Navigating Epistemic Monocultures in AI-Driven Science: A Simulation Study

SINA FAZELPOUR, Northeastern University, United States
JOSEPH O'BRIEN, University of California, San Diego, United States
HANNAH RUBIN, University of Missouri, United States

AI integration into scientific communities promises accelerated discovery but raises concerns about detrimental homogenization. We develop an NK landscape model to explore these promises and risks. We find that non-personalized AI systems that offer uniform guidance yield benefits only under a narrow conjunction of specific problem structure, practices, and baseline research capabilities, becoming harmful otherwise. We implement two proposed mitigations: randomization and personalization. While randomization's utility remains restricted to decomposable problems, personalization preserves or enhances diversity, enabling benefits across conditions. Overall, our results highlight the importance of shifting perspectives in systemic AI evaluations from tool adoption to institutional adaptation.

1 INTRODUCTION

Artificial intelligence (AI) tools are increasingly integrated across the scientific pipeline, from hypothesis generation and data analysis to simulating research participants and automating laboratory procedures. This integration raises both hopes and concerns. On the one hand, AI integration promises to dramatically accelerate the rate of discovery and democratize access to high-level capabilities, allowing diverse teams to perform complex inquiries, and flattening resource or expertise disparities (Gottweis et al. 2025; Hume 2025; Wang et al. 2023). On the other hand, it has been argued that the widespread adoption of AI may lead to the formation of epistemic monocultures, reducing the productive heterogeneity of methodologies with which scientists pursue their work (Burton et al. 2024; Messeri and Crockett 2024).

Authors' addresses: Sina Fazelpour, Northeastern University, Boston, United States, s.fazel-pour@northeastern.edu; Joseph O'Brien, University of California, San Diego, San Diego, United States, j3obrien@ucsd.edu; Hannah Rubin, University of Missouri, Columbia, United States, hmrfyb@missouri.edu.

In this paper, we develop a simulation model to explore when the adoption of AI tools in scientific communities may result in such detrimental homogenization, and what strategies might mitigate this risk. Specifically, we use the NK landscape framework (Kauffman and Levin 1987) to model scientific problem-solving as a search through complex solution spaces, and examine how different AI designs affect epistemic success and diversity across different problem structures.

Our results reveal that the impacts of AI adoption depend critically on problem structure, system design, and institutional state and norms of epistemic communities. Non-personalized AI, which recommends uniform best computational practices without accounting for local context, yields benefits only under highly modular problems with moderate use rates; outside these conditions, it accelerates convergence toward epistemic monocultures and harms longer term collective performance. We implement two mitigation strategies inspired by suggestions in the literature (Fügener et al. 2021; Jain et al. 2024): randomization, which introduces diversity by sampling from top-performing solutions rather than recommending a single best option; and personalization, which tailors recommendations to each agent's specific context. While Randomization's utility remains restricted to the same structural conditions as non-personalized AI, Personalization proves robustly beneficial across problem structures.[1]

Importantly, our findings suggest that productive AI integration requires more than tool adoption. For non-personalized systems, the requisite problem modularity is often an achievement of organizational practices, depending on standardization, established protocols, and divisions of labor, rather than simply an inherent feature of scientific domains. For personalized systems, effectiveness requires established institutional standards for documenting and communicating often tacit human and organizational factors that shape scientific practice, and varies based on how communities adapt their exploratory practices to complement AI capabilities. We discuss these implications for scientific communities navigating AI adoption, along with limitations of our approach and directions for future research.

The rest of the paper is structured as follows. In §2, we review previous work on AI in science, concerns about monoculture formation, and the NK landscape framework as

[1]Terminological note: we use capitalization to distinguish between Randomization and Personalization as implemented in our model from the broader possibilities that those concepts may involve.

an appropriate tool for examining these dynamics. §3 provides the formal details of our model implementation. In §4 and §5, we present simulation results exploring monoculture formation and the effectiveness of various interventions. Finally, §6 discusses implications for scientific practice.

2 BACKGROUND AND RELATED WORK

2.1 AI in Scientific Communities

AI systems are now integrated into every stage of the research pipeline. Recent reviews highlight how AI systems are now employed to synthesize vast bodies of literature, formulate novel hypotheses, automate experimental design through "self-driving" laboratories, generate synthetic data—from digital twins in healthcare to simulated research participants—and even act as automated scientific partners (Gottweis et al. 2025; Lu et al. 2024; Si et al. 2024; Wang et al. 2023; Zhang et al. 2025).[2] This widespread adoption is powered by a diverse set of technical architectures, training paradigms, and operational principles: transformer-based language models assist with knowledge synthesis, graph neural networks predict molecular properties, diffusion models generate novel protein structures, reinforcement learning algorithms optimize experimental protocols, and computer vision systems analyze microscopy images.

Across this variety in design and application, advocates highlight the shared promise of AI-driven science to accelerate scientific discovery. Beyond automating routine tasks, AI systems are seen as facilitating interdisciplinary communication, lowering barriers to entry, and enabling efficient navigation of complexity. For example, AI systems are said to "democratize" science by facilitating access to knowledge or capabilities that previously required deep domain expertise and specialized resources (Dessimoz and Thomas 2024; Hume 2025). Moreover, emerging "cloud laboratories" and AI-driven experimental platforms may allow researchers to remotely execute and reproduce complex experiments (Adam 2024; Lu et al. 2024). By standardizing experimental protocols into executable code, these systems are said to make complex methodologies accessible to and implementable by a broader community. Finally, AI tools offer the promise of navigating

[2]Messeri and Crockett (2024) provide a useful taxonomy of these applications across the research pipeline.

complex problem spaces by identifying patterns across vast, disparate datasets (Sourati and Evans 2023; Wang et al. 2023).

2.2 Epistemic Monocultures

Despite these promises, a growing body of literature warns that the widespread reliance on algorithmic tools may inadvertently reduce the diversity of scientific inquiry in particular, and institutional decision-making more broadly (Burton et al. 2024; Kleinberg and Raghavan 2021; Messeri and Crockett 2024). In particular, Messeri and Crockett (2024) distinguish between two types of such *epistemic monocultures*: *monocultures of knowers* (homogeneity in standpoints from which science is done) and *monocultures of knowing* (homogeneity in how science is done). With respect to the latter, which will be our focus in this paper, they argue that the widespread adoption of AI tools risks producing a monoculture by prioritizing quantitative ways of knowing that appear portable and scalable, but which systematically "strip out the contextual sensitivity and local details" required for deep understanding (Messeri and Crockett 2024, p. 54).

On Messeri and Crockett's account, a key driver of this homogenization is the way in which the technical success of AI tools can obscure their underlying methodological and theoretical commitments. As Messeri and Crockett observe, the construction of any quantitative resource or model requires developers to make various choices that inevitably embed specific assumptions and values into the resulting artifacts. When resource- or expertise-constrained teams adopt these "high-impact" quantitative artifacts to study similar questions, however, they often do so without appreciating the crucial import of those hidden, context-specific choices. This can lead scientific communities to fall prey to an "illusion of exploratory breadth" (Messeri and Crockett 2024): researchers believe they are exploring the full landscape of potential solutions, when in reality they are converging on a narrow subset of hypotheses that are legible to the dominant algorithmic paradigm.

