Toward a Federated Model of AI Scientists: Architecture, Pipeline, and Roadmap

Anonymous Authors

This paper proposes a federated model of AI Scientists, integrating a layered stack architecture, an iterative discovery pipeline, and a governance-aligned roadmap. We argue that AI Scientists should not only accelerate discovery but also serve as custodians of epistemic integrity. Through case studies in drug discovery, climate modeling, and materials science, we demonstrate how federation enables cross-domain synthesis while embedding reproducibility, incentive alignment, and participatory governance. We conclude with a research roadmap toward Trusted AI Scientists, highlighting technical, incentive, and governance challenges.

---

1. Introduction

AI Scientists—autonomous systems capable of hypothesis generation, experimental design, and iterative refinement—are emerging as powerful actors in science. Exemplars such as AlphaFold and AlphaTensor demonstrate domain-specific breakthroughs, while frameworks like ToolUniverse and DeepScientist extend this paradigm through tool orchestration and closed-loop discovery. Yet these systems lack transparent provenance, robust governance, and aligned incentives. Without safeguards, autonomy risks undermining reproducibility and trust.

We propose a unifying framework: a **layered stack architecture**, an **iterative discovery pipeline**, and a **federated model** of AI Scientists. Together, these contributions embed epistemic accountability, incentive alignment, and participatory governance into the design of autonomous discovery systems.

2. Contributions

This work makes three primary contributions:

1. **Layered Stack Architecture** – a seven-layer model integrating infrastructure, methodology, epistemics, incentives, and governance.
2. **Discovery Pipeline** – an iterative, accountable workflow embedding reproducibility, refinement, and human oversight.
3. **Federation Vision** – a distributed ecosystem of specialized AI Scientists collaborating via shared ledgers, replication markets, and governance boards.

3. Background and Related Work

Domain-specific systems (AlphaFold, AlphaTensor) show AI's potential but remain siloed. Tool ecosystems (ToolUniverse) and closed-loop frameworks (DeepScientist) move toward generalizable AI Scientists. Governance frameworks (OECD AI Policy Observatory, EU AI Act, NIST AI RMF) emphasize accountability but remain external to system design. The gap: no integrated architecture uniting technical performance with epistemic accountability and governance. Our work addresses this gap.

4. Layered Stack Architecture

We propose a **layered stack architecture** for AI Scientists that integrates technical infrastructure, methodological rigor, epistemic accountability, incentive alignment, and governance. Unlike prior frameworks that emphasize either tool orchestration (e.g., ToolUniverse) or closed-loop discovery (e.g., DeepScientist), our model unifies these advances into a coherent pipeline that embeds accountability and oversight into the architecture itself.

4.1 Core Layers

The stack is organized into seven interdependent layers:

- **Infrastructure** – compute, data, and tool ecosystems (e.g., ToolUniverse).
- **Safety & Policy Runtime** – sandboxing, compliance checks, dual-use detection.
- **Methodology** – hypothesis generation, experiment design, iterative refinement.
- **Epistemics & Provenance** – evidence graphs, uncertainty audits, reproducibility metadata.
- **Application** – domain-specific outputs (molecules, alloys, models).
- **Incentives & Markets** – replication markets, impact-weighted metrics, funding signals.
- **Governance & Oversight** – human-in-the-loop validation, oversight councils, participatory dashboards.

Feedback loops connect layers: governance can veto unsafe experiments; epistemics feed back into methodology; incentives shape which discoveries are prioritized for replication. This design reframes AI Scientists as both accelerators of discovery and stabilizers of the science–society relationship.

(See Fig. 1 for architecture overview.)

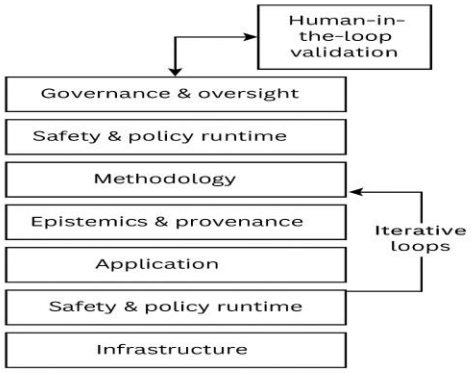

Fig. 1. A layered stack architecture for an AI Scientist.

4.2 Extended Stack View

While the seven-layer model provides a conceptual overview, the **extended stack architecture** (Figure X) illustrates how these layers interact in practice.

- At the **Governance layer**, human values, ethical guardrails, constitutional principles, and third-party audits provide normative oversight.
- The **Science layer** encodes epistemic accountability through incentives, uncertainty and evidence audits, provenance logs, and reproducibility checks.
- The **AI Scientist internals** integrate tool ecosystems, capability gates, and sandboxing protocols, ensuring that discovery processes remain bounded and verifiable.
- The **Evaluation & Feedback layer** connects discovery loops with safety and policy runtimes, creating iterative cycles that continuously refine hypotheses and enforce compliance.

