DOI: 10.1017/psa.2026.10267
This is a manuscript accepted for publication in *Philosophy of Science*.
This version may be subject to change during the production process.

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

Sina Fazelpour[1*], Joseph O'Brien[2] and Hannah Rubin[3]

[1]Northeastern University, Boston, United States; [2]University of California, San Diego, United States;
[3]University of Missouri, Columbia, United States
[*]Corresponding author: Sina Fazelpour; s.fazel-pour@northeastern.edu

Abstract

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 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 can enhance diversity, enabling benefits across a broader range of conditions. Crucially, these benefits are not automatic: they depend on effective institutional adaptation, requiring new standards, protocols, and practices.

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, for some researchers, AI integration promises to accelerate discovery by allowing diverse teams to perform complex inquiries, more efficiently and less hampered by 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).

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.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

Our results reveal that the impacts of AI adoption depend critically on AI system design, problem structure, and institutional norms and behaviors in the relevant communities. We first look at a type of *non-personalized* AI systems, which recommends uniform best practices without accounting for problem-solving context of querying epistemic agents. We find that this yields benefits only under highly specific conditions: when problems are easily decomposed into smaller sub-problems, such that those aspects handled by AI can be tackled somewhat independently, and when scientists rely on AI only moderately for their research. Outside these conditions, AI adoption results in premature convergence toward epistemic monocultures and harms longer term collective performance.

We implement two mitigation strategies inspired by proposals in the literature for counteracting the risks of AI-driven homogenization (Fügener et al., 2021; Jain et al., 2024): *randomization*, which introduces diversity in the selection of good recommendations by sampling from top-performing solutions rather than recommending a single best one; and *personalization*, which tailors recommendations to each agent's specific context. We find that, while useful in some settings, Randomization does not reliably increase performance across different problem-solving contexts; its utility remains restricted to the same structural conditions as Non-personalized AI. In contrast, Personalization proves robustly beneficial, albeit to varying extents, across various use rates and different kinds of problems epistemic communities may be attempting to solve.[1]

Importantly, our findings suggest that productive AI integration requires more than tool adoption. For Non-personalized AI systems, the requisite problem structure 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 AI systems, effectiveness requires established institutional standards for documenting and communicating often tacit human and organizational factors that shape scientific practice, and can vary 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 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.

[1]Terminological note: we use capitalization to distinguish between Randomization and Personalization *as implemented in our model* from the broader real-world phenomena, in order to create some separation between our findings and potential policy implications, which would require additional assumptions (e.g., see Section 6).

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

2. Background and Related Work

2.1. AI in Scientific Communities

AI systems are integrated into every stage of the research pipeline. They are employed to synthesize vast bodies of literature, formulate novel hypotheses, automate experimental designs, generate synthetic data, and perform multiverse analysis (Bertran et al., 2026; Gao et al., 2025; Gottweis et al., 2025; Lu et al., 2024; Messeri and Crockett, 2024; Si et al., 2024; Wang et al., 2023; Zhang et al., 2025).[2] These varied use cases draw on a correspondingly broad set of technical methods and architectures for tasks such as literature synthesis, prediction, data generation, and the optimization of experimental protocols. More recently, agentic and coding-capable AI systems have been used to orchestrate these diverse capabilities into multi-agent workflows that automate ever larger portions of the scientific pipeline with minimal human intervention (Baek et al., 2025; Lu et al., 2026; Schmidgall et al., 2025).

Across this variety in design and application, advocates highlight the shared promise of AI-driven science to accelerate scientific discovery. Beyond automating and speeding up routine tasks, AI systems are seen as broadening both the type of tasks that can be performed and the community that can perform them. For example, AI tools are seen as offering the promise of more efficiently navigating complex problem spaces, by identifying promising, but neglected directions of inquiry or novel patterns across vast, disparate datasets (Sourati and Evans, 2023; Wang et al., 2023). Moreover, these systems are said to "democratize" science by facilitating access to capabilities that previously required deep domain expertise and specialized resources (Dessimoz and Thomas, 2024; Hume, 2025). Similarly, emerging "cloud laboratories" and AI-driven experimental platforms may allow researchers to remotely execute and reproduce experiments, making procedures accessible to and implementable by a broader community (Adam, 2024; Lu et al., 2024).

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 that of epistemic communities more broadly (Burton et al., 2024; Hao et al., 2026; 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

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

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

of any computational tool 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" artifacts to study similar questions, however, they may 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.

Importantly, Messeri and Crockett's concern is not restricted to cases in which researchers consciously defer to particular AI systems on discrete research questions. As these systems become part of standard scientific infrastructure, they can embed consequential choices—e.g., how a research question is operationalized, which sources are consulted, or how results are filtered and ranked—that users may adopt without insights into what those choices are, how they were made, and why. In this way, AI systems can inconspicuously curate the space of methodological options that researchers encounter, adopt, and build upon, without appearing to restrict those choices at all (see also Passi, 2025).

