How Artificial Intelligence Steers the Course of Science

Petr Spelda[1] and Vit Stritecky, Department of Security Studies, Charles University

Abstract

General-purpose AI models used as co-scientists are developed with the help of public evaluations that represent performative feedback from scientists pushing for greater levels of AI research assistance and acceleration with every model improvement. Formal methods that can ensure the risk of mismatches between model capabilities and scientific requests approaches the optimum over time are insufficient to guarantee science will be accelerated evenly, remain diverse and keep widening its scope. Should validity be the sole peer review criterion determining problem acceptance to impactful AI benchmarks when co-scientist models steer science by disproportionately improving in areas covered with this performative feedback? We show that while validity is necessary, it does not always support strategic selection. Peer reviewing problems for AI benchmarks as performative feedback requires a new competence that weighs validity and steering of co-scientist models jointly and with foresight. A failure to meet the challenge could ruin even acceleration by valid feedback that does not exploit performativity for steering AI co-scientists away from a monoculture. The key insight is to maintain benchmark diversity by keeping model scores on a benchmark hard to predict from existing evaluations.

1. Introduction

Specialized AI (Artificial Intelligence) models advanced some parts of science more than others (Jumper et al. 2021; Lin et al. 2023; Romera-Paredes et al. 2024; Avsec et al. 2026). Until recently, they were the focus of attention, which is now shifting toward general purpose AI models and the ways in which they can change science (Bubeck et al. 2025; Gottweis et al. 2025; Novikov et al. 2025; Feng et al. 2026a; Woodruff et al. 2026). Here, we characterize how growing capabilities of AI models used as co-scientists steer the course of science, a question that goes beyond traditional debates asking about AI and scientific discovery (e.g., Wang et al. 2023; Spelda and Stritecky 2021 for a philosophy of science reflection).

We show that public, impactful AI benchmarks and challenges become channels of performative feedback characterizing requested model capabilities that are not met in the current version of the models. Against intuition, improving model capabilities does not lead to saturation but rather repetitions of the cycle that steers the course of science by improving AI

[1] Correspondence to petr.spelda@fsv.cuni.cz, preprint, under review. Date: April 4, 2026.

co-scientist models on performative feedback. On examples of scientific benchmarks, we demonstrate how evaluation channels of performative feedback operate in practice. Most importantly, we show that while algorithms learning from the feedback can steer AI co-scientist models toward the formal, performative optimum (Jagadeesan et al. 2022), model capabilities approaching this optimum might not accelerate science in the way many are hoping for. The reason for this lies in the fact that predicting how performative feedback changes ways in which AI co-scientists are used and subsequently updated, effectively steering the course of science, is an open problem. Even and beneficial acceleration of science under wide adoption of AI co-scientist models requires strategic steering, which represents a significant policy challenge between public interest in science and the most capable privately owned and developed AI co-scientist models.

A first step in addressing this challenge is understanding the necessary changes to peer review of candidate tasks for impactful scientific AI benchmarks that serve as performative feedback, through which AI models steer the course of science. In Section 2, we first explain how performative prediction by model developers can gradually approach the optimal risk of mismatching model capabilities and user requests influenced by the model itself. In Section 3, we focus on the key insight regarding science steering with AI co-scientists:

> Peer review of tasks for scientific AI benchmarks focusing only on validity misses the steering opportunity for beneficial science acceleration. Understanding peer review panels as juries (Arvan et al. 2025) means correct decisions on candidate tasks depend on independent and competent reviewers. Competence to independently evaluate output validity is an integral part of science since its beginning. The ability to determine how valid scientific tasks that become part of performative feedback change capabilities of co-scientist AI models represents a significant departure from how competence was understood in peer review in the past.

The fact that model developers can approach the performative optimum as closely as possible does not guarantee science will be accelerated evenly and not lose diversity. This outcome depends on the new reviewer competence that does not stop at asking about task validity but also solicits tasks and their revisions based on what is considered strategic for steering from understanding the impact of performative feedback on capabilities of downstream AI models. This will be a formidable challenge due to the lack of transparency in the development of closed as well as open-weight frontier models (Wan et al. 2025).

Amidst general concerns for AI-caused homogenization that could lead to a science monoculture (Messeri and Crockett 2024; Hao et al. 2026; Traberg et al. 2026), we contribute to the debate in two crucial ways:

Contribution: 1) We precisely characterize how performative feedback changes capabilities of AI models that contribute to steering science and 2) design a new, *actionable intervention* that makes peer review of candidate problems for scientific AI benchmarks effective in steering science away from the mono to a polyculture. The key insight lies in maintaining benchmark diversity by keeping model scores on a benchmark hard to predict from existing evaluations (precise statement of the mechanism is in Figure 1).

2. Performative Feedback Improves AI Co-Scientists and Steers Science

Our characterization uses the concept of performative prediction (Perdomo et al. 2020; Hardt and Mendler-Dünner 2025) to capture that each time a better AI model becomes available, the improvement of its capabilities influences which tasks people want the model to perform well. The currently available model can be thought of as a prediction by its developer on the probability distribution $\mathcal{D}_{tasks}$ over the task domain users are interested in. This distribution can shift because users realized that the improved capabilities of the current model enable them to execute tasks that were difficult or impossible to accomplish with its earlier versions. This shift then prompts further model improvements following the failure to correctly forecast the users, and the cycle repeats itself.

