven dynamics. This mismatch matters because such conditions increasingly characterize high-stakes domains in which AI is deployed, including climate governance, technological regulation, and public policy (Ilcic et al., 2025). Due to this misalignment, AI can generate systemic risks by interacting with, amplifying, or reshaping underlying system dynamics in unintended ways (Schön et al., 2025). Yet frameworks for anticipating and evaluating these systemic effects remain limited, leaving concepts of systemic risk underdeveloped and inconsistently operationalized in AI governance (Carey, 2025; Stahl et al., 2023).

Climate change illustrates why this systems perspective matters: The scale of global warming is driven not only by human emissions, but by reinforcing feedbacks within the Earth system. Feedbacks, such as water-vapor amplification and cloud responses, roughly double to triple the temperature response to anthropogenic greenhouse gas emissions, accounting for much of the 1.2°C of global warming ob-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

served to date (IPCC, 2013). Similar dynamics arise in technological systems (Galaz et al., 2021; Ilcic et al., 2025), where early design choices, deployment incentives, and governance can entrench behaviors that persist even as cumulative harms become evident. In both cases, effective intervention depends on understanding feedbacks, long-term dynamics, and systemic interactions (Stirling, 2010)—dimensions that current AI paradigms struggle to represent.

We propose Planet-Centered AI (PCAI) as a design philosophy and research agenda that complements the limits of human-centered framings. Human-centered approaches rightly focus on protecting individuals from harm, but often do not consider environmental risks, long-term dynamics, and systemic effects. Planet-centered approaches instead recognize these as constitutive of human and planetary futures. PCAI expands responsibility beyond users, communities and societies to include ecosystems, systemic risks, and Earth-system dynamics, reframing intelligence as a tool for collective understanding and planetary stewardship.

2. The Limits of Contemporary AI Paradigms in the Anthropocene

Across the AI ethics landscape, a strong commitment to protecting humanity has emerged (Jobin et al., 2019). Frameworks such as Human-Centered AI (HCAI) (Shneiderman, 2020) and Responsible AI have expanded research directions beyond narrow definitions of performance (Schmager et al., 2025), introducing desiderata such as explainability, human oversight, robustness, and fairness, as well as a move toward human augmentation. We argue, however, that these paradigms remain insufficient in the Anthropocene (Creutzig et al., 2022). The Anthropocene denotes the geological era in which human activity—increasingly mediated by large-scale technology—has become the dominant force shaping Earth systems. It is characterized by global interconnectedness, nonlinear change, tightly coupled social–ecological dynamics, and amplified systemic risk—dynamics increasingly accelerated by AI (Delannoy et al., 2025; Galaz et al., 2021).

2.1. Wicked problems in the Anthropocene

A central source of AI's limitations lies in the distinction between *tame* and *wicked* problems, originally introduced to explain why scientific and engineering approaches often fail in complex social and policy domains (Rittel & Webber, 1973). Tame problems—such as puzzles or well-defined optimization tasks—have stable objectives, agreed-upon problem formulations, and objective criteria for success. Even when technically complex, they can be decomposed, optimized, and evaluated against fixed goals.

Wicked problems exhibit properties that violate these assumptions[1] (Peters, 2017): Objectives are contested and non-stationary; interventions alter system dynamics; effects propagate across domains; and outcomes unfold over long, uncertain horizons. There is no well-defined global optimum, no safe regime for trial-and-error learning, and no reliable evaluation of success. Examples include climate change mitigation and adaptation, biodiversity conservation, sustainability transitions, and poverty reduction, where interventions interact with social, economic, and ecological systems (Toyama, 2010). Problems are further characterized by deep uncertainty, where key elements of the system—causal structure, feedbacks, objectives, or future conditions—are unknown, contested, or not reliably quantifiable (Marchau et al., 2019). Anthropocene challenges specifically, are frequently described as *super-wicked* because they intensify these features: decisions are high-stakes and potentially irreversible, feedbacks amplify over time, and action must occur under deep uncertainty and time pressure.

2.2. AI Failure Mechanisms in the Anthropocene

We argue that the properties of Anthropocene's wicked challenges stand in tension with the assumptions in AI, and this mismatch drives the unintended and systemic consequences of AI. Next, we examine existing frameworks and practices, highlighting the mechanisms for AI failure in wicked systems.

2.2.1. Technical misalignment

Technical misalignment refers to the failure that arises when AI—built around assumptions of fixed objectives, stationarity, and decomposability—is deployed in wicked systems whose dynamics violate these conditions (Figure 1). Standard AI pipelines e.g. typically assume that objectives can be specified in advance, that historical data provides a reliable basis for learning, and that interventions do not alter the data-generating process. These assumptions enable optimization, benchmarking, and iterative improvement in tame domains. Some real-world sustainability problems can however be treated as tame: coordinating the orientation of wind turbines within a wind farm to maximize energy output involves a clear objective, stable system boundaries, and objective evaluation criteria (Howland et al., 2022).

When similar methods are applied to wicked challenges, however, this technical misalignment is crucial. Deploying AI-based biodiversity monitoring for conservation operates within entangled social–ecological systems where objectives are contested and boundaries are porous. Such interventions can reshape behavior, altering land use, enforcement practices, conflict dynamics, and surveillance

[1] Appendix A provides a way to diagnose the wickedness of a problem, together with a list of wicked system properties and how AI's assumptions violate these.

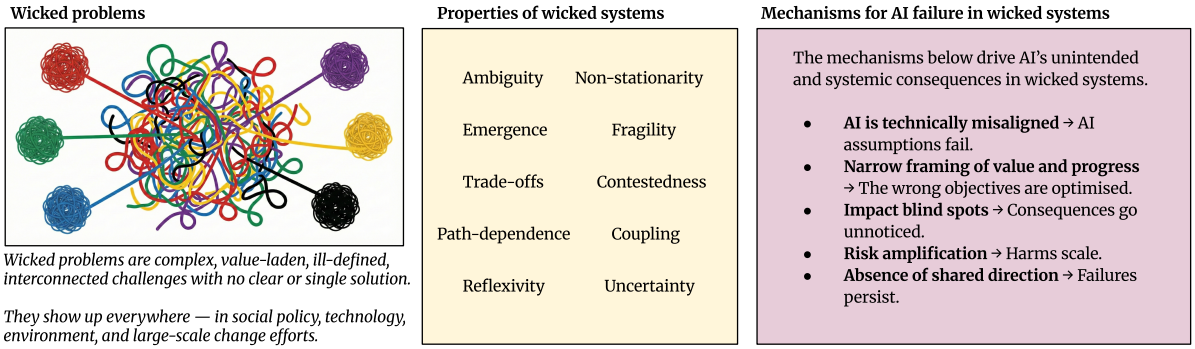

Figure 1. Wicked problems, their structural sources of difficulty, and the mechanisms of AI failure.

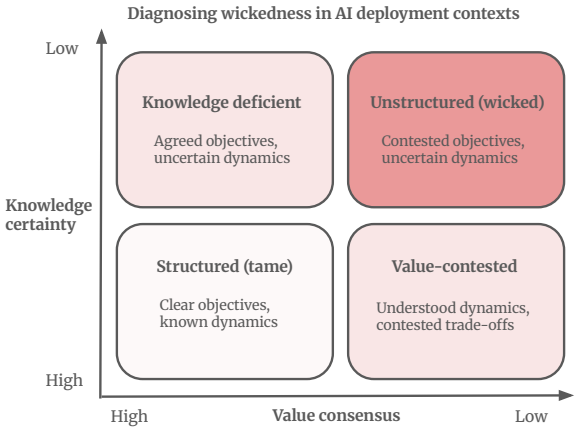

Figure 2. Wicked problem classification adapted from Hisschemöller & Hoppe, 1996. Problems vary in the degree of wickedness along two dimensions: whether the system's causal dynamics are understood (knowledge certainty) and whether there is agreement on objectives and trade-offs (value consensus). As wickedness increases, more stages of the PCAI lifecycle (Figure 3) become relevant. Wickedness is further amplified by coupling to adjacent socio-ecological systems, which can introduce additional uncertainties and contestations in any quadrant. See Appendix A for worked examples and a detailed diagnostic.

relationships, while introducing privacy risks (Duffy et al., 2019; Sandbrook et al., 2021).

Generally, wicked systems lack stable objectives, exhibit deep uncertainty, and evolve in response to intervention (Marchau et al., 2019). This mismatch between AI's technical assumptions and the properties of wicked systems constitutes a failure mechanism: even well-intentioned AI deployments can mis-steer system trajectories because the learning problem becomes ill-posed.

2.2.2. Narrow Framing of Value and Progress

Notions of progress are implicitly encoded in AI: Technological progress is often measured through gains in efficiency, accuracy, scale, speed, or generality (Pansera & Fressoli, 2021; Birhane et al., 2022). Benchmarks reward performance improvements, while deployment prioritizes systems that reduce costs, increase throughput, or enable growth (LaCroix & Luccioni, 2025). Although treated as neutral, these metrics embed a specific conception of progress—one that equates advancement with optimization and expansion. Empirical analyses show that influential AI research overwhelmingly prioritizes quantitative performance gains, while explicit articulation of societal benefits or harms remains rare (Birhane et al., 2022).

