PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mba-9eoh-xj9p
Comments Due: May 28, 2025 **Submission Type:** Web

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0348
Comment on FR Doc # 2025-07332

Submitter Information

Organization: Texas A&M Engineering Experiment Station

General Comment

See attached file(s)

Attachments

NSF-RFI-TEES

Input from Texas A&M Engineering Experiment Station (TEES)

Rapid advances in Artificial General Intelligence (AGI), with recent successes in large language models (LLMs), have raised new questions whether they can help and soon lead scientific discovery to significantly accelerate progress in areas critical for U.S. competitiveness in science, engineering, healthcare, and public policy making.

General-purpose AGI agents can observe multi-modal signals, manipulate samples, launch simulations, and revise hypotheses in a single, self-correcting loop. But current agentic systems largely remain *single* monolithic models interacting one-on-one with users. To tackle society's most complex problems, new **collaborative and organizational AI** – systems composed of multiple specialized AI agents working in concert, and the accompanying research efforts, should move beyond isolated chatbots towards *agent societies* that can **perceive, reason, and act collaboratively**. For scientific discovery, this prompts the following question: How can we transform fragmented, manually steered workflows into self-improving multi-agent ecosystems that learn, reason, and act on behalf of and collaboratively with human scientists?

The following foundational research efforts, currently ongoing at **Texas A&M**, may help achieve such a paradigm-shifting vision.

Integrating Knowledge, Data, Computing, and AI/ML: Scientific discovery naturally is an explorative process, which in principle differs from the data-centered developments of LLMs and AI agents. While exploring in the dark for new scientific findings, lack of relevant data and knowledge is a common challenge to overcome. Integration of relevant knowledge, data, and computational tools is essential to developing collaborative AI agentic systems. Appropriate knowledge representation, modeling, and data fusion into efficient AI/ML models via appropriate optimization formulation targeting directly at ultimate operational objective can help derive effective agent systems to accelerate scientific discovery[1].

Knowledge Representation and Hypothesis Generation: In scientific domains, foundational knowledge (e.g. experimental results, theoretical models, databases of molecules or materials) can be vast and complex[2]. AI agents must represent this knowledge in forms they can reason with – representation learning and knowledge graphs are key enabling tools for developing effective and efficient AI agents that can generate reliable predictions for optimal decision making. By mining and reasoning the learned representations and knowledge relationships, agents can also suggest open-ended hypotheses that a human might miss, combining disparate findings into a novel conjecture. The goal is to move toward AI systems capable of generating and evaluating scientific hypotheses autonomously, rather than just answering questions. Encouraging results have emerged in using LLMs to propose research ideas when guided by structured knowledge. Future agents may patrol the ever-growing literature, identifying gaps or inconsistencies, and proactively suggest experiments to resolve them.

Multi-Agent Scientific Teams: Rather than a single AI doing all tasks, a collaborative approach could have dedicated agents for specialized tasks but working as a team. For instance, a "Theory Agent" could use domain knowledge to ensure new hypotheses are physically plausible, while a "Data Agent" scours through prior datasets for identifying patterns supporting or contradicting a hypothesis. They would communicate in a shared scientific language, perhaps augmenting natural language with formal notations (e.g., chemistry formulas, math equations) as needed. By exchanging information, they can collectively ensure that a proposed experiment is novel, grounded in theory, and likely to produce informative results. Scientific knowledge graphs would function as a common memory for the agents, where each new result

or insight gets added to this structured repository, which agents query when reasoning. Over time, the AI ensemble can build up an increasingly sophisticated model of the science domain.

Autonomous Experimentation: One of the most exciting developments is the rise of self-driving laboratories – automated lab setups where AI plans and executes experiments in a closed loop, possibly collaborating with robots and human experts. In such a setup, a "Planning Agent" might decide on a new experiment (e.g., a new material composition to synthesize or a new variable to test) based on prior results, followed by a "Robot Control Agent" running the experiment with lab robotics, and a "Data Analysis Agent" assessing the outcome, feeding results back for the next cycle. This loop can continue with minimal human intervention, dramatically accelerating discovery. Notably, researchers at Texas A&M have demonstrated that automated experimentation can be achieved in both novel materials discovery and bioinformatics data analysis tasks[1,2]. By integrating simulation models and physical experiments, these agent-driven labs can be generalized to broader science, engineering, and biomedicine applications.

Robustness Under Uncertainty: AI agents must operate robustly under uncertainty, incomplete information, and changing environments. Agents therefore need to handle distributional shift and be aware of what they know and do not know. Approaches include incorporating Bayesian learning, reasoning, experimental design for agents to express uncertainty about their conclusions, and to adapt on the fly to new conditions[1,3,4]. Probabilistic verification and fault tolerance mechanisms from distributed computing can be applied. Research into emerging behavior, uncertainty propagation in complex systems, as well as optimization under uncertainty may help achieve end-to-end robustness – ensuring the multi-agent system can maintain performance when parts of it are uncertain or faced with novelty – will be key to deploying these systems in the real world, where unpredictability is the norm.