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, which can be hampered by the 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 plausibly 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.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

3. General Modeling Framework

In this section, we describe the NK framework and how we extend it to model AI integration in scientific communities. To make the formalisms concrete, we employ a running example 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 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 beliefs about patient adherence and healthcare infrastructure, and more. We can represent each team's overall scientific strategy as a vector of N binary decisions, $d = (d_1, \ldots, d_N)$, 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 its constituent decisions work together to achieve certain scientific goals. If fitness contributions of decisions were independent, optimization would be straightforward: evaluate each decision separately, identify the better option, and combine them. Scientific decision-making is rarely so simple, however, since the value of one choice often depends on others. For instance, adopting certain machine learning tools in healthcare settings may be beneficial, only when coupled with 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.

3.2. Epistemic Agents

In practice, a research team's expertise and resources put constraints on which decisions it can independently evaluate and modify. In our drug discovery example, a computational biology team might specialize in virtual screening and molecular optimization,

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

[4]In our model, the K interdependencies of each decision d_i are randomly generated for each landscape. 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 $\max f(d) = 1$.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

but lack expertise in clinical trial design or manufacturing scale-up. A 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, by assigning each agent, a, a *specialization set*, h_a, that consists of a randomly generated subset of the total N decisions, where $|h_a| = H < N$. While we keep the size of H fixed for a given simulated community, the particular decisions that make up the specialization set differs between agents. 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 external source of guidance are scientific peers. Specifically, agents can learn from others to whom they are socially connected.[5] We model scientific communities as random networks (Erdős and Rényi, 1960) where each pair of agents is connected with a fixed probability p_{edge}. To ensure that (cluster of) agents do not remain isolated, we sample 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 neighbors and adopts the decision vector of the most successful one, provided that the solution improves its fitness.[6]

3.4. The Computational Module and Problem Decomposability

Increasingly, research teams can also depend on AI tools as an external source of guidance. 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 may be amenable to computational tools, other decisions, such as those concerning aspects of animal studies, qualitative study design, or patient-adherence considerations may involve factors that are less amenable to automation.[7] 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.[8]
(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 the two modules raises interesting methodological considerations that are not simply about the number of dependencies

[5]Here, we understand being a "neighbor" in terms of placement on a static, exogenously defined network structure.

[6]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.

[7]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.

[8]We assume that the computational module includes the first M of the N decisions for ease of exposition and without loss of generality.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

(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:

- $k_{\text{in}}(i)$: 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.[9]
- $k_{\text{out}}(i)$: 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* of the two modules:

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

A high $\rho(m)$ indicates that computational and non-computational modules 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 and methodological assumptions about data) or downstream considerations (e.g., qualitative patient study 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 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.[10] In Sections 4 and 5, we consider different designs for AI systems, 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

[9]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.

[10]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.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

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. Inferior peer solutions, unsuccessful explorations, and rejected AI recommendations, all carry opportunity costs, reflecting that evaluating alternatives requires time and resources.

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 that offers uniform best-practice recommendations. This "non-personalized" mode reflects typical deployments, where few 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 (NP AI)

When agent a queries the Non-personalized AI (NP AI) system, the tool returns the computational configuration used by the globally best-performing agent b. In this sense, NP AI acts akin to a social learning mechanism that is *global* in reach but *partial* in scope: it provides guidance about leading computational practices across all agents (not just the querying agent's local neighborhood), but it ignores how those practices interact with agents' non-computational choices.

This formalization captures some of the key 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.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

4.1.2. Dependent Measures

We focus on two measures to evaluate the epistemic effects of adopting NP 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 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.[11]

We compare otherwise identical communities differing only in AI access across a range of parameter settings.[12] Each configuration is simulated 1000 times.

4.2. Results

Our simulation results show that the impact of NP 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.

[11] In all conditions reported, communities reach convergence well within the 100-round simulation window. By the end of the simulation, communities have converged to a single shared solution in nearly all replications (mean unique solutions ≈ 1 and mean pairwise Hamming distance ≈ 0; for 100 agents, over 1000 simulation runs).

[12] Specifically, we keep $N = 20$ and $M = 10$ fixed across all communities, while varying $K = \{5, 9\}$, target modularity $\rho(m)$ from 0.6 to 1, $p_{social} = \{0.1, 0.2\}$, $H = \{5, 10, 15\}$, 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 $\rho \leq .5$

