[Preprint]

Automating Pursuitworthiness: Four Concerns About 'AI Scientists' and the Proper Roles for Machine Learning Systems in Scientific Discovery

Donal Khosrowi

Centre for Ethics and Law in the Life Sciences

Leibniz University Hannover

donal.khosrowi[at]cells.uni-hannover.de

Abstract

Machine learning (ML) systems play increasingly important roles in scientific discovery. Recent efforts seek to build ML systems that predict upcoming discoveries and who is likely to make them, identify emerging research trends, and suggest novel concepts, ideas, questions, hypotheses and experiments to investigators. Notably, unlike other ML systems, these *predictive discovery and recommender systems* (PDRS) seek to augment and automate *agenda-setting* roles currently played by human researchers: determining or shaping the goals and trajectories of scientific discovery, rather than taking given goals and merely executing tasks in their pursuit. This paper argues, first, that PDRS raise novel conceptual and methodological disruptions, creating uncertainty around whether PDRS can and should play such roles. Second, the paper draws out four major questions, and associated concerns, about the roles PDRS are envisioned to play. These issues have not received attention in the literature thus far, leaving unclear what the proper roles of PDRS in science could be and how these roles should be carved out through appropriate designs and divisions of labor. To address these issues, the paper explores concerns about PDRS' potential impacts and limitations, and how PDRS fit with broader views of how science should function.

Keywords: scientific discovery; machine learning; artificial intelligence; performativity; recommender systems; values in science; pursuitworthiness.

1 Introduction

Machine learning (ML) systems play increasingly important roles in scientific discovery across a wide range of fields, including in astrophysics, materials science, and biology (Jumper et al. 2021; Sourati and Evans 2023; Iten et al. 2020; Cranmer et al. 2020; Udrescu et al. 2020; Wu and Tegmark 2019; Boiko et al. 2023; Melnikov et al. 2018), yielding major successes such as AlphaFold 2.0's contributions to predicting protein structures being recognized through the 2024 Nobel prize in chemistry. As ever more significant epistemic tasks are delegated to ML systems, and ML researchers endeavour to build fully automated systems that perform research tasks from start to finish (Gu and Krenn 2024a, 6; Boiko et al. 2023; Melnikov et al. 2018; Lu et al. 2024), significant *conceptual disruptions* (Löhr 2023; Hopster et al. 2023) arise: central concepts we use to understand and organize scientific pursuits come under pressure. A key concept affected is the

role-concept of 'researcher': do emerging ML systems already partly assume functions associated with this role (cf. Clark and Khosrowi 2022)? What abilities and skills are necessary for this role and what limitations do ML systems exhibit in regard to such abilities and skills (Bergamaschi Ganapini 2025)? Answers to these questions are crucial for determining good divisions of labor between human researchers and ML systems (cf. Stuart 2019; Stuart 2025; Barman et al. 2023), but there are no clear answers emerging.

However, this paper argues, making progress on such questions becomes increasingly urgent in light of recent efforts to extend the roles for ML systems beyond *execution-level* roles and towards *agenda-setting* roles in scientific discovery. Specifically, so far, ML systems in scientific discovery are largely used to search large epistemic spaces in contexts where the goals of research projects are determined by human investigators, e.g. to identify new protein structures, compounds that bind to specific targets, or new materials that exhibit certain properties (Jumper et al. 2021; Abramson et al. 2024; Juan et al. 2021; Sen et al. 2022; Notin et al. 2024). However, more recently, there are rapidly accelerating efforts by researchers and industry to use ML approaches to predict upcoming discoveries and who is likely to make them, identify emerging research trends, and suggest novel concepts, ideas, questions, hypotheses and experiments to investigators (e.g. Krenn et al. 2023, Wang et al. 2023; Sourati and Evans 2023; Gottweis et al. 2025). These efforts, often accompanied by ambitious visions of building "AI scientists" or "AI Co-scientists" (Lu et al. 2024; Yamada et al. 2025; Gottweis et al. 2025; Castelvecchi 2024; Luo et al. 2025), aim to augment and automate key aspects of the role of researchers, which include determining or shaping the goals of scientific inquiry, rather than just taking *given* goals and executing tasks in their pursuit.

In this paper, I argue that these emerging ML systems, which I call *predictive discovery and recommender systems* (PDRS), raise major questions, and concerns, centering around the core issue of whether PDRS can and should be used to automate agenda-setting tasks in science. To make progress on this issue, this paper provides a first systematic discussion of four sub-questions that have so far not received sufficient attention in the ML or philosophy of science literatures:

1) Can PDRS *successfully* inform discovery trajectories and how could we evaluate this?
2) By what standards, and underlying values, do PDRS make predictions and recommendations, and which standards are the right ones?
3) Do PDRS have sufficient capacity to facilitate genuinely *novel* discoveries?
4) How may PDRS account for the role of social, moral and political values in science, and what views about the aims of science are they compatible with?

In drawing out these questions, this paper makes several contributions at once. First, PDRS have so far not been identified as a cluster of ML-in-science approaches that, while exhibiting substantial variation in goals and techniques, share important commonalities regarding the epistemic, methodological and ethical concerns they raise. In carving out PDRS as an interesting target of inquiry for philosophers of science, the paper draws out the important, but underappreciated shift in moving from uses of ML systems for execution-level towards agenda-setting tasks. Second, the paper centrally frames the issues that PDRS raise as connected to the phenomenon of *performativity* (van Basshuysen 2023; 2025; Ortmann 2025; Perdomo et al. 2020; [reference blinded]): rather than merely *predicting* research trajectories, PDRS have the capacity to *steer* these trajectories, which raises a range of important epistemic, methodological and value-related concerns – in highlighting these issues, the paper and establishes missing

connections between the philosophy of ML-in-science to the growing literature on performativity. Third, by putting PDRS on the map, the paper highlights and addresses important gaps in the philosophy of science and ML literatures, which have so far not sufficiently clarified what the proper roles of PDRS in science could be and how they should be carved out through appropriate designs and divisions of labor (cf. Markowitz et al. 2024; Leslie 2023). Fourth, the paper applies the emerging framework of conceptual disruptions (Löhr 2023; Hopster et al. 2023) to frame the discussion around PDRS: the issues that PDRS raise can be usefully understood as speaking to larger conceptual and practical questions around the concept of 'researcher' and how emerging ML approaches seek to automate central tasks associated with this role, yet raise questions about their abilities to perform these roles *well*. While not seeking to offer definitive answers to the as-of-yet underexplored issues it raises, the paper prepares the grounds for careful, informed and systematic community-wide efforts to negotiate the emerging roles of PDRS in science, in light of their possible impacts, limitations, and our views of how science should function.

The discussion is organized as follows. *Section 2* provides an overview of ongoing efforts to build PDRS. *Section 3* outlines how PDRS aim at automating tasks associated with the role of 'researcher' and how this yields conceptual and methodological disruptions. *Section 4* draws out four questions that emerge from these disruptions, articulates a series of concerns related to them, and makes suggestions for how to address them. Finally, *Section 5* concludes by stressing the importance of engaging the larger, interdisciplinary project of negotiating the novel roles for ML in scientific discovery.

2 Agenda-Setting Roles for ML in Scientific Discovery

A much-discussed, Nobel prize-yielding success story of ML in scientific discovery concerns AlphaFold's achievements in predicting the three-dimensional structures of proteins with levels of accuracy on par with experimental determinations of protein structures (Jumper et al. 2021; Abramson et al. 2024) but at much greater speed, thus promising to significantly accelerate structure determination and downstream efforts that build on structural knowledge (e.g. drug discovery). Beyond AlphaFold, ML systems are increasingly used across the special sciences to search vast epistemic spaces, e.g. to discover new materials and predict the stability of new compounds, generate new molecules with specific, desired properties, discover planets and black holes, and so on (Juan et al. 2021; Sen et al. 2022; Notin et al. 2024). What connects these efforts is that ML systems are used for *execution-level* tasks: given a concrete epistemic objective (e.g., to find new protein structures), dedicated ML systems are built and incentivized to learn subtle, distributed patterns from existing data, and exploit these patterns for predictive purposes. Such ML approaches are often successful when 1) the boundaries of a discovery problem are well-understood, 2) relevant background knowledge about how to tackle the problem empirically, at least in principle, as well as large, high-quality datasets exist, 3) an epistemic space is too vast to be efficiently searched by humans (e.g. by means of experimental methods), and 4) human investigators can't draw on strong priors for where to search for what. Here, human investigators hence have a clear understanding of the *goals* at issue: they determine the objectives of an inquiry, i.e. what the research problems and questions are, the *class* of hypotheses to be investigated, and the *type* of outcome that is desired and why it is significant – but they may lack the ingredients (e.g. time, resources, access, or abilities) to execute a set of epistemic tasks that promote these *given* goals efficiently.

There is a second, rather different approach to using ML in scientific discovery settings that has been emerging over the last decade and is currently receiving significantly increasing interest from researchers and industry. This approach aims to build ML systems that can 1) find fruitful adjacencies and unexplored relationships among existing scientific knowledge and data (often based on large, science-scale knowledge graphs and semantic networks) (Krenn and Zeilinger 2020), 2) predict upcoming discoveries and collaborations (Krenn et al. 2023), 3) suggest novel ideas, concepts, hypotheses and experiments to investigate (Si et al. 2024; Gu and Krenn 2024a; 2024b; Gottweis et al. 2025), and 4) conduct discovery projects end-to-end, from formulating research questions, over planning experiments, conducting them with self-driving labs, to interpreting and writing up the results as research papers (see Gu and Krenn 2024a 5; Kramer et al. 2023; Kitano 2021; Lu et al. 2024; see Castelvecchi 2024; Strickland 2024 for criticisms).

Among this menu of related pursuits, I will focus here on systems that *predict* discoveries and *recommend* novel research pursuits. I will call these systems, accordingly, *predictive discovery and recommender systems* (PDRS). PDRS notably mark a move away from using ML systems for *execution-level* tasks and instead aim to support and (partly) automate *agenda-setting* roles thus far played by human researchers: assessing what are promising questions, hypotheses or experiments to pursue in order to gain new and significant knowledge about the world. In doing so, PDRS are supposed to mitigate challenges relating to the ever-increasing amount of scientific literature and increasing specialization that make it difficult for human researchers to recognize fruitful connections and adjacencies to be pursued (Hanson et al. 2024; Gu and Krenn 2024a).

Let me offer a brief overview of two families of PDRS to provide a better sense of what these systems are supposed to do and how they work.

