Hierarchical Planning and Execution for Autonomous Scientific Discovery Agents Using Fast–Slow Reasoning Architectures

Kiran Ganguly
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
contactkiran@uc.edu

Abstract

Autonomous scientific discovery agents represent a frontier in artificial intelligence, aiming to accelerate the generation and validation of knowledge across disciplines. This paper introduces a hierarchical planning and execution framework grounded in the cognitive metaphor of fast and slow reasoning, originally articulated by Kahneman and subsequently adapted into computational architectures. The proposed system integrates a rapid, intuitive pattern-matching module for hypothesis generation and anomaly detection with a deliberate, resource-intensive reasoning module for causal inference, experimental design, and theory refinement. A hierarchical controller governs the interplay between these two modes, prioritizing tasks based on epistemic uncertainty, resource constraints, and long-term scientific goals. We examine the architectural trade-offs between speed and accuracy, the governance mechanisms required to ensure robustness and reproducibility, and the infrastructure demands of deploying such agents at scale. The discussion extends to sustainability concerns, including energy consumption and data stewardship, as well as fairness and policy implications arising from automated science. By situating the framework within the broader history of AI and cognitive science, we argue that fast–slow architectures offer a principled path toward trustworthy and efficient autonomous discovery, provided that careful attention is paid to system-level design, oversight, and alignment with human values. This paper contributes a systems perspective that bridges cognitive modeling, engineering, and socio-technical governance.

Keywords

autonomous scientific discovery, hierarchical planning, fast–slow reasoning, cognitive architecture, AI governance, socio-technical systems.

1. Introduction

The pursuit of scientific discovery has long been a hallmark of human intelligence, yet the accelerating complexity of modern research challenges the capacity of individual researchers to synthesize vast quantities of data, generate novel hypotheses, and design rigorous experiments. Autonomous artificial intelligence agents capable of conducting scientific inquiry have emerged as a promising response to this challenge, with notable successes in domains such as drug discovery, materials science, and genomics [1,2]. These agents, however, must navigate a fundamental tension between the need for rapid exploration of large hypothesis spaces and the demand for deliberate, verifiable reasoning that underlies scientific validity. This tension mirrors the classic distinction between fast, intuitive thinking and slow, analytical thinking described by Kahneman in his dual-process theory of cognition [3]. The integration of such dual-process reasoning into computational systems has inspired a growing

body of research on architectures that combine heuristic pattern recognition with systematic deliberation [4,5]. In this paper, we propose a hierarchical planning and execution framework that operationalizes the fast–slow dichotomy for autonomous scientific discovery agents. We emphasize system-level considerations: how the two reasoning modes interact, how hierarchical control allocates computational resources across planning horizons, and how the resulting architecture can be deployed in a manner that is robust, sustainable, and aligned with societal expectations.

The design of autonomous discovery agents raises unique challenges beyond those encountered in traditional AI planning. Scientific reasoning demands causal understanding, not merely correlation; it requires the ability to design controlled experiments, manage uncertainty, and update beliefs in light of evidence [6,7]. Moreover, the scientific process is inherently social—it relies on peer review, reproducibility, and cumulative knowledge building. An autonomous agent that operates in isolation risks generating findings that are brittle, irreproducible, or disconnected from the broader scientific discourse [8]. Therefore, any practical architecture must embed mechanisms for transparency, interpretability, and external validation. The hierarchical fast–slow framework we describe is intended to address these requirements by separating the rapid generation of candidate hypotheses from the slow, resource-intensive process of verification and refinement, while a supervisory planning layer ensures that the overall trajectory of inquiry remains coherent and goal-directed. The following sections elaborate on the cognitive foundations, architectural components, governance strategies, and policy implications of this approach.

2. Background and Related Work

The concept of dual-process reasoning has a rich history in cognitive science, originating from Simon's notion of bounded rationality and the distinction between algorithmic and reflective modes of thought [9]. Kahneman's influential book popularized the labels System 1 and System 2, where System 1 is fast, automatic, and associative, while System 2 is slow, deliberate, and rule-based [3]. In AI, early attempts to emulate such architectures appeared in the work on hybrid systems, which combined neural networks for pattern recognition with symbolic reasoners for logical inference [4,10]. More recently, deep learning has provided powerful System 1–like capabilities for perceptual tasks, while modern planning algorithms, such as Monte Carlo tree search, embody aspects of System 2 deliberation [11]. However, the direct application of these ideas to scientific discovery has been limited. Most existing scientific AI systems, such as those used for hypothesis generation, rely on monolithic models that either emphasize speed (e.g., large language models suggesting plausible conjectures) or depth (e.g., causal discovery algorithms), but rarely integrate both in a unified, hierarchical manner [12,13].