The model developer seeks to be optimal in minimizing the performative risk induced by deploying the model to a population of responding users (Hardt and Mendler-Dünner 2025). The optimality can be defined as unimprovable minimization of performative regret, which measures the difference between the performative risk of the model selected at each time step and the performative minimum[2] (Jagadeesan et al. 2022). By observing which tasks work well against the model, users respond by trying related tasks that could be executed before with only partial success or not executed at all. The model developer observes these unsuccessful

[2] Performative regret at time T is defined as $\mathrm{Reg}(T) = \sum_{t \leq T} \mathrm{PR}(\theta_t) - \mathrm{PR}(\theta_{PO})$, where θ_t is a version of the co-scientist AI model updated with the performative feedback from previous round and PR its performative risk defined as $\mathrm{PR}(\theta) = \mathbb{E}_{z \sim \mathcal{D}_{tasks}(\theta)} \ell(z, \theta)$ (Jagadeesan et al. 2022), a performative optimum is defined as $\theta_{PO} \in \mathrm{argmin}_\theta \mathrm{PR}(\theta)$ (ibid.). ℓ is a loss function measuring success of the model on a user task z drawn from a distribution induced by the model, $\mathcal{D}_{tasks}(\theta)$ (ibid.), here referring to situations in which scientists finding out about improved capabilities of the AI model require it to perform more challenging tasks, representing another round of performative feedback for model updates. As a result, versions of the model, $\theta_1, \ldots, \theta_T$, evolve over time thanks to this shifting probability distribution over the domain of scientific tasks induced by improvements of the model itself.

attempts, which are samples from the shifted distribution over tasks, and updates the model (e.g., using one of the post-training methods such as reinforcement learning with verifiable rewards for large language models, Lambert et al. 2024; Jaech et al. 2024; Guo et al. 2025), shaping its capabilities inherited from the base model acquired by pretraining. Proxies for this observability are problems included in impactful, public scientific AI benchmarks that we deal with here. The failed attempts to run partially supported or unsupported tasks are performative feedback that allows the developer to minimize their performative regret using, for example, a modified multi-armed bandit algorithm from online learning (Jagadeesan et al. 2022) to minimize mismatches between supported and requested tasks. This means that performative feedback shapes capabilities of AI models that in turn influence which new tasks are requested by users. It is not the only source of information that helps the developer update their model but acting on performative feedback ensures that the model will be used as much as possible because its improvements copy the trajectory of tasks required by users in response to previous updates.

If the user population includes scientists who use the model as a co-scientist, assisting in and accelerating their research, their performative feedback steers the course of science by determining what the model will and will not be able to do well enough to continue enabling further research:

> As AI models become widely adopted as co-scientists in everyday research practice, performative feedback steers the course of science by helping to improve the models for next scientific requirements. They will be only partially matched by the next model as in previous rounds, paradoxically thanks to its improved capabilities triggering further, more demanding improvement rounds.

We should ask to what degree model developers exercise their performative power (Hardt et al. 2022) in this improvement cycle. Hardt et al. (2022) showed that performative power, the ability to steer users to more predictable states benefiting the developer, is strongest in monopolistic systems. This means that individual performative power is low under parallel competing AI co-scientist models from which human scientists can choose. While consolidation of closed frontier AI models, which are most useful as co-scientists, could theoretically happen, it is even less likely that geopolitical competition in AI could end soon (Wang and Siler-Evans 2026). This competition will most likely sustain competing AI co-scientist models that could be accessed despite political tensions. However, the main reason for not exercising performative power even in a monopolistic system is the fact that steering is worthless for learning scientific

performative feedback that allows the developer to improve their model and promote its use as the state-of-the-art AI co-scientist.

Multiple AI co-scientist models whose improvements are based on performative feedback can make predicting the course of science difficult. Piliouras and Yu (2023) showed that in a simple reinforcement learning setting, multiple agents with large influence over the data distribution, in our case $\mathcal{D}_{tasks}$, do not converge to the performative optimum but rather transition to instability and chaos. Even though this simple setting does not fit the methods used for updating LLMs (large language models) underpinning AI co-scientists, it highlights the following question:

> While the steering mechanism can be characterized as performative prediction, predicting the course of science influenced by AI co-scientist models, updated with performative feedback, is an open question.

Having a formal convergence guarantee is important but not as important as understanding how scientists become the source of performative feedback for improving AI co-scientist models for next rounds of assisted research. Algorithmic collective action (Hardt et al. 2024) showed how small groups of individuals can achieve goals by modifying their data on digital platforms despite the influence of machine learning models used for classifying the individuals. As a formal mechanism, collective algorithmic action is close in spirit to performative feedback steering science through improvements of AI co-scientist models. The performative feedback we discuss here is *not collected via platform surveillance but is obtained from AI benchmarks*, scientific challenges and similar means, targeting individual disciplines in an in-depth manner or providing wide spectrum AI evaluation suites. We focus on these evaluation channels of performative feedback because they serve as a funnel through which performative feedback reaches the developer and compels them to update the prior distribution over tasks their model should be able to perform well:

> Evaluation channels of performative feedback not only improve the model for everybody but, more importantly, steer the course of science by determining where the model offers interesting opportunities for improvement in next rounds because past performative feedback made it good enough for assisting and accelerating research in those areas. This is partially a result of trading evaluation for learning by letting public impactful benchmarks inspire curricula designed to improve external validity of AI models (cf. Spelda and Stritecky 2025).

This kind of steering, which flows through intermediaries developing the AI co-scientist models, is a new epistemic development in science. Under wide adoption of AI co-scientists, there are key insights regarding performative feedback the scientific community should act upon to ensure even and beneficial acceleration of science with AI. This outcome is not guaranteed without carefully considering which AI evaluations get elevated to the position of high-impact public, scientific benchmarks or challenges. They will become performative feedback that steers the course of science by providing material for updating AI co-scientist models.