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

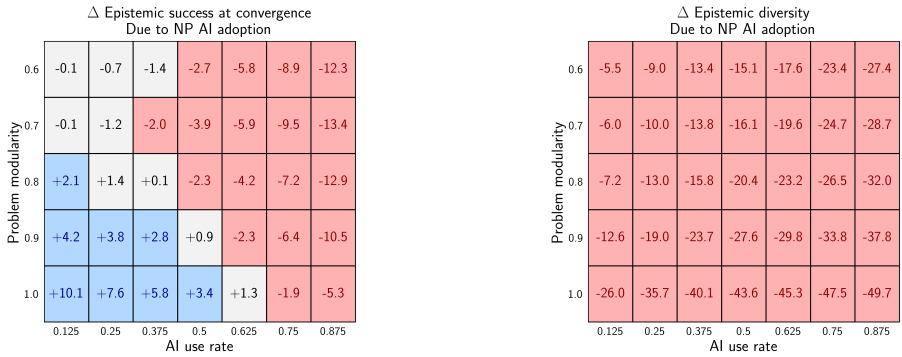


Figure 1: The differential impacts of introducing Non-personalized (NP) AI, expressed as *percentage differences* relative to communities without AI tools (averaged across 1000 simulation runs). Communities consist of 100 agents with specialization set size $H = 10$, navigating a problem with $N = 20$, $K = 9$, $M = 10$, 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 ($p < 0.05$) differences: blue for improvements, red for declines, gray for non-significant ($p >= 0.05$).

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, NP AI provides negligible benefits at best, and becomes actively detrimental at high use rates.

When it comes to transient epistemic diversity, we find that NP AI significantly reduces diversity across all settings by 25% on average (See Figure 1b). Importantly, 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 impacts on diversity and performance. At high modularity, NP AI reduces diversity substantially. Yet, this homogenization accompanies significant epistemic improvements, suggesting productive coordination around superior computational configurations. At low modularity, diversity loss is more modest, but occurs alongside performance stagnation or decline.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

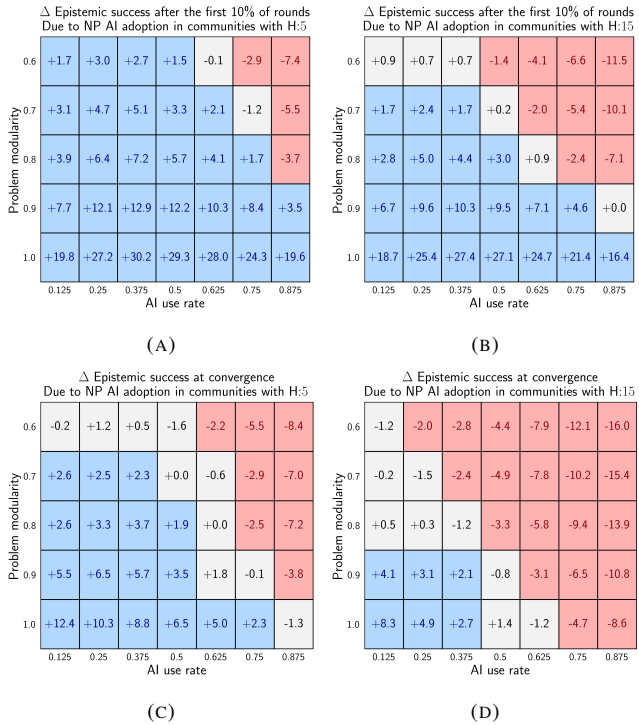


Figure 2: Difference in early versus final average epistemic success between communities with access to Non-personalized (NP) AI and communities without AI tools, expressed as percentage differences (averaged across 1000 simulation runs). Communities consist of 100 agents facing a problem with $N = 20$, $K = 9$, $M = 10$, 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 ($p < 0.05$) differences: blue for improvements, red for declines, gray for non-significant ($p \geq 0.05$).

When, then, is homogenization harmful? Our analyses identify two mechanisms through which NP AI induces diversity loss, with differential epistemic consequences. First, especially in contexts with higher modularity or use rate, which increase uptake of AI recommendations, NP AI produces dynamics similar to those documented in studies of social learning (Lazer and Friedman, 2007; Zollman, 2010). Specifically, NP AI recommendations based on the currently best-performing m_c configuration provide early gains, but can cause communities to converge prematurely on local optima. This mechanism is reflected in the monotonic decrease of NP AI's benefits as use rate increases, even for highly modular problems (Figure 1a).

Second, especially when lower modularity renders NP 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 querying AI, lead to marked reduction in the

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

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 that have more to gain from the potential benefits of transient diversity. When agents have small specialization sets ($H = 5$), they explore limited regions of the decision space, leaving gaps that NP AI can fill. With larger specialization sets ($H = 15$), agents collectively achieve broader coverage, and NP AI's marginal benefits may not justify its homogenizing costs. Figure 2 illustrates this pattern. NP AI offers substantial early gains regardless of specialization (Figures 2a and 2b). After convergence, however, while communities with $H = 5$ retain small but significant benefits (Figure 2c), those with $H = 15$ frequently see early gains reverse into losses (Figure 2d). Accordingly, while coordinating around NP AI recommendations accelerates initial progress, it 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 NP 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, NP AI harms collective performance by reducing valuable transient diversity without providing compensating coordination benefits, supporting Messeri and Crockett's concerns about AI-instigated epistemic monocultures.[13]

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 (TDR 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 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. We will refer to this implementation as Top-Decile Randomized AI, or TDR AI for short. 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.