Hierarchical planning, a classic technique in AI, structures a problem into multiple levels of abstraction, where high-level plans are refined into low-level actions [14]. This approach naturally aligns with the fast–slow dichotomy: high-level strategic decisions (e.g., which research direction to pursue) can be made using slow, deliberative reasoning, while low-level tactical actions (e.g., performing a specific simulation) can be executed rapidly using learned heuristics. In the context of scientific discovery, such hierarchical decomposition is essential because the space of possible experiments is astronomically large, and exhaustive exploration is infeasible [2]. The combination of hierarchical planning with dual-process reasoning offers a means to maintain both breadth and depth in the discovery process. Furthermore, the growing emphasis on AI safety and alignment [15] underscores the need for architectures that

can be audited and controlled. A hierarchical fast–slow design facilitates oversight by providing clear points of intervention at each level of abstraction, enabling human researchers to monitor high-level goals and intervene when necessary. The following section details the cognitive and computational underpinnings of the fast–slow architecture as applied to scientific discovery.

3. Fast–Slow Reasoning Architectures for Scientific Discovery

The core of our proposed system lies in the simultaneous operation of two reasoning modules: a fast, intuitive module and a slow, analytical module. The fast module is responsible for rapidly scanning large volumes of existing scientific literature, experimental data, and simulation outputs to identify patterns, anomalies, and promising hypotheses. It is implemented using deep neural networks, including large language models and variational autoencoders, that have been pre-trained on vast corpora of scientific text and experimental results [5,12]. This module operates with low latency and high throughput, but its outputs are often associative and may lack causal grounding. In contrast, the slow module engages in causal reasoning, formal model construction, and rigorous statistical testing. It employs techniques such as Bayesian inference, causal graph discovery, and symbolic reasoning to evaluate the hypotheses generated by the fast module, design confirmatory experiments, and update the agent's internal knowledge base [7,16]. The slow module is computationally expensive and time-consuming, but it provides the epistemic reliability required for scientific validity.

The interaction between these two modules is governed by a dynamic control mechanism that decides when to trust the fast module's outputs and when to invoke the slow module for deeper analysis. This decision is informed by several factors: the novelty of the hypothesis, the availability of prior evidence, the cost of a false positive, and the current computational budget. For instance, when an agent is exploring a well-understood domain where patterns are stable, the fast module may be sufficient to guide routine experiments. However, when a surprising result emerges that could challenge existing theories, the slow module is triggered to conduct a thorough investigation. This gating mechanism is reminiscent of the "thinking fast and slow" paradigm in decision-making [3,11], but it is adapted to the iterative, hypothesis-driven nature of science. Moreover, the control policy itself can be learned through reinforcement learning, where the reward signal is based on the quality of scientific discoveries produced over time [1]. Such an adaptive controller enables the agent to balance speed and accuracy in a resource-efficient manner.

4. Hierarchical Planning and Execution Framework

To operationalize the fast–slow modules within a coherent discovery process, we embed them in a hierarchical planning and execution architecture. The highest level of the hierarchy defines the long-term scientific goals, such as discovering a new catalyst for carbon capture or elucidating the mechanism of a disease pathway. These goals are decomposed into a sequence of subgoals, each representing a manageable research question or experimental campaign. The decomposition is performed by a strategic planner that uses the slow module to reason about dependencies, feasibility, and expected information gain. At the intermediate level, tactical planners translate each subgoal into a set of concrete actions, such as running a particular simulation, querying a database, or performing a laboratory experiment. These planners leverage the fast module to generate candidate actions quickly, then use the slow module to evaluate and select the most promising ones. At the lowest level, an execution layer

carries out the chosen actions, often relying on robotic hardware or cloud-based computational resources, and feeds observations back up the hierarchy [2,13].