This framing aligns closely with market-driven and modernization theories of development, in which productivity and growth are treated as proxies for social value (Pansera & Fressoli, 2021). Within coupled socio-ecological systems, such assumptions are problematic: AI systems may register as successful according to dominant metrics while contributing to inequality, fragility or environmental degradation (Schön et al., 2025). Alternative development perspectives instead emphasize distributional fairness outcomes, resilience, human capabilities such as political voice or access to education, and compatibility with ecological limits (Pansera & Fressoli, 2021; Kallis et al., 2025)—criteria largely absent from AI evaluation practices.

Social media recommendation illustrates this narrow framing concretely. These systems define progress through engagement metrics—clicks, shares, dwell time—and succeed by those measures. Yet engagement is a proxy that embeds a specific conception of user value: it equates what users interact with to what they want. A preregistered algorithmic audit of Twitter/X found that the platform's engagement-based ranking algorithm amplifies emotionally charged and out-group hostile content, and that users do not prefer the political content selected by the algorithm—engagement optimisation underperforms in satisfying users' own stated preferences (Milli et al., 2025). The system registers as successful according to its dominant metric while actively degrading the outcome it is meant to serve. Moreover, the

feedback loop between algorithmic amplification and user behaviour reshapes the information environment over time, producing system-level consequences—polarisation, erosion of shared epistemic ground—that lie entirely outside the evaluation boundary. This is a structural consequence of equating progress with a narrow, optimisable signal.

At the same time, many AI governance frameworks articulate high-level value aspirations—such as human-centeredness or social good—without specifying how these should be operationalized (Jobin et al., 2019; Whittlestone et al., 2019). This underspecification ultimately defers difficult questions about trade-offs and responsibility (Mittelstadt, 2019). Core concepts such as "the human" or "the good" remain ambiguous: humans may be implicitly treated as users, consumers, workers, or abstract fairness categories, despite the fact that these roles entail incompatible values and consequences (Bucknall & Dori-Hacohen, 2022; Selbst et al., 2019). As a result, moral aspirations are translated into technical objectives through problem formulations, proxies, and metrics that embed implicit value trade-offs without sustained scrutiny (Birhane et al., 2022; LaCroix & Luccioni, 2025). In wicked systems, this narrow and underspecified framing of value obscures broader social and ecological consequences (Whittlestone et al., 2019; Stahl et al., 2023).

2.2.3. Impact Blind Spots in Entangled Systems

In fact, a consistent finding across AI governance and ethics is that impact assessment remains narrowly scoped and weakly integrated across the AI lifecycle (Stahl et al., 2023; UNESCO, 2023). Evaluation is typically conducted at the level of tasks, models, or deployment settings, with limited consideration of downstream effects, cross-sector interactions, or long-term consequences (Ahlborg et al., 2019). Systematic reviews highlight the absence of methods for anticipating indirect, cumulative, long-horizon, and intergenerational effects, even as risk assessment research emphasizes the importance of accounting for these (Stahl et al., 2023; Kondor et al., 2024). These limitations reflect a deeper assumption of decomposability in AI design: the idea that interventions can be isolated, optimized, and evaluated independently of broader system dynamics. In coupled socio-ecological systems, this assumption produces systematic impact blind spots (Schön et al., 2025).

Anthropocentric framing constitutes a particularly important form of such impact blind spots. Prevailing AI governance largely confines ethical concern to humans and social systems (Rigley et al., 2023). Empirical analyses show that only a small fraction of AI ethics guidelines—typically between 16% and 26%—explicitly address non-human life, environmental sustainability, or ecological systems (Sebestyén, 2025; Rigley et al., 2023; Jobin et al., 2019). Where these concerns appear, non-human entities and planetary processes are typically treated as externalities, valued primarily for their instrumental role in human well-being.

However, harms such as biodiversity loss and the destabilization of life-support systems unfold gradually, interact with other stressors, and manifest over long timescales (Rigley et al., 2023; Bucknall & Dori-Hacohen, 2022). While such impacts may not register as direct or near-term risks now, they accumulate and compound, constraining the conditions for both human and non-human communities to persist and flourish (Bucknall & Dori-Hacohen, 2022). Analyses of systemic environmental risks of AI emphasize that frameworks focused on direct, human-centered harms risk underestimating the most consequential forms of risk in socio-ecological systems (Schön et al., 2025).

2.2.4. Risk Amplification

Technical misalignment, obscured by impact blind spots and reinforced by narrow framings, does not merely limit AI's effectiveness—it can actively amplify systemic risk (Schön et al., 2025): A clear example of this mechanism is provided by rebound effects in sustainability. AI for Sustainability has shown that AI can improve efficiency in tasks related to energy, agriculture, and transport (Gohr et al., 2025). However, these gains are typically evaluated within narrow system boundaries that exclude behavioral, economic, and institutional responses. As a result, efficiency improvements can lower costs, accelerate adoption, and expand overall system activity, offsetting—or even reversing—environmental benefits (Wright et al., 2025; Mhlanga, 2025). Autonomous vehicles illustrate this mechanism. Autonomous driving is designed to improve safety and reduce per-mile emissions through optimised routing and driving efficiency, objectives that are human-centered and environmentally motivated. At the vehicle level, autonomy introduces an average 21% decrease in operational emissions through improved fuel economy (Onat et al., 2023). However, by reducing the perceived cost and inconvenience of travel, autonomous vehicles stimulate additional demand—through longer commutes, modal shifts away from public transit, new trips by previously underserved populations, and empty vehicle repositioning miles. Estimates of induced travel demand range from 2% to 47% increases in household vehicle miles travelled (Taiebat et al., 2019). When the full lifecycle is considered (including manufacturing, increased vehicle use, and infrastructure expansion) autonomous electric vehicles may emit approximately 8% more greenhouse gas emissions than their non-autonomous counterparts (Onat et al., 2023). The AI system succeeds on the metrics it is evaluated against, yet the system-level outcome is increased total emissions: a paradigmatic rebound effect in which narrow optimisation amplifies the risk it was designed to reduce.

This pattern is consistent with a broader empirical regularity

documented across decades of technology-for-development research: technology's impact is multiplicative rather than additive with respect to existing conditions, amplifying both positive and negative dynamics in the systems it enters (Toyama, 2011). Where underlying inequalities, misaligned incentives, or resource asymmetries are present, even well-designed technology tends to widen gaps rather than close them — a finding replicated across telecenters, educational computing, and mobile services in both developing and developed contexts (Toyama, 2015).

Similar dynamics appear in Sustainable AI, where reducing the carbon footprint of models does not account for downstream effects such as wider deployment, increased demand, or infrastructure expansion (Mhlanga, 2025). At scale, these feedbacks can lock systems into trajectories that intensify resource use, reinforce dependence on existing infrastructures, and constrain future options. From a strong sustainability perspective (Neumayer, 2010), these dynamics are particularly concerning because certain ecological functions are subject to absolute limits and cannot be substituted through efficiency alone. Thus, when AI interventions focus narrowly on optimization without accounting for rebound effects, lock-in, and system-scale feedbacks, they risk amplifying rather than mitigating planetary risk (Wright et al., 2025).

Beyond rebound effects, AI deployment introduces additional amplification pathways: i) its pervasiveness across critical infrastructures; ii) a pace and scale that outstrip regulatory capacity; iii) its technical opacity, which limits democratic oversight; and iv) propagation risks, where reliance on the same models and datasets allows localized failures to cascade (Galaz et al., 2021).

2.2.5. Absence of Shared Global Direction

Beyond technical and evaluative failures, a deeper structural limitation lies in the absence of a widely shared agenda guiding AI research, deployment, and evaluation (Carey, 2025; Whittlestone et al., 2019; Hagendorff, 2020). In practice, problem selection is shaped by data availability, benchmarkability, and short-term incentives, favoring domains that integrate smoothly into existing pipelines (Gohr et al., 2025). In environmental applications, this is reflected in the dominance of satellite imagery, while less observable but ecologically critical processes—soil health or biodiversity interactions—remain underexplored (Gohr et al., 2025).

ADD material on leverage points! Citing Earth4All policies?

More broadly, a systematic review of 792 articles at the intersection of AI and the SDGs finds that very few studies effectively bridge advanced AI methodologies with deep sustainability expertise, and that the literature remains fragmented into disciplinary silos dominated by forecasting and system optimisation (Gohr et al., 2025). This confirms that the absence of shared direction is not merely a governance gap but a structural feature of the current research landscape.

As a result, AI systems tend to model what is observable and measurable, rather than supporting identifying and acting on the most consequential leverage points in socio-ecological systems. Unlike domains guided by shared global aspirations—such as the Sustainable Development Goals—AI development remains fragmented across sectors and jurisdictions. Existing international efforts (e.g. Global Digital Compact and emerging UN-level AI governance) acknowledge the need for alignment but offer limited guidance on research priorities, acceptable trade-offs, or deployment boundaries (UNESCO, 2023). In the absence of direction, AI development defaults to optimizable objectives, reinforcing the persistence of misalignment, risk amplification, and narrow value framing. This represents a missed opportunity to orient AI research toward the complex, high-stakes challenges of the Anthropocene (Creutzig et al., 2022).

