Balancing Sustainable Development Goals through Orchestrating Autonomous Multi-agent Architecture

Sivaramakrishnan R Guruvayur†
Aquarius & Center for Cybersecurity Systems and Networks Amrita Vishwa Vidyapeetham, Amritapuri, India
sivaramakrishnangr@am.amrita.edu

Krishnasree Achuthan
Center for Cybersecurity Systems and Networks Amrita Vishwa Vidyapeetham, Amritapuri, India
krishnashree@am.amrita.edu

Joao Pita Costa
IRCAI
Institut Jozef Stefan
Ljubljana
joao.pitacosta@ijs.si

Abstract / Povzetek

Sustainable Development Goals (SDGs) motivate a coupled control problem in which interventions targeting one goal may generate synergies or trade-offs with other goals and may also interact with economic, institutional and governance constraints. This paper presents a conceptual design for a hierarchical multi-agent architecture intended to support the analysis of such policy trade-offs while preserving human decision authority. The architecture comprises SDG Goal Guardian and Sub-goal Advocate agents, a Scenario Planner, a Trade-off Balancer, a Macro-Economy Agent, a Governance and Fairness Agent, and a Meta-Orchestrator, with a Human Policy Interface connecting the system to a Policy Leadership Council. Knowledge graphs provide contextual and provenance-aware information, while domain-specific simulation or digital-twin models are proposed as forward models for estimating policy impacts. The proposed decision problem is formulated as a constrained multi-objective optimization problem in which candidate policies are evaluated against SDG objectives and macroeconomic, fairness and governance constraints. Rather than presenting an implemented system or empirical performance results, the paper defines an architecture and an evaluation methodology through which the design could subsequently be tested.

Keywords / Ključne besede

Agentic AI, multi-agent orchestration, Sustainable Development Goals, human-in-the-loop, SDG indicators, AI governance.

1. Introduction

The SDG framework established by the United Nations specifies 17 goals and 169 targets with dense causal interdependencies across economic, environmental, and social subsystems. Interlinkage studies show that interventions targeting a single goal—for example, industrial expansion under SDG 8—can simultaneously drive poverty reduction (SDG 1) and water stress or emissions (SDG 6, SDG 13), depending on local context and macroeconomic conditions. On a positive angle, improvements to the response to water disaster risk reduction (SDG 11) in remote areas allied with the proper education of local communities (SDG 4) can improve climate change resilience of populations (SDG 13), their public health due to fast response (SDG 3), and even food security (SDG 2). It often happens that what affects one SDG will also affect others in that balance. Policy coherence analyses from OECD and the European Commission JRC underline that SDG indicators and macro variables (growth, debt, fiscal space) must be treated jointly; isolated optimization is insufficient for coherent strategy [9]. Agentic AI, where LLM-based agents act as goal-directed entities with tools and memory, has been applied to SDG-related modeling by combining agent-based modeling with autonomous reasoning [2,10]. Multi-agent orchestration surveys [20] and reference architectures identify capability-aligned agents, structured communication protocols (MCP, A2A), and centralized or hierarchical orchestration layers [1,8] as key design elements for scalable, governed multi-agent systems [4,14]. AI governance frameworks and enterprise blueprints further insist on human-in-the-loop (HitL) gates, audit trails, and explicit governance agents for high-stakes decisions [11,18].

We build on this work to propose a hierarchical, capability-aligned multi-agent architecture for SDG balancing. SDG Goal Guardians and Sub-goal Advocates act as strong domain advocates [6,7]; a Trade-off Balancer Agent (TBA) and Macro-Economy Agent compute constrained compromises; a Governance & Fairness Agent (GFA) enforces AI ethics and fairness; and a Human Policy Interface Agent (HPIA) supports a Policy Leadership Council (PLC) that retains final authority under a gated Human-in-the-Loop (HITL) override model. A Meta-Orchestrator Agent (MOA) coordinates these components, grounding reasoning in SDG Knowledge Graphs [3], built on Context Graphs [19] towards a conceptual SDG digital twin through the orchestration of resources [12,15]. In that, this paper addresses the following research question:

Q: What hierarchical multi-agent architecture be designed to address competing SDG policy objectives under fairness and governance constraints while preserving human decision?

To address this, the paper makes a conceptual-design threefold contribution: (i) a hierarchical multi-agent architecture for representing SDG-specific objectives, macroeconomic constraints, governance requirements and stakeholder interests; (ii) a unified constrained multi-objective formulation for evaluating candidate policies across these dimensions; and (iii) an

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
http://doi.org/10.70314/is.2026.sikdd.121

evaluation blueprint specifying the data, simulation, stakeholder and policy-coherence tests required for future empirical validation. The proposed architecture, compliant to the UNESCO Recommendation on the Ethics of Artificial Intelligence [13], hasn't yet been implemented nor its optimization performance has been empirically demonstrated.