2.1 Generating Novel Research Ideas & Assessing Them

A first family of PDRS is designed to generate novel, expert-level research ideas. For instance, Gu and Krenn (2024a) propose SCIMUSE, an LLM-based system that can "[…] suggest compelling research directions and collaborations, revealing opportunities that might not be readily apparent and positioning AI as a source of inspiration in scientific discovery" (2024a, 1). The suggestions made by SCIMUSE are evaluated by a large (110) cross-disciplinary panel of research group leaders from the Max Planck Society, who ranked 4,400 LLM-generated research ideas across various fields in the natural and social sciences according to how interesting they are. Gu and Krenn report that expert raters scored nearly 25% of all suggested ideas at 4 or 5 (out of 5), with 394 suggestions rated as 'very interesting' (5) (2024a, 4). Moreover, based on these data, Gu and Krenn proceed to train two ML systems to predict how interesting human expert raters would find specific research ideas. According to Gu and Krenn, their approach "[…] not only allows us to identify connections between properties of ideas and their interest-level, but also enables us to accurately *predict* the level of interest of new ideas […], which will be important when expensive human-expert data is unavailable." (2024a, 1, emphasis added). Gu and Krenn sketch an optimistic outlook on the utility of PDRS for research idea synthesis, indicating that "[…] large scientific organizations, national funding agencies, and other stakeholders may find value in adopting [PDRS] methodologies […] to foster new highly interdisciplinary and interesting collaborations and ideas that might otherwise remain untapped. This, hopefully, could advance the progress and impact of science at a large scale." (Gu and Krenn 2024a, 6)

More recently, Google Research has presented a related PDRS called "AI co-scientist" (Gottweis et al. 2025). Framed as a collaborator, rather than a tool to fully automate scientific discovery, their PDRS involves a multi-stage agential workflow building on Google's Gemini 2.0 LLM to furnish a pipeline that "is intended to help uncover new, original knowledge and to formulate demonstrably novel research hypotheses and proposals" (Gottweis et al. 2025, 1). The AI co-scientist generates research ideas in response to user prompts (e.g., specifications of research problems and constraints), using a multi-agent workflow, where generated ideas are subjected to internal review and competition by LLM "agents" in order to distil the most promising competitors to suggest to users. Beyond presenting such outputs, the AI co-scientist also offers a conversational layer to allow users to interact with the system in natural language to provide feedback, ask follow-up questions, or make suggestions. In developing and testing their PDRS, Gottweis et al. (2025) involved multiple research groups to trial their system. For instance, a group of researchers from Imperial College London's (ICL) Department of Life Sciences used the AI co-scientist in the context of research on antimicrobial resistance (Penadés et al. 2025), querying the system about a mechanistic question regarding gene transfer in bacterial evolution, which the group had previously investigated in as-of-then unpublished experimental work. They report that "remarkably, AI co-scientist's top-ranked hypothesis matched our experimentally confirmed mechanism […]. We critically assess its five highest ranked hypotheses, showing that some opened new research avenues in our laboratories." (Penadés et al. 2025, 1) These test-bed impressions support Google's framing of their PDRS as a collaborator to accelerate scientific discovery, e.g. by speeding up detailed literature reviews, identifying gaps and feasible, but yet unexplored, hypotheses and experiments. By the lights of the ICL researchers, the AI co-scientist shows that "AI can act not just as a tool but as a creative engine, accelerating discovery and reshaping how we generate and test scientific hypotheses" (ibid.).

2.2 Predicting Future Discoveries

A second family of PDRS aims to predict future discoveries, research trends, and fruitful collaborations. There is a large and growing literature on this sub-programme of literature- and network-based discovery, starting with early work in information science, bibliometrics, scientometrics, and science of science (e.g. Swanson 1986; Swanson and Smalheiser 1997), which draws on statistical and network-based techniques to predict, among others, how many citations specific papers, researchers or bodies of literature are likely to receive in the future, what topics will be popular in a discipline, or what future collaborations are likely between researchers (Wang et al. 2013; Bai et al. 2017, Xia et al. 2023). Recent efforts in this space have turned to ML methods for these tasks. One instance is Krenn et al. (2023), who propose a benchmark competition called Science4Cast, centered around a temporal graph-based semantic network that traces the evolution of concepts and ideas in artificial intelligence research over time. They propose and compare ten predictive systems with respect to whether they can successfully predict the evolution of this network, e.g. whether two nodes representing specific ideas or concepts, e.g. "generative adversarial networks" and "image synthesis", that have not been connected at time t will be connected at time $t + \delta$. In related, subsequent work, Gu and Krenn (2024b) present an ML system that "[…] can predict with high accuracy which concept pairs, that have never been jointly investigated before in any scientific paper, will be highly cited in the future" (2024b, 2). Reporting the predictive successes of different ML systems, Krenn et

al. note that "[…] [b]eing able to *predict* what scientists will work on is a first crucial step for *suggesting* new topics that might have a high impact." (2023, 1327, emphasis added)

A potential drawback to this approach is that the data that PDRS are trained on reflect a highly filtered distribution of research, i.e. research that was eventually conducted and published. This observed distribution naturally differs from the distribution of potential research ideas considered and explored by researchers before being filtered by, e.g., considerations regarding impact and strategy, funding mechanisms, as well as various biasing mechanisms, such as those involved in publication bias. When aiming to turn predictions into recommendations, as Krenn et al. suggest (2023, 1327), it seems plausible to think that recommendations should not *only* be based on predicting research that will be eventually observed with high probability. For instance, it may be important for human investigators to learn which research trajectories have the promise of yielding significant discoveries, but have a *low* rather than *high* probability of being pursued.

A related approach that takes such considerations into account is developed by Sourati and Evans (2023) in the context of materials science and biomedicine, where a key discovery aim is to find desirable compound-property relationships, e.g. whether a material is ferroelectric or whether a compound binds to a specific target. Building on previous work that used natural language processing techniques to identify latent knowledge of future discoveries in existing published literature (e.g., Tshitoyan et al. 2019), Sourati and Evans propose a PDRS that does not only take into account the contents of published research (content-only models), but also the distribution of researchers and the collaboration networks among them. This information can be scraped from publication metadata and is jointly represented in a so-called hypergraph that encodes 1) existing relationships among materials and properties, 2) among researchers to those materials and properties, and 3) collaborative relationships among researchers. Based on this, Sourati and Evans' PDRS seeks to identify discovery pathways that are *cognitively accessible* to researchers. For instance, a researcher A who has worked on material M is more likely to infer and explore a fruitful connection between material M and property P if they have previously collaborated with researcher B, who has experience with P. According to Sourati and Evans, this hybrid approach allows their PDRS not only to identify previously unrecognized adjacencies between existing ideas and concepts (e.g. whether a specific compound may exhibit a certain desirable property such as ferroelectricity) but also predict promising, but so far unrealized, collaborative pathways between researchers that are favourably situated to make certain discoveries. In addition, Sourati and Evans also stress a second important advantage of their approach: it enables the generation of 'alien' predictions and recommendations, i.e. suggesting "[…] complementary hypotheses, which are not only unlikely to be considered by unassisted human experts, but outperform published discoveries." (2023, 1683;) In sum, Sourati and Evans' approach seeks to further augment PDRS capacities beyond merely accelerating discovery trajectories that are likely to be pursued anyway by also helping researchers 'go against the grain' to pursue promising ideas that are unlikely to be explored without intervention (see also Artiles et al. 2026).

2.3 Are PDRS a Serious Target?

How seriously should we take PDRS, then? As with many emerging technologies, it is currently unclear whether and how PDRS will be adopted by scientists. Despite increasing investment into PDRS, including by major players such as Google Research, adoption remains a largely speculative issue for now (though see Luo et al. 2025). Yet, even if PDRS fail at generating

significant adoption anytime soon, there may be independent reasons to challenge the often uncritical technological imaginaries around automated science they figure in, and that motivate their development. There is no shortage of highly visible communications, publications and prizes that promote visions about automated science, such as Nobel prize lectures by Deepmind's Demis Hassabis, position papers (Griffin et al. 2024), and review articles (Wang et al. 2023), or competitions and prizes like the Nobel Turing Challenge, which aims at "developing a highly autonomous AI and robotics system that can make major scientific discoveries, some which may be worthy of the Nobel Prize and even beyond." (Kitano 2021, 1) Even those sceptical of PDRS' promises may hence think that it is useful to critically challenge such imaginaries *precisely because* automating agenda-setting tasks in scientific discovery holds little promise. On such a stance, it may seem especially urgent to point out that resources (e.g. funding) would be misdirected towards such endeavours and that sociotechnical imaginaries about artificial researchers and scientists misguide the attention of funding agencies or policymakers, and perpetuate simplistic views of science on the part of the public.

Across divides between optimists and pessimists, this paper seeks to advance debate about PDRS by drawing on philosophy of science resources to point out underappreciated problems that may arise already at the stage of conceiving PDRS and not only *if and when* PDRS are deployed broadly. Such a project mirrors other philosophical literature discussing emergent technologies that have yet to reach maturity and/or widespread adoption, such as fully autonomous vehicles. Such literatures, while arguably precautionary in flavor, help publicize important philosophical concerns around the responsible development and deployment of novel technologies. While PDRS may not raise attention-grabbing, trolley-caliber issues, the largely uncritical imaginaries they are embedded in present a clear and relevant target to be engaged by philosophy of science. In particular, as early critics of PDRS like Leslie (2023) point out, determining what role ML systems should play in science must involve deliberate reflection, negotiation, and active design- and use-choices by developers and scientists. At this current, critical juncture in the rapidly accelerating PDRS space, philosophy (of science) has plenty to contribute to enabling and shaping these negotiations, beginning with the observation that PDRS put entrenched concepts and divisions of labor under increasing pressure, as I now turn to highlight.

3 PDRS Generate Disruptions and Uncertainties

Efforts to build PDRS for agenda-setting roles in scientific discovery raise significant conceptual disruptions (cf. Löhr 2023; Hopster et al. 2023): central concepts we use to structure and organize scientific discovery enterprises come under pressure. When concepts are disrupted, both conceptual as well as methodological uncertainty arises regarding how we should apply familiar concepts, such as 'researcher', 'scientist', or 'discoverer', and distribute associated expectations or responsibilities (Michel 2020; Clark and Khosrowi 2022). For instance, among the growing literature discussing the roles of ML (and AI more broadly) in the sciences, there are both increasingly frequent suggestions that ML systems may or should take on the role of 'researchers' or 'co-scientists' (Kitano 2021; Markowitz et al. 2024; Griffin et al. 2024; Gottweis et al. 2025) as well as mounting objections to ML systems playing such roles on various grounds (Messeri and Crockett 2024; Leslie 2023; Chawla 2021). These disagreements suggest that there are indeed substantial conceptual and methodological uncertainties regarding how to understand and apply central concepts to emerging ML systems and whether ML systems should be built and understood to perform certain roles. In what follows, I focus specifically on disruptions that

affect the concept of 'researcher'[1], which is a role-concept traditionally applied to humans, to highlight how PDRS may unhelpfully encroach on this role when assuming central functions and tasks associated with it. In light of concerns about PDRS' abilities perform these functions and tasks competently, such encroachment creates practical and methodological uncertainties about how to divide labor between humans and machines in discovery settings, and broader conceptual uncertainties regarding how to frame the roles that PDRS may and should play, including in the wider technological imaginaries that surround their development.