3. Performative Feedback via AI Evaluation Channels

We demonstrate the general mechanism behind evaluation channels of performative feedback on two rolling AI benchmarks that target the most capable, frontier AI models used directly as co-scientists or as models in harnesses or scaffolds that make up AI agents. Rolling benchmarks are sets of AI evaluation problems that are continually updated with performative feedback from the community that is responding to model improvements from past performative feedback. Static benchmarks are a source of performative feedback as well until AI models reach high scores on them. This can take some time depending on the benchmark hardness and the attention that model developers pay to individual benchmarks. The attention can be influenced by quality of the benchmark, and apart from expanding or deepening its scope and focus, updates can also serve as subsequent rounds of peer review that improve quality of the benchmark. In sum, the influence of performative feedback channels based on AI benchmarks depends on their quality and the way in which they respond to model improvements with updates. Task quality and updates of impactful scientific benchmarks are shaped by peer review of candidate problems. The peer review offers an opportunity for intervening in how co-scientist AI models steer science. We look into two recent examples, Humanity's Last Exam (Phan et al. 2026) and First Proof (Abouzaid et al. 2026a), to explain what competence will be required from reviewers and, more broadly, benchmark organizers to use this opportunity for even acceleration of diverse science. These two examples are instructive because HLE (Humanity's Last Exam) is the most comprehensive scientific benchmark of close-ended questions today and FP (First Proof) targets non-competition problems that arise naturally in mathematical research.

HLE is characterized as an AI benchmark that evaluates scientific reasoning across many different fields of natural science, social science, and humanities, consisting of close-ended and multiple-choice problems (Phan et al. 2026). The original dataset consists of 2,500

problems[3], and many frontier model developers self-report performance of their best models[4], which gradually increases[5] and shows that HLE became an impactful performative feedback channel. The original dataset was forked into HLE-Rolling[6] implementing community feedback by adding new problems and updating or removing existing ones identified as faulty by the community (Phan et al. 2026). The peer review for HLE comprised of two rounds of expert review against format and validity requirements from an evaluation rubric (Phan et al. 2026, Supplementary information, Section 2), followed by two rounds of subsample auditing, attempting to increase problem validity by expert agreement, and a 'bug bounty' effort to eliminate as many faulty problems as possible by community feedback (Phan et al. 2026). The only steering involved banning problems concerning virology, chemical, biological, radiological, and nuclear weapons or cyberattacks against critical infrastructure (Phan et al. 2026, Supplementary information, Section 1), which adheres to generally accepted safety alignment of AI models (Mazeika et al. 2024, Li et al. 2024, Götting et al. 2025) and does not represent polyculture steering of science that we have in mind here.

First Proof started as an experiment evaluating capabilities of frontier models to autonomously produce proofs for research-level questions that were identified by mathematicians during their research, solved and not made public to make the challenge unseen for the evaluated models (Abouzaid et al. 2026a). The first batch of FP problems consisted of 10 problems at a difficulty level described as well-specified by their authors because it did not involve finding new research questions but rather only providing answers to the selected problems (ibid.). The challenge attracted attention of frontier model developers, attempting to solve the first batch of problems with some success (Feng et al. 2026; OpenAI 2026) before their solutions were released publicly (Abouzaid et al. 2026b). Before providing details on the second batch of FP problems, Abouzaid et al. (2026b) noted that apart from verifying solutions were produced autonomously and securing that they can be peer reviewed in a sustainable way, the organizers intend to characterize the problem selection process. The First Proof Editorial Board (2026) later specified that they will solicit problems broadly, require that problem authors do not have conflict of interest with AI companies (disclosure of employment or consulting

[3] https://huggingface.co/datasets/cais/hle

[4] https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, https://openai.com/index/introducing-gpt-5-4/, https://www.anthropic.com/news/claude-opus-4-6, https://huggingface.co/deepseek-ai/DeepSeek-V3.2

[5] https://agi.safe.ai

[6] https://huggingface.co/datasets/cais/hle-rolling, changelog: https://github.com/centerforaisafety/hle/blob/main/hle-rolling-changes.txt

during 3 months prior to submission, committed to not consult with them 6 moths afterward, and no submissions from AI companies in conflict with FP authors), and limit problem solutions to 8 pages including references. Thanks to its unique organization, FP attracted attention of frontier AI labs and similarly as HLE became an impactful channel of performative feedback.

As a result, the peer review of HLE and FP is built on the assumption that mean expert competence of determining validity of candidate problems is better than 0.5 because, then, if the validity is determined by majority vote, adding experts to the referee panel increases the probability that the vote will be correct (Arvan et al. 2025). While for epistemic theories of democracy this assumption is among reasons why jury theorems cannot easily justify democracy (Goodin and Spiekermann 2018; Spelda et al. 2024), in science it can be uncontroversial thanks to the fact that in the ideal case reviewers are practicing, truth-seeking scientists. If, however, reviewed problems become part of performative feedback via influential scientific AI benchmarks like HLE or FP, validity is not enough to avoid a science monoculture. New candidate problems are likely to seek improvement of model capabilities in areas already covered by past performative feedback. The exploitation of path-dependencies will prioritize improvement of model capabilities covering some conceptual frameworks and methodologies at the expense of others. This then becomes part of a broader 'AI monoculture feedback loop' proposed by Traberg et al. (2026).