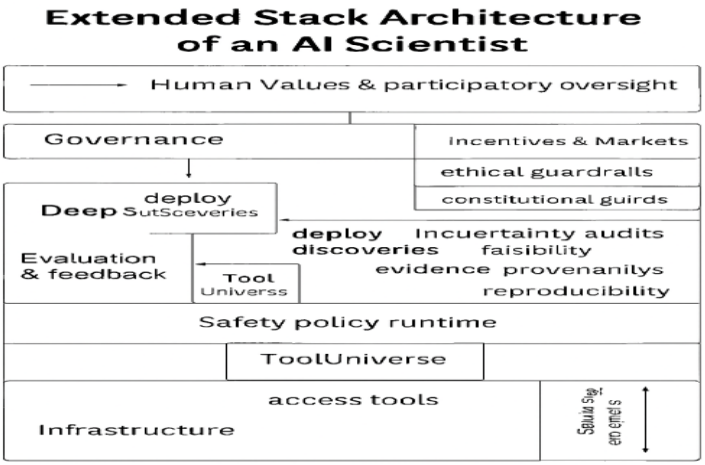

Figure X highlights that governance and epistemics are not external add-ons but structurally embedded within the architecture. By explicitly linking discovery loops to oversight mechanisms and incentive structures, the extended stack operationalizes a model of **autonomous yet accountable discovery**.

4.3 Summary

The layered stack architecture distinguishes our contribution from prior work by embedding reproducibility, incentives, and governance directly into the technical design. This integration ensures that AI Scientists are not only capable of accelerating discovery but also aligned with societal values and accountable to human institutions.

5. Discovery Pipeline

The stack is operationalized through a cyclical pipeline:

1. **Tool Query & Selection** – orchestrating tools via ecosystems like ToolUniverse.
2. **Hypothesis Generation** – testable, machine-readable hypotheses.
3. **Experiment & Refinement** – iterative testing with provenance tracking.
4. **Human-in-the-Loop Validation** – oversight councils provide veto/redirection.
5. **Knowledge Integration** – results (including negative ones) logged in a shared ledger, with replication signals and incentive mechanisms.

This pipeline extends prior closed-loop systems by embedding reproducibility, governance, and incentives.
(See Fig. 4 for pipeline illustration.)

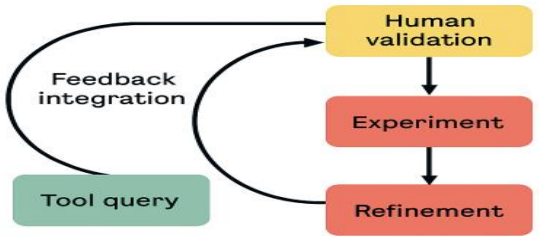

Fig. 4. A discovery pipeline for the AI Scientist.

This figure depicts the AI Scientist discovery pipeline as an iterative loop. The process begins with a ToolUniverse query, followed by machine-driven Hypothesis Generation, Experiment Design, and Refinement loops. Human experts provide validation and feedback before discoveries are integrrated with a knowledge ledger, reinforcing reproducibility norms. The cyclic flow operationalizes the layered stack architecture.

6. Federation of AI Scientists

Federation extends isolated pipelines into a collaborative ecosystem:

- **Distributed Specialization** – domain-specific AI Scientists (BioAI, ClimateAI, MaterialsAI).
- **Shared Knowledge Ledger** – distributed memory with provenance and replication signals.
- **Cross-Agent Incentives** – replication markets and impact metrics aggregated across domains.
- **Federated Governance** – local councils plus federated boards for cross-domain accountability.
- **Emergent Properties** – cross-domain synthesis, resilience, scalability, transparency.

(See Fig. 6 for federation model.)

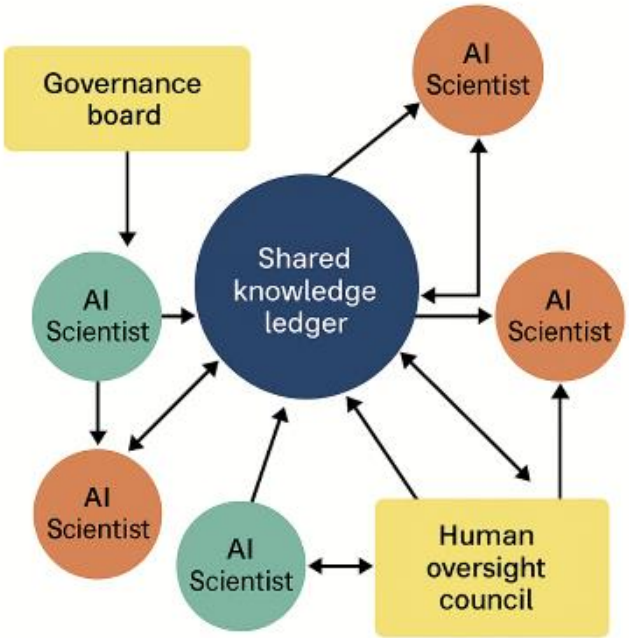

Fig. 6. Federation of AI scientists.