2. Technical Architecture

A. Layered Design

The architecture (described in Figure 1, image GPT generated) follows a layered multi-agent reference design with clear separation of concerns.

1. *Orchestration Layer:* The Orchestration Layer consists of the Meta-Orchestrator Agent (MOA), Scenario Planner Agent (SPA), Trade-off Balancer Agent (TBA), Governance & Fairness Agent (GFA), and Macro-Economy Agent.
2. *Agent Layer:* The Agent Layer consists of SDG Goal Guardians and Sub-goal Advocates, each encapsulating domain objectives and tools.
3. *Knowledge Layer:* The Knowledge Layer consists of the SDG Knowledge Graph (KG), Context Graph (CG), and Macro-Context Graph (MCG).
4. *Simulation Layer:* The Simulation Layer consists of domain-specific digital twins (DTs) for urban, water, energy, and agriculture.
5. *Human Interface Layer:* The Human Interface Layer consists of the Human Policy Interface Agent (HPIA) and Policy Leadership Council (PLC), implementing human-in-the-loop (HITL) override.

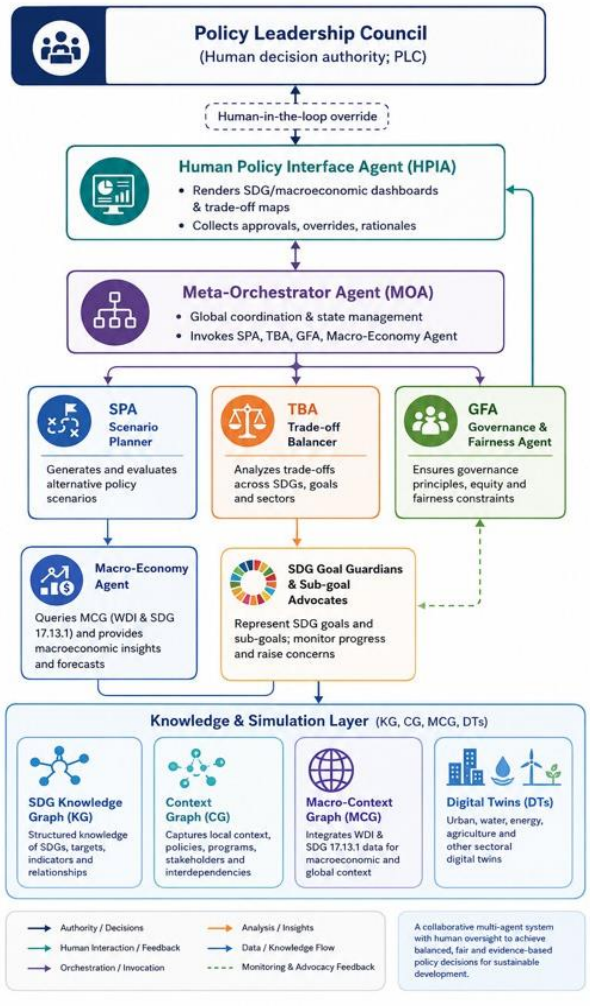

Figure 1: SDG Multi-Agent Orchestration Architecture

B. Knowledge and Macro-Context Representation

The SDG Knowledge Graph (KG) captures goals, targets, indicators, policies, legal norms, and stakeholder entities, built via LLM-assisted extraction and SDG knowledge graph resources. The Context Graph (CG) adds temporal validity, provenance, confidence scores, and local data (e.g., spatial re-source maps and infrastructure), enabling time-aware, context-sensitive reasoning. The Macro-Context Graph (MCG) aggregates World Development Indicators [16] (GDP growth, inflation, debt ratios, employment) and SDG macro-coherence indicators [17] (e.g., 17.13.1), together with policy guidance from the World Bank and IMF resilience and sustainability facilities [5]. A Macro-Economy Agent exposes MCG queries to SPA and TBA, allowing them to impose macro-stability constraints and align scenarios with global economic guidelines.