There is, to my knowledge, no comprehensive account in the philosophy of science that systematically analyses what it means to be a researcher, and, specifically, what qualities, abilities, duties and tasks are distinctively associated with this role (though see Michel 2020; Neta 2025 for related projects). This is perhaps not surprising as there was not much need for such an account so far: for the most part, we think that we know a researcher when we see one[2]. But as increasing efforts aim to build PDRS that are supposed to augment and automate crucial parts of agenda-setting activities, there is a need to be clearer. Absent a mature account to lean on, we can make some progress by considering some plausible qualities, abilities, skills, duties and tasks. Roughly, researchers appear to have two types of jobs: *planning* and *executing* scientific research. On the planning side, researchers must exhibit qualities, abilities, and skills such as curiosity, imagination, and creativity (Stuart 2019; Currie 2019; Sánchez-Dorado 2023; Boden 2009); they must be able to form concrete epistemic and practical goals; draw on values to inform these goals (Douglas 2023; Sullivan 2025); and must be strategic in making plans to reach these goals. On the doing side, researchers must exert epistemic agency, autonomy, and leadership to pursue their goals effectively, e.g. by running experiments, collecting and analysing data, building models, and recruiting instruments; they must seek and acquire epistemic goods such as knowledge, understanding, or explanations; and draw on existing epistemic resources as well as skills, such as perceptiveness, imagination, or interpretive abilities, in pursuing these goods (Stuart 2025).

Across these clusters, execution-level ML systems such as AlphaFold have been built to assist with, and partly automate, a range of *doing*-tasks, specifically 1) executing epistemic scripts, such as by conducting physical or virtual experiments, 2) extracting correlational and causal information from data, 3) furnishing/selecting (often implicit, connectionist) models that represent and compress data, and 4) providing predictions (and sometimes explanations) of phenomena.

Notably, PDRS instead focus on automating *planning*-related tasks, such as identifying novel questions, hypotheses or experiments to pursue. But it is unclear whether they can and should be used for these purposes. Specifically, should we think that it is possible and promising to automate agenda-setting roles using PDRS if they only possess *some* of the qualities and abilities that we usually associate with researchers? If not, can labor in discovery be neatly separated along the distribution of such qualities, e.g. using ML systems to *highlight* patterns in large datasets and having human investigators *interpret* these patterns? Can PDRS and human investigators work together seamlessly, without sacrifice, and perhaps substantial advantage to

[1] I use 'researcher' as shorthand to refer to 'scientific researcher' and take this concept to be sufficiently overlapping with the concept of 'scientist', to which the arguments offered here extend.

[2] Though citizen science and activist research routinely raise questions about the distinction between 'scientists', 'researchers' and other epistemic agents (see Koskinen 2023), as do thorny demarcation issues between science, non-science, and pseudo-science more generally (see Hansson 2021; Michel 2020).

discovery outcomes? Or should we think that the role for PDRS must be carved out, and restricted, more clearly, in light of principled, conceptual, or practical concerns we may have about their functioning and their impacts on discovery enterprises? The subsequent discussion seeks to make progress on these larger questions.

4 Questions and Concerns about PDRS

On the heels of the disruptions created by PDRS, a whole range of under-investigated philosophical and methodological questions arises about how to carve out the proper roles for PDRS in discovery. Centrally, we should ask: is it possible and desirable to augment and automate agenda-setting roles in scientific discovery using PDRS? From this general, guiding question, four important sub-questions emerge that have received little attention so far:

1) Can PDRS *successfully* inform discovery trajectories and how could we evaluate this?
2) By what standards, and underlying values, do PDRS make predictions and recommendations, and which standards are the right ones?
3) Do PDRS have sufficient capacity to facilitate genuinely *novel* discoveries?
4) How may PDRS account for the role of social, moral and political values in science, and what views about the aims of science are they compatible with?

In what follows, I draw out a series of four interconnected concerns that are prompted by these questions, demonstrating that carving out and negotiating the proper roles for PDRS in science is an important and equally delicate project that has been undertheorized so far, and requires joint attention by philosophers of science, methodologists of the special sciences, and ML researchers on multiple fronts.

4.1 Performativity

The first major concern that PDRS raise is about *performativity*, the phenomenon by which putatively predictive systems do not just predict things in the world, but where these predictions *causally affect* the very things that they predict (van Basshuysen 2025). Performativity raises severe challenges for attempts to answer the first question: what it means for PDRS to *successfully* inform discovery trajectories and how this could be evaluated.

Currently, PDRS are widely framed as predictive tools, whose success should be evaluated by considering their predictive accuracy. However, as I argue in this section, this framing is misleading. If PDRS work as intended, they must have causal powers to *steer* the very discovery trajectories they purport to predict. Realizing this reveals fundamental ambiguities around what kinds of things PDRS are, what they are meant to do, how they should be evaluated, and whether their use is desirable. In particular, PDRS are designed to promote two potentially conflicting aims: 1) accelerating research that would be done anyway, and 2) augmenting discovery trajectories towards projects that would otherwise remain inaccessible. Pursuing these goals, however, also raises important risks, including steering discovery trajectories away from things that would be discovered without PDRS, e.g. towards discovery trajectories with higher speed and volume, but less potential for ground-breaking novelty (cf. Hao et al. 2026). Crucially, as I will argue throughout the following subsections 4.2-4.4, we currently lack both the evidence and the independent normative standards required to tell whether PDRS may steer discovery trajectories in *desirable* directions, which I take to be essential to telling whether PDRS are

successful. Let me begin by explaining what performativity is and how it challenges current approaches to building and validating PDRS.

4.1.1 What is performativity and why does it matter?

In recent years, social scientists, philosophers of science, and computer scientists have increasingly emphasised that many predictive tools, such as analytical or computational (scientific) models or decision-support systems, can causally influence the outcomes to be predicted – their predictions are *performative* (van Basshuysen et al. 2021; van Basshuysen 2023; 2025; Ortmann 2025; Khosrowi and van Basshuysen 2024; Perdomo et al. 2020; [reference blinded] forthcoming).

Broadly, performativity obtains when a model predicts an outcome X and the prediction has the capacity to causally influence X. The existing literature distinguishes two basic types of performativity: self-fulfilling and self-undermining. In the first case, a model might predict $X = x$ and this causes $X = x$, whereas X would have been $X = x'$ otherwise. For instance, an economic model might predict liquidity problems of a bank, and, in response, customers withdraw assets, causing the very problems that were predicted. Conversely, predictions might also be self-undermining when a model predicts $X = x$ and this causes the value of X to be $X = x'$, rather than $X = x$. For instance, an epidemiological model might predict high infection numbers and deaths in a population, but, in response, individuals may become more cautious, avoiding travel and contacts, thus decreasing infections and deaths relative to predicted values (van Basshuysen et al. 2021). In each of these cases, a model is not merely predicting a static target quantity, but there is a causal coupling between the models' predictions and that quantity.

What these literatures highlight is that it is easy to misconstrue the role of predictive tools as being *merely* predictive, failing to recognize the way in which they can causally shape outcomes ([reference blinded]). When recognizing that models and other epistemic instruments can be performative, novel epistemological and ethical questions about the responsibilities of modelers arise, and about the proper roles of putatively predictive tools in various real-world contexts (Khosrowi 2023; [reference blinded]).

4.1.2 Understanding Performativity in PDRS

PDRS raise acute concerns about performativity, which have not been anticipated and discussed in depth in the literature to date (though see Evans 2013 for an early version of this concern; Chawla 2021; [reference blinded]). Specifically, predicting discoveries can *causally affect* the research pursuits that human researchers will undertake. Two general ways in which researchers may respond to PDRS predictions and recommendations are: 1) to follow them and invest in projects that are predicted as likely or promising, or 2) strategically eschew them by focusing on other, including unrelated or orthogonal, projects, e.g. in the hope of minimizing competition and achieving discovery priority (see also Sikimić and Radovanović 2022). This range of responses suggests that it is largely uncertain how, exactly, PDRS may impact discovery trajectories.

Such performative effects are not problematic per se. After all, if PDRS are supposed to contribute anything towards scientific discovery, e.g. by highlighting promising ideas, boosting human researchers' imagination or creativity, or making the identification of fruitful questions,

hypotheses and experiments more efficient, their predictions and recommendations should have *some* causal bearing on eventual discovery trajectories, if only by accelerating discoveries that would have been made anyway. However, current attempts to build PDRS have failed to recognize and discuss these performative potentials, often casting PDRS as mere predictive systems, and their task as making *accurate* predictions (Sourati and Evans 2023; Krenn et al. 2023; Gu and Krenn 2024a; 2024b).

This framing is a mistake. PDRS are unlikely to be merely *predictive* tools: if they work as intended, they *must* be able to causally affect actual discovery pursuits. However, acknowledging this raises difficult questions around *where* PDRS should steer discovery trajectories, which the literature so far has ignored.

To better understand PDRS' potential performative impacts and why it matters which occur, consider a simple model of an epistemic landscape with a set of in-principle possible discoveries D, as shown in Fig. 1. D is subdivided into two partitions, the upper partition D^+, encoding discoveries worth pursuing and the lower partition D^-, those that are not worth pursuing.[3] Within this space, we envision two discovery envelopes: M (for 'machines') which contains discoveries that can be achieved with PDRS and H (for 'humans') that can be achieved by human investigators without PDRS.

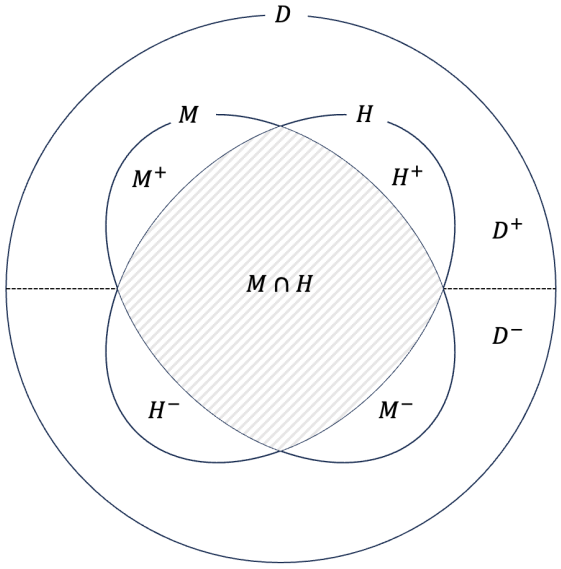

Figure 1. The space of possible discoveries D, and the two discovery envelopes M and H, with and without machines.