3. Planet-Centered AI

Our previous analysis showed that AI paradigms struggle in Anthropocene contexts. These limitations indicate that existing notions of responsible AI are insufficient under current planetary conditions. Planet-Centered AI (PCAI) is proposed as a response: **PCAI is a design philosophy and research agenda that integrates systems thinking and ecological responsibility across the AI lifecycle, aligning AI development with the demands of planetary-scale challenges.** By ecological responsibility, PCAI means treating the integrity, resilience, and limits of ecological systems as first-class design constraints—rather than as externalities inferred indirectly through human outcomes. Rather than asking how AI can become more capable, efficient or ethical in isolation, PCAI poses a different orienting question: ***How can AI support societies in understanding, navigating, and responsibly shaping Earth-system futures?***. Figure 3 summarizes PCAI design principles. The remainder of this section shows how these principles reshape the AI lifecycle across: (i) foundational research priorities, and (ii) applied design requirements for real-world AI systems in planetary contexts. Importantly, PCAI does not imply uniform requirements across all research; similarly than with HCAI, the depth of responsibility should scale with the potential for irreversible or systemic harm (Shneiderman, 2020).

3.1. Problem Setting & Diagnosis

Given the urgency and complexity of planetary challenges, **under PCAI both foundational and applied ML research should be oriented toward shared goals**, such as those articulated in national or international agendas. This re-

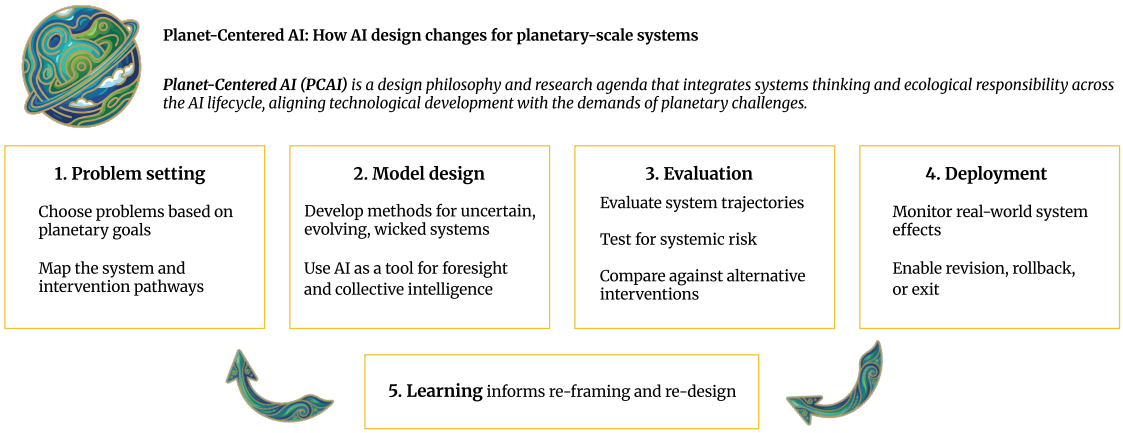

Figure 3. PCAI key design commitments across the lifecycle. Each stage emphasizes anticipating system effects and trajectories.

quires an expert-led assessment of where AI can provide the greatest leverage relative to alternative interventions, what solutions may multisolve challenges[2], and which technical constraints prevent AI from contributing meaningfully. These constraints may be methodological (e.g., uncertainty quantification, causal reasoning, human–AI interaction), transdisciplinary (e.g., system-level evaluation frameworks that capture cross-domain effects), or pre-conditional for applied science (e.g., missing data, inadequate monitoring infrastructure, weak interfaces through which model outputs influence decisions). Such assessments will guide application choice, but also, importantly, shape technological development agendas, directing research toward the foundations for AI to contribute meaningfully to planetary goals.

For planetary goals, **PCAI introduces system mapping and theories of change as design preconditions for applied AI research and real-world deployment**[3]. PCAI requires AI interventions to be situated within the complex systems in which they operate. System mapping makes explicit the relevant system boundaries and dynamics, decision-makers, affected communities, and temporal horizons over which impacts unfold—supporting a diagnosis of which wicked characteristics the problem exhibits. This mapping reduces the risk that consequential dynamics are omitted during problem formulation and evaluation. PCAI requires that criteria for success be articulated *before* modeling begins, grounding optimization targets and evaluation metrics in system-level effects. Researchers are expected to document a causal account of how model output should influence decisions and system dynamics—that is, a theory of change. This will identify intended leverage points, plausible feedbacks, and potential failure modes—such as behavioral adaptation or rebound effects—that could undermine intended benefits. This approach mirrors anticipatory reasoning in public policy design and provides a basis for evaluating the wickedness of the problem (Peters, 2017).

[2]For example, Project Drawdown (Hawken, 2017)—one of the most comprehensive, evidence-based assessments of climate mitigation strategies—ranks educating girls among the most effective interventions for reducing global greenhouse gas emissions, highlighting the impact of social factors in environmental challenges.

[3]We illustrate this framing with a conservation use case in Appendix C.

3.2. Model Design

PCAI strengthens systems-aware technical foundations. PCAI aims to reorient AI research in light of the failure mechanisms identified in Section 2. Many relevant challenges are already recognized within the AI community, but are typically addressed in isolation or without explicit attention to system dynamics. Consider non-stationarity. While work on continual learning and distribution shift focuses on maintaining model performance as data changes, planetary systems often exhibit endogenous change, abrupt regime shifts, and tipping points that invalidate assumptions of gradual or reversible dynamics. Under PCAI, non-stationarity therefore raises questions not only of adaptation, but of evaluation: models must be stress-tested against plausible alternative system regimes and structural breaks (Beucler et al., 2024), rather than optimized for a single expected distribution. This shift motivates greater emphasis on distributionally robust and minimax-regret formulations that aim to avoid catastrophic failure under deep uncertainty, rather than maximizing average-case performance. Uncertainty provides a second example. AI focuses on aleatoric and epistemic uncertainty, yet planetary contexts are frequently characterized by deep uncertainty. Under PCAI, identifying, communicating, and reasoning under such uncertainty becomes a core technical requirement. Related research on open-endedness begins to address novelty and unanticipated conditions (Stanley, 2019), but remains underdeveloped as a foundation for high-stakes, systemic intervention. At the foundational level, PCAI thus motivates a research agenda that prioritizes robustness, adaptability,

and uncertainty-aware reasoning. At the applied level, researchers are expected to draw on prior diagnosis of system wickedness to inform model selection, training objectives, evaluation protocols, and deployment strategies.

PCAI emphasizes AI systems as epistemic infrastructure. Given the properties of wicked systems, automated decision-making is brittle, and prediction—while useful for short-horizon forecasting—can exacerbate the failures identified in Section 2 by collapsing deep uncertainty into a single expected projection. In complex systems, such overconfident forecasts obscure alternative futures and intervention pathways (Amoore, 2023; Pérez-Ortiz, 2024; Søgaard Jørgensen et al., 2024). Instead, PCAI builds on a longstanding tradition of using computation to augment human reasoning in complex systems (Meadows et al., 1972; Selin et al., 2023; Van Beek et al., 2020)—through simulation, exploratory modeling, and scenario analysis (Lavin et al., 2021)—reframing AI's role from prediction and control toward foresight (Geurts et al., 2022; Bankes, 1993). The goal of foresight is epistemic: to generate understanding about how complex systems behave, evolve, and respond to intervention (Selin et al., 2023), rather than to optimize or control outcomes. In this role, AI supports sensemaking rather than automation, helping humans reason about robustness, trade-offs, and risks. Foresight must then empower human agency, translating exploration into interpretable insights[4] that guide collective reasoning and action. Human–AI collaboration is therefore central. Interfaces and workflows are oriented toward deliberation, contestation, and coordinated judgment. A concrete instantiation of this epistemic role is the use of world models or digital twins—simulation-based representations that combine data-driven learning with domain knowledge to explore system behavior under alternative interventions, trajectories and scenarios (Rudd-Jones et al., 2025). Their value lies not in identifying the "best" policy, but in revealing fragile trajectories and conditions under which intended interventions could fail catastrophically.

Integrated Assessment Models (IAMs) illustrate both the need for and the feasibility of this reorientation. IAMs have long supported climate governance by linking physical, economic, and social dynamics to explore transformation pathways (Van Beek et al., 2020). Yet most still rely on opaque optimisation solvers that converge on a single expected trajectory, collapsing the deep uncertainty and contested values that characterise climate governance into a fixed objective. Emerging work is shifting IAMs toward foresight-oriented architectures: interpretable multi-agent reinforcement learning enables the exploration of cooperative strategies across heterogeneous actors with competing interests (Rudd-Jones et al., 2025), while mixture-of-experts frameworks couple IAMs with agent-based and Earth-system models to test policy robustness across scales and structural assumptions (Filatova et al., 2025). These developments exemplify the PCAI vision of AI as epistemic infrastructure: computation that generates understanding about how systems behave under alternative interventions, rather than optimising toward a single projected outcome (Pérez-Ortiz, 2024).

[4]These insights could include trends, emerging risks, trade-offs, black swans, gray rhinos, tipping points, value mappings, ethical preferences, causal loops, cross-impact relations, and other systemic features that deepen understanding, support reflection, and open pathways for responsible action (Selin et al., 2023).