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

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

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 (SBP 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. We will refer to this implementation as Single-Bit Personalized AI, or SBP AI for short.[14]

This design can be seen as a *personalized myopically 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 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. The randomization implemented in TDR AI should address premature convergence when NP 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 NP AI in low modularity settings. In contrast, the personalization afforded by SBP AI should address both mechanisms: 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.

[14] We also consider an alternative personalized design which evaluate each community member's computational module in the context of the querying agent's non-computational decisions, and recommends the one that yields the highest fitness for the querying agent. See the online-only appendix for details.

[15] We also conducted sensitivity test with an error prone version of SBP AI, which suggests the best bit to change with a certain level of accuracy $(1 - \varepsilon)$, suggesting a random bit otherwise (ε). The system performs slightly worse, but did not change the qualitative patterns discussed below.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

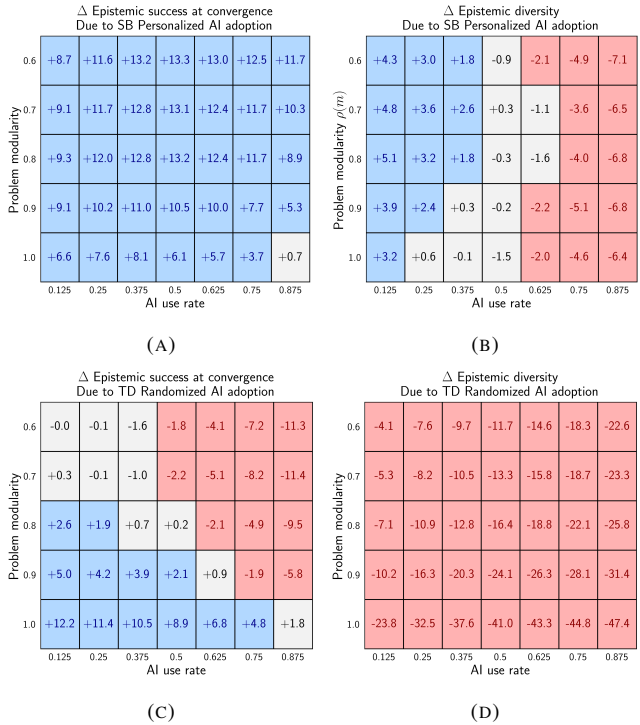


Figure 3: Percentage differences in average outcomes between communities with access to AI tools and communities without AI (averaged across 1000 simulation runs). Communities consist of 100 agents with $H = 10$, facing a problem with $N = 20$, $K = 9$, $M = 10$, and $p_{social} = 0.1$. Panel (A) and (B) show *Single-Bit (SB) 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 *Top-Decile (TD) Randomized AI* across the same conditions. Colored cells indicate statistically significant ($p < 0.05$) differences: blue for improvements, red for declines, gray for non-significant ($p \geq 0.05$).

5.2. Results

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

SBP AI's context-sensitivity has implications for epistemic diversity as well. We find that, especially at lower use rates, SBP 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* (compared to non personalized AI) and *more varied* (compared to non personalized AI as well as baseline communities

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

dependent only on their own specialization sets) promising solutions. This diversity-preserving property stands in stark contrast to our non-personalized approaches, which reduce diversity across all conditions (see Figures 1b and 3d).

As shown in Figure 3d, TDR AI offers no such benefits at low modularity. In this context, randomization performs comparably to SBP 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 due to a lack of variety in recommendations, but additionally their insensitivity to individual contexts. Randomizing among top performers merely varies *which* context-mismatched recommendation each agent receives without addressing the underlying structural challenge.

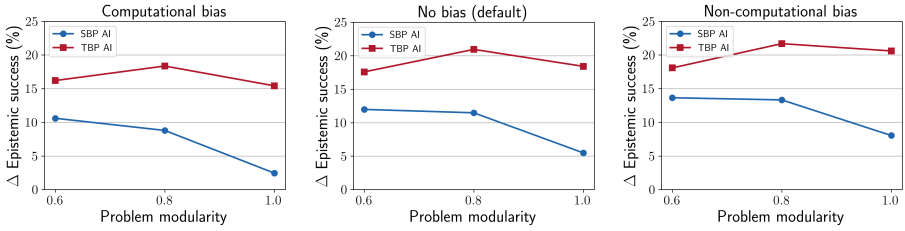
Interestingly, however, the comparative effectiveness of these mitigation strategies somewhat changes, under high modularity. As Figures 3a and 3c show, when $\rho(m) = 1$, TDR AI not only maintains the benefits of non-personalized AI at higher query rates; it even slightly outperforms SBP AI across 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 SBP AI's benefits in fully decomposable problems with $\rho(m) = 1$? We can think of two related 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 SBP 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 SBP AI. This may be 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 SBP 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 SBP 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 $\neg m_c$. SBP AI recommendations thus reveal little about the non-computational module, 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 enhancement strategies. If the issue is due to the myopic implementation of SBP AI, then it can be addressed by a less myopic