2.3 Modeling the Impact of AI Adoption in Scientific Communities

How should we evaluate the systemic impacts of AI integration in scientific communities in light of these promises and risks? To address this, we employ the NK landscape framework as a generalized model of scientific problem-solving (Kauffman and Levin 1987). NK landscapes have been fruitfully employed in philosophical and social scientific research

to examine many of the aforementioned factors that arise in the context of AI integration in science: the complexity of a scientific problem; potential interdependence between different aspects of a problem (e.g., those amenable to quantitative methods and those that are not); relative (in)efficiencies of different methodologies; the promises and failure modes of communication; and the value of diversity and the risks of homogenization in scientific communities (Gomez and Lazer 2019; Grim et al. 2013; Huang 2024; Lazer and Friedman 2007; Muldoon 2013; Reijula et al. 2023; Wu 2024; Wu and O'Connor 2023). Appropriately augmented, we suggest, this framework can offer a productive lens for understanding the systemic effects of widespread AI adoption on scientific search at a useful level of abstraction.

Importantly, the abstraction afforded by the framework allows us to draw more generalizable lessons than implementation-specific approaches that can be hampered by the extraordinary diversity of AI applications and architectures. At the same time, it allows us to move beyond general concerns about monocultures to explore specific trade-offs, and the conditions under which the benefits of AI adoption could possibly outweigh its risks. In the following section, we formally describe the model and how we extend it to capture issues surrounding AI integration in science. In Section 4, we return to the dynamics of monoculture formation described by Messeri and Crockett, and use it to motivate our formalization of Non-Personalized AI systems.

3 GENERAL MODELING FRAMEWORK

In this section, we describe the NK framework and how we extend it to model AI integration in scientific communities. In doing so, we highlight particular considerations that inform our investigations of the conditions under which AI adoption can be epistemically harmful or beneficial. To make the formalisms concrete, we employ a running case study of drug discovery in public health contexts.

3.1 The NK Landscape and Problem Complexity

Consider research teams developing treatments for a neglected disease. Such teams face a range of decisions: which molecular features to prioritize in virtual screening, which machine learning architecture to use for prediction, which animal model to use for efficacy testing, how to design community engagement for clinical trials, which drug

administration to prioritize given expectations about patient adherence and healthcare infrastructure, and more. We can represent each team's overall scientific strategy as a binary vector $d = (d_1, \ldots, d_N)$, where each element encodes a choice, such as whether to adopt a particular methodological assumption ($d_i = 1$) or not ($d_i = 0$).[3]

NK landscapes formalize how such decisions jointly determine the "fitness" of a research strategy, or how well they work together to achieve certain scientific goals. If fitness contributions of decisions were independent, optimization would be straightforward: evaluate each decision in isolation, identify the better option, and combine them. Scientific decision-making is rarely so simple, however. The value of one choice often depends on others. For instance, adopting certain machine learning tools in healthcare settings may be beneficial, but only when coupled with appropriate training data and justified assumptions about the biological mechanisms and patient populations involved (Chen et al. 2021).

In the NK framework, the parameter K formalizes this *interdependence*. When $K = 0$, decisions do not interact, and contribute independently to overall fitness. Higher values of K represent increasingly complex landscapes, where the contribution of one choice depends (on average) on K other choices. In this case, for any focal decision d_i, its fitness contribution changes as a function of those other K decisions.[4]

The NK model thus captures a central challenge of scientific problem-solving: researchers must navigate a landscape where altering a single theoretical or methodological commitment can help or hinder depending on their other choices. This explains why communities may become stuck on local optima: when K is large, local improvements can mislead researchers about the location of global optima.

[3]Following works on NK modeling, we treat these decisions as binary, without loss of generality.

[4]Formally, the overall fitness of a given research strategy $d = \{d_1, \ldots, d_N\}$ is given by:

$$f(d) = \frac{1}{N} \sum_{i=1}^{N} \phi_i\big(d_i; S(d_i)\big)$$

where $S(d_i)$ represents the set of K other decisions in d whose values influence the contribution of d_i, ϕ_i, and ϕ_i assigns a normalized payoff in $[0, 1]$ to each combination of d_i and decisions in $S(d_i)$. In our simulations, landscapes are regenerated for each run, and transformed and normalized so that $\max f(d) = 1$.

3.2 Epistemic Agents

In practice, research teams' expertise and resources put constraints on which decisions they can independently evaluate and modify. In our drug discovery example, a computational biology team might specialize in virtual screening and molecular optimization, but lack expertise in clinical trial design or manufacturing scale-up. A global health policy team might specialize in community engagement and health system integration, but require external guidance on computational methods. We model this variance in expertise and resources, by assigning each agent a a *specialization set*, h_a, that consists of a random subset of H out of the total N decisions.[5] The specialization set h_a thus represents the decisions that agent a can explore without external assistance. As H decreases relative to N, agents face greater pressure to rely on external sources for navigating the full decision space.

3.3 Social Learning in Networks

One source of external information are others. Specifically, agents can learn from others to whom they are socially connected.[6] We model scientific communities as random networks (Erdős et al. 1960) where each pair of agents is connected with a fixed probability p_{edge}.[7] To ensure that (cluster of) agents do not remain isolated, we create these random networks under the constraint that they must be connected. Within this network, each agent engages in social learning with probability p_{social}. When doing so, it considers its

[5]While we keep the size of H fixed for a given simulated community, the particular specialization sets differ between agents.

[6]The notion of a "neighbor" in epistemic landscapes can be understood differently. Some authors understand this dynamically in terms of agents' current location on the landscape. We understand this here in terms of placement on a static, exogenously defined network structure (Lazer and Friedman 2007).

[7]Though we do not systematically vary network structure, because our interest is not in the effects of social learning, we also tested a scale-free Barábasi-Albert network constructed via preferential attachment (Barabási and Albert 1999), which approximates the skewed connectivity patterns observed in scientific communities (e.g., in citation networks Price 1965; Redner 2005; Zhong and Liang 2024). However, we find no observable differences between communities established on random versus Barábasi-Albert graphs. To simplify the parameter space, we focus on random graphs.

neighbors and adopts the decision vector of the most successful one, provided that the solution improves its fitness.[8]

3.4 The Computational Module and Problem Decomposability

Increasingly, research teams can also depend on AI tools as an external source of information. However, not all decisions facing a research team are equally amenable to AI assistance. In our drug discovery example, while decisions pertaining to data-analysis pipelines, chemical-property prediction, and simulation of biological processes are increasingly amenable to computational tools, other decisions, such as those concerning aspects of animal studies, qualitative study design, patient-adherence considerations, and manufacturing constraints, may involve factors that are less amenable to automation.[9] To model this, we partition decisions into two modules:

(1) *Computational module* $m_c = \{d_1, \ldots, d_M\}$, consisting of decisions where AI tools can effectively operate.[10]
(2) *Non-computational module* $\neg m_c = \{d_{M+1}, \ldots, d_N\}$ consisting of decisions outside AI's purview.

The relationship between decisions across different modules raises interesting methodological considerations that are not simply about the number of dependencies (i.e., K), but the *pattern* of those dependencies (Ethiraj and Levinthal 2004; Ganco and Hoetker 2009; Reijula et al. 2023): to what extent can the two modules be treated as separate or decomposable sub-problems? To make this point precise, for each decision $d_i \in d$, let $S(d_i)$ denote the set of K other decisions whose values influence the fitness contribution of d_i. We can distinguish two sets of dependencies:

[8] In our reported results, we consider $p_{social} = \{0.1, 0.2\}$. Sensitivity analyses varying this parameter reproduce the core findings of Lazer and Friedman (2007): higher p_{social} accelerates short-term gains but reduces long-run exploration by pushing communities more quickly toward local optima.
[9] Of course, AI systems might provide suggestions—potentially hallucinatory or unjustified ones—about a very many things. Here we are taking AI assistance to consist of something stronger akin to providing agents with the know-how to deliberate about certain decisions and enact change therein, even if those decisions are not part of their specialization sets.
[10] We assume that the computational module includes the first M of the N decisions for ease of exposition and without loss of generality.