The envelopes M and H may overlap significantly, captured by their intersection $M \cap H$, indicating discoveries that are made regardless of whether PDRS are involved or not (although perhaps at different speed, or in different volume). But they can also come apart importantly. Specifically, M^+ and M^- indicate the sets of pursuitworthy and non-pursuitworthy discovery trajectories that may be taken with PDRS but couldn't (or wouldn't) be taken without them.

[3] I ask the reader to imagine, for the sake of argument, that pursuitworthiness is a tractable measure we can agree on, which of course is rarely true (see e.g. Shaw 2023 and further contributions in the same issue; Duerr and Fischer 2025)

Likewise, H^+ and H^- indicate the sets of pursuitworthy and non-pursuitworthy discovery trajectories that may be taken without PDRS but couldn't (or wouldn't) be taken with PDRS.

The epistemic landscape model helps draw out both benefits and risks in how PDRS may steer discovery trajectories. For instance, PDRS may promote pursuitworthy trajectories that humans would fail to consider at all, or fail to consider as sufficiently promising, without PDRS (M^+). But equally, there are risks of foreclosing discovery trajectories that, perhaps for now, are inaccessible to PDRS-based discovery but may be accessible to human investigators (H^+), e.g. because PDRS do not have the creative or imaginative capabilities to conceive of them (I expand more on this concern in Section 4.3). More systematically, we can hence distinguish at least five ways in which PDRS may steer discovery efforts:

1) Acceleration: PDRS may support discoveries in $M \cap H$ that would be pursued anyway, but more quickly.
2) Expansion: PDRS may promote the pursuit of promising trajectories in M^+ that would not be pursued without PDRS.
3) Misdirection: PDRS may steer discovery efforts towards M^-; efforts that are accessible with PDRS, but not worth pursuing.
4) Foreclosure: PDRS may hinder the pursuit of worthwhile trajectories in H^+ that humans would pursue absent PDRS.
5) Prevention: PDRS may (usefully) prevent the pursuit of efforts not worth pursuing in M^- that humans would pursue absent PDRS.

Crucially, current PDRS research remains ambiguous on which of these effects are intended, how pertinent risks are managed, and does not provide suitable frameworks and relevant evidence for systematically evaluating how PDRS may affect discovery trajectories across these spaces. But performing such evaluations is crucial, since different induced distributions of discovery trajectories may have radically different implications for how PDRS should be designed and whether their use is desirable, and these effects hinge on how human researchers respond to their predictions and recommendations, and on larger visions and narratives of what PDRS' proper roles in supporting/automating agenda-setting should be. However, questions about the effects of PDRS have not only received insufficient attention in existing literature and remain under-evidenced, but also face a fundamental evaluation challenge, which I now turn to.

4.1.3 An Evaluation Challenge for PDRS

Following broader tendencies in ML research, most PDRS approaches understand the project of identifying promising new discovery trajectories as a *predictive* challenge. Key to demonstrating the effectiveness of PDRS is to show that they can *accurately* predict important features of actual discovery trajectories undertaken by human researchers (e.g. Gu and Krenn 2024a; 2024b; Sourati and Evans 2023). Importantly, however, in existing proof-of-concept work that demonstrates such successes, PDRS predict 'off policy': they predict past distributions of discovery trajectories observed in existing literature, i.e. discoveries that could not be influenced by the predictions made. In such an off-policy regime, predictive success may indicate that PDRS must somehow have latched onto features of the mechanisms by which human researchers form new ideas and select questions and hypotheses to investigate, as well as features of the

world that are latent in existing literature, e.g. whether a new material might exhibit a useful property. In essence, PDRS that can boast such predictive successes hence seem promising tools to accelerate the pursuit of *some* discovery trajectories in $M \cap H$, i.e. discoveries made anyway.

However, what we can learn from predictive success fundamentally changes once PDRS are deployed and predictions/recommendations get take-up by researchers. Here, predictions are made 'on-policy', i.e. PDRS predict discovery trajectories that are chosen, partly, *in light of* their predictions and recommendations. In this on-policy regime, *accurate* prediction tells a much murkier story: predictive accuracy does not only measure how good PDRS are at latching onto relevant features of the world and the mechanisms by which human researchers form new ideas and select questions and hypotheses to investigate, but is rather, at the same time, a measure of the *performative power* (Hardt et al. 2022) they exert on the distribution of discovery trajectories, i.e. their power to *induce* trajectories that are pursued *in virtue* of being predicted or recommended (cf. Chawla 2021).

This creates a deep ambiguity: on-policy predictive success is compatible with both PDRS successfully *predicting* research that is genuinely pursuitworthy and with PDRS merely successfully *inducing* researchers to pursue *whatever* is predicted, regardless of its pursuitworthiness. A PDRS may hence achieve high 'predictive' accuracy, without offering genuine epistemic benefits, or even foreclosing valuable alternative discovery trajectories.

Success is, therefore, difficult to validate and remains highly ambiguous without careful study. But careful study is also crucially limited: in an on policy-regime, there is no easy access to relevant counterfactuals needed to assess the epistemic costs (e.g. foreclosure and misdirection risks) and benefits (e.g. speed, expansion) of PDRS relative to pursuits undertaken without them. Evaluating PDRS may hence, at the very least, require sophisticated experimental studies to shed light on these important counterfactual questions about the causal effects of PDRS on discovery pursuits.

4.1.4 Taking Stock

The analysis offered here yields three interconnected insights. First, while PDRS are often framed as mere predictive tools, there are good reasons to believe that they can be *performative*, not only predicting discovery trajectories, but *steering* them through their predictions and recommendations. In particular, PDRS involve dual, but potentially conflicting aims of accelerating research that gets done anyway and expanding the scope of discovery towards projects that would otherwise remain inaccessible. Against the background of these ambitions, however, we can envision a whole range of both beneficial and harmful effects, and it remains largely unclear, and understudied, which effects we should expect. The predictive framing is hence misleading and there are significant empirical questions about what effects PDRS may have, which must be answered to assess whether PDRS are successful and whether their use is desirable.

Second, evaluating PDRS success by looking at predictive accuracy faces a significant evaluation challenge. While in an off-policy-regime it may be possible to demonstrate that PDRS can predict discoveries that would be made anyway, that is not the same as showing that they can also successfully steer discovery towards *pursuitworthy* endeavours (I expand on this concern in Section 4.2). Moreover, in an on-policy regime, where predictions can shape actual discovery

trajectories, predictive success leaves unclear whether PDRS successfully predict research that would get done anyway (acceleration), contribute genuine epistemic benefits by augmenting discovery (expansion; prevention), or whether they merely successfully change the distribution of discovery trajectories to *whatever* is predicted, raising risks of misdirection and foreclosure.

Finally, third, the PDRS literature currently lacks normative frameworks to assess and weigh different forms of impact that PDRS may have. The epistemic landscape model sketched here conceptually distinguishes between pursuitworthy and non-pursuitworthy discovery trajectories; but such a distinction is of course extremely difficult to make in practice at a level required for large-scale evaluation. I will not attempt to draw such a distinction here, but rather highlight that 1) any case for or against PDRS must, at least implicitly, take *some* stances on pursuitworthiness, e.g. by committing to or rejecting the idea that research that would get done anyway ($M \cap H$) reliably tracks pursuitworthy endeavours. Since there may also be tradeoffs between different kinds of performative effects (e.g. acceleration/expansion vs. foreclosure; cf. Hao et al. 2026), any case for or against PDRS must also take some stance on the *relative* desirability of these effects. Finally, any *framework* for evaluating PDRS and grounding cases for or against their deployment, or ways of designing them, must be flexible enough to accommodate different stances on pursuitworthiness issues – otherwise, no meaningful evaluation and debate can take place.

Jointly, these concerns give rise to an evaluative vacuum where we lack both pertinent evidence and relevant normative frameworks to appraise PDRS, leaving us in a position without compelling reasons to either increase or decrease PDRS usage, maintain or change how they are designed and deployed, or be more optimistic or pessimistic about their potential to radically advance science, as their proponents suggest. In view of these significant difficulties, it is clear that the PDRS literature, researchers and methodologists in the special sciences targeted by PDRS efforts, and philosophers of science, must actively interrogate and negotiate PDRS development and use.

The epistemic landscape model offered here helps advance these efforts. It allows sceptics and critics of PDRS to point to misdirection and foreclosure risks to articulate their worries. Proponents, meanwhile, should offer supporting evidence that PDRS successfully expand discovery into M^+ and accelerate progress in $M \cap H$, while addressing sceptics' concerns about foreclosure of $H^+$and misdirection towards M^-. Without such evidence, and without clearer evaluative frameworks for what counts as *successful* steering rather than mere performative influence, the case for PDRS deployment remains unsubstantiated.

With the need for clearer evaluative frameworks for appraising PDRS' successes in view, let me proceed to examining three subsequent issues that connect to the broader concerns regarding performativity: by what standards and values PDRS make predictions and recommendations (Section 4.2); whether PDRS have sufficient capacity for novelty (Section 4.3), and how PDRS may account for the role of values in science and what general views regarding the aims of science are they compatible with (Section 4.4).

4.2 What, Exactly, do PDRS Predict?

Following on the heels of concerns about PDRS steering, rather than merely predicting, discovery trajectories, a second major concern is that it is often unclear what, exactly, PDRS

predict, and on grounds of which (implicit) values they make predictions/recommendations (cf. Chawla 2021). So far, I have described PDRS as providing two kinds of outputs: *predictions* of future discoveries and *recommendations*, e.g. of ideas, questions, hypotheses or experiments. Both are, usually, the result of training an ML system on existing observational data, usually published papers and publication metadata (e.g. about citations, authors, networks, collaborations etc.), as well as additional data, such as human experts' ratings and preferences regarding feasibility, surprise, etc. As outlined earlier, PDRS, such as those developed by Gu and Krenn (2024a) or Sourati and Evans (2023), are trained by incentivizing a learning system to predict certain relevant characteristics, e.g. the likelihood of a new connection between previously unconnected concepts A and B (i.e. that a new publication at time t will contain A and B), or whether a specific idea or connection will be rated as 'feasible' or 'surprising' by human raters.