The performative optimum pursued by model developers seeking to match model capabilities to user requests over time is independent of the content of performative feedback itself. The performative optimum is merely a solution concept which is filled with model capabilities co-determined by the competence required from experts during peer review of new candidate problems for scientific benchmarks like HLE or FP. It is unlikely that the competence selecting for validity alone will be able to sustain science diversity. The question is how to measure that peer review of candidate problems actually steers co-scientist AI models apart from separating valid problems from invalid ones? Perceived diversity of individual candidate problems could be deceiving due to small scale comparisons or the inability of reviewers to set apart problems that are foundationally different from those that frame only a small number of issues in multiple complex ways. Increasing diversity of problem proposers and reviewers helps to avoid homogenization but does not offer a way of checking whether despite the effort co-scientist AI models remain diverse and do not contribute to a monoculture. To help address this issue, we provide the following mechanism based on the hardness of predicting model scores on a benchmark from existing evaluations:

A benchmark's diversity can be expressed as unpredictability of held-out model scores from the rest of the score matrix across other models and benchmarks (Figure 2). The more unpredictable models are on a benchmark by an aggregated prediction error the less likely it is that the benchmark contributed to a science monoculture by recreating existing evaluations. By a similar token, a continually updated benchmark sustains diversity if it is difficult to predict held-out model scores after some time t from the score matrix before t across other models and the benchmark versions (Figure 3). The more unpredictable models are after t by an aggregated prediction error the less likely it is that the updates contributed a science monoculture by recreating already covered evaluations. Predictors of model performance need to be continually updated as the matrix of models and benchmark performance grows to adapt to out-of-population scores introduced by new models, benchmarks or their updates.

Figure 1: AI benchmark diversity as unpredictability of model scores from existing evaluations.

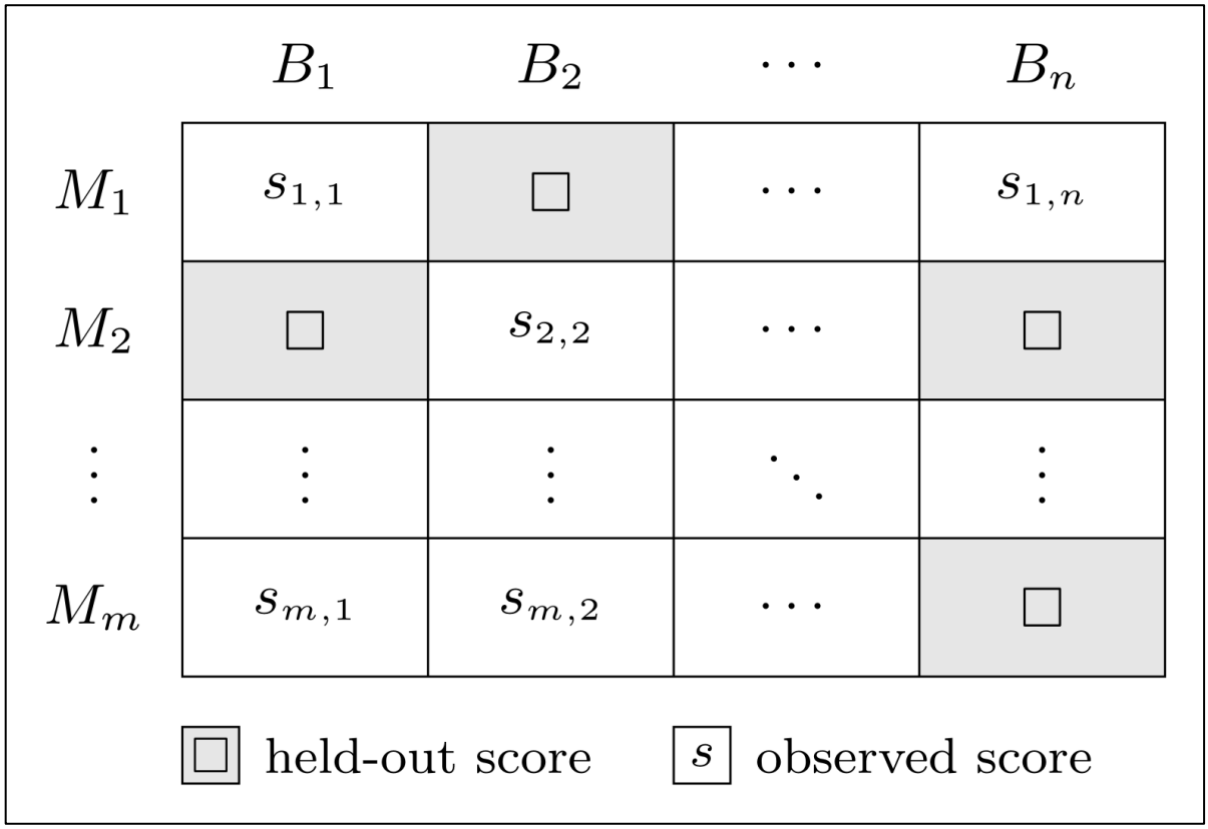

Figure 2: Cross-benchmark diversity, $M_1, \ldots, M_m$ are co-scientist AI models, $B_1, \ldots, B_n$ benchmarks.