- $k_{\text{in}}(i)$: Number of *within-module* dependencies, given by $|S(d_i) \cap m_c| + 1$, if $d_i \in m_c$; and $|S(d_i) \cap \neg m_c| + 1$, otherwise.[11]
- $k_{\text{out}}(i)$: Number of *out-of-module* dependencies, given by$|S(d_i) \cap \neg m_c|$, if $d_i \in m_c$; and $|S(d_i) \cap m_c|$, otherwise.

Within-module dependencies capture how a decision's value is influenced by all the decisions in the same module, whereas out-of-module dependencies capture how a decision's value is influenced by decisions in the other module.

Tracking the proportion of within- and out-of-module dependencies for each module $m \in \{m_c, \neg m_c\}$ offers a way of tracking the extent of *modularity* or *decomoposability*:

$$\rho(m) := \frac{\sum_{d_i \in m} k_{\text{in}}(i)}{\sum_{d_i \in m} k_{\text{in}}(i) + \sum_{d_i \in m} k_{\text{out}}(i)} \in [0, 1].$$

A high $\rho(m)$ indicates that computational and non-computational decisions form relatively self-contained subproblems. That is, in finding the optimal configuration of decisions in each module, one can largely abstract away from the configuration of decisions in the other. In contrast, a low $\rho(m)$ indicates that the contributions of decisions within the module depend heavily on decisions outside of it. In our running example, for instance, this can happen when the appropriateness of computational decisions depends critically on non-computational upstream (e.g., theoretical assumptions and data quality from qualitative studies) or downstream considerations (e.g., manufacturing feasibility and community acceptance). Here, one might expect that optimizing only over the set of computational decisions without regards for the non-computational context can create solutions that are likely to be misaligned with the overall optimum.

In our simulations, therefore, we vary $\rho(m)$ to explore how modularity moderates AI effectiveness, assuming symmetric modularity across both modules to isolate the effects of decomposability itself.[12] In Sections 4 and 5, we consider different designs for AI systems,

[11]Notice that we add one to include the dependency of the fitness contribution of a decision on the decision itself. This allows $\rho = .5$ to be a meaningful middle point in problem modularity.

[12]Of course, the dependency pattern need not be symmetric. For example, the dependencies in the non-computational module may form a relatively self-contained cluster, while the fitness contribution of decisions in the computational module may be impacted by out-of-module decisions (e.g., depending on technological and organizational factors influencing data collection). That said, the symmetry assumption enables us to focus on the extent (as opposed to potentially varied patterns) of decomoposability between modules.

and describe implementation details about how they provide recommendations about the computational module.

3.5 Collective Epistemic Search

We can now formalize collective problem-solving. Each agent a is initialized with a random decision vector $d^a \in \{0,1\}^N$ and specialization set h_a. Agents know their current fitness, $f(d^a)$, and can evaluate alternatives. To obtain such alternatives, each round, agents pursue one of three strategies:

(1) **Social learning:** With probability p_{social}, the agent adopts the decision vector of their highest-performing neighbor, if it improves fitness.
(2) **Explore:** Otherwise, they can explore the landscape independently, by flipping a randomly selected decision from their specialization set h_a, adopting the change if it improves fitness.
(3) **Query AI:** For agents with access to AI tools, if they do not engage in social learning, instead of independent exploration, they can choose to query an AI tools, and receive recommendations for decisions within the computational module m_c, adopting them if they improve overall fitness. Agents' query behavior is determined by a community-level *AI use rate* $\in [0,1]$. For example, in communities with an AI use rates of 0.25, agents will, on average, seek AI recommendations in 25% of rounds not spent on social learning.

In sum, each round: agents engage in social learning with probability p_{social}. When they aren't social learning, they query AI with a probability equal to the AI use rate, otherwise they explore. Importantly, each action consumes one round regardless of outcome. Rejected AI recommendations, unsuccessful explorations, and declined peer solutions all carry opportunity costs, reflecting that evaluating alternatives requires time and resources even when they prove inferior.

4 SIMULATION 1: NON-PERSONALIZED AI AND THE FORMATION OF EPISTEMIC MONOCULTURES

Our first simulations examine the systemic impacts of adopting a single AI system offering uniform best-practice recommendations. This "non-personalized" mode reflects typical

deployments, where pretrained systems provide potentially useful but context-insensitive guidance. When, and under what conditions, do such tools enhance collective problem-solving, and when do they instead produce detrimental epistemic monocultures?

4.1 Experimental design

4.1.1 Formalizing Non-personalized AI. When agent a queries the non-personalized AI system, the tool returns the computational configuration used by the globally best-performing agent b:

$$d^{NP}_{m_c}(D_t) = d^b|_{m_c}$$

where $d^{NP}_{m_c}(D_t)$ refers to the recommendation of non-personalized AI system, D_t denotes the state of all decision vectors in the community at time t, and $d^b|_{m_c} = \{d^b_1, \ldots, d^b_M\}$ is the computational portion of b's strategy. The system tracks fitness across *all* agents, not just the querying agent's local neighborhood. In this sense, non-personalized AI acts as social learning that is *global* in reach but *partial* in scope: it provides guidance about leading computational practices while ignoring how those practices interact with agents' non-computational choices.

This formalization captures the dynamic underlying Messeri and Crockett's concern about monocultures of knowing (§2): agents incorporate the "best" computational practices of successful peers without recognizing that those practices may be contingent on non-computational assumptions that do not necessarily hold in their own context.

4.1.2 Dependent Measures. We focus on two measures to evaluate the epistemic effects of introducing such a non-personalized AI system.

Epistemic success. We operationalize epistemic success as the average fitness of agents in the community, measured by the mean of $f(d^a)$ across all agents. Higher mean fitness indicates that the community, on average, has identified superior research strategies. We examine epistemic fitness at different points in a simulation run.

Transient epistemic diversity. To capture diversity in research practices, we measure the *mean pairwise Hamming distance* between agents' decision vectors at each round, and compute the area under this curve (AUC) over the course of a simulation. The Hamming distance between two binary decision vectors counts the number of positions on which

they differ; a higher average distance thus indicates greater heterogeneity in the strategies pursued by agents. For example, two agents differing on three of ten decisions have a Hamming distance of 0.3. Tracking the mean Hamming distance across time provides a measure of *transient diversity*—how varied the community's approaches are during its search for better solutions. Since our simulations converge when the community reaches a consensus, the mean Hamming distance in later rounds approaches zero. The AUC thus offers a useful summary statistic for how much diversity the community maintained before convergence. A larger AUC indicates that the community preserved its transient diversity for longer, while a smaller AUC reflects faster homogenization in research practices.

We compare otherwise identical communities differing only in AI access across a range of parameter settings.[13] Each configuration is simulated 1000 times.