What this literature has so far largely ignored is that the observed joint distributions of ideas, hypotheses, concepts, researchers, their collaborations, and so on, as well as how human evaluators rate these, e.g. in terms of novelty, are high-level realizations of a whole array of underlying mechanisms and norms that govern discovery processes. In such a complex joint distribution, discovery outcomes correlate with various features, some of which may be intrinsically desirable (e.g. the 'novelty' of a discovery) whereas others are not (that a discovery is more likely, only because there is a lot of funding available for it). Because of this, even though it may be clear what PDRS predict outright, it may often be unclear what underlying features are co-predicted and what underlying values are encoded and promoted by the choice of predictive targets, including non-epistemic values that may have significant epistemic consequences (cf. Elliott and McKaughan 2009).

To see these complications more clearly, let me distinguish between *observable predictive targets*, i.e. what we explicitly train PDRS to predict and what we use to evaluate their predictive skill, and *latent features*, i.e. the sorts of things that are difficult to measure (and hence predict) directly (e.g. the intrinsic pursuitworthiness of an idea). Observable predictive targets may include, for instance, whether two nodes, e.g., concepts, attributes, entities, researchers, are connected at time t; how frequently a paper is cited and by who; or how human raters evaluate an output, e.g., regarding such features as feasibility, plausibility or surprise. These observable predictive targets are hoped to correlate with other desirable features, i.e. latent features such as the pursuitworthiness of an idea, and the quality, novelty, impact, or significance of the research relating to it.

However, in many cases, observable predictive targets will correlate not just with *desirable* latent features but also with a range of (possibly undesirable) confounding factors. For instance, whether two ideas A and B unconnected at t will be connected at $t + \delta$ may depend on 1) the conceptual and theoretical continuity between A and B, e.g. whether B, such as a material property, can be theoretically conceived to be connected to A, given available theories at t, 2) the dynamics of funding mechanisms that govern whether discovery trajectories regarding A-B are deemed pursuitworthy by expert panels, 3) how visible research regarding A and B is to researchers, 4) how much agenda-setting power and funding researchers potentially interested in the A-B relationship have; 5) the social networks among researchers with expertise regarding A and B, and so on.

A PDRS that is trained to successfully predict the linkage of A-B must hence also, invariably, incorporate information about these features. For instance, *predicting* a connection between A and B when there are no social network links permitting exchange of information between

researchers with expertise on A and B respectively, will unlikely turn out accurate (even though recommending this pursuit might be fruitful). Similarly, when a community does not consider exploring the connection between A and B as interesting, regardless of its actual pursuitworthiness, then it is less likely that a positive prediction regarding the A-B connection will be borne out. PDRS that pursue predictive goals, such as Gu and Krenn (2024a) or Sourati and Evans (2023), are hence bound to take into account a host of potentially undesirable confounding factors that govern the evolution of actual discovery trajectories when making predictions. However, it is unclear whether it is *desirable* to take such factors into account. For instance, predicting a connection between A and B mainly on the grounds that researchers at well-connected institutions I and J are generally likely to explore adjacencies between their individual research foci does not imply that pursuing the A-B connection is pursuitworthy – it is just *likely* to be explored, but this likelihood does not necessarily confer information about its (relative) pursuitworthiness over projects that explore other connections, such as A-C or B-D, and are less likely to occur without intervention.

This is a major problem for PDRS programmes. Developers of PDRS frequently slip between describing PDRS as making predictions and as making recommendations, for instance when Krenn et al. 2023 articulate their vision of building "[…] a computer program that can automatically read, comprehend and act on AI literature. It can *predict* and *suggest* meaningful research ideas that transcend individual knowledge and cross-domain boundaries." (2023, 1326). Such slippage is problematic since it is prone to fall into a *naturalistic fallacy trap*: predicting that discovering a new relationship, say, between a material A and a property B, is *likely* and concluding that, *therefore*, discovering the A-B relationship is *good* and hence should be recommended. For PDRS not aimed at prediction but rather *only* aimed at making recommendations to steer discovery trajectories towards outcomes that wouldn't otherwise be achieved (e.g. Sourati and Evans 2023), a similar conclusion holds: just because A-B *wouldn't* be discovered otherwise, doesn't imply, all by itself, that discovery pursuits *should* be steered towards A-B. An independent standard for judging pursuitworthiness is needed, but the extant PDRS literature has not theorized what such a standard would look like or in what it could be grounded.

The upshot of the naturalistic fallacy trap is this: neither the natural evolution of discovery trajectories is uncontroversially good, in and of itself, nor is it ever clear that intervening on these trajectories is good. For a PDRS prediction/recommendation to be considered a *good* intervention on the evolution of discovery trajectories, it is plausible to think that 1) it must be clear by which (implicit) values such interventions proceed, including particularly which predictive targets are chosen and what their correlates are, 2) a community must agree that these values, and the targets that supposedly track them, are appropriate, and 3) the intervention must successfully promote the achievement of those values. Already the first of these requirements is not met by current PDRS, since it remains unclear what exactly they (co-)predict and by which values they make recommendations. Existing distributions of discovery efforts across the sciences are known to be the result of mechanisms that exhibit well-known and crucial dysfunctions: publication and funding bias, hype, power relations that exploit vulnerabilities of peer-review systems, etc. (Brown et al. 2017; Intemann 2022). Predicting and recommending future research endeavours by extrapolating features of this distribution hence clearly does not

ensure that PDRS recommend *pursuitworthy* research – it may merely amount to predicting and recommend pursuits that are *likely*, regardless of their pursuitworthiness.

What is more, even if specific predictive targets could be isolated more precisely from their potentially undesirable correlates, the choice between them is non-trivial and value-laden. For instance, aiming to predict results that are highly surprising may yield outcomes that are discontinuous with existing theory and conceptual schemas and hence less likely to attract attention, citations or funding. Conversely, optimizing predictions for the latter metrics may fail to steer discovery towards surprising and strongly novel results. Finally, there are various qualities that are difficult to predict when training PDRS on observed publication data, as is common. For instance, for lack of being represented in such data, *subliminal* research, i.e. pursuits that may in principle be fruitful but whose results don't get published and hence are not included in PDRS' training data, is more difficult to take into account in making predictions (see e.g. Krenn et al. 2023, 1327), hence potentially further aggravating the effects of the 'file-drawer problem' (Rosenthal 1979) and other forms of publication and citation bias (Song et al. 2013). Similarly, ideas published in remote places, approaches articulated in a conceptual language that is discontinuous with existing theoretical orthodoxies, and so on, may lie beyond the predictive horizon of PDRS.

In sum, 1) PDRS often leave unclear what exactly is predicted and on grounds of what values recommendations are made; 2) the choice between predictive targets is delicate; 3) it is unclear what downstream consequences different choices of predictive targets may have on the distribution of discovery trajectories undertaken. Taken together, these concerns aggravate the main issues raised in Section 4.1, inducing further ambiguities around whether PDRS-based predictions and recommendations are, overall, beneficial or rather harmful interventions on the distribution of discovery trajectories. These challenges have so far not received systematic attention among PDRS developers and philosophers of science but require scrutiny before PDRS can be responsibly deployed with a clearer understanding of what they do.

4.3 Novelty, Creativity, Conservatism and Homogenization

A third class of concerns about PDRS regards *novelty, creativity*, *epistemic conservatism*, and *homogenization*. In a nutshell, the concern here is that PDRS may be substantially restricted in regard to how novel the connections, questions, hypotheses, and suggestions they predict and recommend can be: they might perform well in making *conservative* predictions, e.g. combining familiar ideas in novel ways, but may be less able to anticipate more creative and transformative moves, e.g. predicting or recommending altogether new ways of conceptualizing or approaching phenomena (cf. Weisberg and Muldoon 2009; Ratti 2020). Relying on PDRS may hence carry the risk of conservatism, and possibly inferior epistemic outcomes (e.g. oversteering towards $M \cap H$ and M^-, foreclosing $H^+$and failing to push significantly towards M^+). This relates directly to ongoing concerns about the (re)homogenization of discovery (cf. Messeri and Crockett 2024; Anderson et al. 2024; Griffin et al. 2025; Hao et al. 2026). Many of the special sciences have only recently made sincere efforts to increase the epistemic diversity among knowledge-producers, partly in light of impactful feminist critiques of the past decades (e.g. Longino 1990; cf. Oreskes 2019). Conservatism in PDRS hence raises the risk that they may partly wash out this diversity, starting from concerns about the comprehensiveness of the data they are trained on, the assumptions and goals that go into building them, and extending to

worries about the PDRS landscape being dominated by few systems that have sufficient reach to streamline the efforts of large knowledge-seeking communities in ways that unhelpfully homogenize the distribution of discovery pursuits (e.g. owed to incentives not to depart from PDRS' predictions and recommendations).

Novelty-related concerns about ML in science are not new: the existing literature has articulated related worries regarding execution-level ML systems in scientific discovery and paints a mixed picture of whether ML systems used for execution-level tasks have the capacity to achieve what Ratti calls *strong novelty* (Ratti 2020; see also Champion 2025; Boge 2022; see Boden 2009 in regard to creativity), i.e. to facilitate predictions and explanations of phenomena that are significantly novel, e.g. to correctly predict the three-dimensional structure of never-before-synthesised proteins, or to correctly identify novel physics equations or state variables that best describe a new phenomenon or system without any knowledge of the physics that govern that system.

There is a wealth of proof-of-concept work demonstrating purported successes of ML systems in achieving strong novelty. For instance, in materials discovery, Szymanski et al. (2023) claim to demonstrate how their automated ML-driven discovery pipeline A-Lab has "[…] realized 41 novel compounds from a set of 58 targets." (2023, 86). Similarly, Chen et al. (2022) propose an ML-system that can purportedly identify fundamental state-variables of physical systems from observational data (see Eva et al. 2023 for a related project; see Langley et al. 1987 for a historical precursor). The goal behind this and other, related projects is to identify new physics variables and relationships *from scratch*, i.e. without drawing on any prior knowledge about what fundamental variables and relationships best describe the behavior of a system.

Yet, several such purported demonstrations of strong novelty have been met with scepticism. For instance, Leeman et al. (2024) point out doubts about whether Szymanski et al.'s (2023) materials discovery ML pipeline A-Lab has truly discovered *any* new materials. Similarly, Hillar and Sommer (2012) put pressure on purported demonstrations by Schmidt and Lipson (2009) of symbolic regression algorithms learning Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation from experimental data without any prior physics knowledge (Schmidt and Lipson 2009, 81). As Hillar and Sommer (2012) argue, there are reasons to believe that these successes are driven by strong inductive biases and physics knowledge leaking into the algorithms (cf. Battaglia et al. 2018). More broadly, some authors caution that we should remain sceptical about ML systems' potential for strong novelty. For instance, Ratti (2020) argues that ML systems are not able to produce strongly novel knowledge in molecular biology and genomics but are constrained to *weakly* novel discoveries.