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press



(A) (B) (C)

Figure 4: Percentage differences in epistemic success at convergence between communities with access to two variants of personalized AI tools, compared to communities without AI (averaged across 1000 simulation runs and across AI use rates): blue lines show the difference due to adopting *Single-Bit Personalized (SBP) AI*, while red lines show the difference due to adopting *Two-Bit Personalized (TBP) AI*. 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 $p_{social} = 0.1$. Each panel shows one exploration bias condition: agents' post-AI-adoption search practice is biased toward (A) the *computational* module, (B) unbiased (our default thus far), or (C) biased toward the *non-computational* module.

version of personalization that better deals with complexity, at least in contexts 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 a restructuring of existing practices.

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 (TBP AI), 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 (reinforcing AI's focus), toward non-computational decisions (complementing AI's scope), or explore uniformly without bias (our default condition in prior simulations).[16]

The results, shown in Figure 4, provide support for the mechanisms suggested above. Unsurprisingly, the more capable TBP AI not only outperforms the SBP AI, across different exploration biases; it can also perform *better* with increased modularity (compare $\rho(m) = 0.6$ with $\rho(m) = 1$ in Figures 4b 4c). This makes sense given that increased modularity negatively impacts the performance of the baseline, no-AI community. What is more, for both SBP and TBP implementations, exploration bias moderates their effectiveness. When agents bias exploration toward non-computational decisions (Figure 4c), performance improvements increase markedly across all modularity levels compared to computational bias (Figure 4a), and no bias (Figure 4b) with more pronounced effects in fully decomposable problems.

[16]We implement exploration bias by adjusting how agents' specialization sets h_a are drawn: computational bias places $\lfloor 0.75\,H \rfloor$ of the H bits in m_c, subject to module-size constraints, and the remainder in $\neg m_c$ (non-computational bias does the reverse), while unbiased exploration—our default condition thus far—draws uniformly from all N decisions.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

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 can also depend on how communities adapt their norms and practices to AI adoption. Productive AI-human division of labor 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.

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 can 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—our implementation of 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—our implementation of another proposed mitigation—enables robust benefits across structural conditions, while often preserving or enhancing diversity. Yet, realizing its full benefits may depend on organizational adaptation and coordinated adjustments in shared epistemic practices.

Overall, the model does not only help assess claims about AI-driven homogenization; It also allows us to explore the effectiveness of mitigation strategies proposed in the AI literature, namely randomization among high-performing recommendations and personalization. These interventions are most naturally interpreted as design features of AI systems, rather than of other drivers of homogenization. Our findings about them are therefore especially relevant to the sociotechnical design and governance of AI tools in science.

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 indicate that the effectiveness of the non-personalized AI systems of

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

the sort modeled here may be restricted to highly modular problems with moderate use rates. Outside these conditions, context-insensitive recommendations may 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 such AI systems. In particular, our results suggest that institutions may want to evaluate whether target problems are sufficiently decomposable for uniform AI guidance to succeed. Such assessments would require determining abstraction boundaries—identifying which factors fall within the purview of a problem, 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 objectives in AI-based decision-making can be framed computationally, particularly in ways that are meaningfully decomposable from human, organizational, and social factors, and thus illustrate the challenge of decomposing complex social problems into clearly separable computational and non-computational modules.

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 TD 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.[17] Moreover, our results suggest an important role for 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 SBP and TBP AI systems have 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. Indeed, recent ethnographic works show how visions of AI in science are driving precisely this type of explication of previously tacit procedural knowledge and expertise at an unprecedented rate and scale, whether carried out by scientists themselves (in order to be able to use systems that require such context) or by institutionally hired automation engineers (Nelson and Mruk, 2025). Nonetheless, work on 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 potential difficulties.

[17] 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.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

Institutional considerations also extend to factors moderating personalized AI's impacts. Beyond governance tools for preventing overuse, depending on the design of such systems, coordinated changes in exploratory practices may be required for better leveraging their capabilities. Our results suggest this might be particularly salient in decomposable problems: when certain AI systems become effective in autonomously dealing with or assisting in self-contained computational aspects of the scientific practice, this effectiveness would shift the bottlenecks of epistemic practice to the non-computational module. This, in turn, calls 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.