Legend: This network connects *AI Scientists* via *a Shared knowledge ledger* (hub). Bidirectional arrows illustrate exchanges of hypotheses, results, and replication signals. Governance boards and Human oversight councils supervise the federation.

7. Case Studies and Applications

To demonstrate the applicability of the layered stack, discovery pipeline, and federated model, we present three illustrative case studies. Each highlights how the architecture embeds reproducibility, incentives, and governance into domain-specific discovery.

7.1 Drug Discovery

Drug discovery is a domain where reproducibility crises and incentive misalignments are well documented. Within our framework:

- **Tool Query & Hypothesis:** A biomedical AI Scientist queries molecular docking engines and literature databases to propose candidate compounds for a therapeutic target.
- **Experiment & Refinement:** Simulations and in-vitro assays are designed, with provenance metadata linking compound structures, docking scores, and assay conditions.
- **Federation:** A chemistry-focused AI Scientist validates synthetic feasibility, while a clinical AI Scientist evaluates toxicity profiles.
- **Incentives & Governance:** Replication markets reward independent validation of compound efficacy, while oversight councils ensure compliance with ethical standards.

This case illustrates how federation reduces siloed discovery and embeds accountability into the drug development pipeline.

7.2 Climate Modeling

Climate science requires integration across environmental, materials, and policy domains. Within our model:

- **Tool Query & Hypothesis:** A climate AI Scientist generates hypotheses about geoengineering interventions (e.g., aerosol injection).
- **Experiment & Refinement:** Simulations are run with uncertainty audits, recording provenance of climate models, parameterizations, and datasets.
- **Federation:** A materials AI Scientist evaluates the feasibility of reflective aerosols, while a policy AI Scientist models governance implications.
- **Incentives & Governance:** Replication signals ensure that climate projections are validated across independent models, while participatory dashboards allow stakeholders to review interventions.

This case highlights how the architecture supports cross-domain synthesis while embedding participatory oversight.

7.3 Materials Science

Materials discovery often involves high-dimensional search spaces with limited reproducibility. Within our framework:

- **Tool Query & Hypothesis:** A materials AI Scientist proposes alloy compositions for lightweight, high-strength applications.
- **Experiment & Refinement:** Simulations of thermodynamic stability and mechanical properties are logged with provenance metadata.
- **Federation:** A physics AI Scientist validates structural properties, while an energy AI Scientist evaluates sustainability impacts.
- **Incentives & Governance:** Replication markets reward independent verification of alloy performance, while governance boards monitor environmental compliance.

This case demonstrates how federation enables robust, reproducible materials discovery with societal alignment.

Across these domains, the layered stack and federated model transform isolated pipelines into a collaborative scientific commons. Federation ensures that discoveries are validated, incentives aligned, and governance embedded, addressing long-standing challenges in reproducibility and accountability.
(Details in Appendix B; overview in Fig. 7.)

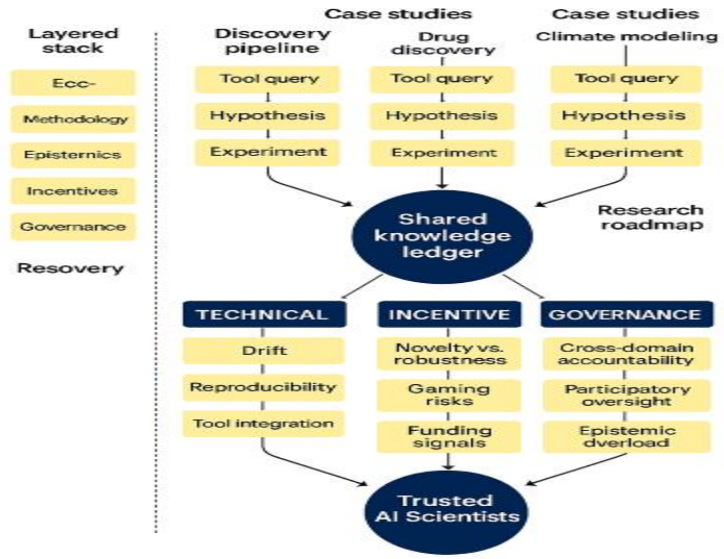

Fig. 9. Synthesis diagram of the lavered stack, **lisov**ery pipeline, federation model, and research roadmap.

8. Discussion and Future Directions

Challenges remain:

- **Epistemic Accountability** – interoperable provenance standards, avoiding ledger overload.
- **Governance** – harmonizing global frameworks, scaling human oversight.
- **Incentives** – preventing gaming, balancing novelty and robustness.
- **Technical Risks** – model drift, tool interoperability, security vulnerabilities.