The PDRS literature has mostly ignored to what extent these wider concerns may carry over to PDRS, instead focusing on street-level demonstrations, such as surveying domain experts to rate LLM-generated ideas for novelty (Si et al. 2024) or demonstrating the potential for novelty through highly curated validation studies (Gottweis et al. 2025). Yet, following extensive concerns about execution-level ML systems, we may be equally sceptical whether PDRS are, or will be, able to predict and recommend discovery trajectories that yield strongly novel discoveries: in the language of the epistemic landscape sketched earlier, it remains unclear whether the envelope of PDRS' predictions and recommendations M overlaps sufficiently with, while also reaching significantly beyond, H, i.e. towards M^+.

4.4 Values

The final concern that existing PDRS programmes raise is that they largely ignore the role that social, moral and political values play in making choices about scientific pursuits. The philosophy of science as well as science and technology studies (STS) literatures have long recognized and elaborated that science is not a value-free fact-finding endeavour simplistically aiming at truth or knowledge, full stop (Longino 1990). Rather, prominent views insist that science is laden with social, moral, and political values (often summarized as *non-epistemic* values), and that, in many cases, this is unavoidable and/or even desirable (Douglas 2009; Elliott 2017; 2022; Sullivan 2025). While disagreement persists around the appropriate roles for non-epistemic values in science, there is widespread agreement that so-called external stages of scientific research, in particular, the selection of problems, questions and hypotheses to investigate and pursue, are necessarily and appropriately informed by non-epistemic values without thereby compromising desirable attributes of the subsequent research, such as objectivity (e.g. Longino 1990; Anderson 1995; Koskinen 2023; Sullivan 2025). Values are needed to tell which real-world problems are important, e.g. diseases, climate change, or cybersecurity risks, and in carving out what role science should play in addressing them, e.g. studying what drugs may help cure a disease, acquiring knowledge about how the earth's climate is likely to change, or how to make IT systems robust to attacks. While the selection of problems to study and ways to study them must importantly be shaped by members of society and their representatives (e.g. political decision-makers), scientists, too, must draw on value-judgments in deciding what issues and questions to focus on and how.

Debates around the proper role of values in science also center around the larger question what science, as a general epistemic enterprise, should aim at. Historically, views have been divided between those who insist that science should aim at producing truth or knowledge simpliciter, and, more recently, those who stress that science should aim at *significant* truth: it should function in the service of society, engaging and addressing the epistemic and practical needs of humans by pursuing epistemic goods that cater to these needs (Kitcher 2011).

If we subscribe to this latter picture, at least one important role played by human researchers may be difficult to delegate to PDRS: determining by what values scientific inquiry should proceed and what goals it should cater to. Two types of concerns arise here. First, contingently, we might worry that ML systems might simply not be very good at inferring values and dealing with value-related issues: much research doesn't wear its values on its sleeves, and general concerns about PDRS capacity for novelty and creativity (Section 4.3) may make us sceptical about their ability to predict major discontinuities in what sorts of non-epistemic values scientific research prioritizes. A second, more principled concern arises from a view according to which discoveries are *made* significant by virtue of the fact that humans played the right role in deciding that this is the case. On such a view, it follows trivially that PDRS cannot play value-setting roles: even if ML systems may better than humans at determining efficient discovery trajectories that promote finding significant truth, determining what is significant in the first place, is, trivially, up to humans. I do not wish take a stance on this view here, but rather want to highlight that both the contingent and the principled concern suggest that determining a suitable division of agenda-setting labor between humans and PDRS is important, but it is unclear how such labor can and should be effectively divided.

One simple proposal could insist that humans determine the values and goals of an inquiry, and PDRS then help steer discovery efforts *conditionally* on these values and goals. But this

conditioning view faces problems. First, it is not clear whether conditioning is always possible as, often, values and goals are not fully settled in advanced, but rather discovered and negotiated as inquiry proceeds (cf. Brown 2020). Consider the many ethical issues, e.g. regarding bias, fairness, or explainability, that are raised by ongoing ML research only as such research progresses, and how these issues *iteratively* shape research trajectories, e.g. towards efforts to mitigate algorithmic bias, study different fairness criteria, or build systems that are inherently explainable. It can hence not be assumed that a sequential division of labor is generally or typically possible. Moreover, second, as argued earlier, it is often not clear what additional values are encoded in PDRS' predictions and recommendations, so it is difficult for human investigators to tell whether these values align with those deemed relevant. Third, it is unclear how to design institutions that could effectively enforce a conditioning-based workflow: given the promises of PDRS to rapidly accelerate discovery trajectories, there may be strong incentives for human investigators to forego thorough reflection of the value-related aspects of PDRS' predictions and recommendations and 'run with' their predictions and recommendations in the accelerated pursuit of, e.g., publications or funding. Finally, fourth, existing PDRS do not *explicitly* encode the role of values and do not allow human investigators to prompt these systems for *conditional* predictions or recommendations.[4]

In sum, whether we deem certain roles as suitable for automation through PDRS will depend on our answers to the larger question what science, as an enterprise, should aim at. If machines are better at finding efficient trajectories to produce new knowledge, and the aim of science is the pursuit of truth or knowledge *simpliciter*, then machines perhaps should, as much as possible, play agenda-setting roles. But if the aim of science, as per Kitcher (2011), should be to produce truth or knowledge that is significant as judged by human interests and values, and these values are not fixed but rather discovered as science progresses, then humans must remain heavily involved in agenda-setting roles and PDRS should not encroach on this role, either because we worry they might not be good at it, or because they shouldn't do so on principle. This, of course, does not imply that ML systems more broadly cannot be helpful in values discovery and negotiation; indeed, some think there is significant scope for augmenting value elicitation and reflection (Awad et al. 2022; Giubilini et al. 2024). The point, however, is that at least some views insist that the *determination* of goals and values central to an inquiry should heavily and authentically (and perhaps exclusively) involve human agents, which may suggest new conceptual emphases in our understanding of what it means to be a researcher.

4.5 What's new and What's Next?

Jointly, the concerns elaborated in this section form a set of key obstacles in the way of building and deploying PDRS to play the kinds of agenda-setting roles that larger imaginaries around automated science envision. So how should one respond to them?

One option is to brush the problems off and insist that several of the concerns articulated here apply to exclusively human-led discovery, too, suggesting there is nothing particularly novel or acute about them. However, this response misses that automation and standardization streamline

[4] One exception to this may be Google's "AI co-scientist" (Gottweis et al. 2025), which allows users to specifically prompt the system with detailed descriptions of concrete research problems and questions. Doing so may fix values relevant to determining the scope of an inquiry. However, since the AI co-scientists also involves extensive internal iteration through its multi-agent, asynchronous design, this may open yet further opportunities for difficult-to-detect forms of value-encroachment that remain inscrutable to investigators.

and scale undesirable attributes in a way that significantly amplifies concerns about them. This is not a new insight itself: across a range of domains where algorithmic systems are used to automate tasks, e.g. hiring, loan approval, or recidivism risk prediction, critics have argued that automation not only amplifies familiar concerns about, e.g., bias, fairness, and discrimination, which apply to humans, too, but also raises novel issues, e.g., explainability and transparency, homogenization, automation bias, the atrophying of important skills, or responsibility gaps, only few of which have so far been investigated in connection to PDRS (see e.g. Luo et al. 2025). What is more, unlike for human decision-makers or investigators, it often seems possible, at least in principle, to mitigate these issues. This suggests a more active stance on PDRS, following calls by critics like Leslie (2023), to explore ways in which PDRS development as well as the broader negotiations of their proper roles in discovery may be facilitated. The discussion offered here yields five recommendations to promote this project.

1) In light of concerns about performativity, PDRS should not be framed as mere predictive tools. Rather, their performative potentials should be recognized already at the stage of forming and communicating imaginaries and narratives around PDRS' role in advancing science and scientific discovery, as well as in designing them. If it is indeed inevitable that PDRS *steer* rather than just predict discovery, then we should be explicit about this feature, and manage it accordingly (cf. Khosrowi 2023). This goes hand in hand with making clear distinctions between the aims of prediction and recommendation, clarifying on what grounds, and in what ways, predictions, e.g. of the likely impact or surprise of ideas, may, or should, inform recommendations of research pursuits related to these ideas.

2) The predictive targets of PDRS should be made explicit and it should be studied how families of concepts, ideas, or questions hang together with observable predictive targets (e.g. publication and citation outcomes) and latent features (e.g. disparities in access to funding) and how correlations of predictive targets with undesirable latent features could be disentangled and broken.

3) Building on the first two points, systematic efforts, such as through simulation studies, should be undertaken to study potential performative effects (e.g., by simulating a network of agents that responds to PDRS' predictions and recommendations). Likewise, systematic sensitivity studies are needed to explore how the distribution of predictions and recommendations made by PDRS changes in response to changes to the predictive targets they track. Both types of studies may help epistemic communities better understand the directions in which PDRS may steer discovery pursuits.

4) Developers and proponents of PDRS should articulate more clearly what roles, exactly, PDRS are envisioned to play, and in virtue of what features PDRS can and should play such roles. Such more explicit framings may help promote critical debate among experts around whether PDRS are indeed suitable for these roles, and regarding what divisions of labor and what forms of interaction between human researchers and PDRS are desirable.

5) Relevant stakeholders, including developers and prospective users, should work towards creating new benchmarks and experiments to study 1) relevant features of human-machine

interaction with PDRS and 2) the effects of PDRS on research trajectories regarding salient features such as speed, diversity (of ideas, approaches, etc.), originality, surprise, and others.

6) To facilitate negotiation of PDRS' proper roles in discovery settings, it is necessary to furnish evaluative frameworks. These may, for instance, draw on the blueprint provided by the epistemic landscape model sketched earlier, and use both simulation and empirical evidence to shed light on the distributions of discovery trajectories that specific kinds of PDRS may induce. Importantly, as outlined earlier, such frameworks must be expressive enough to accommodate different views on the relative desirability of different effects that PDRS may have, and facilitate the inclusion of different stances on how to distinguish, at a general level, between pursuitworthy and non-pursuitworthy endeavours. Both features are required in order for evaluative frameworks to help proponents and sceptics of PDRS locate sources of disagreement and characterize the evidential needs involved in appraising PDRS.

Together, such advances may help put PDRS research programs on track to deal with the significant ambiguities concerns raised in this paper, and facilitate reflection, debate, and negotiation among stakeholders around the proper roles for PDRS in scientific discovery.