3.3. Evaluation

PCAI reframes evaluation as a tool for anticipating system-level consequences rather than ranking models. While standard metrics remain useful for task performance, they are insufficient for understanding how AI shapes system behavior once deployed in wicked contexts. PCAI therefore adopts an umbrella evaluation approach that complements task metrics with analyses of trade-offs, stability, and systemic risk. Rather than collapsing performance into a single score, PCAI emphasizes evaluation practices that make competing objectives explicit across possible system trajectories. Pareto frontiers are used to surface tensions between efficiency, equity, resilience, or environmental impact, supporting transparent deliberation over trade-offs. We refer to this approach as trajectory-oriented evaluation: these trade-offs may help assess how models may shape integrative system trajectories over time.

PCAI introduces systemic risk probes. Evaluation explicitly tests for amplification mechanisms e.g. rebound effects and correlated failures. Where possible, simulation-based analyses (Guliyeva et al., 2025) are used to explore how deployment may alter broader system dynamics.

PCAI encourages counterfactual baselines. Models are compared against state-of-the-art, but also against no-AI baselines, simple heuristics, or alternative non-ML interventions. This makes opportunity costs visible and avoids justifying deployment solely on benchmark gains.

3.4. Deployment

PCAI reframes deployment around monitorability. Deployment is understood as a sustained intervention in an evolving system. Monitoring therefore extends beyond model-level signals—such as prediction error or data drift—to encompass system-level responses, including behavioral adaptation, rebound dynamics, early warning signals of emerging patterns (e.g., black swans or gray rhinos), and distributional effects that may indicate the reinforcement of fragile trajectories.

PCAI treats deployed systems as revisable. Continued operation is provisional and contingent on observed system-level impacts. Deployment includes predefined escalation,

modification, and rollback pathways, triggered by monitored risk indicators. This requires that system boundaries, assumptions, contexts of use, and intervention pathways are specified in advance, so that observed changes can be attributed, contested, and acted upon.

3.5. A Falsifiable Claim

Recent work on Anthropocene traps (Søgaard Jørgensen et al., 2024) characterizes many contemporary crises as self-reinforcing trajectories in which short-term gains or delayed feedbacks erode long-term system resilience (Steffen et al., 2018). These traps—such as growth dependence, infrastructure lock-in, and rebound dynamics—are not caused by any single technology, but are frequently intensified by technologies that accelerate scale, efficiency, or coordination without attention to system-level feedbacks. In this context, technology functions as a powerful modulator of system trajectories, capable of stabilizing or destabilizing social–ecological systems. Against this background, PCAI sets a contestable and empirically examinable hypothesis:

> *In domains governed by wicked dynamics, AI systems optimized for efficiency or narrow objectives—without explicit consideration of systemic feedbacks and long-horizon effects—are more likely to exacerbate systemic instability than to mitigate it.*

This is, AI systems optimized for speed, scale, or narrow task performance, can reshape system dynamics, accelerating feedback loops, triggering rebound effects, and reinforcing Anthropocene traps that mask accumulating risk while narrowing future options (Søgaard Jørgensen et al., 2024). Apparent short-term improvements may coincide with declining system resilience and shrinking safe operating space. This claim is falsifiable. It predicts that AI systems designed and evaluated under PCAI principles should exhibit measurably different system-level effects than conventional deployments. This claim should be empirically testable through stress-testing and systems monitoring: evidence that PCAI-aligned systems reduce rebound effects or expand the safe operating space of a system would support the paradigm; evidence that they do not would falsify it.

4. Alternative Views

4.1. Human-Centered AI (HCAI) as Sufficient

A common view holds that HCAI provides a sufficient framework for addressing planetary concerns, since environmental instability ultimately harms humans. From this perspective, the challenge is not anthropocentric framing but incomplete implementation, such as extending temporal horizons or improving impact assessment. PCAI agrees that human well-being depends on planetary stability, but argues that anthropocentric framings treat environmental and systemic risks only indirectly, as downstream proxies for human harm, which can delay risk recognition and weaken responses to emerging dynamics. Appendix B compares PCAI to HCAI along different relevant dimensions.

Influential assessments of AI's relationship to the SDGs illustrate this limitation. Vinuesa et al. (2020) find that AI may act as an enabler on 79% of SDG targets but may inhibit 35%, yet the authors themselves acknowledge that the interactions between targets—where progress on one may undermine another—remain poorly characterised, and that "novel methodologies are required to ensure that the impact of new technologies are assessed from the points of view of efficiency, ethics, and sustainability." PCAI responds to this call by providing the system-level reasoning that target-by-target assessment cannot capture.

4.2. AI Safety as Dominant Risk Framework

A second view argues that AI safety and alignment research already addresses long-term and catastrophic risks, rendering additional planetary framing unnecessary. AI safety has indeed developed powerful tools for analyzing misalignment, loss of control, and other AI internal failure modes. PCAI agrees that this work is essential, but argues that it targets a different class of risks. AI safety focuses primarily on whether AI systems pursue intended objectives and remain controllable. By contrast, many relevant risks in the Anthropocene arise from the deployment of AI within socio-ecological systems, where even well-aligned systems can amplify existing crises. From a planetary perspective, the central concern is thus how AI reshapes system dynamics over time. PCAI therefore complements AI safety by shifting attention from internal alignment to system-level embedding: safety asks whether AI systems behave as intended, while PCAI asks whether those intentions, when enacted at scale, contribute to stable, resilient, and sustainable planetary trajectories (Steffen et al., 2018).

4.3. Systemic Reasoning Outside the Scope of AI

A related objection holds that planetary-scale dynamics, system mapping, and theories of change lie outside the scope of AI, belonging instead to policy or Earth system science. From this view, AI should focus on general-purpose tools, leaving system-level reasoning to downstream users, with incremental improvements—such as better benchmarks, audits, or regulation—seen as sufficient. PCAI does not claim that AI researchers should model entire planetary systems or replace policy judgment. Rather, it argues that AI design choices inevitably encode assumptions about the system modeled itself. When left implicit, these assumptions can lead to failure modes which amplify risk. System diagnosis

Table 1. Call to Action: Operationalizing PCAI across the AI ecosystem

Actor	Concrete Actions
Foundational ML Researchers	**Develop epistemic AI for foresight.** Advance methods that support collective sensemaking, e.g. exploratory modeling, scenario generation, stress-testing or world models. **Build AI paradigms for evolving complex systems.** Prioritize learning under non-stationarity, regime shifts, and deep uncertainty, including robustness to structural change. **Advance hybrid and collective intelligence.** Design human–AI interaction paradigms that support deliberation, co-creation, and coordinated judgment.
Applied ML Researchers	**Mandate explicit system mapping.** Situate AI interventions within complex systems. **Require explicit theories of change.** Articulate how outputs influence system dynamics. **Demonstrate comparative value to alternatives.** Specify opportunity costs and risks.
Research Institutions & Funders	**Institutionalize planetary agenda-setting.** Translate existing international and national goals into AI-relevant research priorities through expert-led, interdisciplinary panels. **Tie funding to agenda.** Require proposals to specify which targets within planetary goals they impact, why AI is an appropriate lever, and what technical constraints exist. **Support high-uncertainty, long-horizon planetary research.** Fund work with exploratory outcomes, long validation cycles, and non-benchmarkable success criteria. **Require comparative and counterfactual baselines.** Make funding contingent on comparisons to no-AI and alternative interventions to avoid default technological solutionism.
Conferences, Journals, and the ML Community	**Redefine scientific contribution.** Recognize system-level evaluation, robustness analysis, scenario stress-testing, and systemic risk probes as first-class research outputs. **Encourage transparency of assumptions and impacts.** Require reporting of system boundaries, theories of change, and anticipated downstream effects. **Create durable venues for planetary reasoning.** Establish tracks, review criteria, and workshops for AI-augmented foresight, long-term trajectories, and planetary stewardship. **Extend AI ethics standards to planetary responsibility.** Evolve ethical frameworks to include ecological impacts, planetary limits, and systemic risk. **Reduce reliance on single-score rankings.** Encourage Pareto-front, trade-off, and robustness-oriented evaluation. **Standardize system mapping artifacts,** e.g. developing an "Impact Datasheet".
Education and Training	**Treat systems reasoning as a core ML competency.** Integrate complex systems dynamics and systems impact assessment into AI curricula. **Teach system mapping for AI practice.** Train students and practitioners to map system boundaries, feedbacks, and causal pathways linking model outputs to systemic effects.
Governance and Deployment Contexts	**Extend existing impact assessment regimes.** Embed systemic risk analysis into lifecycle governance processes and develop tools for failure assessment. **Mandate system-level monitoring and independent verification.** Require observation of behavioral, ecological, and institutional responses to deployment. **Treat deployment as provisional and revisable.** Require predefined escalation, rollback, and decommissioning triggers tied to systemic risk indicators.

is therefore set as a design precondition for applied science rather than a modeling task: it constrains objective specification, evaluation, and deployment, often in collaboration with domain experts. In this sense, PCAI shifts the burden from individual researchers to interdisciplinary processes, motivating concrete technical commitments.

Field evidence supports this position. A recent review of 25 GenAI deployments across LMICs found that a common thread was the shift from a tech-solutionist paradigm to a more socio-technical approach, as "problems tended to be nested, with one issue revealing others, requiring adaptability and the application of a holistic perspective" (Adams et al., 2026). These practitioners—working in health, agriculture, education, and gender-based violence—did not set out to do system mapping, but discovered that deployment in wicked contexts demanded it. This suggests that systemic reasoning is not an optional add-on to AI design but an

emergent requirement of deployment in complex settings.