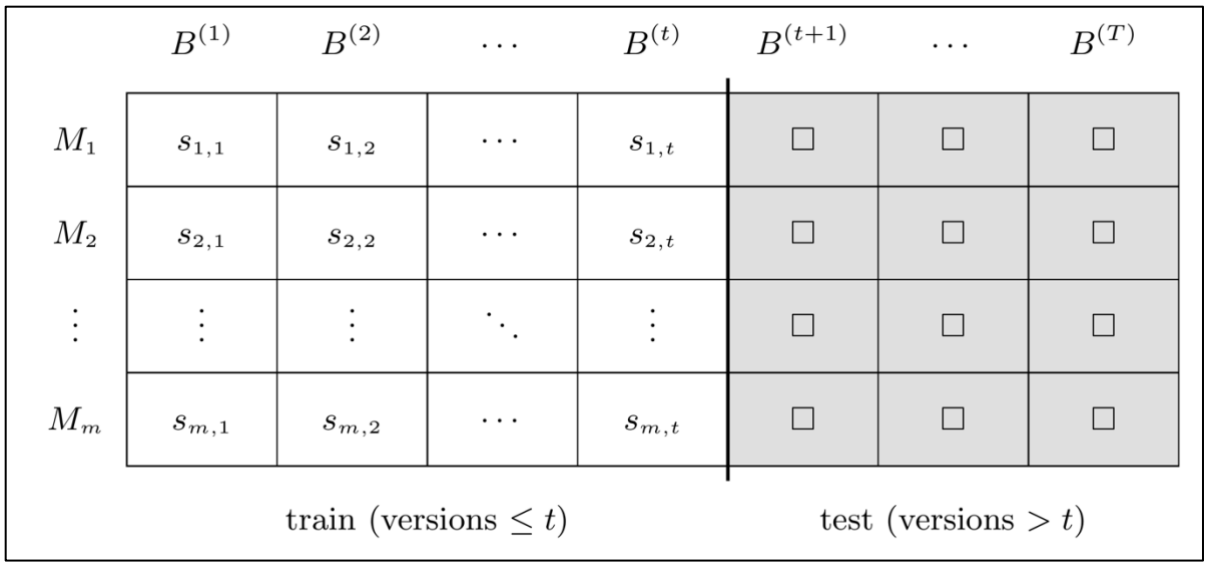

Figure 3: Benchmark diversity across updates, $M_1, \ldots, M_m$ are co-scientist AI models, $B^{(1)}, \ldots, B^{(T)}$ updates of a benchmark.

For simplicity, it is assumed that in the rolling case (Figure 3) the benchmark is picked up as a source of performative feedback from round 1 (or not picked up at all) and remains selected to rule out unpredictability caused by delayed or intermittent selection. In the rolling case, an unpredictable improvement of a model's score that does not translate to robust generalization to other problems from the domain suggests that the model developer merely correctly anticipated the evaluation round. On the other hand, an unpredictable improvement that generalizes well across the problem domain might represent an early sign of successful domain generalization and the benchmark saturation.

The prediction problem in Figure 3 is different from the one in Figure 2 because the former involves time series prediction on a single benchmark whereas the latter represents a matrix completion problem on many benchmarks. They afford different perspectives on the evaluation landscape. Measurements of diversity should remain diverse as well but homogeneous in how they treat predictability of model scores as an indicator of science monocultures.

If predictors trained on observed scores can infer held-out scores across benchmarks or their updates, the diversity of AI evaluations is low and the risk of a science monoculture high. Recent experiments showed that unpredictability of held-out model scores can identify a small number of benchmarks that are most informative about model capabilities because predictors can infer the rest of the scores across selected models and benchmarks with decent accuracy (Papailiopoulos 2026). Keeping the predictors up to date is important because improvements

of model capabilities can make benchmark prediction extrapolative and more prone to failure (Zhang et al. 2025).

Key for steering science away from a monoculture with co-scientist AI models is diversity of performative feedback from impactful scientific benchmarks. Peer review of candidate problems prioritizes validity because it is modelled on traditional peer review that did not have to explicitly address steering the course of science during rapid acceleration caused by AI. Strategic solicitation or selection of problems for scientific benchmarks can be guided by unpredictability of model scores from existing evaluations. Checking this unpredictability during peer review of candidate problems can help reviewers and organizers develop a new competence that jointly prioritizes validity and steering of co-scientists AI models with foresight.

3.1 Some Policy Options on Diverse Scientific AI Benchmarks

Steering science with co-scientist models and performative feedback from impactful scientific AI benchmarks needs to remain as democratic as possible. This means not only that peer review of candidate problems should be inclusive but also transparent and independent. Inclusiveness is key for diversity which can in turn drive science away from a monoculture by avoiding recreation of existing evaluations, measured, as we suggested, by unpredictability of model scores from existing evaluations. Transparency enables auditability of the steering process and independence should prevent steering that benefits only a handful of interests at the expense of the whole scientific community. There are proprietary AI benchmarks, such as FrontierMath (Glazer et al. 2024) including research-level mathematics problems with closed-form solutions (Tier 4), that lack in openness as well as independence. For FrontierMath, only example problems were released publicly, and the problem collection was supported by OpenAI, which has access to a portion of the private problems including those from Tier 4 others do not have access to[7]. While the intention is to evaluate if models can generalize to unseen problems by ruling out training data contamination, it is hard to determine whether OpenAI models have not been indirectly steered by the private problems the company has access to. OpenAI models has been consistently ahead of the competition on FrontierMath Tier 4[8], and it is impossible to determine if they generalize better or were steered appropriately thanks to the privileged access.

Steering models with evaluation problems improves their external validity if the problems have real-world utility and training on curricula inspired by them allows the models

[7] https://epoch.ai/frontiermath/tiers-1-4/about#:~:text=Conflict%20of%20interest%20statement

[8] https://epoch.ai/frontiermath/tiers-1-4, Tier 4 Leaderboard.

to generalize to adjacent problems. There is, therefore, no issue with a company supporting creation of a private benchmark they have privileged access to unless it significantly contributes to science steering via co-scientist AI models. This is why, even if driven by best intentions, it is unclear whether small groups of experts can achieve the necessary level of foresight ensuring that science is not drifting toward a monoculture thanks to the performative feedback created by AI evaluations accepted to the benchmarks. For this reason, we believe that broadly democratic principles of transparency, inclusiveness and independence are jointly needed as a supporting structure for anti-monoculture peer review of AI evaluations built not only on problem validity but also on diversity ensured by unpredictability of model scores on a benchmark from existing evaluations. Otherwise, it will be difficult to reach the desired level of foresight required for steering science with AI away from a monoculture.