5 Conclusions: The Proper Roles for ML in Science

Predictive discovery and recommender systems (PDRS) are machine learning systems built to predict scientific discoveries and recommend ideas, questions, hypotheses, and experiments for human investigators to pursue. The growing PDRS literature casts these systems as crucial instruments in accelerating scientific discovery and scientific progress more broadly, enabling otherwise inaccessible discoveries by helping researchers and other stakeholders (such as funding agencies) identify promising connections and adjacencies and pursue fruitful research agendas more efficiently. It also paints more ambitious visions of "[…] the entire scientific process becoming fully automated – from the generation of an interesting idea […] to its automated execution and implementation" (Gu and Krenn 2024a, 6), suggesting significant reallocations of labor from humans towards PDRS. Notably, this includes assigning PDRS roles that do not merely support human researchers but more fully automate central agenda-setting tasks, such as identifying novel, unrealized connections between existing knowledge items, formulating new research hypotheses and ideas and gauging what capacity for novelty and surprise they harbor, predicting the impact and success of research agendas, or planning which collaborations are likely to be fruitful.

In view of these visions, PDRS may have the capacity to significantly intervene on the trajectories of humanity's largest epistemic project. Yet, despite much uncritical enthusiasm and interesting proof-of-concept demonstrations, ongoing efforts to build PDRS have so far failed to anticipate and engage a range of important questions and concerns about what roles PDRS can and should play in science. This paper has drawn out four such questions, including 1) whether PDRS merely predict or causally influence discovery and whether their use is desirable, given this ambiguity; 2) by what values PDRS make predictions and recommendations, 3) whether PDRS have sufficient capacity for novelty, and 4) whether PDRS can account for the ineliminable role that values play in scientific inquiry.

By highlighting a series of concerns related to these questions, I have argued that it currently remains largely unclear how the use of PDRS may affect the sciences, at the broadest level, and

in what ways their impacts may be beneficial or harmful. In light of reasonable doubts about PDRS' abilities, and open questions about their impacts, it is crucial for PDRS developers, researchers and methodologists in the special sciences, and philosophers of science to more fully reflect two central questions: 1) can and should PDRS assume roles distinctive of researchers, and 2) what are good divisions of labor between humans and machines in discovery settings, given our views on what science is supposed to do and how it should function? The questions, concerns and arguments provided in this paper, I hope, will stimulate a wider debate among philosophers of science, ML researchers involved in building PDRS, as well as domain experts and stakeholders across the special sciences targeted by PDRS efforts. In line with the arguments offered here, the goals and values that PDRS research pursues are far from settled, and negotiating where PDRS development and use should go next is an important task that we shouldn't ignore, and cannot delegate.

Declarations

This research was funded by the European Union (ERC, MAPS, 101115973). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Abramson, J., Adler, J., Dunger, J. et al. (2024). "Accurate structure prediction of biomolecular interactions with AlphaFold 3." *Nature* 630, 493–500. https://doi.org/10.1038/s41586-024-07487-w

Anderson, E. (1995). "Knowledge, Human Interests, and Objectivity in Feminist Epistemology." *Philosophical Topics*, 23(2): 27-58.

Anderson, B. R., J. H. Shah, and M. Kreminski. (2024). "Homogenization Effects of Large Language Models on Human Creative Ideation". In *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*. Association for Computing Machinery, New York, NY, USA, 413–425. https://doi.org/10.1145/3635636.3656204

Artiles, A.H., M. Weiss, L. Brinkmann, A. Goyal, and N. Rahaman. (2026). "Alien Science: Sampling Coherent but Cognitively Unavailable Research Directions from Idea Atoms". arXiv preprint. https://doi.org/10.48550/arXiv.2603.01092

Awad, E; Levine, S; Anderson, M; Anderson, SL; Conitzer, V; Crockett, MJ; Everett, JAC; Evgeniou, T; Gopnik, A; Jamison, JC; Kim, TW; Liao, SM; Meyer, MN; Mikhail, J; Opoku-Agyemang, K; Borg, JS; Schroeder, J; Sinnott-Armstrong, W; Slavkovik, M; Tenenbaum, JB. (2022). "Computational Ethics". *Trends in Cognitive Sciences*, 26(5), 388–405. https://doi.org/10.1016/j.tics.2022.02.009

Bai, X., H. Liu, F. Zhang, Z. Ning, X. Kong, I. Lee, and F. Xia. (2017). "An Overview on Evaluating and Predicting Scholarly Article Impact". *Information* 8(3):73. https://doi.org/10.3390/info8030073

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, Á., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, Ç., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., & Pascanu, R. (2018). "Relational inductive biases, deep learning, and graph networks". *arXiv*. https://arxiv.org/abs/1806.01261

Barman, K.G., Caron, S., Claassen, T. et al. (2024). "Towards a Benchmark for Scientific Understanding in Humans and Machines". *Minds & Machines* 34(6). https://doi.org/10.1007/s11023-024-09657-1

Bergamaschi Ganapini, M. (2025). "Can AI make scientific discoveries?". *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02299-8

Boden, M. 2009. "Computer Models of Creativity". *AI Magazine* 30(3): 23. https://doi.org/10.1609/aimag.v30i3.2254

Boge, F. J. 2022. "Two Dimensions of Opacity and the Deep Learning Predicament." *Minds & Machines* 32: 43– 75. https://doi.org/10.1007/s11023-021-09569-4

Boiko, D.A., MacKnight, R., Gomes, G. (2023). "Emergent Autonomous scientific research capabilities of large language models." arXiv preprint arXiv:2304.05332. https://arxiv.org/abs/2304.05332

Brown, Andrew W., Tapan S. Mehta, and David B. Allison. (2017). "Publication Bias in Science: What Is It, Why Is It Problematic, and How Can It Be Addressed?', in Kathleen Hall Jamieson, Dan M. Kahan, and Dietram A. Scheufele (eds.), *The Oxford Handbook of the Science of Science Communication*, Oxford Library of Psychology, https://doi.org/10.1093/oxfordhb/9780190497620.013.10

Brown, M. (2020). *Science and moral imagination*. Pittsburgh, University of Pittsburgh Press.

Castelvecchi, D. (2024). "Researchers built an 'AI Scientist' — what can it do?" *Nature* 633: 266. https://doi.org/10.1038/d41586-024-02842-3

Champion, H. (2025). "Strong novelty regained: high-impact outcomes of machine learning for science". *Synthese*, 206, 134. https://doi.org/10.1007/s11229-025-05228-8

Chawla, D. S. (2021). "Frosty reception for algorithm that predicts research papers' impact". *Nature*. Doi: 10.1038/d41586-021-01358-4.

Chen, B., Huang, K., Raghupathi, S., Chandratreya, I., Du, Q., and Lipson, H. (2022). "Automated discovery of fundamental variables hidden in experimental data." *Nature Computational Science*, 2(7), 433-442. https://doi.org/10.1038/s43588-022-00281-6

Clark, E., and D. Khosrowi. (2022). "Decentring the discoverer: how AI helps us rethink scientific discovery". *Synthese* 200: 463. https://doi.org/10.1007/s11229-022-03902-9

Cranmer, M., A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, S. Ho. (2020). "Discovering Symbolic Models from Deep Learning with Inductive Biases". arXiv preprint: https://doi.org/10.48550/arXiv.2006.11287

Currie, A. (2019). "Existential risk, creativity & well-adapted science". *Studies in History and Philosophy of Science Part A*, 76:39-48. https://doi.org/10.1016/j.shpsa.2018.09.008.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh, University of Pittsburgh Press.

Douglas, H. (2023). "The importance of values for science". *Interdisciplinary Science Reviews*, 48(2): 251–263. https://doi.org/10.1080/03080188.2023.2191559

Duerr, P., Fischer, E. (2025). "Rationally warranted promise: the virtue-economic account of pursuit-worthiness". *Synthese* 206, 68. https://doi.org/10.1007/s11229-025-05077-5

Elliott, K.C. (2022). *Values in Science*. Elements in the Philosophy of Science. Cambridge, Cambridge University Press. https://doi.org/10.1017/9781009052597

Elliott, K. (2017). *A Tapestry of Values: An Introduction to Values in Science*. New York, NY: Oxford University Press.

Elliott, K. C., & McKaughan, D. J. (2009). How Values in Scientific Discovery and Pursuit Alter Theory Appraisal. *Philosophy of Science*, *76*(5), 598–611. https://doi.org/10.1086/605807

Evans, J. A. (2013). "Future Science" *Science* 342, 44-45. DOI :10.1126/science.1245218

Eva, B., Ried, K., Müller, T. et al. (2023). "How a Minimal Learning Agent can Infer the Existence of Unobserved Variables in a Complex Environment." *Minds & Machines* 33: 185–219. https://doi.org/10.1007/s11023-022-09619-5

Giubilini, A., Porsdam Mann, S., Voinea, C. *et al.* (2024). "Know Thyself, Improve Thyself: Personalized LLMs for Self-Knowledge and Moral Enhancement". *Science and Engineering Ethics* 30, 54. https://doi.org/10.1007/s11948-024-00518-9

Gottweis, Juraj, Wei-Hung Weng, Alexander Daryin, et al. (2025) "Towards an AI co-scientist". Preprint. https://doi.org/10.48550/arXiv.2502.18864

Griffin, C., D. Wallace, J. Mateos-Garcia, H. Schieve, and P. Kohli. (2024). *A new golden age of scientific discovery*. Technical report, Google Deepmind. Available at: https://deepmind.google/public-policy/ai-for-science/ (accessed 25 April 2025)

Gu, X., and M. Krenn. (2024a). "Interesting Scientific Idea Generation Using Knowledge Graphs and LLMs: Evaluations with 100 Research Group Leaders". Cambridge University Press. https://doi.org/10.48550/arXiv.2405.17044

Gu, Xuemei, and Mario Krenn. (2024b). "Forecasting high-impact research topics via machine learning on evolving knowledge graphs". arXiv:2402.08640.

Hansson, S. O. (2021). "Science and Pseudo-Science", *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2021/entries/pseudo-science.

Hanson M. A., P. Gómez Barreiro, P. Crosetto, and D. Brockington (2024). "The strain on scientific publishing". https://doi.org/10.48550/arXiv.2309.15884

Hao, Q., Xu, F., Li, Y. et al. (2026) "Artificial intelligence tools expand scientists' impact but contract science's focus". *Nature* 649, 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, 22969-22981.

Hillar, C., and F. Sommer. (2012). "Comment on the article 'Distilling free-form natural laws from experimental data'" arXiv preprint arXiv:1210.7273.

Hopster, J. et al. 2023. "Conceptual Disruption and the Ethics of Technology". In I. van de Poel et al. (eds.) *Ethics of Socially Disruptive Technologies: An Introduction*, pp. 141-162. Open Book Publishers.