Digital twins implement forward models for socio-environmental systems; SDG agents query DTs to estimate indicator trajectories and utilities under candidate policies. When f_θ can be mechanistic, data-driven or an ensemble, we get

$$x^t+1=f_\theta(x_t,u_t,z_t)$$

with uncertainty defined as

$$p(x_t+1|x_t u_t)$$

from where we therefore introduce uncertainty-aware policy optimization:

$$\max_\pi \mathbb{E}[J(\pi)] - \lambda \mathrm{Risk}\big(J(\pi)\big)$$

or, more concretely,

$$P\big(U_g(x,u) \geq \tau_g\big) \geq 1 - \alpha_g.$$

C. Agent Roles

Each SDG Goal Guardian Agent is defined by a goal-specific utility function, a set of hard constraints (regulatory and physical limits), and read access to KG, CG, DTs, and selected macro signals. Sub-goal Advocates evaluate specific targets or indicators and feed structured reports to their Goal Guardian. These agents never attempt cross-goal balancing; they act as strong advocates. The Scenario Planner Agent (SPA) decomposes policy questions into a set of scenarios by querying KG/CG/MCG for affected goals and feasibility bounds. The Trade-off Balancer Agent (TBA) aggregates utilities and constraints from SDG and macro agents and computes candidate compromises via multi-objective optimization , introduced as the state transition:

$$x_{t+1} = f_\theta(x_t, u_t, c_t) + \epsilon_t$$

where:

x_t: SDG and macroeconomic state at time t
u_t: policy intervention
c_t: contextual variables
f_θ: learned/simulated transition model
ϵ_t: model uncertainty/noise.

Then define each SDG objective as a trajectory-level utility:

$$J_g(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t U_g(x_t, u_t) \mid \pi\right]$$

where $\pi(u \mid x)$is the policy-selection strategy.

The Governance & Fairness Agent (GFA) applies AI ethics, fairness, and legal constraints, rejecting or flagging non-compliant policies. The Meta-Orchestrator Agent (MOA) manages global state and coordination, while HPIA renders outputs into decision briefs and dashboards for the PLC, which operates under HITL override. Moreover, when x denotes the vector of SDG indicator states and macroeconomic variables, and let u represent policy actions such as investment al-locations, regulatory changes, or infrastructure deployments. The planner's goal is to choose u to optimize a set of objective functions {Ug(x, u)} corresponding to SDG goals g ∈ {1, . . . , 17} and macroeconomic utility Umacro(x, u), subject to constraints C(x, u) capturing feasibility, stability, fairness, and governance. In practice, these objectives and constraints are high-dimensional, non-linear, and context-dependent, making manual planning and isolated models in-sufficient to anticipate emergent behavior and trade-offs.

3. Orchestration Logic & Human-in-the-Loop

A. Trade-off Computation

For scenario s, SDG agents produce a utility vector $\mathbf{U}_{\text{SDG}}(s)$ and constraints $\mathbf{C}_{\text{SDG}}(s)$; the Macro-Economy Agent produces a macro utility $U_{\text{macro}}(s)$ and constraints $\mathbf{C}_{\text{macro}}(s)$. TBA seeks compromise policies $\mathbf{u}^*$ that maximize the combined objectives subject to: (i) SDG floors for each goal; (ii) macro-stability constraints, such as debt, growth, and inflation bounds; (iii) fairness constraints on distributional outcomes; (iv) governance constraints from GFA.TBA approximates Pareto frontiers using evolutionary algorithms, weighted sums, or lexicographic orderings, possibly combined with negotiation among Goal Guardians. It returns a small set of candidate compromise scenarios annotated with utilities and trade-off diagnostics.

B. Orchestrated Decision Cycle

1) Formal Trade-off Balancing: For a given scenario s, let

$$\mathbf{U}(s) = [U_1(s), U_2(s), \ldots, U_{17}(s)] \quad (1)$$

be the vector of goal utilities, and let $U_{\text{macro}}(s)$ denote a composite macro utility derived from MCG, such as growth, debt, and stability indices. TBA's task is to find a policy $\mathbf{u}$ and scenario configuration that solve

$$\mathbf{u}^* = \arg\max_{\mathbf{u}} [U_1(\mathbf{x}, \mathbf{u}), \ldots, U_{17}(\mathbf{x}, \mathbf{u}), U_{\text{macro}}(\mathbf{x}, \mathbf{u})]. \quad (2)$$

subject to:

$$U_g(\mathbf{x}, \mathbf{u}) \geq \tau_g, \qquad g = 1, \ldots, 17, \quad (3)$$

$$C_{\text{macro}}(\mathbf{x}, \mathbf{u}) \leq 0, \quad (4)$$

$$C_{\text{fair}}(\mathbf{x}, \mathbf{u}) \leq 0, \quad (5)$$

$$C_{\text{gov}}(\mathbf{x}, \mathbf{u}) \leq 0. \quad (6)$$

Here, τ_g represents the minimum acceptable floor for SDG goal g. In practice, TBA approximates Pareto frontiers using evolutionary algorithms such as NSGA-II, weighted sums, lexicographic orderings, or market-style negotiation among Goal Guardians, with parameters chosen to reflect policy priorities. TBA produces a small set of candidate compromise policies, each annotated with an SDG utility vector, macro metrics, and trade-off diagnostics, which GFA and HPIA then process.