Preventing alignment of co-scientist AI models with a handful of interests is also a governance issue whose solution will determine whether a monoculture can be avoided and science accelerated in a beneficial way. There are proposals for national initiatives that see AI benchmarks in a similarly strategic light (Krier and Wang 2025), as we consider them here, and similarly as we do here predicate science breakthroughs and acceleration on carefully produced AI evaluations. We are, however, unsure that in the current geopolitical competition in AI research and development (cf. Hendrycks et al. 2025) it will be beneficial to create a similar incentive structure for producing scientific AI evaluations when they are beginning to play the role of one of the progress drivers.

Rather, science societies and associations should recognize the opportunity and exercise their leadership by maintaining AI benchmarks and incentivizing members to participate in the process of their creation. Such AI evaluations could carry a significant weight and have a potential to become an impactful source of performative feedback. They can also become a source of empowerment and agency because if the authoritative AI evaluations were created by the community, they would to a degree lend it meaningful control over the tools, co-scientist AI models, that are set to steer science hopefully away from a monoculture. Model developers could hardly afford to ignore such channels of performative feedback because that would make their models obsolete in assisting with frontier research. Societies and associations have appropriate governance structures already in place. Acting on this opportunity would be then mostly a matter of setting the right priority by realizing the potential of performative feedback and peer reviewing candidate problems with foresight as suggested here. Above all, societies and associations operating on democratic principles are more likely to uphold transparency, inclusiveness and independence than state-sponsored or private projects that could become

dominated by goals deprioritizing universal advancement of science and dissemination of its results.

There have been considerable concerns over the state of AI evaluation (Raji et al. 2021) which led to the introduction of the validity framework designed to clarify claims about results of AI evaluations (Salaudeen et al. 2025). While crucial for better understanding real-world relevance and robustness of evaluation results, the individual types of validity cannot directly assist with maintaining benchmark diversity that will be key for avoiding a science monoculture if scientific benchmarks remain impactful sources of performative feedback. This gap suggests that the performative feedback loop is an underappreciated part of AI benchmarking that carries out not only evaluation but also high-level AI alignment in science. Unlike in AI safety where alignment specifications are discussed for some time now (Rauh et al. 2024), in this case the goal is still rather general and defined negatively as avoiding epistemic homogenization leading to monocultures (Marchal et al. 2026, pp. 9-10). Benchmark diversity as unpredictability of model scores from existing evaluations could serve as one of the alignment mechanisms that make the goal empirically clearer, ensuring that co-scientist AI models do not contribute to homogenization and monocultures.

4. Conclusion

We showed that even acceleration of science under wide adoption of co-scientist AI models requires treating impactful AI benchmarks as performative feedback that can shape model capabilities. To ensure beneficial steering and acceleration of science with AI, peer review of candidate problems for impactful scientific benchmarks should seek not only validity but also diversity, maintained as model scores on a benchmark that are hard to predict from existing evaluations. Science societies and associations built on democratic principles of transparency, inclusiveness, and independence should act on this opportunity by maintaining diverse benchmarks to help avoid science being steered toward a monoculture by improper performative feedback.

References

Abouzaid M, Blumberg AJ, Hairer M, Kileel J, Kolda TG, Nelson PD, Spielman D, Srivastava N, Ward R, Weinberger S, Williams L (2026a) First Proof. arXiv:2602.05192 [cs.AI]. https://doi.org/10.48550/arXiv.2602.05192.

Abouzaid M, Blumberg AJ, Hairer M, Kileel J, Kolda TG, Nelson PD, Spielman D, Srivastava N, Ward R, Weinberger S, Williams L (2026b) First Proof solutions and comments. https://1stproof.org/documents/FirstProofSolutionsComments.pdf.

Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, Bycroft C, Nicolaisen L, Arvaniti E, Pan J, Thomas R, Dutordoir V, Perino M, De S, Karollus A, Gayoso A, Sargeant T, Mottram A, Wong LH, Drotár P, Kosiorek A, Senior A, Tanburn R, Applebaum T, Basu S, Hassabis D, Kohli P (2026) Advancing regulatory variant effect prediction with AlphaGenome. Nature 649, pp. 1206-1218. https://doi.org/10.1038/s41586-025-10014-0.

Arvan M, Bright LK, Heesen R (2025) Jury Theorems for Peer Review. The British Journal for the Philosophy of Science 76(2), pp. 319-344. https://doi.org/10.1086/719117.

Bubeck S, Coester C, Eldan R, Gowers T, Lee YT, Lupsasca A, Sawhney M, Scherrer R, Sellke M, Spears BK, Unutmaz D, Weil K, Yin S, Zhivotovskiy N (2025) Early science acceleration experiments with GPT-5. arXiv:2511.16072 [cs.CL]. https://doi.org/10.48550/arXiv.2511.16072.

Feng T, Trinh T, Bingham G, Kang J, Zhang S, Kim S-h, Barreto K, Schildkraut C, Jung J, Seo J, Pagano C, Chervonyi Y, Hwang D, Hou K, Gukov S, Tsai C-C, Choi H, Jin Y, Li W-Y, Wu H-A, Shiu R-A, Shih Y-S, Le QV, Luong T (2026a) Semi-Autonomous Mathematics Discovery with Gemini: A Case Study on the Erdős Problems. arXiv:2601.22401 [cs.AI]. https://doi.org/10.48550/arXiv.2601.22401.

Feng T, Jung J, Kim S-h, Pagano C, Gukov S, Tsai C-C, Woodruff D, Javanmard A, Mokhtari A, Hwang D, Chervonyi Y, Lee JN, Bingham G, Trinh TH, Mirrokni V, Le QV, Luong T (2026b) Aletheia tackles FirstProof autonomously. arXiv:2602.21201 [cs.AI]. https://doi.org/10.48550/arXiv.2602.21201.