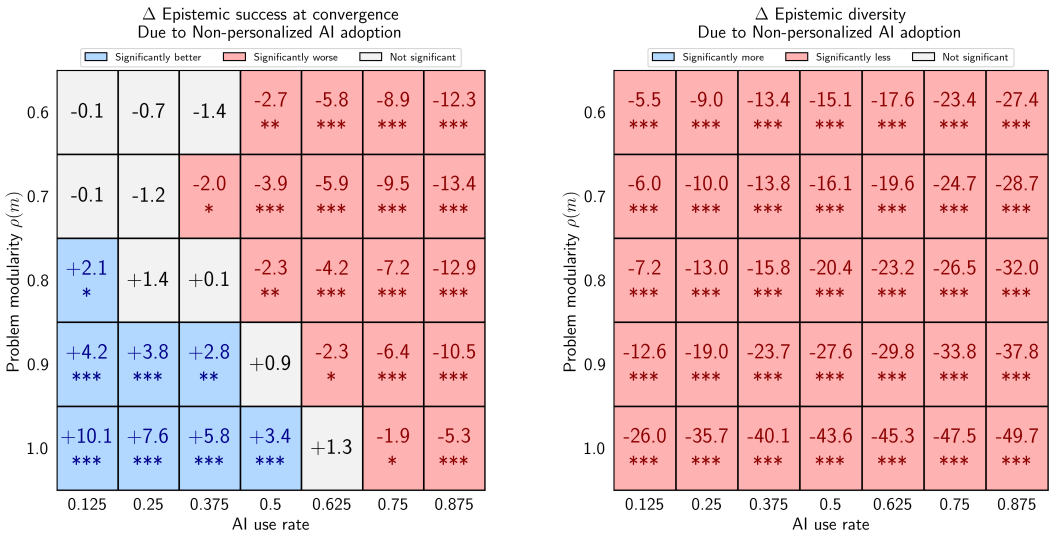
4.2 Results

Our simulation results show that the impact of non-personalized AI recommendations on epistemic success crucially depends on the extent of problem modularity. As shown in Figure 1a, when the underlying problem is highly modular ($\rho(m) \geq 0.8$), and can be decomposed into two relatively self-contained sub-problems, AI use—particularly at low to moderate levels—can yield reliable gains in community-level fitness relative to the no-AI baseline. In these settings, the problem structure allows AI's global "best practice" recommendations about m_c configuration to efficiently inform agents' decisions within that module.

These advantages disappear, however, as modularity decreases. In this case, the insensitivity of the AI tool to the heterogeneous contexts of agents—that is, their differing configurations in $\neg m_c$—adversely impacts its effectiveness. As Figure 1a shows, when $\rho(m) \leq 0.7$, at convergence, non-personalized AI provides negligible benefits at best, and becomes actively detrimental at high use rates.

When it comes to transient epistemic diversity, we find that non-personalized AI significantly reduces diversity across all settings by 25% on average (See Figure 1b). Crucially,

[13]Specifically, we keep $N = 20$ and $M = 10$ fixed across all communities, while varying $K = \{5, 9\}$, $\rho(m) \in [0.6, 1]$ with 0.1 increments, $p_{social} = \{0.1, 0.2\}$, $H = \{5, 10, 15\}$, AI use rate $\in [0.125, 0.875]$ with 0.125 increments. Note that since scientific problems are generally modular to some extent, we do not consider $\rho \leq .5$



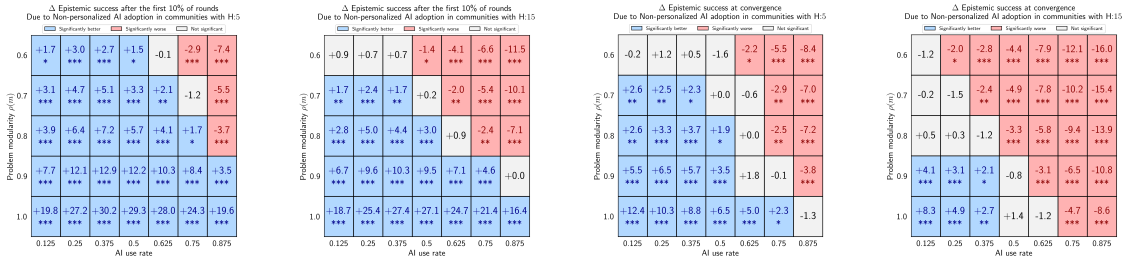
(A) (B)

FIG. 1. The differential impacts of introducing non-personalized AI, expressed as *percentage differences* relative to communities without AI tools (averaged across 1000 simulation runs). Communities consist of 100 agents with specialization set size $H = 10$, navigating a problem with $N = 20$, $K = 9$, $M = 10$, while engaging in social learning with $p_{social} = 0.1$. Panel (A) shows differences in epistemic *success* at convergence across problem modularity and AI use frequency. Panel (B) shows differences in transient epistemic diversity (measured by AUC of mean Hamming distance). Colored cells indicate statistically significant differences: blue for improvements, red for declines, gray for non-significant ($p >= 0.05$). Asterisks denote significance levels ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

however, this diversity loss is not *necessarily* epistemically harmful. Rather, diversity reduction can reflect productive convergence when AI successfully coordinates the community toward superior solutions. This is illustrated by examining modularity's differential effects on diversity and performance. At high modularity, non-personalized AI reduces diversity substantially. Yet, this homogenization accompanies significant epistemic improvements (Figure 1a), suggesting productive coordination around superior computational configurations. At low modularity, diversity loss is more modest, but occurs alongside performance stagnation or decline (Figure 1b).

When, then, is homogenization harmful? Our analyses identify two mechanisms through which non-personalized AI induces diversity loss, with differential epistemic consequences.

First, especially in contexts with higher modularity or use rate—both of which increase uptake of AI recommendations—non-personalized AI produces dynamics similar to those documented in studies of social learning (Lazer and Friedman 2007; Zollman 2010). AI recommendations based on the currently best-performing m_c configuration provide early gains, but high use frequency causes communities to converge prematurely on local optima. This mechanism is reflected in the monotonic decrease of AI's benefits as use rate increases, even for highly modular problems (Figure 1).



(A) (B) (C) (D)

FIG. 2. Difference in early versus final average epistemic success between communities with access to non-personalized AI and communities without AI tools, expressed as percentage differences (averaged across 1000 simulation runs). Communities consist of 100 agents facing a problem with $N = 20, K = 9, M = 10$, while engaging in social learning with $p_{social} = 0.1$. Panels (A) and (B) show percentage differences in *early epistemic success* (first 10% of rounds) across problem modularity and use rate for communities specialization set sizes $H = 5$ and $H = 15$, respectively; panels (C) and (D) shows percentage differences in final epistemic success after convergence for those communities. Colored cells indicate statistically significant differences: blue for improvements, red for declines, gray for non-significant ($p \geq 0.05$). Asterisks denote significance levels ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Second, especially when lower modularity renders non-personalized AI recommendations ineffective, the reliance on such systems can involve significant *opportunity costs* due to context-mismatch of recommendations. In this case, the missed rounds of individual exploration, which were instead spent on querying AI, lead to significant reduction in the diversity of practices in the community. This in turn limits the quality of solutions from which agents can socially learn.

The marginal risk of both of these mechanisms of homogenization increases in more capable scientific communities. When agents have small specialization sets ($H = 5$), they explore limited regions of the decision space, leaving gaps that non-personalized AI can

fill. With larger specialization sets ($H = 15$), agents collectively achieve broader coverage, and AI's marginal benefits may not justify its homogenizing costs. Figure 2 illustrates this pattern. Non-personalized AI offers substantial early gains regardless of specialization (panels A, B show improvements exceeding 20% at $\rho(m) = 1$). After convergence, however, while communities with $H = 5$ often retain small but significant benefits (panel C), those with $H = 15$ frequently see early gains reverse into losses (panel D). Early coordination accelerates initial progress but leads to premature convergence—a speed-performance tradeoff mediated by diversity loss, familiar from prior work in philosophy of science (Lazer and Friedman 2007; Zollman 2010).

Overall, our findings reveal that non-personalized AI's effectiveness depends on a specific alignment of problem structure (high modularity), practices (moderate use), and existing research capabilities (limited specialization creating coverage gaps). Outside these conditions, non-personalized AI harms collective performance by reducing valuable transient diversity without providing compensating coordination benefits—precisely the dynamic underlying Messeri and Crockett's concerns about epistemic monocultures.[14]

5 SIMULATION 2: MITIGATING EPISTEMIC MONOCULTURES WITH RANDOMIZATION AND PERSONALIZATION

Our second set of simulations explores two design interventions for mitigating the homogenization documented in Section 4: *randomizing* recommendations and *personalizing* them. Randomization introduces diversity by sampling from multiple high-performing options rather than always recommending the single best, while personalization tailors recommendations to each agent's context. Below, we detail the implementations, and describe our findings about the conditions under which each approach succeeds. As before, we keep the broader discussion for Section 6.