4.4. Technological Progress as Primary Solution

A final view holds that technological innovation will mitigate planetary harm through decoupling, whereby efficiency gains allow growth without increasing environmental impact. PCAI acknowledges real efficiency gains enabled by AI, but argues that in coupled systems these gains often trigger rebound effects, scale expansion, and lock-in, amplifying systemic risk rather than reducing it. It is therefore crucial for AI paradigms to consider these effects.

Empirical evidence challenges this assumption. Decades of research on information technology for development have shown that technology acts as a multiplier on pre-existing conditions: it accelerates progress where the underlying dynamics are already favourable, but widens disparities where they are not (Toyama, 2011). The US poverty rate, for instance, has not declined since 1970 despite four decades of intensive digital innovation (Toyama, 2010). Scaling this observation to AI, PCAI argues that without deliberate reorientation of objectives and evaluation, more capable AI systems will amplify the trajectories societies are already on — including unsustainable ones.

5. Call to Action and the Path to PCAI

PCAI calls for action across the AI ecosystem—spanning research practice and incentives, evaluation norms, and governance—to better align AI development with planetary challenges. PCAI is intentionally ambitious. It does not offer a complete solution, nor does it claim that AI could fully model or control planetary systems. Instead, it delineates an initial scope for a long-term research agenda—one that may unfold over decades—aimed at equipping societies with AI and impact assessment tools that support foresight, deliberation, and responsible intervention in complex challenges. In this sense, PCAI invites the AI community to contribute its distinctive technical expertise to the defining challenges of the Anthropocene. Table 1 summarizes the core commitments of PCAI and outlines how they translate into concrete shifts across the AI lifecycle.

References

Ahlborg, H., Ruiz-Mercado, I., Molander, S., and Masera, O. Bringing technology into social-ecological systems research—motivations for a socio-technical-ecological systems approach. *Sustainability*, 11(7):2009, 2019.

Alford, J. and Head, B. W. Wicked and less wicked problems: a typology and a contingency framework. *Policy and society*, 36(3):397–413, 2017.

Amoore, L. Machine learning political orders. *Review of International Studies*, 49(1):20–36, 2023.

Bankes, S. Exploratory modeling for policy analysis. *Operations research*, 41(3):435–449, 1993.

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'Gorman, P. A., et al. Climate-invariant machine learning. *Science Advances*, 10(6):eadj7250, 2024.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 173–184, 2022.

Bucknall, B. S. and Dori-Hacohen, S. Current and near-term ai as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 119–129, 2022.

Carey, S. Regulating uncertainty: Governing general-purpose ai models and systemic risk. *European Journal of Risk Regulation*, pp. 1–17, 2025.

Creutzig, F., Acemoglu, D., Bai, X., Edwards, P. N., Hintz, M. J., Kaack, L. H., Kilkis, S., Kunkel, S., Luers, A., Milojevic-Dupont, N., et al. Digitalization and the anthropocene. *Annual review of environment and resources*, 47 (1):479–509, 2022.

Delannoy, L., Sampieri, J., Jansen, R. E., Jørgensen, P. S., Nyström, M., and Galaz, V. Artificial intelligence in the polycrisis: fueling or fighting flames? 2025.

Duffy, R., Massé, F., Smidt, E., Marijnen, E., Büscher, B., Verweijen, J., Ramutsindela, M., Simlai, T., Joanny, L., and Lunstrum, E. Why we must question the militarisation of conservation. *Biological conservation*, 232:66–73, 2019.

Galaz, V., Centeno, M. A., Callahan, P. W., Causevic, A., Patterson, T., Brass, I., Baum, S., Farber, D., Fischer, J., Garcia, D., et al. Artificial intelligence, systemic risks, and sustainability. *Technology in society*, 67:101741, 2021.

Geurts, A., Gutknecht, R., Warnke, P., Goetheer, A., Schirrmeister, E., Bakker, B., and Meissner, S. New perspectives for data-supported foresight: The hybrid ai-expert approach. *Futures & Foresight Science*, 4(1):e99, 2022.

Gohr, C., Rodríguez, G., Belomestnykh, S., Berg-Moelleken, D., Chauhan, N., Engler, J.-O., Heydebreck, L., Hintz, M. J., Kretschmer, M., Krügermeier, C., et al. Artificial intelligence in sustainable development research. *Nature Sustainability*, 8(8):970–978, 2025.

Guliyeva, N., Bhardwaj, E., and Becker, C. Exploring the viability of the updated world3 model for examining the impact of computing on planetary boundaries. *arXiv preprint arXiv:2510.07634*, 2025.

Hagendorff, T. The ethics of ai ethics: An evaluation of guidelines. *Minds and machines*, 30(1):99–120, 2020.

Hawken, P. *Drawdown: The most comprehensive plan ever proposed to reverse global warming*. Penguin, 2017.

Hisschemöller, M. and Hoppe, R. Coping with intractable controversies: the case for problem structuring in policy design and analysis 1. In *Knowledge, power, and participation in environmental policy analysis*, pp. 47–72. Routledge, 2018.

Howland, M. F., Quesada, J. B., Martínez, J. J. P., Larrañaga, F. P., Yadav, N., Chawla, J. S., Sivaram, V., and Dabiri, J. O. Collective wind farm operation based on a predictive model increases utility-scale energy production. *Nature Energy*, 7(9):818–827, 2022.

Ilcic, A., Fuentes, M., and Lawler, D. Artificial intelligence, complexity, and systemic resilience in global governance. *Frontiers in Artificial Intelligence*, 8:1562095, 2025.

IPCC, C. C. Climate change 2013: The physical science basis. *Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 2013.

Jobin, A., Ienca, M., and Vayena, E. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9): 389–399, 2019.

Kallis, G., Hickel, J., O'Neill, D. W., Jackson, T., Victor, P. A., Raworth, K., Schor, J. B., Steinberger, J. K., and Ürge-Vorsatz, D. Post-growth: the science of wellbeing within planetary boundaries. *The lancet planetary health*, 9(1):e62–e78, 2025.

Kondor, D., Hafez, V., Shankar, S., Wazir, R., and Karimi, F. Complex systems perspective in assessing risks in artificial intelligence. *Philosophical Transactions A*, 382 (2285):20240109, 2024.

LaCroix, T. and Luccioni, A. S. Metaethical perspectives on 'benchmarking'ai ethics. *AI and Ethics*, pp. 1–19, 2025.

Lavin, A., Krakauer, D., Zenil, H., Gottschlich, J., Mattson, T., Brehmer, J., Anandkumar, A., Choudry, S., Rocki, K., Baydin, A. G., et al. Simulation intelligence: Towards a new generation of scientific methods. *arXiv preprint arXiv:2112.03235*, 2021.

Lawrence, M., Homer-Dixon, T., Janzwood, S., Rockström, J., Renn, O., and Donges, J. F. Global polycrisis: the causal mechanisms of crisis entanglement. *Global Sustainability*, 7:e6, 2024.

Levin, K., Cashore, B., Bernstein, S., and Auld, G. Overcoming the tragedy of super wicked problems: constraining our future selves to ameliorate global climate change. *Policy sciences*, 45(2):123–152, 2012.

Marchau, V. A., Walker, W. E., Bloemen, P. J., and Popper, S. W. *Decision making under deep uncertainty: from theory to practice*. Springer Nature, 2019.

Meadows, D. *Thinking in systems: International bestseller*. chelsea green publishing, 2008.

Meadows, D. H., Meadows, D. H., Randers, J., and Behrens III, W. W. The limits to growth: a report to the club of rome (1972). *Technical Report*, 91(2), 1972.

Mhlanga, D. Ai beyond efficiency, navigating the rebound effect in ai-driven sustainable development. *Frontiers in Energy Research*, 13:1460586, 2025.

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., and Dragan, A. D. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062, 2025.

Mittelstadt, B. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507, 2019.

Neumayer, E. Weak versus strong sustainability: exploring the limits of two opposing paradigms. In *Weak versus Strong Sustainability*. Edward Elgar Publishing, 2010.

Onat, N. C., Mandouri, J., Kucukvar, M., Sen, B., Abbasi, S. A., Alhajyaseen, W., Kutty, A. A., Jabbar, R., Contestabile, M., and Hamouda, A. M. Rebound effects undermine carbon footprint reduction potential of autonomous electric vehicles. *Nature Communications*, 14(1):6258, 2023.

Pansera, M. and Fressoli, M. Innovation without growth: Frameworks for understanding technological change in a post-growth era. *Organization*, 28(3):380–404, 2021.

Pérez-Ortiz, M. From prediction to foresight: The role of ai in designing responsible futures. *Journal of Artificial Intelligence for Sustainable Development*, 1(1):1–9, 2024.

Peters, B. G. What is so wicked about wicked problems? a conceptual analysis and a research program. *Policy and Society*, 36(3):385–396, 2017.

Rigley, E., Chapman, A., Evers, C., and McNeill, W. Anthropocentrism and environmental wellbeing in ai ethics standards: A scoping review and discussion. *AI*, 4(4): 844–874, 2023.

Rittel, H. W. and Webber, M. M. Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169, 1973.

Rudd-Jones, J., Thendean, F., and Pérez-Ortiz, M. Crafting desirable climate trajectories with reinforcement learning explored socio-environmental simulations. *Environmental Data Science*, 4:e41, 2025.