Safety and Value Alignment: When multiple agents autonomously make decisions, the risk of unintended or unsafe outcomes can be amplified. Ensuring alignment – that agents' actions remain within human-approved bounds and aligned with our objectives – is a paramount challenge[5]. In multi-agent contexts, novel failure modes appear: miscoordination (agents share goals but fail to cooperate effectively), conflict (agents work at odds due to misaligned goals), or even undesirable collusion (agents cooperating in ways that harm humans, e.g. price-fixing behaviors in market simulations). Avoiding these outcomes requires serious research efforts. For example, multi-agent reinforcement learning with safety-aware reward shaping can steer agents away from unsafe joint behaviors. Techniques from adversarial training can be applied where possible: e.g., training some agents as adversaries to probe the weaknesses of the system (red teaming). Additionally, alignment in the multi-agent case isn't just about *human-AI* alignment, but also agent-agent alignment with each other's sub-goals to avoid inner conflicts. Developing reward structures or oversight mechanisms that ensure a *team of agents remains aligned to the overarching human objective* (and doesn't get sidetracked optimizing a proxy goal) is an open problem in cooperative AI.

Transparency and Explainability: For high-stakes applications of AI agents, it's vital that the multi-agent system can explain its decisions and actions in a way humans can understand. This is challenging because an outcome might be the result of a complex chain of inter-agent dialogues and decisions. Explainability also relates to user trust. Achieving a balance between the raw efficiency of autonomous agents and the necessary transparency for human oversight is a key design consideration. Techniques from XAI (explainable AI) need to be extended to multi-agent contexts, possibly by having some agents explicitly tasked with monitoring and explaining the rest of the system's behavior to humans.

Collaborative and organizational AI agent systems hold immense promise for scientific advancement, engineering applications, modern healthcare, evidence-based policy-making and societal governance. Yet, their **safe deployment and verification** is an equally important frontier. Unlike single algorithms, multi-agent ecosystems exhibit *emergent behaviors* and complex dynamics that can be difficult to predict or control. Ensuring these systems are **reliable, aligned, and verifiable** calls for a multi-disciplinary approach drawing on control theory, optimization, game theory, and rigorous software engineering. From a *verification and systems engineering* perspective, formal methods will also play a key role. Additionally, **testing and simulation at scale** is crucial – before deploying a healthcare multi-agent system, one might simulate thousands of patient scenarios to statistically verify that errors are below an acceptable threshold. **Co-design** is another important principle: building safe AI is not just about post hoc verification, but designing the AI ecosystem (hardware, software, data, and algorithms), and its operating environment *together* so that safety is inherently assured. This means involving domain experts, ethicists, and engineers in the design loop – effectively encoding domain guidelines (e.g., medical protocols, legal regulations) into the system's requirements from day one, rather than retrofitting them later. The fusion of advanced AI with rigorous mathematics and principled engineering will be critical as we embark on deploying multi-agent AI in high-stakes domains, ensuring that these powerful new tools truly serve and benefit society.

References

1. Xiaoning Qian, Byung-Jun Yoon, Raymundo Arroyave, Xiaofeng Qian, Edward R. Dougherty. "Knowledge-driven learning, optimization, and experimental design under uncertainty for materials discovery," ***Patterns***, doi: 10.1016/j.patter.2023.100863, 2023.
2. Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, Keir Adams, Maurice Weiler, Xiner Li, Tianfan Fu, Yucheng Wang, Haiyang Yu, YuQing Xie, Xiang Fu, Alex Strasser, Shenglong Xu, Yi Liu, Yuanqi Du, Alexandra Saxton, Hongyi Ling, Hannah Lawrence, Hannes Stark, Shurui Gui, Carl Edwards, Nicholas Gao, Adriana Ladera, Tailin Wu, Elyssa F. Hofgard, Aria Mansouri Tehrani, Rui Wang, Ameya Daigavane, Montgomery Bohde, Jerry Kurtin, Qian Huang, Tuong Phung, Minkai Xu, Chaitanya K. Joshi, Simon V. Mathis, Kamyar Azizzadenesheli, Ada Fang, Alan Aspuru-Guzik, Erik Bekkers, Michael Bronstein, Marinka Zitnik, Anima Anandkumar, Stefano Ermon, Pietro Lio, Rose Yu, Stephan Gunnemann, Jure Leskovec, Heng Ji, Jimeng Sun, Regina Barzilay, Tommi Jaakkola, Connor W. Coley, Xiaoning Qian, Xiaofeng Qian, Tess Smidt, Shuiwang Ji. "Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems," ***Foundations and Trends in Machine Learning***, under review, 2025.
3. Yucheng Wang, Mingyuan Zhou, Xiaoning Qian. "Hashing with Uncertainty Quantification via Sampling-based Hypothesis Testing," ***Transactions on Machine Learning Research (TMLR)***, 2024.
4. Randy Ardywibowo, Zepeng Huo, Zhangyang Wang, Bobak Mortazavi, Shuai Huang, Xiaoning Qian. "VariGrow: Variational Architecture Growing for Task-Agnostic Continual Learning based on Bayesian Novelty," ***The 39th International Conference on Machine Learning (ICML)***, 2022.
5. Guang Zhao, Byung-Jun Yoon, Gilchan Park, Shantenu Jha, Shinjae Yoo, Xiaoning Qian. "Pareto Prompt Optimization," ***The 13th International Conference on Learning Representations (ICLR)***, 2025.