5.1 Experimental Design

5.1.1 Formalizing Randomized Non-personalized AI. We implement a design inspired by randomization as a strategy for combating outcome homogenization (Jain et al. 2024). The idea is that instead of recommending the same "best" option to everyone, the system

[14]We observe qualitatively similar patterns concerning the results discussed in the section for $K = 5$, and $p_{social} = 0.2$ as well.

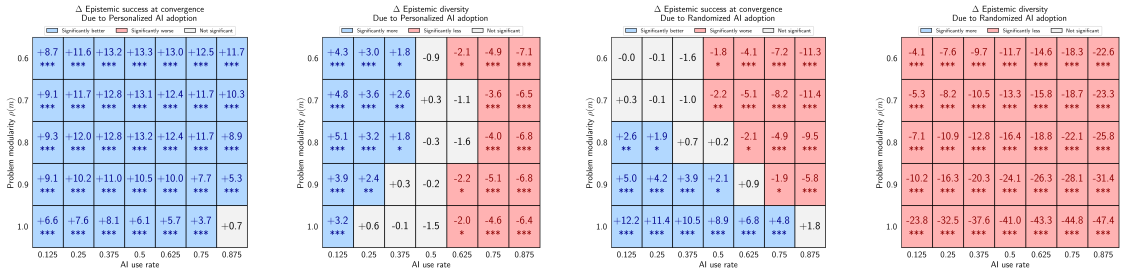
randomizes over high-performing options, potentially preserving heterogeneity, while maintaining utility. We operationalize this as follows: when an agent queries the AI tool, the tool randomly selects an agent from the top decile performers in the community and returns the computational portion of its decision vector. This approach is non-personalized, insofar as it only considers the state of the community at the time of the query, ignoring the querying agent's context. In contrast with non-personalized AI design above, however, it samples from high performers (as opposed to only the single best performer), thus promoting diversity, while ensuring a high baseline of quality.

5.1.2 Formalizing Personalized AI. Personalized AI has been proposed as a way to mitigate the homogenizing effects of AI adoption (Fügener et al. 2021). We operationalize this idea as follows. The AI tool identifies which *single bit* in the computational subspace, m_c, the agent i should flip in its current research practice d^i to yield the greatest improvement.

This design can be seen as a *personalized bit-greedy* procedure: It models systems that guide researchers toward the next most promising experiment or methodological change *for them*. This is unlike the non-personalized designs above, which simply recommend the computational portion of best or top-performing agents' practices without taking into account the specific AI querying user's existing non-computational decisions. This approach thus models a highly tailored, context-sensitive use of AI that optimizes its recommendations about where to intervene next for each specific user's context. This is a demanding and highly idealized design, and we return to the assumptions that are implicit in this conception in Section 6.[15]

If the two mechanisms of homogenization identified in Section 4 are correct, then we should have differential expectations about the effectiveness of these mitigation strategies. Randomization should address premature convergence when non-personalized AI is effective (at high modularity) by diversifying which recommendations agents receive, and preventing universal adoption of a single configuration. However, it cannot resolve context-mismatch problems and the opportunity cost of using non-personalized AI in low modularity settings. Personalization, in contrast, should address both mechanisms:

[15]We also conducted sensitvity test with an error prone version of Personalized AI, which suggests the best bit to change with a certain level of accuracy, but suggests a random bit with an error rate. The system performs slightly worse, but did not change the qualitative patterns discussed below.



(A) (B) (C) (D)

FIG. 3. Percentage differences in average outcomes between communities with access to AI tools and communities without AI (averaged across 1000 simulation runs). Communities consist of 100 agents with $H = 10$, facing a problem with $N = 20$, $K = 9$, $M = 10$, and $p_{social} = 0.1$. Panel (A) and (B) show *Personalized AI* effects on epistemic success at convergence and transient diversity, respectively, across problem modularity and AI use rates; panel (C) and (D) show these effects for *Randomized AI* across the same conditions. Colored cells indicate statistically significant differences: blue for improvements, red for declines, gray for non-significant ($p \geq 0.05$). Asterisks denote significance levels ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

by tailoring recommendations to individual contexts, it increases recommendation effectiveness (addressing opportunity costs), and by providing different recommendations to different agents, it naturally generates diverse search trajectories (preventing premature convergence). We examine the systemic effects of these designs through the same experimental parameters as Section 4.

5.2 Results

Our simulations reveal that personalization fundamentally alters AI's impact on collective epistemic performance, enabling benefits across structural conditions where non-personalized approaches fail. Figure 3a demonstrates Personalized AI's most striking advantage: in conditions where non-personalized AI proves ineffective or harmful, Personalized AI provides substantial improvements. This pattern validates our theoretical prediction that, by tailoring computational recommendations to each agent's specific non-computational configuration, Personalized AI avoids the cross-module interference that undermines the effectiveness of non-personalized recommendations in tightly-coupled problems.

Personalized AI's context-sensitivity has implications for epistemic diversity as well. We find that, especially at lower use rates, Personalized AI maintains or even slightly increases transient diversity, compared to no-AI baselines (see Figure 3b). This occurs because tailored recommendations guide agents toward *different* promising solutions suited to their distinct contexts, creating productive heterogeneity. This diversity-preserving property stands in stark contrast to non-personalized approaches, which reduce diversity across all conditions (see Figures 1b and 3d).

As shown in Figure 3d, Randomized AI offers no such benefits at low modularity. In this context, randomization performs comparably to the basic non-personalized AI, with both approaches ranging from ineffective to actively harmful as use rates increase. This similarity confirms that the fundamental limitation of non-personalized recommendations is not just a lack of variety in recommendations, but rather their insensitivity to individual contexts. Randomizing among top performers merely varies *which* context-mismatched recommendation each agent receives, failing to address the underlying structural challenge.

Interestingly, however, the comparative effectiveness of these mitigation strategies somewhat changes, under high modularity. As Figures 3a and 3c show, when $\rho(m) = 1$, Randomized AI not only maintains the benefits of non-personalized AI at higher query rates; it even slightly outperforms Personalized AI across all use rates. Notably, this advantage emerges despite randomization's non-personalized nature. At high modularity, module independence renders context-sensitivity less critical, and diverse computational configurations can succeed paired with various non-computational choices. Under these conditions, the diversity introduced by randomization sustains the benefits of simple non-personalized AI, even at higher use rates.

What accounts for the relative reduction in Personalized AI's benefits in fully decomposable problems with $\rho(m) = 1$? We can think of two complementary mechanisms. First, at $\rho(m) = 1$, each module operates as a highly complex but independent subproblem with multiple high-quality local optima. This can be because Personalized AI recommendations myopically guide agents along greedy paths toward whichever local optimum is nearest to their current position within this module. Given the higher internal complexity of the computational module in modular problems, this can more easily result in agents getting stuck in different local optima. Indeed, we find that increased modularity has a detrimental

impact on the performance of baseline (no AI) community. But, as Figure 3a shows, the drop in performance is even more substantial for the Personalized AI. This is because in the no AI community, the alternative computational decisions that the agents adopt through exploration are simply *better* than their current ones, preserving some transient diversity, whereas with Personalized AI agents adopt the myopically *best* decision change for the entire computational module (For a similar phenomenon see Wu 2024).