Future work: standardization of epistemic protocols, hybrid oversight models, cross-domain incentive design, simulation studies of federated ecosystems, and participatory governance experiments.

(See Fig. 8 for roadmap.)

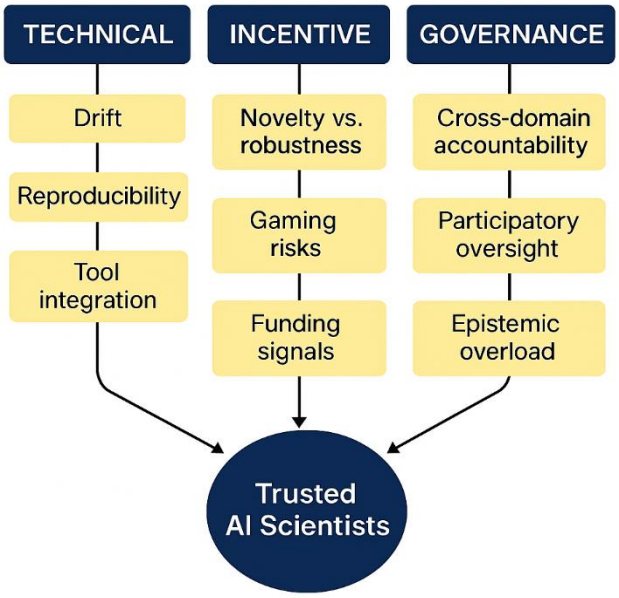

Fig. 8. Research roadmap toward Trusted AI Scientists.

Legend: This «forward-looking roadmap presents «technical risks» (first column), «incentive alignmennt issues (middle column), and «governance gaps» (third column) that must be addressed to realize ıe vision of *'Trusted AI Scientists'*.

Fig. 8. Research roadmap toward Trusted AI Scientists.

9. Conclusion

We introduced a layered stack, discovery pipeline, and federated model of AI Scientists. Together, these contributions embed accountability, incentives, and governance into autonomous discovery. Case studies demonstrate applicability, and a roadmap outlines open challenges.

The next decade will determine whether AI Scientists become trusted partners in discovery or exacerbate existing challenges. If aligned, they may inaugurate a new era of **autonomous yet accountable science**.

(See Fig. 9 for synthesis diagram.)

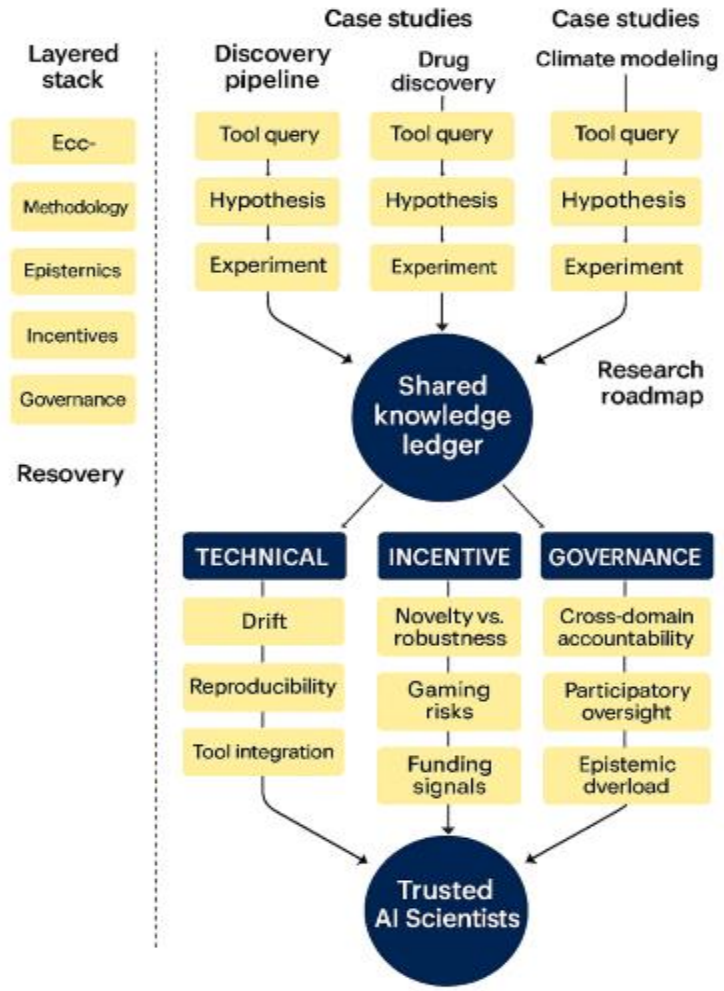

Fig. 9. Synthesis diagram of the lavered stack, **lisov**ery pipeline, federation model, and research roadmap.