Sandbrook, C., Clark, D., Toivonen, T., Simlai, T., O'Donnell, S., Cobbe, J., and Adams, W. Principles for the socially responsible use of conservation monitoring technology and data. *Conservation Science and Practice*, 3(5):e374, 2021.

Schmager, S., Pappas, I. O., and Vassilakopoulou, P. Understanding human-centred ai: a review of its defining elements and a research agenda. *Behaviour & Information Technology*, pp. 1–40, 2025.

Schön, J., Hoffmann, L., and Becker, N. Expert assessment: The systemic environmental risks of artficial intelligence, 2025. URL https://dl.gi.de/handle/20.500.12116/47924.

Sebestyén, M. Focal points and blind spots of human-centered ai: Ai risks in written online media. *Humanities and Social Sciences Communications*, 12(1):1–20, 2025.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68, 2019.

Selin, N. E., Giang, A., and Clark, W. C. Progress in modeling dynamic systems for sustainable development. *Proceedings of the National Academy of Sciences*, 120(40): e2216656120, 2023.

Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

Søgaard Jørgensen, P., Jansen, R. E., Avila Ortega, D. I., Wang-Erlandsson, L., Donges, J. F., Österblom, H., Olsson, P., Nyström, M., Lade, S. J., Hahn, T., et al. Evolution of the polycrisis: Anthropocene traps that challenge global sustainability. *Philosophical Transactions of the Royal Society B*, 379(1893):20220261, 2024.

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., et al. A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, 56(11):12799–12831, 2023.

Stanley, K. O. Why open-endedness matters. *Artificial life*, 25(3):232–235, 2019.

Steffen, W., Rockström, J., Richardson, K., Lenton, T. M., Folke, C., Liverman, D., Summerhayes, C. P., Barnosky, A. D., Cornell, S. E., Crucifix, M., et al. Trajectories of the earth system in the anthropocene. *Proceedings of the national academy of sciences*, 115(33):8252–8259, 2018.

Stirling, A. Keep it complex. *Nature*, 468(7327):1029–1031, 2010.

Taiebat, M., Stolper, S., and Xu, M. Forecasting the impact of connected and automated vehicles on energy use: a microeconomic study of induced travel and energy rebound. *Applied Energy*, 247:297–308, 2019.

Toyama, K. Can technology end poverty. *Boston review*, 36 (5):12–29, 2010.

UNESCO. Ethical impact assessment: A tool of the recommendation on the ethics of artificial intelligence. Technical report, United Nations Educational, Scientific and Cultural Organization (UNESCO), 2023. URL https://unesdoc.unesco.org/ark:/48223/pf0000386276.

Van Beek, L., Hajer, M., Pelzer, P., van Vuuren, D., and Cassen, C. Anticipating futures through models: the rise of integrated assessment modelling in the climate science-policy interface since 1970. *Global Environmental Change*, 65:102191, 2020.

Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. The role and limits of principles in ai ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200, 2019.

Wright, D., Igel, C., Samuel, G., and Selvan, R. Efficiency is not enough: A critical perspective on environmentally sustainable ai. *Communications of the ACM*, 68(7):62–69, 2025.

A. Diagnosing wickedness and its tensions with standard AI assumptions

This appendix provides (i) a diagnostic framework for assessing the degree of wickedness a problem exhibits, and (ii) a detailed enumeration of the structural properties of wicked systems that conflict with standard AI assumptions.

A.1. From Binary to Spectrum: Diagnosing Wickedness

The concept of wicked problems originates with (Rittel & Webber, 1973), who identified ten properties that distinguish wicked from tame problems: no definitive formulation, no stopping rule, solutions that are good-or-bad rather than true-or-false, no immediate or ultimate test of a solution, every attempt counting significantly (no safe trial-and-error), no enumerable set of potential solutions, essential uniqueness, each problem being a symptom of another, the framing determining the resolution, and the planner having no right to be wrong. Subsequent work has extended this characterisation to *super-wicked* problems—those where time is running out, the actors causing the problem also seek to solve it, central authority is weak, and irrational discounting defers action (Levin et al., 2012)—a description that closely fits many Anthropocene challenges.

However, treating wickedness as a binary category limits its practical utility: researchers need to assess *how* wicked their problem is, and in what respects, in order to calibrate the depth of methodological response. Subsequent scholarship has therefore moved toward characterising wickedness as a spectrum (Peters, 2017; Alford & Head, 2017). A particularly useful diagnostic is the two-dimensional framework of (Hisschemöller & Hoppe, 2018), which classifies problems along two axes:

- **Knowledge certainty**: Is the causal structure of the system understood? Are the relevant variables, feedbacks, and future dynamics identifiable and quantifiable?
- **Value consensus**: Do stakeholders agree on objectives, on what counts as success, and on acceptable trade-offs?

These two dimensions yield four problem types (Figure 4). Importantly, wickedness also scales with the degree of *coupling to other socio-ecological systems*: each coupling introduces additional knowledge uncertainties (how does the intervention propagate across system boundaries?) and additional value contestations (whose interests in adjacent systems are affected?). Structured problems tend to be self-contained, with tight system boundaries and weak couplings. Fully wicked problems are deeply entangled with other domains—food security, livelihoods, biodiversity, climate, trade—such that intervening in one system inevitably reshapes dynamics in others.

These two dimensions yield four problem types (Figure 4):

1. **Structured problems** (high certainty, high consensus): Well-understood dynamics and agreed objectives. Standard AI methods apply with low risk. *Example*: AI-based coordination of wind turbine orientation to maximise energy output (Howland et al., 2022). The objective is unambiguous, the physics are well modelled, interventions are reversible, and success is directly measurable. The system boundary is tight and couplings to other domains are weak.
2. **Moderately structured (value-contested)**: Causal dynamics are reasonably understood, but stakeholders disagree on goals or acceptable trade-offs. Multi-objective formulations and Pareto-based evaluation become necessary. *Example*: camera-based wildlife monitoring and AI-assisted patrol optimisation for anti-poaching (Duffy et al., 2019; Sandbrook et al., 2021). The technology functions as intended and the causal pathways are well documented. However, camera traps also record people—particularly Indigenous peoples and local communities—and monitoring introduced for ecological purposes can become instruments of surveillance, linked to the militarisation of conservation and the criminalisation of subsistence practices. What registers as improved conservation performance on task-level metrics may simultaneously erode the social legitimacy on which durable conservation depends. The wickedness lies in irreducible value contestation—who defines the problem and whose interests are embedded in system design—rather than in causal opacity.
3. **Moderately structured (knowledge-deficient)**: Objectives are broadly agreed upon, but the system's causal structure or future behaviour is poorly understood. Robustness under deep uncertainty and stress-testing become essential. *Example*: ML-based prediction of ecosystem responses to climate interventions such as reforestation or coral reef restoration. The goal of ecosystem recovery is broadly shared, but the relevant dynamics—species interactions, tipping points, lag effects, responses to novel climate regimes—are characterised by deep uncertainty. Models trained on historical data may fail under conditions with no historical precedent. The wickedness lies in knowledge deficiency: the system's future behaviour cannot be reliably inferred from past observations.

4. **Unstructured (fully wicked)**: Both dynamics and objectives are uncertain, contested, or evolving, compounded by deep entanglement with adjacent systems. The full suite of PCAI principles applies. *Example*: AI-driven precision agriculture for sustainable food production. At first glance, this appears unambiguously beneficial: AI optimises fertiliser, water, and pest management, and performs well on metrics such as yield and input efficiency. However, agriculture is coupled to food security, rural livelihoods, land tenure, water systems, biodiversity, trade, and climate simultaneously—and each coupling introduces both knowledge uncertainties and value contestations. On the knowledge dimension, how AI-optimised farming interacts over decades with soil health, landscape-level biodiversity, and market dynamics driving farm consolidation is poorly understood. Recent evidence suggests that precision agriculture overwhelmingly benefits large commercial operations, with limited demonstrated environmental benefits for the small farms that produce roughly a third of the world's food (Haggerty et al., 2026); the systemic effect may be to accelerate consolidation toward industrial monoculture. On the value dimension, agribusiness sees scalable efficiency; smallholders see displacement; ecologists see biodiversity erosion; food sovereignty movements see corporate control of the food system. What makes this *fully* wicked is the interconnectedness: intervening in agricultural efficiency reshapes dynamics across all coupled domains simultaneously. An AI system evaluated at the field level cannot detect the systemic trajectory it helps produce.

   Cite https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2025.1460586/full

This diagnostic operationalises the scaling principle articulated in Section 3: the depth of PCAI engagement should scale with the degree of wickedness. Researchers can use these two dimensions to assess where their problem sits and, correspondingly, which subset of the structural challenges below—and which PCAI responses—are most relevant. (Alford & Head, 2017) offer a complementary typology that adds stakeholder divergence as a third consideration, useful when institutional or political complexity is a primary source of difficulty.

A.2. Structural Properties of Wicked Systems in Tension with AI

The properties below detail the specific ways in which low knowledge certainty and low value consensus—the two sources of wickedness identified above—manifest as tensions with standard AI assumptions. Each item identifies a specific way in which common AI methods, evaluation practices, or deployment strategies can fail when applied to wicked systems. The focus is on violations of core technical assumptions rather than downstream ethical or governance outcomes. Together, these properties help explain why AI systems that perform well in controlled or well-specified settings can produce brittle, misleading, or harmful behaviour when embedded in complex systems.