First Proof Editorial Board (2026) Second Batch Announcement. https://1stproof.org/documents/First_Proof_March_14_Announcement.pdf.

Glazer E, Erdil E, Besiroglu T, Chicharro D, Chen E, Gunning A, Olsson CF, Denain J-S, Ho A, de Oliveira Santos E, Järviniemi O, Barnett M, Sandler R, Vrzala M, Sevilla J, Ren Q, Pratt E, Levine L, Barkley G, Stewart N, Grechuk B, Grechuk T, Enugandla SV, Wildon M (2024) FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. arXiv:2411.04872 [cs.AI]. https://doi.org/10.48550/arXiv.2411.04872.

Goodin RE, Spiekermann K (2018) An Epistemic Theory of Democracy. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198823452.001.0001.

Götting J, Medeiros P, Sanders JG, Li N, Phan L, Elabd K, Justen L, Hendrycks D, Donoughe S (2025) Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark. arXiv:2504.16137 [cs.CY]. https://doi.org/10.48550/arXiv.2504.16137.

Gottweis J, Weng W-H, Daryin A, Tu T, Palepu A, Sirkovic P, Myaskovsky A, Weissenberger F, Rong K, Tanno R, Saab K, Popovici D, Blum J, Zhang F, Chou K, Hassidim A, Gokturk B, Vahdat A, Kohli P, Matias Y, Carroll A, Kulkarni K, Tomasev N, Guan Y, Dhillon V, Vaishnav ED, Lee B, Costa TRD, Penadés JR, Peltz G, Xu Y, Pawlosky A, Karthikesalingam A, Natarajan V (2025) Towards an AI co-scientist. arXiv:2502.18864 [cs.AI]. https://doi.org/10.48550/arXiv.2502.18864.

Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, Xu R, Zhang R, Ma S, Bi X et al. (2025) DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 645, pp. 633-638. https://doi.org/10.1038/s41586-025-09422-z.

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

Hardt M, Jagadeesan M, Mendler-Dünner C (2022) Performative Power. Advances in Neural Information Processing Systems 35 (NeurIPS 2022). https://proceedings.neurips.cc/paper_files/paper/2022/hash/90e73f3cf1a6c84c723a2e8b7fb2b2c1-Abstract-Conference.html.

Hardt M, Mazumdar E, Mendler-Dünner C, Zrnic T (2024) Algorithmic Collective Action in Machine Learning. Proceedings of the 40th International Conference on Machine Learning, PMLR 202, pp. 12570-12586. https://proceedings.mlr.press/v202/hardt23a.html.

Hardt M, Mendler-Dünner C (2025) Performative Prediction: Past and Future. Statistical Science 40(3), pp. 417-436. https://doi.org/10.1214/25-STS986.

Hendrycks D, Schmidt E, Wang A (2025) Superintelligence Strategy: Expert Version. arXiv:2503.05628 [cs.CY]. https://doi.org/10.48550/arXiv.2503.05628.

Jaech A, Kalai A, Lerer A, Richardson A, El-Kishky A, Low A, Helyar A, Madry A, Beutel A, Carney A et al. (2024) OpenAI o1 System Card. arXiv:2412.16720 [cs.AI]. https://doi.org/10.48550/arXiv.2412.16720.

Jagadeesan M, Zrnic T, Mendler-Dünner C (2022) Regret Minimization with Performative Feedback. Proceedings of the 39th International Conference on Machine Learning, PMLR 162, pp. 9760-9785. https://proceedings.mlr.press/v162/jagadeesan22a.html.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596, pp. 583-589. https://doi.org/10.1038/s41586-021-03819-2.

Krier S, Wang Z (2025) Benchmarking for Breakthroughs. https://ifp.org/benchmarking-for-breakthroughs/.

Lambert N, Morrison J, Pyatkin V, Huang S, Ivison H, Brahman F, Miranda LJV, Liu A, Dziri N, Lyu S, Gu Y, Malik S, Graf V, Hwang JD, Yang J, Le Bras R, Tafjord O, Wilhelm C, Soldaini L, Smith NA, Wang Y, Dasigi P, Hajishirzi H (2024) Tulu 3: Pushing Frontiers in Open Language Model Post-Training. arXiv:2411.15124 [cs.CL]. https://doi.org/10.48550/arXiv.2411.15124.

Li N, Pan A, Gopal A, Yue S, Berrios D, Gatti A, Li JD, Dombrowski A-K, Goel S, Mukobi G, Helm-Burger N, Lababidi R, Justen L, Liu AB, Chen M, Barrass I, Zhang O, Zhu X, Tamirisa R, Bharathi B, Herbert-Voss A, Breuer CB, Zou A, Mazeika M, Wang Z, Oswal P, Lin W, Hunt AA, Tienken-Harder J, Shih KY, Talley K, Guan J, Steneker I, Campbell D, Jokubaitis B, Basart S, Fitz S, Kumaraguru P, Karmakar KK, Tupakula U, Varadharajan V, Shoshitaishvili Y, Ba J, Esvelt KM, Wang A, Hendrycks D (2024) The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, pp. 28525-28550. https://proceedings.mlr.press/v235/li24bc.html.

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379(6637), pp. 1123-1130. https://doi.org/10.1126/science.ade2574.

Mazeika M, Phan L, Yin X, Zou A, Wang Z, Mu N, Sakhaee E, Li N, Basart S, Li B, Forsyth D, Hendrycks D (2024) HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, pp. 35181-35224. https://proceedings.mlr.press/v235/mazeika24a.html.

Marchal N, Chan S, Franklin M, Revel M, Keeling G, Fischli R, Chandra B, Gabriel I (2026) Architecting Trust in Artificial Epistemic Agents. arXiv:2603.02960 [cs.AI]. https://doi.org/10.48550/arXiv.2603.02960.