This consideration supports broader efforts toward designing AI systems that *complement* rather than replicate human expertise (Rastogi et al., 2023; Steyvers et al., 2022). This is particularly importance, since, in contrast to our default simulations where agents' capabilities are randomly drawn and jointly cover all problem dimensions equally well, in practice, scientific communities may exhibit systematic strengths or weaknesses. When such patterns exist, AI systems that address gaps and augment existing capabilities may provide further 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 are thus in line with calls for a shift in development priorities: rather than replicating existing human strengths, AI systems should be designed to target capability gaps to enable a productive division of labor.[18]

The Value-Laden Nature of AI Preparedness

More broadly, our results suggest that epistemic communities cannot simply adopt AI tools and expect benefits. Rather, successful AI adoption requires prior investment in shared practices and institutional infrastructure. Context legibility and complementary expertise are such potential preconditions. Even, problem modularity, which enables portability of solutions, is rarely a fixed property of scientific problems, but rather an achievement of institutional practices around standardization, protocol development, workflow redesign, and divisions of disciplinary labor (Bowker and Star, 2000; Fujimura, 1987; Kitcher, 1990; Leonelli, 2016; Reijula et al., 2023).[19]

But, if realizing visions of AI-driven science is contingent on such institutional investment, then the choice to pursue AI preparedness becomes a *value-laden* and *context-dependent* question. How we prepare and what we choose to invest in depends on balancing many competing values and making judgments about which types of questions are most pressing, given resource constraints and existing capabilities. Even assuming a common aim of epistemic progress, the way forward is not automatically clear. If in a given domain, the tasks facing the most acute resource constraints are

[18]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).

[19]Unsurprisingly, recent empirical findings show the failure of AI initiatives that ignore those prerequisite organizational factors (Van Noorden and Perkel, 2023; Yee et al., 2025).

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

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]

6.2. Assumptions, Limitations, and Future Directions

Like any other model, the framework proposed here relies on abstractions and 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 neutral nor static. Understanding how such boundaries are constructed, and how they evolve alongside AI capabilities is therefore an important direction for future work.

Consider, moreover, two assumptions underlying the Personalized AI recommendations. First, our model treats an agent's current research practice (their decision configuration) as fully explicit and directly observable, whereas in practice such context needs to be inferred and made explicit—whether by AI systems, automation engineers, or users themselves (Nelson and Mruk, 2025)—from incomplete, noisy, or previously tacit information. Second, in our model, the Personalized AI perfectly predicts consequences of hypothetical incremental changes to m_c in virtue of perfect access to the true payoffs of the resulting configurations, rather than relying on inferred patterns in observed configurations to predict unseen ones.[21] In practice, both forms of knowledge—of user context and of outcome payoffs—are inferred, partial, and error-prone, which could undermine the benefits of personalization as modeled here.

Our model also abstracts away from factors that are necessary for understanding 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). Yet, in our model, agents are assumed to have direct access to fitness values of the landscape, and so cannot be said to possess belief-like representational states—that is, states that could *misrepresent* their target. As a result, our model does not capture how the conduct of epistemic agents may be biased by their particular pre-existing beliefs. For this reason, given that our model does not capture those risks, our results should not be interpreted as a naively optimistic endorsement of personalization.

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 an 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 or model, precisely in

[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.

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

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

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 dynamics in which the value of a contribution or discovery depends on when or by whom it is made (Kitcher, 1990; Strevens, 2003). This abstraction allows us to examine the impacts of AI adoption on communities as a whole under different structural conditions, but it can 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 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 disparities. 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 and other forces of homogenization. Such work could reveal whether selective AI access disrupts existing epistemic hierarchies or creates new forms of 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 the aforementioned idealized and simplifying assumptions. In real scientific communities, where expertise and resources are unevenly distributed, incentives are complex, these constraints will likely be more acute. Avoiding the trap of epistemic monoculture requires shifting our focus from the passive adoption of new technologies to the active design of the institutions that wield them.

Acknowledgments. We thank the editor and two anonymous reviewers for their careful and generous engagement and valuable comments.

Funding Statement. This work was supported by an AI2050 Early Career Fellowship from Schmidt Sciences.

Declarations. The authors declare no competing interests.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

References

Adam, D. (2024) The automated lab of tomorrow. *Proceedings of the National Academy of Sciences*. *121*(17), e2406320121. https://doi.org/10.1073/pnas.2406320121.

Baek, J., Jauhar, S. K., Cucerzan, S. and Hwang, S. J. (2025) Researchagent: Iterative research idea generation over scientific literature with large language models. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 6709–6738. https://doi.org/10.18653/v1/2025.naacl-long.342.

Bertran, M., Fogliato, R. and Wu, Z. S. (2026) Many ai analysts, one dataset: Navigating the agentic data science multiverse. *arXiv preprint arXiv:2602.18710*. https://doi.org/10.48550/arXiv.2602.18710.

Blagec, K., Kraiger, J., Frühwirt, W. and Samwald, M. (2023) Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*. *137*, 104274. https://doi.org/10.1016/j.jbi.2022.104274.