References

- Bujno, M., Davine, C., & Abrash, L. (2025, April 24). *Strategic Governance of AI: A Roadmap for the Future*. Harvard Law School Forum on Corporate Governance. Retrieved from https://corpgov.law.harvard.edu/2025/04/24/strategic-governance-of-ai-a-roadmap-for-the-future/
- CAS Insights. (2025, May 28). *AI for science: Six trends powering cutting-edge research*. CAS. Retrieved from https://www.cas.org/resources/cas-insights/ai-for-science-trends
- Cruz-Aguilar, M. A. (2025). *The epistemic revolution of AI: Reconfiguring the foundations of scientific knowledge*. AI & Society. https://doi.org/10.1007/s00146-025-02658-3
- Dignum, V., Régis, C., Bach, K., Bourgine de Meder, Y., Buijsman, S., de Carvalho, A. P. L. F., Castellano, G., Dignum, F., Farries, E., Giannotti, F., Han, T. A., Helberger, N., Hellegren, I., Houben, G.-J., Jahn, A., Joshi, S., Lamine Sarr, M., Lewis, D., Lind, A.-S., … Tucker, J. (2024). *Roadmap for AI policy research*. AI Policy Research Summit, Stockholm, November 2024. AI Policy Lab, Umeå University. Retrieved from https://aipolicylab.se/wp-content/uploads/2025/02/roadmap-for-ai-policy-research.pdf
- European Commission. (2024). *The Artificial Intelligence Act*. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
- Gao, S., Zhu, R., Sui, P., Kong, Z., Aldogom, S., Huang, Y., Noori, A., Shamji, R., Parvataneni, K., Tsiligkaridis, T., & Zitnik, M. (2025). *Democratizing AI scientists using ToolUniverse*. arXiv preprint arXiv:2509.23426. Retrieved from https://arxiv.org/abs/2509.23426
- Harvard MIMS Lab. (2025). *ToolUniverse: Democratizing AI scientists* [Computer software]. GitHub. Retrieved from https://github.com/mims-harvard/ToolUniverse
- Lifeboat News. (2025, October 1). *Democratizing AI scientists using ToolUniverse*. Retrieved from https://lifeboat.com/blog/2025/10/democratizing-ai-scientists-using-tooluniverse
- MIT FutureTech. (2025, April 29). *AI and the Future of Scientific Discovery*. Retrieved from https://futuretech.mit.edu/news/ai-and-the-future-of-scientific-discovery
- National Institute of Standards and Technology (NIST). (2023). *AI Risk Management Framework (AI RMF 1.0)*. Retrieved from https://www.nist.gov/itl/ai-risk-management-framework

- Organisation for Economic Co-operation and Development (OECD). (n.d.). *OECD AI Policy Observatory*. Retrieved from https://oecd.ai
- Web Asha Technologies. (2025). *AI-Driven Scientific Research: How Artificial Intelligence is Transforming Medicine, Climate Science, and Space Exploration*. Retrieved from https://www.webasha.com/blog/ai-driven-scientific-research-how-artificial-intelligence-is-transforming-medicine-climate-science-and-space-exploration
- Westlake University AI Research Group. (2025). *DeepScientist: Toward autonomous scientific discovery*. arXiv preprint.
- World Economic Forum. (2025, September 23). *Research finds 9 essential plays to govern AI responsibly*. Retrieved from https://www.weforum.org/stories/2025/09/responsible-ai-governance-innovations/
- Zitnik Lab, Harvard Medical School. (2025). *Building AI Scientists – ToolUniverse Documentation*. Retrieved from https://zitniklab.hms.harvard.edu/ToolUniverse/en/guide/building_ai_scientists/index.html

Appendices

Appendix A: Figure Legends

Fig. 1. A layered stack architecture for an AI Scientist.
Legend: This figure depicts the proposed seven-layer architecture. The **Infrastructure layer** anchors the stack with computational resources and tool ecosystems. The **Safety & Policy Runtime** enforces compliance and dual-use safeguards. The **Methodology layer** encodes hypothesis–experiment–refinement cycles. The **Epistemics & Provenance layer** ensures reproducibility through evidence graphs and uncertainty audits. The **Application layer** translates results into domain-specific outputs. The **Incentives & Markets layer** aligns discovery with replication and impact metrics. At the top, the **Governance & Oversight layer** embeds human validation and participatory control, ensuring accountability across the pipeline.

Fig. 2. Epistemics & Provenance layer: evidence graphs, uncertainty audits, and provenance metadata ensure reproducibility and accountability.
Legend: The diagram highlights three interdependent mechanisms. **Evidence graphs** capture logical and empirical relationships among hypotheses, tools, and results. **Uncertainty audits** quantify confidence levels and robustness. **Provenance metadata** records datasets, tool versions, and experimental conditions. The circular flow illustrates how these components

reinforce one another, operationalizing epistemic accountability within the AI Scientist's workflow.