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

Novikov A, Vũ N, Eisenberger M, Dupont E, Huang P-S, Wagner AZ, Shirobokov S, Kozlovskii B, Ruiz FJR, Mehrabian A, Kumar MP, See A, Chaudhuri S, Holland G, Davies A, Nowozin S, Kohli P, Balog M (2025) AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv:2506.13131 [cs.AI]. https://doi.org/10.48550/arXiv.2506.13131.

OpenAI (2026) First Proof? https://openai.com/index/first-proof-submissions/.

Papailiopoulos D (2026) LLM Benchmark Matrix Completion. https://github.com/anadim/llm-benchmark-matrix.

Perdomo J, Zrnic T, Mendler-Dünner C, Hardt M (2020) Proceedings of the 37th International Conference on Machine Learning, PMLR 119, pp. 7599-7609. https://proceedings.mlr.press/v119/perdomo20a.html.

Phan L, Gatti A, Li N, Khoja A, Kim R, Ren R, Hausenloy J, Zhang O, Mazeika M, Hendrycks D et al. (2026) A benchmark of expert-level academic questions to assess AI capabilities. Nature 649, pp. 1139-1146. https://doi.org/10.1038/s41586-025-09962-4.

Piliouras G, Yu F-Y (2023) Multi-agent Performative Prediction: From Global Stability and Optimality to Chaos. Proceedings of the 24th ACM Conference on Economics and Computation, pp. 1047-1074. https://doi.org/10.1145/3580507.359775.

Raji ID, Denton E, Bender EM, Hanna A, Paullada A (2021) AI and the Everything in the Whole Wide World Benchmark. The 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). https://openreview.net/forum?id=j6NxpQbREA1.

Rauh M, Marchal N, Manzini A, Hendricks LA, Comanescu R, Akbulut C, Stepleton T, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel I, Rieser V, Isaac W, Weidinger L (2024) Gaps in the Safety Evaluation of Generative AI. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7(1), pp. 1200-1217. https://doi.org/10.1609/aies.v7i1.31717.

Romera-Paredes B, Barekatain M, Novikov A, Balog M, Kumar MP, Dupont E, Ruiz FJR, Ellenberg JS, Wang P, Fawzi O, Kohli P, Fawzi A (2024) Mathematical discoveries from program search with large language models. Nature 625, pp. 468-475. https://doi.org/10.1038/s41586-023-06924-6.

Salaudeen OE, Reuel A, Ahmed AM, Bedi S, Robertson Z, Sundar S, Domingue BW, Wang A, Koyejo S (2025) Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling. https://openreview.net/pdf?id=2Bw6uC49QF.

Spelda P, Stritecky V (2021) What Can Artificial Intelligence Do for Scientific Realism? Axiomathes 31, pp. 85-104. https://doi.org/10.1007/s10516-020-09480-0.

Spelda P, Stritecky V, Symons J (2024) No-Regret Learning Supports Voters' Competence. Social Epistemology 38(5), pp. 543-559. https://doi.org/10.1080/02691728.2023.2252763.

Spelda P, Stritecky V (2025) Benchmark-Driven Selection of AI: Evidence from DeepSeek-R1. arXiv:2508.10173 [cs.LG]. https://doi.org/10.48550/arXiv.2508.10173.

Traberg CS, Roozenbeek J, van der Linden S (2026) AI is turning research into a scientific monoculture. Communications Psychology 4, 37. https://doi.org/10.1038/s44271-026-00428-5.

Wan A, Klyman K, Kapoor S, Maslej N, Longpre S, Xiong B, Liang P, Bommasani R (2025) The 2025 Foundation Model Transparency Index. arXiv:2512.10169 [cs.AI]. https://doi.org/10.48550/arXiv.2512.10169.

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 CP, 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 CW, Bengio Y Zitnik M (2023) Scientific discovery in the age of artificial intelligence. Nature 620, pp. 47-60. https://doi.org/10.1038/s41586-023-06221-2.

Wang AHE, Siler-Evans K (2026) U.S.-China Competition for Artificial Intelligence Markets. Rand Research Report, https://www.rand.org/pubs/research_reports/RRA4355-1.html.

Woodruff DP, Cohen-Addad V, Jain L, Mao J, Zuo S, Bateni MH, Branzei S, Brenner MP, Chen L, Feng Y, Fortnow L, Fu G, Guan Z, Hadizadeh Z, Hajiaghayi MT, JafariRaviz M, Javanmard A, C. S. K, Kawarabayashi K-i, Kumar R, Lattanzi S, Lee E, Li Y, Panageas I, Paparas D, Przybocki B, Subercaseaux B, Svensson O, Taherijam S, Wu

X, Yogev E, Zadimoghaddam M, Zhou S, Matias Y, Manyika J, Mirrokni V (2026) Accelerating Scientific Research with Gemini: Case Studies and Common Techniques. arXiv:2602.03837 [cs.CL]. https://doi.org/10.48550/arXiv.2602.03837.

Zhang G, Dorner FE, Hardt M (2025) How Benchmark Prediction from Fewer Data Misses the Mark. The 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025). https://openreview.net/forum?id=o3bftqj17e.

Ethics Declarations

Competing Interests

The authors declare no competing interests.

Funding

This research was supported by the project 'Human-Centered AI for a Sustainable and Adaptable Society', registration number CZ.02.01.01/00/23_025/0008691, co-funded by the European Union.

Author Information

Contributions

P.S.: Conceptualization, Formal Analysis, Methodology, Visualization, Investigation, Writing – original draft, Writing – review & editing.

V.S.: Investigation, Writing – original draft, Writing – review & editing.