This hierarchical design offers several advantages for autonomous scientific discovery. First, it provides a natural mechanism for abstraction: high-level plans are robust to minor fluctuations in experimental outcomes, while low-level actions can be adapted in real time based on sensor feedback. Second, it enables human scientists to inspect and intervene at multiple levels of granularity. For example, a human supervisor may override a high-level goal if ethical concerns arise, while allowing the fast module to continue generating low-level experiments within acceptable bounds. Third, the hierarchy facilitates efficient allocation of computational resources. Expensive slow reasoning is reserved for strategic decisions and critical hypothesis evaluations, while routine operations are handled by the fast module, thereby reducing overall energy consumption and latency [14,15]. The framework also incorporates a memory component that stores past discoveries, experimental protocols, and failure cases, which can be retrieved by the fast module to avoid repeating errors. Over time, the agent's knowledge base grows, making the fast module more accurate and reducing the need for slow deliberation in familiar contexts.

5. System Architecture and Governance

The successful deployment of hierarchical fast–slow discovery agents requires careful attention to system architecture and governance. At the architectural level, the agent must be designed as a modular, loosely coupled system to allow independent updates to the fast and slow modules without disrupting the overall workflow. Each module should have well-defined interfaces for data exchange, with standardized formats for hypotheses, experimental designs, and results [17]. Furthermore, the controller that orchestrates the modules must maintain a log of all decisions, including which hypotheses were generated, how they were prioritized, and which verification steps were performed. This log serves as an audit trail, enabling reproducibility and post-hoc analysis by human researchers. Given the potential for autonomous agents to generate large numbers of findings, a curation layer is needed to filter, rank, and summarize discoveries for human consumption, akin to the role of a senior research scientist [6,8].

Governance of such agents extends beyond technical architecture to encompass institutional policies and oversight mechanisms. Because autonomous discovery agents operate within a sociotechnical system that includes funding agencies, publishers, and regulatory bodies, their design must incorporate principles of transparency, accountability, and fairness [18,19]. For instance, the agent should be required to share its experimental protocols and raw data in open repositories, following FAIR (Findable, Accessible, Interoperable, Reusable) data principles. The decision-making process of the controller should be interpretable, allowing human overseers to understand why a particular research direction was chosen or abandoned. Additionally, governance frameworks must address the risk of bias in the fast module, which may inadvertently favor hypotheses that align with existing literature while neglecting unconventional but promising avenues [20]. To mitigate this, the controller can allocate a fixed fraction of computational resources to exploration of low-probability hypotheses, effectively implementing a form of "slow" oversight over the fast module's tendencies. Such governance mechanisms are essential for ensuring that autonomous agents contribute to science in a responsible and equitable manner.

6. Deployment, Sustainability, and Robustness

Deploying autonomous scientific discovery agents at scale presents significant engineering and sustainability challenges. The computational demands of both fast and slow modules are substantial: training large neural networks for hypothesis generation requires vast amounts of energy, and running causal inference algorithms can be computationally intensive [5,15]. A single discovery agent may consume as much electricity as a small data center, raising concerns about carbon footprint and environmental impact. To address these concerns, the hierarchical framework can incorporate resource-aware scheduling that prioritizes the use of energy-efficient hardware for routine tasks (fast module) and reserves high-performance computing for critical analyses (slow module). Additionally, the agent can be designed to share intermediate results across multiple parallel instances, reducing redundant computation through a distributed knowledge graph [1,2]. Sustainability also involves data stewardship: the agent must manage large volumes of experimental data, ensuring that only high-quality, well-annotated data are retained, while obsolete or low-value data are purged to minimize storage costs.

Robustness is another key consideration. Scientific discovery agents must be resilient to noisy data, equipment failures, and unexpected environmental conditions. The hierarchical structure contributes to robustness by providing fallback mechanisms: if the fast module generates an unstable recommendation, the slow module can re-evaluate the situation. Moreover, the intermediate tactical planners can incorporate redundancy by maintaining multiple experimental approaches for the same subgoal, so that if one fails, another can be tried without restarting the entire process [16]. The controller can also monitor the performance of each module over time and trigger recalibration or retraining if drift is detected. For real-world deployment in fields like materials science or biology, physical robotic platforms must be integrated with the software stack, requiring careful synchronization and error handling. The fast module can be used for anomaly detection to anticipate hardware failures, while the slow module plans maintenance schedules [13]. Overall, the architecture must be designed with a robustness-first mindset, acknowledging that scientific progress often involves unexpected setbacks.