A.2.1. Objective & Optimization Challenges

These properties arise primarily from low *value consensus*: contested, evolving, or incompatible objectives.

1. **No stable objective function**: Objectives are contested and evolve over time, preventing fixed problem formulation or convergence to a single solution.
2. **Incompatible objectives with no global optimum**: Improvements along one dimension (e.g., efficiency) often degrade others (e.g., equity or resilience), yielding irreducible trade-offs.

A.2.2. Environment & Dynamics Challenges

These properties arise primarily from low *knowledge certainty*: poorly understood causal structure, feedbacks, and dynamics.

3. **Non-stationary environments**: The data-generating process changes over time, often endogenously in response to the model's own deployment, invalidating assumptions of stable or slowly shifting distributions.
4. **Path dependence and lock-in**: Early interventions constrain future options and are difficult or impossible to reverse, violating assumptions of reversible or correctable decisions.
5. **Nonlinear effects and tipping points**: Small changes can trigger large, abrupt, or irreversible system shifts, undermining local performance guarantees.
6. **Emergent behaviour**: System-level outcomes arise from interactions and feedbacks and cannot be inferred from component-level performance or isolated task metrics.

7. **Deep uncertainty**: Key elements of the system—such as causal structure, relevant variables, future regimes, or outcome priorities—are unknown, contested, or not reliably quantifiable.

A.2.3. Action & Learning Constraints

These properties reflect the interaction of both dimensions: acting under combined knowledge and value uncertainty.

8. **No safe exploration regime**: Trial-and-error learning risks real-world, lasting, or irreversible harm, undermining standard exploration assumptions.

9. **Open-ended system evolution**: The relevant state space, action space, objectives, and failure modes cannot be exhaustively specified in advance.

A.2.4. Data & Generalisation Challenges

These properties arise primarily from low *knowledge certainty*, compounded by reflexive system dynamics.

10. **Historical data poorly represents the future**: Past data fails to capture emerging regimes, constraints, or feedbacks.

11. **Feedback-contaminated data**: Post-deployment data is shaped by the model's own influence on the system, biasing learning and evaluation.

A.2.5. Evaluation & Benchmarking Challenges

These properties arise from the combined effect of both dimensions: evaluating interventions whose goals are contested in systems whose responses are uncertain.

12. **No definitive success metric**: There is no agreed-upon way to determine whether an intervention succeeded in contested, long-horizon contexts.

Table 2 operationalises these challenges by linking structural properties of wicked systems to the standard assumptions they violate in mainstream AI, and to the research directions motivated under PCAI. Importantly, many of these challenges have partial analogues in existing AI subfields. For example, non-stationarity is studied in continual and online learning; distribution shift and tail risk are addressed in distributionally robust optimisation; conflicting objectives and value trade-offs appear in multi-objective optimisation and reinforcement learning (RL); delayed consequences and irreversibility are explored in long-horizon and risk-sensitive RL; and endogenous feedbacks are examined in causal modelling and strategic or multi-agent learning.

The claim here is not that such tools do not exist, nor that they are irrelevant. On the contrary, these techniques already constitute some of the most promising foundations for AI in complex, high-stakes settings. From a systems perspective, they can be understood as addressing key facets of wicked dynamics albeit often in partial or domain-specific ways. PCAI highlights the need to further integrate these approaches with explicit system-level reasoning. Doing so reframes existing methods as complementary components of a system-aware design and evaluation paradigm. This position paper underscores the importance of sustained research investment in these areas.

	High consensus	**Low consensus**
High certainty	Structured (tame)	Value-contested
Low certainty	Knowledge-deficient	Unstructured (wicked)

Figure 4. Diagnostic framework for assessing wickedness, adapted from Hisschemoller & Hoppe (1996). Wickedness increases along both dimensions. PCAI measures scale accordingly.

B. Structural Challenges of Wicked Systems and Tensions with Standard AI Assumptions

This appendix summarizes structural properties of wicked systems that directly conflict with standard assumptions in AI. Each item identifies a specific way in which common AI methods, evaluation practices, or deployment strategies can fail when applied to such systems. The focus is on violations of core technical assumptions rather than downstream ethical or governance outcomes. Together, these properties help explain why AI systems that perform well in controlled or well-specified settings can produce brittle, misleading, or harmful behavior when embedded in complex systems.

B.1. Objective & Optimization Challenges

1. **No stable objective function**: Objectives are contested and evolve over time, preventing fixed problem formulation or convergence to a single solution.

2. **Incompatible objectives with no global optimum**: Improvements along one dimension (e.g., efficiency) often degrade others (e.g., equity or resilience), yielding irreducible trade-offs.

B.2. Environment & Dynamics Challenges

3. **Non-stationary environments**: The data-generating process changes over time, often endogenously in response to the model's own deployment, invalidating assumptions of stable or slowly shifting distributions.

4. **Path dependence and lock-in**: Early interventions constrain future options and are difficult or impossible to reverse, violating assumptions of reversible or correctable decisions.

5. **Nonlinear effects and tipping points**: Small changes can trigger large, abrupt, or irreversible system shifts, undermining local performance guarantees.

6. **Emergent behavior**: System-level outcomes arise from interactions and feedbacks and cannot be inferred from component-level performance or isolated task metrics.

7. **Deep uncertainty**: Key elements of the system—such as causal structure, relevant variables, future regimes, or outcome priorities—are unknown, contested, or not reliably quantifiable.

B.3. Action & Learning Constraints

8. **No safe exploration regime**: Trial-and-error learning risks real-world, lasting, or irreversible harm, undermining standard exploration assumptions.

9. **Open-ended system evolution**: The relevant state space, action space, objectives, and failure modes cannot be exhaustively specified in advance.

B.4. Data & Generalization Challenges

10. **Historical data poorly represents the future**: Past data fails to capture emerging regimes, constraints, or feedbacks.

11. **Feedback-contaminated data**: Post-deployment data is shaped by the model's own influence on the system, biasing learning and evaluation.

B.5. Evaluation & Benchmarking Challenges

12. **No definitive success metric**: There is no agreed-upon way to determine whether an intervention succeeded in contested, long-horizon contexts.

Table 2 operationalizes these challenges by linking structural properties of wicked systems to the standard assumptions they violate in mainstream AI, and to the research directions motivated under PCAI. Importantly, many of these challenges have partial analogues in existing AI subfields. For example, non-stationarity is studied in continual and online learning; distribution shift and tail risk are addressed in distributionally robust optimization; conflicting objectives and value trade-offs appear in multi-objective optimization and reinforcement learning (RL); delayed consequences and irreversibility are

explored in long-horizon and risk-sensitive RL; and endogenous feedbacks are examined in causal modeling and strategic or multi-agent learning. The claim here is not that such tools do not exist, nor that they are irrelevant. On the contrary, these techniques already constitute some of the most promising foundations for AI in complex, high-stakes settings. From a systems perspective, they can be understood as addressing key facets of wicked dynamics albeit often in partial or domain-specific ways. PCAI highlights the need to further integrate these approaches with explicit system-level reasoning. Doing so reframes existing methods as complementary components of a system-aware design and evaluation paradigm. This position paper underscores the importance of sustained research investment in these areas.

C. Comparison between Human-Centered and Planet-Centered AI Frameworks

Table 3 situates Planet-Centered AI (PCAI) relative to Human-Centered AI (HCAI), highlighting both continuity and divergence.

D. Mini Use Case: System Mapping for AI-Assisted Conservation Enforcement

Domain: Biodiversity conservation in protected areas.
Task-level AI intervention: AI-enabled surveillance and patrol optimization for anti-poaching.

Recent conservation efforts increasingly deploy AI systems—combining remote sensing and computer vision—to detect poaching activity and optimize ranger patrol routes. These systems are typically evaluated on metrics such as detection accuracy, patrol efficiency, or reductions in poaching incidents. However, as documented extensively in conservation social science, anti-poaching operates within coupled socio-ecological systems characterized by feedbacks between wildlife populations, local communities, rangers, armed groups, and political-economic incentives. In many contexts, AI-enabled enforcement becomes embedded within militarised conservation strategies, with documented long-term consequences for ecological integrity, social legitimacy, and system stability (Duffy et al., 2019; Sandbrook et al., 2021).

D.1. System Mapping: AI as a Coupled System Intervention

The system boundary includes:

- **Ecological components:** target species populations, habitat quality, and trophic interactions;
- **Human actors:** rangers, local communities, poachers, conservation NGOs, donors, and state agencies;
- **Institutional dynamics:** governance structures, funding mechanisms, and performance metrics;
- **Technological infrastructure:** surveillance hardware, models, data pipelines, and operational protocols.

Key feedback pathways identified through system mapping include:

1. **Enforcement–adaptation feedback:** improved detection alters poacher behavior, potentially increasing displacement, sophistication, or violence;
2. **Militarisation–legitimacy feedback:** Increased surveillance and force reduce community trust, undermining long-term conservation cooperation.;
3. **Resource allocation feedback:** visible enforcement success attracts funding that may crowd out community-based strategies;
4. **Institutional lock-in:** Militarised enforcement increases attacks on rangers, reinforcing justification for further militarisation.

These feedbacks are extensively documented in the literature on militarised conservation (Duffy et al., 2019). This mapping highlights that the primary impact of the AI system lies in shaping how the conservation system evolves together with its coupled impact on human actors, rather than solely in detecting individual poaching events.