Intemann, K. (2022). "Understanding the Problem of "Hype": Exaggeration, Values, and Trust in Science". *Canadian Journal of Philosophy*, *52*(3), 279–294. doi:10.1017/can.2020.45

Iten, R., T. Metger, H. Wilming, L. del Rio, and R. Renner. (2020). "Discovering physical concepts with neural networks". *Physical Review Letters*, 124: 010508. https://doi.org/10.1103/PhysRevLett.124.010508

Juan, Yongfei, Yongbing Dai, Yang Yang, and Jiao Zhang. (2021). "Accelerating materials discovery using machine learning". *Journal of Materials Science & Technology* 79: 178-190. https://doi.org/10.1016/j.jmst.2020.12.010.

Jumper, J. et al. (2021). "Highly accurate protein structure prediction with AlphaFold". *Nature*, 596: 583-589. https://doi.org/10.1038/s41586-021-03819-2

Khosrowi, D. (2023). "Managing Performative Models", *Philosophy of the Social Sciences*, 53(5): 371-395. https://doi.org/10.1177/00483931231172455

Khosrowi, D., and P. van Basshuysen (2024). "Making a Murderer: How Algorithmic Risk Assessment Tools may Produce Rather Than Predict Criminal Behavior". *American Philosophical Quarterly*, 61 (4): 309–325. https://doi.org/10.5406/21521123.61.4.02

Kitano, H. (2021). "Nobel Turing Challenge: creating the engine for scientific discovery". *Npj Systems Biology and Applications* 7, 29. https://doi.org/10.1038/s41540-021-00189-3

Kitcher, P. (2011). *Science in a Democratic Society*. Amherst, NY: Prometheus Press.

Koskinen, I. (2023) "Participation and Objectivity', *Philosophy of Science*, 90(2): 413–432. Doi:10.1017/psa.2022.77.

Kramer, S., M. Cerrato, S. Džeroski, and R. King. (2023). "Automated Scientific Discovery: From Equation Discovery to Autonomous Discovery Systems". arXiv:2305.02251

Krenn, M., L. Buffoni, B. Coutinho, et al. (2023). "Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network". *Nature Machine Intelligence*, 5: 1326-1335. https://doi.org/10.1038/s42256-023-00735-0

Krenn, M., and A. Zeilinger. (2020). "Predicting research trends with semantic and neural networks with an application in quantum physics." *Proceedings of the National Academy of Sciences*, 117(4): 1910-1916.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Leeman J, Liu Y, Stiles J, Lee S, Bhatt P, Schoop L, et al. (2024). "Challenges in high-throughput inorganic material prediction and autonomous synthesis. ChemRxiv, doi:10.26434/chemrxiv-2024-5p9j4

Leslie, D. (2023). "Does the sun rise for ChatGPT? Scientific discovery in the age of generative AI". *AI and Ethics.* https://doi.org/10.1007/s43681-023-00315-3

Löhr, G. (2023). "Conceptual disruption and 21st century technologies: A framework". *Technology in Society*, 74: 102327. https://doi.org/10.1016/j.techsoc.2023.102327

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, Princeton University Press. https://doi.org/10.1515/9780691209753

Lu, C., C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. (2024). "The ai scientist: Towards fully automated open-ended scientific discovery". arXiv preprint arXiv:2408.06292, 2024

Luo, Ziming, Atoosa Kasirzadeh, and Nihar B. Shah. (2025). "The More You Automate, the Less You See: Hidden Pitfalls of AI Scientist Systems". arXiv preprint. https://doi.org/10.48550/arXiv.2509.08713

Markowitz, D. M., Boyd, R.L., and Blackburn K. (2024). "From silicon to solutions: AI's impending impact on research and discovery". *Frontiers in Social Psychology*. 2:1392128. Doi: 10.3389/frsps.2024.1392128

Melnikov, A. A., Nautrup, H. P., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A., Briegel H.J. (2018). "Active learning machine learns to create new quantum experiments", *Proceedings of the National Academy of Sciences* 115:6, 1221.

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

Michel, J. G. (2020). "Could Machines Replace Human Scientists? Digitalization and Scientific Discoveries." In Benedikt Paul Göcke and Astrid Rosenthal-von der Pütten (eds.), *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences*, pp. 361–376. Leiden, NL: Brill | mentis.

Neta, R. (2025). "Inquiry, research, and articulate free agency". *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02337-5

Notin, P., N. Rollins, Y. Gal, C. Sander, and D. Marks. (2024). "Machine learning for functional protein design". *Nature Biotechnology* 42: 216–228. https://doi.org/10.1038/s41587-024-02127-0

Ofer, D., H. Kaufman, and M. Linial. (2024). "What's next? Forecasting scientific research trends." *Heliyon* 10(1). https://doi.org/10.1016/j.heliyon.2023.e23781.

Oreskes, N. (2019). *Why Trust Science?* Princeton, Princeton University Press.

Ortmann, J. (2025). "Performative paternalism". *European Journal for Philosophy of Science.* 15, 25. https://doi.org/10.1007/s13194-025-00651-7

Penadés, J. R., Gottweis, J., He, L., Patkowski, J. B., Daryin, A., Weng, W.-H., … Costa, T. R. D. (2025). "AI mirrors experimental science to uncover a mechanism of gene transfer crucial to bacterial evolution". *Cell*, *188*(23), 6654–6665.e2. https://doi.org/10.1016/j.cell.2025.08.018

Perdomo, Juan C., Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. (2020). "Performative Prediction". In *International Conference on Machine Learning* (ICML) 119: 7599-7609. PMLR.

Ratti, E. (2020). "What kind of novelties can machine learning possibly generate? The case of genomics", *Studies in History and Philosophy of Science Part A*, 83: 86-96. https://doi.org/10.1016/j.shpsa.2020.04.001.

Rosenthal, R. (1979). "The file drawer problem and tolerance for null results". *Psychological Bulletin*, 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Sánchez-Dorado, J. (2023). "Creativity, pursuit and epistemic tradition". *Studies in History and Philosophy of Science*, 100:81-89. https://doi.org/10.1016/j.shpsa.2023.05.003.

Schmidt, M, and H. Lipson. (2009). "Distilling free-form natural laws from experimental data". *Science*, 324: 81–85.

Sen, S., S. Agarwal, P. Chakraborty, and K. Pratap Singh. (2022). "Astronomical big data processing using machine learning: A comprehensive review". *Experimental Astronomy* 53:1–43. https://doi.org/10.1007/s10686-021-09827-4

Shaw, J. 2022. "On the very idea of pursuitworthiness*". Studies in History and Philosophy of Science* 91:103-112. https://doi.org/10.1016/j.shpsa.2021.11.016.

Si, C., D. Yang, and T. Hashimoto. (2024). "Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers." https://doi.org/10.48550/arXiv.2409.04109

Sikimić, V., and S. Radovanović. (2022). "Machine Learning in Scientific Grant Review: Algorithmically Predicting Project Efficiency in High Energy Physics". *European Journal for Philosophy of Science*, 12: 50. https://doi.org/10.1007/s13194-022-00478-6

Song, F., Hooper, L., & Loke, Y. K. (2013). "Publication bias: what is it? How do we measure it? How do we avoid it?" *Open Access Journal of Clinical Trials*, *5*, 71–81. https://doi.org/10.2147/OAJCT.S34419

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

Strickland, E. (2024). "Will the "AI Scientist" Bring Anything to Science?" IEEE Spectrum, retrieved 1 October: https://spectrum.ieee.org/ai-for-science-2

Stuart, M.T. (2019). "The Role of Imagination in Social Scientific Discovery: Why Machine Discoverers Will Need Imagination Algorithms." In M. Addis et al., eds., *Scientific Discovery in the Social Sciences* (pp. 49-66). Springer. DOI: 10.1007/978-3-030-23769-1_4.

Stuart, M.T. (2025). “A New Account of Pragmatic Understanding, Applied to the Case of AI-Assisted Science”. *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02336-6

Sullivan, E. (2025). “Value encroachment on scientific understanding and discovery”. *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02383-z

Szymanski, N.J., Rendy, B., Fei, Y. et al. (2023). “An autonomous laboratory for the accelerated synthesis of novel materials”. *Nature* 624, 86–91. https://doi.org/10.1038/s41586-023-06734-w

Swanson, D.R. (1986). “Undiscovered Public Knowledge”, *The Library Quarterly: Information, Community, Policy*, Vol. 56, No. 2 (Apr., 1986), pp. 103-118. https://www.jstor.org/stable/4307965

Swanson, D.R., and N.R. Smalheiser. (1997). “An interactive system for finding complementary literatures: a stimulus to scientific discovery”, *Artificial Intelligence*, 91 (2): 183-203. https://doi.org/10.1016/S0004-3702(97)00008-8

Tshitoyan, V., Dagdelen, J., Weston, L. et al. (2019). “Unsupervised word embeddings capture latent knowledge from materials science literature.” *Nature* 571, 95–98. https://doi.org/10.1038/s41586-019-1335-8

Udrescu, S. M., A. Tan, J. Feng, et al. 2020. "AI Feynman 2.0: Pareto-Optimal Symbolic Regression Exploiting Graph Modularity." *Advances in Neural Information Processing Systems* 33 (NeurIPS 2020): 4860–4871.

van Basshuysen, P., L. White, D. Khosrowi, and M. Frisch (2021). “Three Ways in Which Pandemic Models May Perform a Pandemic”, *Erasmus Journal of Philosophy and Economics*, 14(1). https://doi.org/10.23941/ejpe.v14i1.582

van Basshuysen, P. (2023). "Austinian model evaluation". *Philosophy of Science* 90 (5), 1459-1468.

van Basshuysen, P. (2025). “Performativity in Science: Past and Future”. *Philosophy Compass*: e70062. https://doi.org/10.1111/phc3.70062

Wang, D., C. Song, and A Barabási. (2013). "Quantifying Long-Term Scientific Impact", *Science* 342,127-132. DOI:10.1126/science.1237825

Wang, H., T. Fu, Y. Du, et al. (2023). “Scientific discovery in the age of artificial intelligence”. *Nature*, 620: 47-60. https://doi.org/10.1038/s41586-023-06221-2

Weisberg, M. and R. Muldoon. (2009). "Epistemic Landscapes and the Division of Cognitive Labor". *Philosophy of Science*:76(2):225-252.

Wu, T., and M. Tegmark. (2019). “Toward an artificial intelligence physicist for unsupervised learning”. *Physical Review E*, 100(3): 033311. https://doi.org/10.1103/PhysRevE.100.033311

Xia, W., T. Li, and C. Li. (2023). "A review of scientific impact prediction: tasks, features and methods." *Scientometrics* 128(1): 543-585.

Yamada, Y., R. T. Lange, C. Lu, S. Hu, C. Lu, J. Foerster, J. Clune, and D. Ha. 2025. "The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search". arXiv Preprint. https://doi.org/10.48550/arXiv.2504.08066