Second, problem decomposability can fundamentally alter the efficiency of division of labor in collectives relying on Personalized AI. At lower modularity, tight coupling between modules means that finding good configurations in m_c also helps in exploring the fitness contributions of decisions in $\neg m_c$. In perfectly decomposable problems, however, these benefits disappear. Module independence means agents with identical m_c configurations always receive identical recommendations, regardless of differences in $\neg m_c$. Personalized AI recommendations thus reveal nothing about the non-computational module,[16] and time agents allocate to querying AI—or to exploring computational decisions on their own—represents opportunity cost for optimizing the independently-complex non-computational module. This results in a defective division of cognitive labor across the two subproblems.

These two mechanisms suggest different mitigation strategies. If the issue is due to the myopic implementation of Personalized AI, then it can be addressed by a less myopic version of personalization that better deals with complexity—at least in context when this is technologically viable. If the issue is due to a defective division of cognitive labor, then the solution is not merely technological, but requires organizational restructuring.

To further explore this idea, we examined two further variants of the model above. First, we consider a two-bit version of the Personalized AI system, which recommends to agents the optimal *two bits* to change in the computational subspace (as opposed to the single bit case above). Intuitively, this means the system has a higher capability of guiding agents through complex problems. Second, we also explored what happens when communities *alter* their exploratory practices following AI adoption. Specifically, we compared communities where agents bias their individual exploration toward computational decisions

[16] Aside from their general irrelevance.

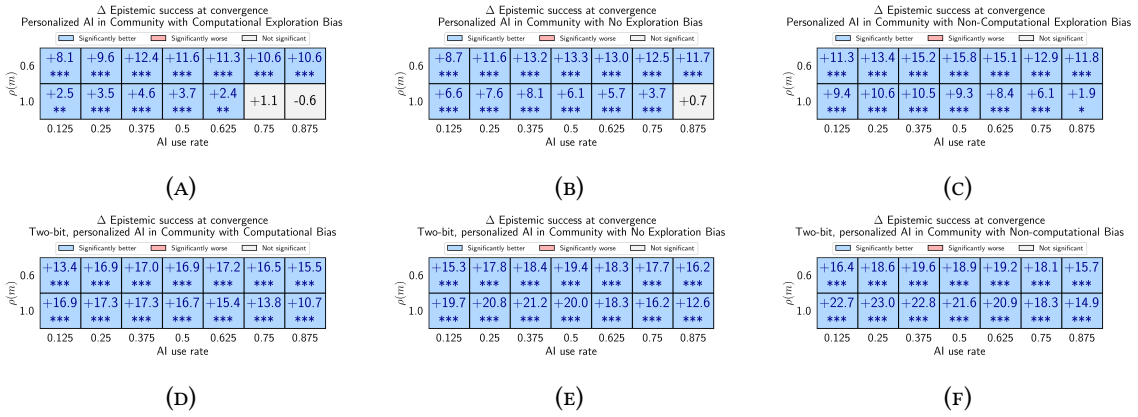


FIG. 4. Percentage differences in epistemic success at convergence between communities with access to personalized AI tools and communities without AI (averaged across 1000 simulation runs). Communities consist of 100 agents with specialization set size $H = 10$, navigating a problem with $N = 20$, $K = 9$, $M = 10$, while engaging in social learning with $p_{social} = 0.1$. Panels (A)-(C) show differences in communities with *single bit* Personalized AI (our default thus far), where agents' post-AI adoption search practice is either biased to *computational* (A), same (B), or biased to *non-computational* module (C). Panels (D)-(F) show the same in communities with *Two-bit* Personalized AI. Colored cells indicate statistically significant differences: blue for improvements, red for declines, gray for non-significant ($p >= 0.05$). Asterisks denote significance levels ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

(reinforcing AI's focus), toward non-computational decisions (complementing AI's scope), or explore uniformly without bias (our default condition in prior simulations).[17]

The results, shown in Figure 4, provide support for the mechanisms suggested above. Unsurprisingly, the more capable (Two-bit) Personalized AI not only outperforms the single bit Personalized AI (compare, for example, Figures 4e with 4b); it also performs *better* with increased modularity (compare $\rho(m) = 0.6$ with $\rho(m) = 1$ in Figures 4e). This makes sense given that increased modularity negatively impacts the performance of the baseline, no-AI community. What is more, for both single and two-bit implementations, exploration bias moderates their effectiveness statistically significant ways. When agents bias exploration toward non-computational decisions (panel 4c for single bit and panel 4f for two-bit version), performance improvements increase markedly across all modularity

[17]We implement exploration bias by adjusting agents' specialization sets H: computational bias draws 75% of bits in H from m_c, non-computational bias draws 75% from $\neg m_c$, and unbiased exploration—our default condition thus far—draws uniformly from all N decisions.

levels compared to computational bias (Panels 4a and 4d), and no bias (Panels 4b and 4e) with pronounced effects in fully decomposable problems.

Importantly, these results demonstrate that, apart from existing technological capabilities that can inform the effectiveness of AI systems in dealing with longer term problem complexity, Personalized AI's effectiveness also depends critically on how communities adapt their norms and practices to AI adoption. Productive AI-human collaboration emerges not from passive AI adoption, but from organizational adjustments that direct human effort toward complementary problems AI cannot address. We now turn to these broader implications of our findings for understanding AI's role in collective problem-solving.

6 DISCUSSION

Our simulations revealed four high-level results: First, Non-personalized AI's epistemic value is fundamentally conditional. Uniform recommendations yield benefits only under a narrow conjunction of highly modular problem structure, moderate use rates, and limited baseline capabilities. Outside these boundaries, non-personalized AI ranges from ineffective to actively harmful. Second, non-personalized AI systems trade coordination gains against transient diversity, with varied epistemic consequences. At high modularity, substantial diversity loss accompanies performance improvements, reflecting productive convergence. At low modularity, diversity loss occurs without compensating longer term gains, indicating premature convergence on and opportunity costs due to context-inappropriate solutions. Third, Randomization—a proposed technical mitigation—fails to address the underlying issue of context-mismatch. We find its utility is restricted to the same narrow structural conditions that enable non-personalized AI, offering no benefit for complex, tightly coupled problems. Fourth, Personalization enables robust benefits across structural conditions, while often preserving or enhancing diversity. Yet, realizing its full benefits can depend on organizational adaptation and coordinated adjustments in shared epistemic practices.

6.1 From Tool Adoption to Institutional Preparedness and Adaptation

Perhaps the key theme emerging from our experiments is that AI's impact on collective inquiry depends critically on how epistemic communities *adapt* their norms and practices to exploit AI's capabilities, while compensating for its limitations. This distinction between passive adoption and active adaptation emerges most clearly when considering the structural prerequisites for different AI designs and the institutional interventions that moderate their risks and benefits.

Non-personalized AI: Modularity, Randomization, and Use Moderation. Consider first the factors that determine and moderate the impacts of non-personalized AI. Our findings reveal that non-personalized AI effectiveness is restricted to highly modular problems with moderate use rates. Outside these conditions, context-insensitive recommendations provide little utility or, worse, accelerate premature consensus around inferior solutions. This restriction carries important organizational implications.

It underscores the need for careful structural assessment before deploying non-personalized AI systems. Institutions must evaluate whether target problems are sufficiently decomposable for uniform AI guidance to succeed. Such assessments require determining abstraction boundaries—identifying which decisions to focus on and which aspects can be safely treated as independent. This determination is complex and often contested. Recent debates about AI integration in high-stakes decision-making illustrate these challenges (Fazelpour and Lipton 2020; Selbst et al. 2019). These debates reflect deeper disagreements about whether socially desirable qualities in AI-based decision-making can be framed computationally, particularly in ways that are meaningfully decomposable from human, organizational, and social factors.