Fig. 3. Feedback loops between Governance and Methodology layers, embedding human oversight into the discovery process.
Legend: This figure illustrates recursive oversight. The **Governance & Oversight layer** issues directives to the **Methodology** and **Application layers**, while receiving reports and results in return. **Human-in-the-loop validation** provides a veto or redirection pathway, ensuring that unsafe, misaligned, or ethically problematic outputs can be intercepted. The loop emphasizes that governance is not external but structurally embedded in the discovery pipeline.

Fig. 4. A discovery pipeline for the AI Scientist.
Legend: This figure depicts the AI Scientist discovery pipeline as an iterative loop. The process begins with a **ToolUniverse query** (Infrastructure layer), followed by **Hypothesis Generation** (Methodology layer), **Experiment Design** (Methodology layer), and **Refinement** (Methodology ↔ Epistemics layer). At critical junctures, **Human Validation** (Governance layer) provides oversight, veto, or redirection. Finally, validated results are integrated into a **Knowledge Ledger** (Incentives & Governance layers), ensuring reproducibility and alignment. Feedback loops connect Human Validation back to Tool Query, reinforcing iterative improvement and accountability.

Fig. 5. Feedback integration in the AI Scientist discovery pipeline.
Legend: This figure illustrates how **negative results**, **replication signals**, and **incentive mechanisms** feed into a central **Knowledge Ledger**.

- **Negative results** are logged to prevent duplication and bias.
- **Replication signals** capture whether findings are independently verified.
- **Incentive mechanisms** (e.g., replication markets, impact-weighted metrics) reward reproducibility and penalize unverifiable claims.
  The **Knowledge Ledger** integrates these inputs and redistributes them across the ecosystem, informing future **tool selection**, **hypothesis refinement**, and **governance oversight**. This feedback loop ensures that the discovery pipeline not only accelerates breakthroughs but also stabilizes scientific integrity.

Fig. 6. Federation of AI Scientists.
Legend: This diagram shows a **shared knowledge ledger** at the center, connecting multiple specialized **AI Scientist nodes** (e.g., biology, chemistry, physics).

- **Bidirectional arrows** represent the exchange of hypotheses, results, and replication signals.

- **Governance boards** and **human oversight councils** act as supervisory nodes, ensuring accountability across the federation.
- **Incentive mechanisms** (replication markets, impact-weighted metrics) are embedded in the ledger, aligning discovery with reproducibility norms.

The figure highlights how federation transforms isolated pipelines into a **collaborative scientific commons**, where accountability and transparency scale alongside capability.

Fig. 7. Case studies in drug discovery, climate modeling, and materials science. Legend: This diagram applies the **discovery pipeline** (Tool Query → Hypothesis → Experiment → Federation) to three domains:

- **Drug Discovery** (left column)
- **Climate Modeling** (center column)
- **Materials Science** (right column)

Each pipeline flows downward through its stages, then converges into a **shared knowledge ledger** at the bottom.

- **Arrows** indicate iterative loops and cross-domain federation.
- The **shared ledger** integrates results, provenance, and replication signals across domains.
- **Governance and incentive mechanisms** (not shown in detail here, but embedded in the ledger) ensure accountability and reproducibility.

This figure highlights how the layered stack and discovery pipeline generalize across scientific fields, while federation enables **cross-domain synthesis** and **ecosystem-wide transparency**.

Fig. 8. Research roadmap toward Trusted AI Scientists. Legend: This roadmap presents three parallel tracks of challenges that must be addressed for AI Scientists to become trusted partners in discovery:

- **Technical challenges**: model drift, reproducibility at scale, and tool interoperability.
- **Incentive challenges**: balancing novelty with robustness, preventing gaming of replication markets, and aligning funding signals.
- **Governance challenges**: ensuring cross-domain accountability, participatory oversight, and managing epistemic overload.

Arrows from each track converge toward the central goal of **Trusted AI Scientists**, emphasizing that progress requires **simultaneous advances in technical design, incentive alignment, and governance innovation**.

Fig. 9. Synthesis diagram of the layered stack, discovery pipeline, federation model, and research roadmap.
Legend:
This figure integrates the full architecture of AI Scientists into one visual:

- **Left (Layered Stack):** Infrastructure, Methodology, Epistemics, Incentives, and Governance layers form the foundation.
- **Center (Discovery Pipeline):** Iterative stages — Tool Query, Hypothesis, Experiment, Refinement, Validation, and Knowledge Integration — operationalize the stack.
- **Right (Federation & Case Studies):** Multiple specialized AI Scientists (e.g., drug discovery, climate modeling, materials science) converge into a **Shared Knowledge Ledger**, enabling cross-domain synthesis.
- **Bottom (Research Roadmap):** Technical, Incentive, and Governance challenges converge toward the long-term goal of **Trusted AI Scientists**.