Bowker, G. C. and Star, S. L. (2000) *Sorting things out: Classification and its consequences*. MIT press. https://doi.org/10.7551/mitpress/6352.001.0001.

Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., Flek, L., Herzog, S. M., Huang, S., Kapoor, S., Narayanan, A., Nussberger, A.-M., Yasseri, T., Nickl, P., Almaatouq, A., Hahn, U., Kurvers, R. H. J. M., Leavy, S., Rahwan, I., Siddarth, D., Siu, A., Woolley, A. W., Wulff, D. U. and Hertwig, R. (2024) How large language models can reshape collective intelligence. *Nature Human Behaviour*. *8*(9), 1643–1655. https://doi.org/10.1038/s41562-024-01959-9.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K. and Ghassemi, M. (2021) Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*. *4*(1), 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757.

Dessimoz, C. and Thomas, P. D. (2024) AI and the democratization of knowledge. *Scientific Data*. *11*(1), 268. https://doi.org/10.1038/s41597-024-03099-1.

Ehsan, U., Singh, R., Metcalf, J. and Riedl, M. (2022) The algorithmic imprint. 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 1305–1317. https://doi.org/10.1145/3531146.3533186.

Erdős, P. and Rényi, A. (1960) On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*. *5*, 17–61.

Ethiraj, S. K. and Levinthal, D. (2004) Modularity and innovation in complex systems. *Management Science*. *50*(2), 159–173. https://doi.org/10.1287/mnsc.1030.0145.

Fazelpour, S. and Lipton, Z. C. (2020) Algorithmic fairness from a non-ideal perspective. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 57–63. https://doi.org/10.1145/3375627.3375828.

Fujimura, J. H. (1987) Constructingdo-able'problems in cancer research: Articulating alignment. *Social Studies of Science*. *17*(2), 257–293. https://doi.org/10.1177/030631287017002003.

Fügener, A., Grahl, J., Gupta, A. and Ketter, W. (2021) Will humans-in-the-loop become borgs? merits and pitfalls of working with AI. *MIS Quarterly*. *45*(3), 1527–1556. https://doi.org/10.25300/MISQ/2021/16553.

Ganco, M. and Hoetker, G. (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. pp. 237–268. https://doi.org/10.1108/S1479-8387(2009)0000005010.

Gao, Y., Lee, D., Burtch, G. and Fazelpour, S. (2025) Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*. *122*(24), e2501660122. https://doi.org/10.1073/pnas.2501660122.

Gomez, C. J. and Lazer, D. M. J. (2019) Clustering knowledge and dispersing abilities enhances collective problem solving in a network. *Nature Communications*. *10*(1), 5146. https://doi.org/10.1038/s41467-019-12650-3.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R. *et al.* (2025) Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*. https://doi.org/10.48550/arXiv.2502.18864.

Grim, P., Singer, D. J., Fisher, S., Bramson, A., Berger, W. J., Reade, C., Flocken, C. and Sales, A. (2013) Scientific networks on data landscapes: Question difficulty, epistemic success, and convergence. *Episteme*. *10*(4), 441–464. https://doi.org/10.1017/epi.2013.36.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

Hao, Q., Xu, F., Li, Y. and Evans, J. (2026) Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*. *649*(8099), 1237–1243. https://doi.org/10.1038/s41586-025-09922-y.

Huang, A. C. W. (2024) Landscapes and bandits: A unified model of functional and demographic diversity. *Philosophy of Science*. *91*(3), 579–594. https://doi.org/10.1017/psa.2023.169.

Hume, J. (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.

Jain, S., Creel, K. and Wilson, A. C. (2024) Position: scarce resource allocations that rely on machine learning should be randomized. Forty-first International Conference on Machine Learning. https://doi.org/10.48550/arXiv.2404.08592.

Kauffman, S. and Levin, S. (1987) Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*. *128*(1), 11–45. https://doi.org/10.1016/S0022-5193(87)80029-2.

Kirk, H. R., Vidgen, B., Röttger, P. and Hale, S. A. (2024) The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*. *6*(4), 383–392. https://doi.org/10.1038/s42256-024-00820-y.

Kitcher, P. (1990) The division of cognitive labor. *The journal of philosophy*. *87*(1), 5–22. https://doi.org/10.2307/2026796.

Kleinberg, J. and Raghavan, M. (2021) Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*. *118*(22), e2018340118. https://doi.org/10.1073/pnas.2018340118.

LaCroix, T., Geil, A. and O'Connor, C. (2021) The dynamics of retraction in epistemic networks. *Philosophy of Science*. *88*(3), 415–438. https://doi.org/10.1086/712817.

Lazer, D. and Friedman, A. (2007) The network structure of exploration and exploitation. *Administrative Science Quarterly*. *52*(4), 667–694. https://doi.org/10.2189/asqu.52.4.667.

Leonelli, S. (2016) *Data-centric biology: A philosophical study*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226416502.001.0001.

Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D. and Clune, J. (2026) Towards end-to-end automation of AI research. *Nature*. *651*(8107), 914–919. https://doi.org/10.1038/s41586-026-10265-5.

Lu, Y., Wang, H., Zhang, L., Yu, N., Shi, S. and Su, H. (2024) Unleashing the power of AI in science-key considerations for materials data preparation. *Scientific Data*. *11*(1), 1039. https://doi.org/10.1038/s41597-024-03821-z.

Messeri, L. and Crockett, M. J. (2024) Artificial intelligence and illusions of understanding in scientific research. *Nature*. *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0.

Muldoon, R. (2013) Diversity and the division of cognitive labor. *Philosophy Compass*. *8*(2), 117–125. https://doi.org/10.1111/phc3.12000.

Nelson, N. C. and Mruk, K. (2025) Automating the scientific frontier: Emerging laboratory forms and the redistribution of tacit knowledge. Talk presented at a Data & Society workshop.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R. and Yarkoni, T. (2015) Promoting an open research culture. *Science*. *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374.

Passi, S. (2025) Agentic AI has a human oversight problem. *Available at SSRN 5529058*. https://doi.org/10.2139/ssrn.5529058.

Pratt, J. and Tanjaya, A. (2025) Documenting the impacts of foundation models. Partnership on AI. https://partnershiponai.org/paper/documenting-the-impacts-of-foundation-models/.

Rastogi, C., Leqi, L., Holstein, K. and Heidari, H. (2023) A taxonomy of human and ML strengths in decision-making to investigate human-ML complementarity. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. pp. 127–139. https://doi.org/10.1609/hcomp.v11i1.27554.

Reijula, S., Kuorikoski, J. and MacLeod, M. (2023) The division of cognitive labor and the structure of interdisciplinary problems. *Synthese*. *201*(6), 214. https://doi.org/10.1007/s11229-023-04193-4.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z. and Barsoum, E. (2025) Agent laboratory: Using LLM agents as research assistants. *Findings of the Association for*

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press

Computational Linguistics: EMNLP 2025. pp. 5977–6043. https://doi.org/10.18653/v1/2025.findings-emnlp.320.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. and Vertesi, J. (2019) Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 59–68. https://doi.org/10.1145/3287560.3287598.

Si, C., Yang, D. and Hashimoto, T. (2024) Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*. https://doi.org/10.48550/arXiv.2409.04109.

Sourati, J. and Evans, J. A. (2023) Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*. *7*(10), 1682–1696. https://doi.org/10.1038/s41562-023-01648-z.

Steyvers, M., Tejeda, H., Kerrigan, G. and Smyth, P. (2022) Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*. *119*(11), e2111547119. https://doi.org/10.1073/pnas.2111547119.

Stodden, V., Leisch, F. and Peng, R. D. (2014) *Implementing reproducible research*. vol. 546. Crc Press Boca Raton, FL. https://doi.org/10.1201/b16868.

Strevens, M. (2003) The role of the priority rule in science. *The Journal of philosophy*. *100*(2), 55–79. https://doi.org/10.5840/jphil2003100224.

Van Noorden, R. and Perkel, J. M. (2023) AI and science: what 1,600 researchers think. *Nature*. *621*(7980), 672–675. https://doi.org/10.1038/d41586-023-02980-0.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C. W., Bengio, Y. and Zitnik, M. (2023) Scientific discovery in the age of artificial intelligence. *Nature*. *620*(7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2.

Winecoff, A. and Bogen, M. (2025) Improving governance outcomes through AI documentation: Bridging theory and practice. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–18. https://doi.org/10.1145/3706598.3713814.

Wu, J. (2024) Better than best: Epistemic landscapes and diversity of practice in science. *Philosophy of Science*. *91*(5), 1189–1198. https://doi.org/10.1017/psa.2023.129.

Wu, J. and O'Connor, C. (2023) How should we promote transient diversity in science? *Synthese*. *201*(2), 37. https://doi.org/10.1007/s11229-023-04037-1.

Yee, L., Chui, M., Roberts, R. and Xu, S. (2025) One year of agentic AI: Six lessons from the people doing the work. McKinsey & Company. `https://www.mckinsey.com/capabilities/quantumblack/our-insights/one-year-of-agentic-ai-six-lessons-from-the-people-doing-the-work`.

Zhang, Y., Khan, S. A., Mahmud, A., Yang, H., Lavin, A., Levin, M., Frey, J., Dunnmon, J., Evans, J., Bundy, A., Dzeroski, S., Tegner, J. and Zenil, H. (2025) Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*. *1*(1), 14. https://doi.org/10.1038/s44387-025-00019-5.

Zollman, K. J. S. (2010) The epistemic benefit of transient diversity. *Erkenntnis*. *72*(1), 17–35. https://doi.org/10.1007/s10670-009-9194-6.

https://doi.org/10.1017/psa.2026.10267 Published online by Cambridge University Press