To be sure, modularity, which involves decomposability of problems and portability of solutions, is rarely a fixed property of scientific problems, but rather an achievement of organizational and disciplinary practices (Bowker and Star 2000; Fujimura 1987; Kitcher 1990; Leonelli 2019; Reijula et al. 2023). Standardization of workflows, data formats and experimental protocols, agreed-upon measurement and calibration frameworks, and established divisions of labor across disciplines create the sort of modular structure required by Non-personalized AI systems.

Even in sufficiently decomposable problems, the benefits of non-personalized AI guidance crucially depends on appropriately navigating the speed-performance (or efficiency-diversity) trade-offs. Our results suggest both technical and institutional interventions. The success of Randomized AI systems in modular settings, for example, provides supports for both technical interventions, such as injecting structured diversity, and social interventions, such as fostering a pluralistic model ecosystem.[18] Moreover, our results underscore the important role of institutionalizing appropriate usage norms (e.g., via query budgets) to limit excessive reliance on non-personalized systems and avoid premature lock-in.

Personalized AI: Context Legibility and Complementarity. Personalized AI systems introduce distinct requirements. Even beyond the technical prerequisites which may not be realistic in many domains, a necessary *institutional* precondition for personalized AI is *context legibility*. In our model, the Personalized system has perfect access to an agent's decision vector. In practice, however, making a research team's context legible to AI systems (or even to themselves and other teams) requires massive institutional work, including the creation of explicit protocols for identifying, documenting, and communicating many, typically tacit, aspects of scientific practice. Recent work on demands and challenges in developing effective transparency documentation for open science in general (Nosek et al. 2015; Stodden et al. 2014) and social applications of AI systems in particular (Pratt and Tanjaya 2025; Winecoff and Bogen 2025) illustrate these difficulties.

Institutional considerations also extend to factors moderating personalized AI's impacts. Beyond governance tools for preventing overuse, coordinated changes in exploratory practices enable better leveraging of Personalized AI capabilities. Our results show this is particularly salient in decomposable problems: when AI systems become effective in autonomously dealing with or assisting in computational aspects of the scientific practice, this effectiveness shifts the bottlenecks of epistemic practice to the non-computational module, calling for institutional restructuring, such as patterns of re-skilling and division of labor that emphasize non-computational expertise while maintaining sufficient computational knowledge and literacy to oversee AI interaction effectively.

[18] Of course, these approaches can differ along other dimensions of benefits and challenges. So, the overall choice, if one needs to be made, depends on those other considerations.

This consideration supports broader efforts toward designing AI systems that *complement* rather than replicate human expertise (Rastogi et al. 2023; Steyvers et al. 2022). In contrast to our default simulations where agents' capabilities are randomly drawn and jointly cover all problem dimensions equally well, in practice, scientific teams may exhibit systematic strengths or weaknesses. When such asymmetries exist, AI systems that address gaps and augment existing capabilities can provide additional gains. Notably, this is in contrast to prevalent AI development practices. For example, recent analyses of clinical AI benchmarks reveal misalignment between AI capabilities and practitioner needs, with benchmarks prioritizing tasks clinicians already handle well, while neglecting those where AI could provide the greatest complementary value (Blagec et al. 2023). Our results thus suggest a shift in development priorities: rather than replicating existing human strengths, AI systems should be designed to target capability gaps, thereby enabling a productive division of labor.[19]

The Value-Laden Nature of AI Preparedness. More broadly, our results highlight that successful AI adoption requires prior investment in shared practices and institutional infrastructure. Epistemic communities cannot simply adopt AI tools and expect benefits; they must often construct necessary preconditions—modularity, context legibility, complementary expertise—through standardization, protocol development, and workflow redesign. This aligns with recent empirical findings that many AI initiatives fail to scale, precisely because success hinges less on computational capabilities than on those prerequisite organizational restructuring (Van Noorden and Perkel 2023; Yee et al. 2025).

But, if realizing visions of AI-driven science is contingent on such institutional investment, then the choice to pursue AI preparedness becomes a thoroughly *value-laden* and *context-dependent* question. If in a given domain, the tasks facing the most acute resource constraints are non-computational, then, rather than depending on promised benefits of widespread AI adoption, scientific communities must seriously consider whether greater returns might not come from investing directly in those non-computational dimensions.[20]

[19]Recent proposals about "human-aware" AI that are tuned to generate promising hypotheses that are likely to be neglected by current epistemic communities also align with this suggestion (Sourati and Evans 2023).
[20]This is not to say that judicious use of AI, such as to address existing capability gaps, as discussed above, cannot be beneficial even in under-resourced settings. Rather, the point concerns the broader allocation of resources and attention.

6.2 Assumptions, Limitations, and Future Directions

Like any other model, the framework proposed here relies on idealizing assumptions that limit its direct applicability. Below we clarify some of these assumptions as a way to both better situate the implications of our findings, and identify directions for future research. One such assumption concerns the boundary between computational and non-computational decisions. In our model, this partition is fixed and exogenously specified. As mentioned in the previous subsection, however, in real scientific practice this boundary is neither natural nor static. It is shaped by methodological conventions, available infrastructure, and disciplinary norms. Understanding how such boundaries are constructed, and how they evolve alongside AI capabilities is therefore an important direction for future work.

A related idealization concerns the assumptions involved in modeling the AI recommendations. Consider, for example, two assumptions related to the Personalized AI recommendations. First, whereas in practice personalization typically depends on inferring user context—often understood as preferences—from incomplete and noisy data, our model treats an agent's current research practice (their decision configuration) as fully transparent and directly observable. Second, the Personalized AI has perfect access to the true payoff consequences of hypothetical changes to m_c, rather than learning outcome structure from data or generalizing uncertainty to unseen configurations.[21] In practice, both forms of knowledge—of user state and of outcome payoffs—are inferred, partial, and error-prone.

Moreover, our model abstracts away from additional risks and negative externalities associated with different AI designs. Personalization, for example, can result in potential amplification of confirmation bias at the individual level or polarization at the collective level (Kirk et al. 2024). These idealizations are nonetheless helpful, because they allow us to isolate the *structural* limits of personalization under the most favorable informational conditions. If even a fully informed and perfectly context-sensitive system exhibits reduced benefits in certain settings (e.g., under high modularity or high reliance), then these constraints apply with even greater force to real personalized AI systems whose access to user context and payoff structure is inherently limited.

[21] Though we also implement an imperfect and error-prone version in Section 5.

Similarly, our model does not capture the full costs associated with misleading or useless recommendations. In the simulations, agents immediately and accurately evaluate the fitness of any proposed change, and thus quickly reject recommendations that do not improve their epistemic position. Useless recommendations, therefore, incur only a minimal opportunity cost—agents forego a round of social learning or individual exploration. In real scientific practice, by contrast, researchers can lack the expertise or information needed to independently assess the quality of a suggested method, model, or assumptions, precisely in the domains where AI assistance is most attractive. In such cases, seemingly promising recommendations may have subtle or delayed adverse effects, and the harms of adopting a misguided approach—from misallocated resources and effort to retracted findings—may only become visible much later and are not easily mitigated (Ehsan et al. 2022; LaCroix et al. 2021). As a result, the opportunity costs of relying on inappropriate or context-mismatched AI suggestions are likely to be far greater in real scientific settings than our idealized model can reveal.

Another simplifying assumption of our model is that the incentives of individual agents are static and independent: an agent's payoff depends solely on the fitness of its own research practice, and not on the discoveries or timing of discoveries made by others. As a result, the model does not capture competitive dynamics in which the epistemic value of a discovery depends on when or by whom it is made. This abstraction allows us to examine the impacts of AI adoption on communities as a whole under different structural conditions, but it may obscure important tensions between individual and collective incentives.