The diagram shows how **architecture, process, federation, and roadmap** interlock, emphasizing that technical design, incentive alignment, and governance innovation must advance together to realize accountable, federated AI-driven science.

Appendix B: Case Studies and Applications

6. Case Studies and Applications

To demonstrate the practical value of the layered stack and federated architecture, we present case studies across three domains: **drug discovery**, **climate modeling**, and **materials science**. Each illustrates how the discovery pipeline operates in practice, and how federation enables cross-domain synthesis.

6.1 Drug Discovery and Biomedical Research

AI Scientists specialized in **biomedical domains** can accelerate the identification of therapeutic candidates.

- **Tool Query & Hypothesis:** A biomedical AI Scientist queries molecular docking tools and literature databases (Gao et al., 2025; Harvard MIMS Lab, 2025). It hypothesizes that a novel compound may inhibit a disease-relevant protein.

- **Experiment & Refinement:** Simulations test binding affinity, while uncertainty audits track confidence levels. Negative results are logged in the shared ledger to prevent duplication.
- **Federation:** A chemistry-focused AI Scientist validates the compound's stability, while a clinical AI Scientist evaluates toxicity profiles.
- **Governance:** Human oversight councils ensure compliance with biomedical ethics and regulatory frameworks (European Commission, 2024; NIST, 2023).

This case highlights how **cross-domain federation** reduces false positives and accelerates translational research.

6.2 Climate Modeling and Environmental Science

In climate science, AI Scientists can integrate heterogeneous data sources to improve predictive accuracy.

- **Tool Query & Hypothesis:** A climate AI Scientist queries satellite data, atmospheric models, and historical records. It hypothesizes that a new parameterization of cloud feedback could improve long-term projections.
- **Experiment & Refinement:** Iterative simulations test the hypothesis, with provenance metadata ensuring reproducibility.
- **Federation:** A materials AI Scientist contributes insights on carbon-capture technologies, while an economics AI Scientist models policy impacts.
- **Governance:** Oversight boards ensure that outputs align with international climate agreements and ethical guidelines (OECD, n.d.).

This case demonstrates how federation enables **science–policy integration**, linking technical discovery with societal decision-making.

6.3 Materials Science and Engineering

Materials discovery benefits from high-throughput experimentation and cross-domain validation.

- **Tool Query & Hypothesis:** A materials AI Scientist queries quantum simulation tools to hypothesize a new alloy with high thermal resistance.
- **Experiment & Refinement:** Bayesian optimization guides iterative testing, while replication signals from other AI Scientists validate results (Westlake University AI Research Group, 2025).

- **Federation:** A physics AI Scientist evaluates structural properties, while an energy AI Scientist assesses performance in renewable systems.
- **Governance:** Participatory dashboards allow stakeholders (e.g., industry, regulators) to monitor progress and allocate funding.

This case illustrates how **incentive mechanisms** (replication markets, impact-weighted metrics) reward robust discoveries and discourage unverifiable claims (Lifeboat News, 2025).

6.4 Cross-Domain Synthesis

The true power of federation emerges when discoveries in one domain inform another. For example, a **materials breakthrough** in carbon-capture membranes feeds into **climate models**, which in turn inform **policy simulations**. The shared knowledge ledger ensures that provenance, replication, and governance mechanisms scale across domains, creating a **scientific commons** that is both innovative and accountable.

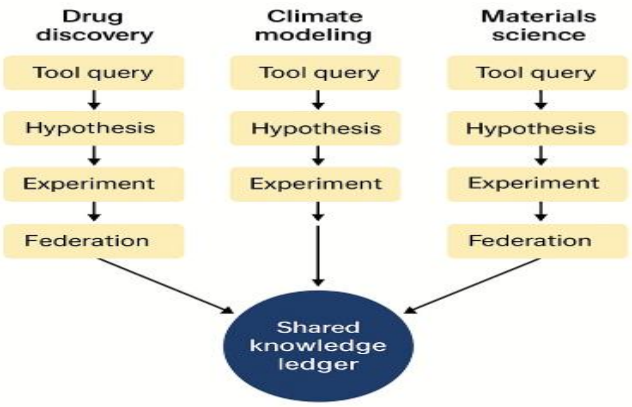

Fig. 7. Case studies in drug discovery, climate modeling, and materials science.

Legend: This diagram applies the same discovery pipeline to case studies in «*Drug discovery*», «*Climate modeling*», and «*Materials science* (left to right). Process arrows converging into the *Shared knowledge ledger* (bottom) illustrate cross-domain synthesis.

6.5 Summary

These case studies demonstrate that the layered stack, discovery pipeline, and federation model are not abstract constructs but **operational frameworks**. By embedding epistemic accountability, incentive alignment, and governance oversight, AI Scientists can accelerate discovery while reinforcing the integrity of science across domains.