The orchestrated decision cycle consists of four stages.

1) Problem intake and scenario generation (MOA + SPA): MOA receives a policy question; SPA identifies impacted SDGs and macro dimensions via KG/CG/MCG and generates feasible scenarios.
2) Distributed evaluation (SDG + Macro agents): Goal Guardians and Sub-goal Advocates evaluate scenarios using KG/CG and DTs, computing utilities and constraint violations; Macro-Economy Agent evaluates macro feasibility from MCG.
3) Balancing and governance (TBA + GFA): TBA aggregates reports and computes compromise scenarios subject to SDG, macro, fairness, and governance constraints. GFA checks AI ethics, fairness, and legal compliance and passes compliant scenarios onward.
4) HITL override (HPIA + PLC): HPIA presents candidate policies with SDG trade-off maps, macro dash-boards, and explanations. PLC approves, modifies, or rejects scenarios, with all actions logged and fed back into KG/CG/MCG as institutional memory.

4. Scenario: Water–Energy–Food Nexus

The following illustrative scenario demonstrates how the proposed architecture could address a water–energy–food policy trade-off. It is not an empirical case study and does not use a specific region, observed dataset, calibrated digital twin, or measured outcomes. Consider a semi-arid region where policymakers propose expanding irrigation to increase agricultural production and rural incomes. The Scenario Planner Agent (SPA) identifies SDGs 2, 6, 7, 8 and 13, together with macroeconomic dimensions including agricultural output, rural employment, energy-import dependence and public debt. It then generates alternative scenarios, such as conventional irrigation expansion, drip irrigation with solar-powered pumps, and phased deployment.

The SDG Goal Guardians and Sub-goal Advocates evaluate the scenarios using relevant hydrological, agricultural and emissions models. Conceptually, conventional irrigation could improve agricultural production and rural income while increasing groundwater depletion and emissions, potentially causing SDG 6 and SDG 13 to fall below predefined floors. The Macro-Economy Agent evaluates infrastructure costs, energy-import exposure and debt implications. The Trade-off Balancer Agent (TBA) then identifies a set of feasible, non-dominated alternatives. In this illustrative example, a combination of drip irrigation, solar-powered pumping and phased deployment could provide a feasible compromise across SDG and macroeconomic objectives. These outcomes are hypothetical and are intended only to demonstrate the proposed decision process.

The Governance and Fairness Agent (GFA) evaluates candidate policies against predefined legal, governance and distributional constraints, while the Human Policy Interface Agent (HPIA)

presents their trade-offs to the Policy Leadership Council (PLC). The PLC retains authority to approve, modify or reject the alternatives, with the decision and rationale potentially recorded as institutional memory in the relevant knowledge graphs. The scenario therefore illustrates the proposed sequence of scenario generation, distributed evaluation, multi-objective trade-off analysis and human decision-making without claiming empirical effectiveness.

5. Evaluation Metrics and Limitations

A pilot development of the multi-agent framework integrates domain digital twins, execute case studies, and calibrate optimization algorithms, culminating in evaluation and dissemination. Evaluation metrics span the axes: (i) **Policy Coherence:** Reduction in conflicting SDG im-pacts, quantified via trade-off curves and conflict counts; (ii) **SDG Co-benefits:** Number and magnitude of SDG indi-cators improved per policy decision; (iii) **Equity:** Fairness of outcomes across regions and stake-holder groups, assessed via indices such as Gini and participatory scoring; (iv) **Resilience:** Robustness of recommended policies under simulated shocks and perturbations in digital twin sce-narios; (v) **Explainability:** Clarity and traceability of agent ratio-nales and Orchestrator decisions, evaluated by domain experts; (vi) **Stakeholder Acceptance:** Qualitative feedback from policymakers and practitioners regarding plausibility, utility, and trustworthiness of system recommendations.

Open datasets (SDG indicators, GIS layers, census and environmental data) and standard AI libraries (e.g., PyTorch or TensorFlow for LLM components) form the technical basis, with cloud platforms such as AWS or GCP providing scalable compute for simulation and optimization workloads.

The architecture faces several technical and institutional limitations.

Scalability. As the number of agents, indicators, macro variables, and DTs increases, MOA, SPA, and TBA must handle larger state spaces and message volumes, raising latency and context-window concerns. Hierarchical orchestration, semantic rout-ing, and dynamic agent registries will be needed to maintain scalability.