7. Fairness, Ethics, and Policy Implications

The automation of scientific discovery raises profound ethical and policy questions that cannot be addressed solely through technical design. Autonomous agents may amplify existing biases in the scientific literature, such as the overrepresentation of certain populations in biomedical research or the neglect of research topics relevant to low-resource communities [18,19]. The fast module, trained on published literature, will inevitably inherit these biases unless proactive measures are taken. The slow module, while more rigorous, also depends on the data and assumptions encoded by its programmers. Fairness therefore requires that the training data for both modules be carefully curated to include diverse perspectives and that the controller explicitly incorporates fairness constraints into its optimization objectives. For example, the agent could be programmed to allocate a portion of its discovery efforts to problems that are understudied but have high potential for social benefit, as determined by a socially aware reward function [20].

Policy implications extend to ownership and credit. Who should be considered the author of a discovery made by an autonomous agent? Current norms in science assign credit to human researchers, but as agents become more autonomous, new frameworks for attribution are needed. One proposal is to treat the agent as a collaborator, with the research team acknowledging its contributions in publications. Governments and funding agencies must also

develop regulatory guidelines for the use of AI in science, particularly in high-stakes domains such as drug development and climate modeling [14,17]. The hierarchical architecture supports policy compliance by enabling transparent logging and auditing, which can be used by regulators to verify that the agent operated within ethical boundaries. Additionally, the fast–slow design allows human oversight to be layered: quick decisions can be automatically vetted against a set of ethical rules, while slow decisions can be escalated to a human review board. This layered oversight mirrors existing frameworks for algorithmic decision-making and can help build public trust in autonomous science.

8. Conclusion

This paper has presented a hierarchical planning and execution framework for autonomous scientific discovery agents that leverages the cognitive metaphor of fast and slow reasoning. By integrating a rapid, heuristic module for hypothesis generation with a deliberate, analytical module for verification and refinement, and by organizing the discovery process into multiple abstraction levels, the proposed architecture balances the competing demands of speed, accuracy, and resource efficiency. We have discussed the system-level trade-offs inherent in such a design, including the governance mechanisms required to ensure reproducibility, robustness, and fairness. The framework also addresses deployment challenges related to sustainability, data management, and ethical oversight. As autonomous scientific agents move from research prototypes to operational systems, the hierarchical fast–slow approach offers a principled foundation that aligns with both cognitive science and engineering best practices. Future work should focus on empirical validation of the architecture in real-world scientific domains, development of standardized benchmarks for evaluating discovery agents, and refinement of governance protocols to adapt to evolving policy landscapes. Ultimately, the goal is to create systems that augment human scientific creativity while maintaining the rigor and integrity that define the scientific enterprise.

References

1. Simon, H. A. (1955). A behavioral model of rational choice. The Quarterly Journal of Economics, 69(1), 99-118.
2. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
3. Russell, S. J., & Norvig, P. (2020). Artificial intelligence: A modern approach (4th ed.). Pearson.
4. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.
5. Chen, X. (2024, November). Cloud Storage User Behavior Analysis and Dynamic Replica Strategy Optimization Based on Improved RFM and Fuzzy Clustering. In International Conference on Cognitive based Information Processing and Applications (pp. 425-434). Singapore: Springer Nature Singapore.
6. Kitano, H. (2016). Artificial intelligence to win the Nobel Prize and beyond: Creating the engine for scientific discovery. AI Magazine, 37(1), 39-49.
7. Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. Science, 346(6206), 171-172.
8. Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., ... & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. Nature, 620(7972), 47-60.

9. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.

10. Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. In Advances in Neural Information Processing Systems (Vol. 31, pp. 2450-2462).

11. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf: Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

12. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

13. Tegmark, M. (2017). Life 3.0: Being human in the age of artificial intelligence. Knopf.

14. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

15. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

16. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

17. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. Nature, 568(7753), 477-486.

18. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-16). ACM.

19. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59-68). ACM.

20. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a 'right to explanation'. AI Magazine, 38(3), 50-57.