D.2. Theory of Change: AI as a Trajectory-Shaping Intervention

Prior to explicit system mapping, AI-assisted anti-poaching interventions are often guided by a simplified and largely implicit theory of change: improved detection leads to fewer poaching events, which in turn leads to species recovery. While intuitively appealing, this linear causal pathway abstracts away the social, institutional, and political dynamics through which conservation interventions operate, and treats enforcement effectiveness as a sufficient proxy for long-term ecological success. System mapping reveals that this intended causal pathway is neither guaranteed nor exhaustive. Instead of a single dominant mechanism, AI-assisted enforcement activates multiple interacting processes that may reinforce or counteract one another. Under PCAI, AI-assisted enforcement is instead modeled as activating multiple interacting mechanisms, whose relative influence determines long-run system trajectories.

Core intervention effect. The AI system reallocates attention, authority, and resources by intensifying surveillance, reshaping ranger practices, and producing data that informs governance decisions and donor priorities.

Mechanism set A: Short-term deterrence (context-dependent). In the short run, increased detection may reduce observable poaching activity or displace it spatially. These effects are contingent on limited adaptive capacity and do not address structural drivers of illegal hunting.

Mechanism set B: Escalation and coercive reinforcement. As actors adapt, AI-enabled enforcement can justify heightened force, expand surveillance of local populations, and shift ranger roles toward paramilitary functions, reinforcing self-amplifying militarisation dynamics.

Mechanism set C: Political–economic lock-in. By privileging quantifiable enforcement outcomes, AI systems shape institutional success criteria and funding flows, entrenching enforcement-centric strategies even when they undermine ecological resilience or social legitimacy.

Mechanism set D: Social legitimacy erosion. Expanded surveillance and coercive practices may deepen historical grievances, reduce community cooperation, and weaken informal conservation governance, increasing long-term system fragility.

D.3. Evaluation Dimensions for Trajectory-Oriented Assessment

Following this system mapping and theory of change, PCAI evaluation focuses on how AI interventions influence long-term system trajectories, rather than point-in-time task performance. Indicative evaluation dimensions include:

- **Ecological resilience:** species recovery under environmental variability and stress;
- **Social legitimacy:** community cooperation, conflict incidence, and trust in conservation institutions;
- **Violence dynamics:** escalation or de-escalation of armed encounters affecting rangers and civilians;
- **Institutional adaptability:** diversity of conservation strategies retained over time;
- **Lock-in risk:** dependence on enforcement-centric approaches and difficulty of reversal;
- **Equity impacts:** distribution of benefits and harms across affected populations.

While many of these dimensions may not be readily quantifiable or reducible to single metrics, making them explicit is nevertheless essential. Explicit articulation of these objectives can shape the choice of modeling paradigm, the form of human–AI interaction, and even the class of interventions considered appropriate. By foregrounding trajectory-level concerns, PCAI encourages discussion of modeling approaches that are developed in close collaboration with domain experts, and supports more meaningful comparison across alternative interventions, including non-AI and hybrid approaches.

Importantly, the need to make such objectives explicit does not imply the existence of a single correct modeling solution. On the contrary, the domains in which PCAI is most relevant are themselves characterized by wicked dynamics, in which no single modeling paradigm can be expected to capture all relevant dynamics. Explicit articulation of trajectory-level concerns therefore serves to structure interdisciplinary modeling processes, constraining assumptions, surfacing uncertainties, and identifying trade-offs that require joint deliberation among ML researchers, domain scientists, and affected stakeholders.

Table 2. Wicked System Challenges, AI Limitations, and PCAI-Motivated Research Directions

Wicked Challenge	How Standard AI Struggles	PCAI-Motivated Research Directions
Objective & Optimization Challenges		
No stable objective function	Assumes objectives can be specified in advance and optimized consistently. This is the case even with AI methods that embed dynamic rewards or loss functions.	Domain generalisation and adaptation, preference learning and adaptive and pluralistic objective representations; methods that support deliberation over objectives and further research on dynamic reward or loss specification.
Incompatible objectives with no global optimum	Encodes trade-offs through fixed scalarization or static Pareto formulations, masking irreducible value conflicts.	Trade-off–aware learning and exploration, evolving objective sets, exploratory modeling with multiple objectives, and decision-support tools that surface value conflicts.
Environment & Dynamics Challenges		
Non-stationary environments	Assumes stable or slowly shifting data distributions; struggles with endogenous change driven by deployment, behavioral adaptation, or policy response.	Continual and regime-aware learning, shift detection, stress-testing under structural breaks, and robustness across plausible future distributions.
Path dependence and lock-in	Evaluates actions within bounded horizons, making delayed, irreversible consequences or loss of future options difficult to observe or attribute.	Methods that explore system dynamics and reason about irreversibility, option value, and long-term trajectory selection through simulation and epistemic infrastructures.
Nonlinear effects and tipping points	Performance degrades near thresholds where small errors can trigger large or irreversible system responses.	Worst-case analysis, early-warning indicators, and robustness to threshold and tail-risk behavior.
Emergent behavior	Task-level validation does not predict system-level outcomes produced by interactions among models, users, and institutions.	System-level evaluation, multi-agent game-theoretic learning, and collective or population-level behavior modeling.
Deep uncertainty	Assumes known system structure and probabilistic uncertainty; cannot represent unknown, contested, or unmodellable dynamics.	Decision-making under deep uncertainty, robust satisficing, and epistemic infrastructures for uncertainty aggregation and communication (e.g., expert elicitation, superforecasting).
Action & Learning Constraints		
No safe exploration regime	Exploration assumes low-cost or reversible errors, inappropriate for high-stakes real-world interventions.	High-fidelity differentiable simulators (digital twins, world models) with foresight, constrained learning, generative AI for scenario generation, and pre-deployment risk analysis.
Open-ended and evolving action space	Assumes a fixed and enumerable action space, while real interventions reshape available options and constraints.	Open-ended learning, adaptive action spaces, and methods for reasoning over expanding or evolving intervention sets.
Data & Generalization Challenges		
Historical data poorly represents the future	Trains on past regimes that omit emerging constraints, feedbacks, or structural change.	Out-of-distribution generalisation, robustness beyond historical fit and hybrid models (e.g. physics-informed neural networks).
Feedback-contaminated data	Post-deployment data is endogenous to model, biasing learning and evaluation.	Causal modeling and representation learning, and feedback-aware evaluation.
Evaluation & Benchmarking Challenges		
No definitive success metric	Presumes agreed-upon success criteria, absent in contested, long-horizon settings.	Multi-criteria, trajectory-oriented, and deliberative/speculative evaluation frameworks.

Table 3. Comparison of Human-Centered AI (HCAI) and Planet-Centered AI (PCAI)

Dimension	Human-Centered AI (HCAI)	Planet-Centered AI (PCAI)
Core guiding question	How can AI systems be designed and deployed to serve people responsibly, safely, and fairly?	How can AI systems support the long-term stability and viability of coupled Earth systems within which human societies operate?
Ethical baseline	Anthropocentric: moral priority assigned to human rights, dignity, autonomy, welfare, and social justice.	Relational and ecological: moral concern extends to ecosystems, non-human life, future generations, and life-support systems as conditions for human flourishing.
Primary stakeholders	Human users, affected communities, workers, institutions, and rights-holders.	Humans, non-human species, ecosystems, future generations, and planetary commons (e.g., climate, biodiversity, biogeochemical cycles).
Scope of responsibility	Identify, prevent, and mitigate harms to humans arising from AI use, including sociotechnical and institutional impacts.	Account for systemic social–ecological effects, intergenerational impacts, feedbacks, and risks of irreversible or path-dependent change.
Core design objective	Enable AI systems that are usable, fair, safe, accountable, and aligned with human values and oversight.	Support planetary stability, resilience, and long-term livability by shaping system trajectories toward desirable futures.
Unit of analysis	AI systems and their interaction with users, organizations, and sociotechnical contexts.	Coupled social–ecological systems, intervention pathways, and multi-scale dynamics over time.
Definition of progress	Improvements in task performance, usability, fairness, accountability, transparency, and human capabilities.	Reduced systemic risk, preserved ecological integrity, resilience under change, and maintenance of future options.
Temporal horizon	Primarily near- to medium-term impacts surrounding design, deployment, and use.	Explicitly long-term and intergenerational, accounting for cumulative, delayed, and irreversible impacts.
Risk framing	Algorithmic and sociotechnical risks affecting humans, such as bias, misuse, privacy loss, exclusion, and unsafe automation.	Systemic and planetary risks, including feedback loops, rebound effects, correlated failures, and tipping points.
Evaluation logic	Combination of task-level benchmarks, user studies, and post hoc impact assessments.	Trajectory-oriented evaluation, including stress-testing under non-stationarity, scenario analysis, and system-level probes.
Role of AI	Human-centered decision support, augmentation, and automation under meaningful human control.	Epistemic and stewardship-oriented: sensemaking, foresight, scenario exploration, and support for collective judgment under deep uncertainty.
Knowledge foundations	Computer science, human-computer interaction, ethics, law, psychology, and social sciences.	Systems science, Earth-system science, ecology, sustainability science, political economy, and plural epistemologies.
Governance model	Institutional compliance, accountability mechanisms, and oversight focused on specific deployments.	Planetary stewardship: governance oriented toward long-horizon coordination, boundary-setting, reversibility, and adaptive control across scales.