Data and Uncertainty. KG/CG/MCG quality depends on SDG and macro data that may be incomplete or contested, especially in lower-capacity contexts. Without uncertainty modeling and expert in-put, optimization outcomes may be brittle; robust or Bayesian techniques and continuous data curation are necessary.

Digital Twin Fidelity. DT-based utilities hinge on calibration and maintenance; simplified models can mislead SDG agents and TBA. Validation pipelines, sensitivity analysis, and hybrid human–model oversight must be implemented.

6. Conclusion

The paper defines a common constrained multi-objective formulation through which heterogeneous SDG and macroeconomic objectives can be evaluated, while treating the resulting Pareto set as input to a human policy decision rather than allowing the AI system to determine a final policy autonomously. The architecture also provides explicit locations for representing stakeholder preferences, institutional constraints, uncertainty, provenance and governance requirements.

The water–energy–food example demonstrates the intended orchestration process but is deliberately presented as an illustrative scenario rather than an empirical case study. Consequently, no claim is made regarding the effectiveness, accuracy or policy superiority of the proposed architecture. Establishing these properties requires implementation and systematic evaluation against established SDG decision-support, including system-dynamics and SDG interaction methods.

Future work should therefore focus on a constrained prototype with a limited number of SDGs and stakeholders, validated data sources, explicit provenance, calibrated simulation models and reproducible optimization procedures. Particular attention should be given to stakeholder conflict, uncertainty propagation, LLM reliability, fairness and human decision-making. Such evaluation will determine whether multi-agent orchestration provides measurable advantages over existing approaches for analysing complex SDG policy trade-offs while maintaining appropriate human and institutional control.

Acknowledgments / Zahvala

This research was partially funded by the European Commission's Horizon research and innovation program under grant agreement 101135800 (RAIDO).

References / Literatura

[1] A. Adimulam, et al. The Orchestration of Multi-Agent Systems: Architectures, Protocols, and Enterprise Adoption, arXiv 2601.13671, 2026

[2] I. Chughtai, "AI Agents and Sustainable Development Goals: Bridging the Gap Through Innovation," SBS Journal of App. Business Research, 2026.

[3] E. Fotopoulou et al., "SustainGraph: A Knowledge Graph for Sustainable Development," Front. Environ. Sci., 26, 2022.

[4] R. Kodali, "Governing Multi-Agent AI Systems: An Enterprise Blueprint for Scalable Autonomy," Architecture & Governance, 2026.

[5] IMF, "Resilience and Sustainability Facility—Updated Operational Guidance Note," 2025.

[6] J. Liu et al., "Agentic AI for Sustainable Development: Leveraging LLM-Enhanced Agent-Based Modeling for Complex Policy Strategies", Emerging Media 3.3 (2025): 401-413., 2025.

[7] D. Maheshwari et al., "AI's Global South Pivot: Equity, Ethics and Ecology," CIGI Policy Brief 225, 2026.

[8] Microsoft, "Multi-agent Reference Architecture," 2025.

[9] A. Miola et al, "Interlinkages and Policy Coherence for the Sustainable Development Goals," European Commission JRC 115163, 2019.

[10] OECD, Policy Coherence for Sustainable Development 2019, OECD Publishing, 2019.

[11] SDG Knowledge Hub, "Tools Support Policy Makers to Prioritize Multiple Goals," SDG Knowledge Hub 30, 2020.

[12] T. A. Syed et al., "Agentic AI-enhanced digital twins for Smart City civil infrastructure: A secure, autonomous and auditable management framework" PLOS ONE 21(7), 2026.

[13] UNESCO, "Recommendation on the Ethics of Artificial Intelligence," UNESCO Library, 2021.

[14] J. deVadoss., "Designing Intelligent Enterprise Agents: A Capability-Aligned Multi-Agent Architecture," arXiv 2605.08258, 2026.

[15] L. Wan et al., "From Urban Modelling to City Digital Twins," Environment and Planning B, Urban Analytics & City Science, 53(4), 2024.

[16] World Bank, "World Development Indicators," 2025.

[17] World Bank, "SDG Indicators and Metadata," 2021.

[18] J. Yang et al., "Automating the Construction of Environmental Policy Knowledge Graph with LLMs," Sustainability, vol. 17, no. 22, 10282, 2025.

[19] C. Xu et al. "Context graph." arXiv:2406.11160 (2024).

[20] Y. Zhu et al., "LLM-Based Multi-Agent Orchestration: A Survey of Frameworks, Communication Protocols, and Emerging Patterns," Future Internet, vol. 18, no. 6, 326, 2026.