Exploring such disconnects between individuals or groups and communities in light of AI adoption offers interesting directions for future research. For instance, across all AI types, we observe an inverted-U relationship between community-level performance and the rate of AI use. Future work could explore whether AI adoption creates a tragedy of the commons, where individual incentives to query AI for short-term efficiency diverge from the community's long-term need for diverse exploration. Such dynamics would constitute an additional mechanism of homogenization that our current model does not capture, and would highlight the need for institutions and governance mechanisms to realign individual benefits with collective epistemic success.

Finally, our simulations assume uniform potential for access across the community. Real scientific environments, however, are often characterized by significant resource stratification (see, e.g. Leonelli 2019). Future research can investigate the emergent dynamics that arise when AI tools are available only to specific subgroups. By modeling these disparities within structures where access is correlated with network position, for example, researchers can examine how AI adoption interacts with antecedent power asymmetries. Such work could reveal whether selective AI access amplifies existing influence hierarchies or creates new forms of epistemic inequality.

Taken together, our analysis underscores that the epistemic impact of AI integration depends as much on the social organization of inquiry as on the design of the tools themselves. The risks we identify arise even under idealized assumptions of costless coordination and perfect information. In real scientific communities, where expertise and resources are unevenly distributed and incentives are complex, these constraints will likely be even more acute. Avoiding the trap of epistemic monoculture, therefore, requires shifting our focus from the passive adoption of new technologies to the active design of the institutions that wield them.

REFERENCES

David Adam. 2024. The automated lab of tomorrow. *Proceedings of the National Academy of Sciences* 121, 17 (2024), e2406320121.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. 2023. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics* 137 (2023), 104274.

Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, et al. 2024. How large language models can reshape collective intelligence. *Nature human behaviour* 8, 9 (2024), 1643–1655.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical machine learning in healthcare. *Annual review of biomedical data science* 4, 1 (2021), 123–144.

Christophe Dessimoz and Paul D Thomas. 2024. AI and the democratization of knowledge. *Scientific data* 11, 1 (2024), 268.

Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The algorithmic imprint. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1305–1317.

Paul Erdős, Alfréd Rényi, et al. 1960. On the evolution of random graphs. *Publications of the* (1960).

Sendil K Ethiraj and Daniel Levinthal. 2004. Modularity and innovation in complex systems. *Management science* 50, 2 (2004), 159–173.

Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 57–63.

Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2021. Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly (MISQ)-Vol* 45 (2021).

Joan H Fujimura. 1987. Constructingdo-Able'problems in Cancer research: Articulating alignment. *Social studies of science* 17, 2 (1987), 257–293.

Martin Ganco and Glenn Hoetker. 2009. NK modeling methodology in the strategy literature: Bounded search on a rugged landscape. In *Research methodology in strategy and management*. Emerald Group Publishing Limited, 237–268.

Charles J Gomez and David MJ Lazer. 2019. Clustering knowledge and dispersing abilities enhances collective problem solving in a network. *Nature communications* 10, 1 (2019), 5146.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* (2025).

Patrick Grim, Daniel J Singer, Steven Fisher, Aaron Bramson, William J Berger, Christopher Reade, Carissa Flocken, and Adam Sales. 2013. Scientific networks on data landscapes: Question difficulty, epistemic success, and convergence. *Episteme* 10, 4 (2013), 441–464.

Alice CW Huang. 2024. Landscapes and bandits: A unified model of functional and demographic diversity. *Philosophy of Science* 91, 3 (2024), 579–594.

Jasmin Hume. 2025. *How AI-powered innovation can democratize breakthrough science*. World Economic Forum. https://www.weforum.org/stories/2025/06/ai-innovation-democratizes-breakthrough-science/ Accessed: 2025-11-27.

Shomik Jain, Kathleen Creel, and Ashia Camage Wilson. 2024. Position: scarce resource allocations that rely on machine learning should be randomized. In *Forty-first International Conference on Machine Learning*.

Stuart Kauffman and Simon Levin. 1987. Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology* 128, 1 (1987), 11–45.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* 6, 4 (2024), 383–392.

Philip Kitcher. 1990. The division of cognitive labor. *The journal of philosophy* 87, 1 (1990), 5–22.

Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118.

Travis LaCroix, Anders Geil, and Cailin O'Connor. 2021. The dynamics of retraction in epistemic networks. *Philosophy of Science* 88, 3 (2021), 415–438.

David Lazer and Allan Friedman. 2007. The network structure of exploration and exploitation. *Administrative science quarterly* 52, 4 (2007), 667–694.

Sabina Leonelli. 2019. *Data-centric biology: A philosophical study.* University of Chicago Press.

Yongchao Lu, Hong Wang, Lanting Zhang, Ning Yu, Siqi Shi, and Hang Su. 2024. Unleashing the power of AI in science-key considerations for materials data preparation. *Scientific Data* 11, 1 (2024), 1039.

Lisa Messeri and Molly J Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature* 627, 8002 (2024), 49–58.

Ryan Muldoon. 2013. Diversity and the division of cognitive labor. *Philosophy Compass* 8, 2 (2013), 117–125.

Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. 2015. Promoting an open research culture. *Science* 348, 6242 (2015), 1422–1425.

Jacob Pratt and Albert Tanjaya. 2025. Documenting the Impacts of Foundation Models. (2025).

Derek J De Solla Price. 1965. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science* 149, 3683 (1965), 510–515.

Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 2023. A taxonomy of human and ML strengths in decision-making to investigate human-ML complementarity. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 11. 127–139.

Sidney Redner. 2005. Citation statistics from 110 years of physical review. *Physics today* 58, 6 (2005), 49–54.

Samuli Reijula, Jaakko Kuorikoski, and Miles MacLeod. 2023. The division of cognitive labor and the structure of interdisciplinary problems. *Synthese* 201, 6 (2023), 214.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109* (2024).

Jamshid Sourati and James A Evans. 2023. Accelerating science with human-aware artificial intelligence. *Nature human behaviour* 7, 10 (2023), 1682–1696.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.

Victoria Stodden, Friedrich Leisch, and Roger D Peng. 2014. *Implementing reproducible research*. Vol. 546. Crc Press Boca Raton, FL.

Richard Van Noorden and Jeffrey M Perkel. 2023. AI and science: what 1,600 researchers think. *Nature* 621, 7980 (2023), 672–675.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature* 620, 7972 (2023), 47–60.

Amy Winecoff and Miranda Bogen. 2025. Improving governance outcomes through AI documentation: Bridging theory and practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

Jingyi Wu. 2024. Better than Best: Epistemic landscapes and diversity of practice in science. *Philosophy of Science* 91, 5 (2024), 1189–1198.

Jingyi Wu and Cailin O'Connor. 2023. How should we promote transient diversity in science? *Synthese* 201, 2 (2023), 37.

Lareina Yee, Michael Chui, Roger Roberts, and Stephen Xu. 2025. One year of agentic AI: Six lessons from the people doing the work. *McKinsey & Company* (12 September 2025). https://www.mckinsey.com/capabilities/quantumblack/our-insights/one-year-of-agentic-ai-six-lessons-from-the-people-doing-the-work

Yanbo Zhang, Sumeer A Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, et al. 2025. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence* 1, 1 (2025), 14.

Xiaoshi Zhong and Huizhi Liang. 2024. On the Scale-Free Property of Citation Networks: An Empirical Study. In *Companion Proceedings of the ACM Web Conference 2024*. 541–544.

Kevin JS Zollman. 2010. The epistemic benefit of transient diversity. *Erkenntnis* 72, 1 (2010), 17–35.