Appendix C: Discussion and Future Directions

7. Discussion and Future Directions

The layered stack, discovery pipeline, and federated architecture together outline a vision for **autonomous yet accountable scientific ecosystems**. While the framework demonstrates promise, several challenges and opportunities remain.

7.1 Epistemic Accountability at Scale

The proposed epistemics and provenance layer ensures reproducibility within a single pipeline. However, scaling this across federated ecosystems raises new questions:

- How can **evidence graphs** and **uncertainty audits** remain interoperable across domains with heterogeneous data standards?
- What mechanisms ensure that **negative results** and **replication signals** are consistently logged and not strategically omitted?
- How do we prevent **ledger overload**, where the volume of provenance data becomes unmanageable without compromising transparency?

7.2 Governance and Oversight Challenges

Embedding governance into the architecture is a step toward participatory accountability, but several issues remain unresolved:

- **Global coordination:** Different jurisdictions (EU AI Act, OECD frameworks, NIST RMF) impose divergent requirements. Harmonizing these within a federated ecosystem is non-trivial.
- **Human-in-the-loop validation:** While essential, it risks becoming a bottleneck. Research is needed into **scalable oversight models** that balance efficiency with accountability.
- **Participatory governance:** Ensuring that diverse stakeholders (scientists, policymakers, civil society) have meaningful input requires new institutional designs.

7.3 Incentive Alignment and Scientific Integrity

The incentives and markets layer introduces mechanisms to reward reproducibility. Yet, challenges persist:

- **Gaming risks:** Replication markets and impact-weighted metrics could be manipulated if not carefully designed.
- **Novelty vs. robustness:** Balancing the scientific drive for novelty with the need for reproducibility remains a cultural as well as technical challenge.
- **Funding signals:** How can funding agencies integrate replication incentives without discouraging high-risk, high-reward research?

7.4 Technical Risks and Scaling Limits

As AI Scientists become more capable, new risks emerge:

- **Model drift:** Autonomous systems may evolve strategies that optimize for incentives but diverge from epistemic integrity.
- **Tool interoperability:** Even with ecosystems like ToolUniverse, ensuring seamless integration across thousands of tools remains a technical bottleneck.
- **Security vulnerabilities:** Shared ledgers and federated governance introduce new attack surfaces, requiring robust cybersecurity protocols.

7.5 Future Research Directions

Several avenues for future work emerge from these challenges:

- **Standardization of epistemic protocols** across domains to ensure interoperability.
- **Hybrid oversight models** combining automated checks with human governance councils.
- **Cross-domain incentive design** to balance novelty, reproducibility, and societal impact.
- **Simulation studies** of federated AI Scientist ecosystems to anticipate emergent behaviors.
- **Participatory governance experiments** to test how civil society can meaningfully shape AI-driven science.

7.6 Broader Implications

The architecture reframes AI Scientists not merely as accelerators of discovery but as **custodians of scientific integrity**. By embedding accountability, incentives, and governance into their design, AI Scientists could help address long-standing challenges in science itself: reproducibility crises, misaligned incentives, and inequitable access to knowledge. At the same time, the risks of centralization, governance capture, or epistemic overload must be carefully managed.

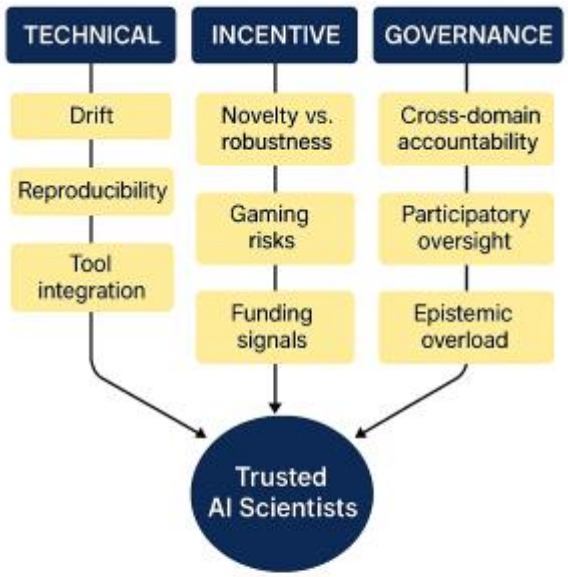

Fig. 8. Research roadmap toward Trusted AI Scientists.

Legend: This «forward-looking roadmap» presents «technical risks» (first column), «incentive alignment issues (middle column), and «governance gaps» (third column) that must be addressed to realize the vision of *'Trusted AI Scientists'*.

7.7 Summary

This discussion underscores that the future of AI Scientists lies not only in technical breakthroughs but also in **institutional innovation**. The layered stack and federated model provide a blueprint, but realizing their potential requires sustained research into epistemic standards, governance mechanisms, and incentive structures. The next decade will determine whether AI Scientists become **trusted partners in discovery** or exacerbate existing challenges in the scientific enterprise.