Preprint

Automating Pursuitworthiness: Four Concerns About The Proper Roles for Machine Learning Systems in Scientific Discovery

Donal Khosrowi

Institute of Philosophy, Leibniz University Hannover

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, questions, hypotheses and experiments to investigators. These *predictive discovery and recommender systems* (PDRS) hence aim to augment and automate key activities that are central to the roles currently played by human researchers. 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, their likely limitations, and how PDRS fit into our broader views of how science should function.

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

1 Introduction

Machine learning (ML) systems play increasingly important roles in scientific discovery across a wide range of fields, including astrophysics, materials science, bioinformatics, and others (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). 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 concept? What abilities are necessary for this role and what limitations do ML systems exhibit in regard to such abilities? Answers to these questions are crucial for determining good divisions of labor between human researchers and ML systems in view of their respective abilities (cf. Stuart 2019; 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 desirable 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 increasing efforts 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). These efforts aim to augment and (partly) automate key activities relating to science's self-governance that are central to the role of 'researcher', which includes determining or shaping the goals of scientific inquiry, rather than just taking *given* goals and performing tasks in their pursuit.

In this paper, I argue that these emerging ML systems, which I call *predictive discovery and recommender systems* (PDRS), raise four major questions, and concerns, about whether machines can and should perform such roles.

1) Do PDRS merely *predict* scientific discoveries or do they also *causally influence* discovery trajectories? If the latter, is it desirable to use PDRS, given it is unclear in which directions they may steer discovery trajectories?
2) What, exactly, is predicted when PDRS predict new discoveries and suggest new ideas, concepts, questions, hypotheses or experiments; and what underlying values (e.g. impact, surprise, feasibility) are encoded in their predictions and recommendations?
3) Do PDRS have sufficient capacity to predict strongly novel discoveries or are they limited to more conservative predictions/recommendations?
4) How may PDRS account for the role of social, moral and political values in shaping research agendas, and what views about the aims of science are they compatible with?

This paper provides a first systematic discussion of these important questions. In doing so, it makes several relevant contributions at once. First, PDRS have so far not been identified as cluster of ML approaches that, while exhibiting substantial variation in goals pursued and techniques applied, also 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, by putting PDRS and the questions and concerns they raise on the map, the paper highlights and addresses important gaps in the philosophy of science and machine learning literatures, which have neither discussed nor addressed the issues this paper highlights, leaving unclear what the proper roles of PDRS in science could be and how they should be carved out through appropriate designs and divisions of labor (see e.g. Markowitz et al. 2024; Leslie 2023). Third, 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 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 question clusters that emerge from these disruptions and articulates a series of concerns related to them. Finally, *Section 5* concludes with suggestions for how to engage these concerns and stresses the importance 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, or access) 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 increased interest. 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, according to what they are intended to do, *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 published 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 provide 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 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-scientists 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 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, Google involved multiple research groups to test their system. For instance, a group of researchers from ICL's Department of Life Sciences used the AI co-scientist in the context of their research on antimicrobial resistance (Penadés et al. 2025), querying the system about a mechanistic question regarding gene transfer in bacterial evolution that the group had previously investigated in so-far 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 are built to predict future discoveries, research trends, and fruitful collaborations. There is a large and growing literature on this sub-programme, starting with early work in information science, bibliometrics, scientometrics and science of science, 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 research 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 such as "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 out 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. Comparing their PDRS against earlier models (Tshitoyan 2019), Sourati and Evans report significant gains in predictive accuracy (2023, 1693). However, beyond improving on benchmarks, they 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 the pursuit of discovery trajectories that are likely to be pursued anyway by also helping researchers 'go against the grain' to pursue fruitful adjacencies that are promising but unlikely to be explored without intervention.

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. While, in light of familiar 'publish-or-perish' pressures, it is not implausible to think that scientists may experience incentives to partly automate the generation of new research ideas with the help of (more advanced, future) PDRS, adoption remains a largely speculative issue for the moment. So, keeping these uncertainties in mind, are there other reasons to take PDRS seriously?

Even if PDRS fail at generating adoption, there may still 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 think that it is useful to critically challenge such imaginaries *precisely because* they think that 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 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 put to use. 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. 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-researchers' (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

[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.

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 for a related project). This is perhaps not surprising as there was not much need for such an account so far: for the most part, we might think that we know a researcher when we see one[2]. But as increasing efforts seek 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, duties and tasks. Roughly, *researchers* appear to have two types of jobs: *planning* and *executing* research. On the planning side, researchers must exhibit curiosity, imagination and creativity (Stuart 2019; Boden 2009); they must be able to form concrete epistemic and practical goals, draw on values to inform these goals (Douglas 2023); 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 or interpretive abilities, in pursuing these goods.

Across these concept clusters, execution-level ML systems have sought 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) collecting data, 3) extracting correlational and causal information from data, 4) furnishing (often implicit, connectionist) models that represent and compress data, and 5) providing predictions or explanations of phenomena.

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 delegate agenda-setting roles to 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 *identify* 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 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 suite of philosophical and methodological questions arises about how to carve out the proper roles for PDRS in discovery

[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 Hanson 2021; Michel 2020).

settings. Here, I focus specifically on four related question-clusters that have so far received little attention in the philosophical and ML literatures: First, can PDRS (successfully) inform discovery trajectories and, if so, is it desirable to use them towards this end? Second, by what standards, and underlying values, do PDRS make predictions and recommendations, and which standards are the right ones? Third, do PDRS have sufficient capacity for predicting paths towards genuinely novel and significant discoveries? Fourth, how may PDRS account for the role of social, moral and political values in determining research agendas, and what general views about the aims of science are they compatible with?

In what follows, I draw out a series of 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 requires joint attention by philosophers of science, methodologists of the special sciences, and ML researchers on multiple fronts. Before engaging these concerns one by one, let me stress two caveats: first, the concerns do not all apply equally to different kinds of PDRS, given substantial differences in their respective aims and techniques. I index the concerns, where needed, to specific kinds of PDRS. Second, some of the concerns I elaborate here build on larger, familiar epistemological and methodological insights from the philosophy of science and other literatures. Any familiarity, however, should not detract from the novelty of the concerns I highlight, however, given some have not been articulated in relation to PDRS so far, or have been anticipated but not elaborated in detail.

4.1 Performativity

The first major concern that PDRS raise is about *performativity*. In recent years, social scientists, philosophers of science and computer scientists have increasingly emphasised that many predictive tools, such as analytical and computational models as well as decision-support systems, do not only (aim to) predict events but that their predictions can also causally influence the outcomes to be predicted – their predictions are *performative* (van Basshuysen et al. 2021; van Basshuysen 2023; Perdomo et al. 2022; Khosrowi et al. 2025). 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, as a result of the prediction, the value of $X = x$, whereas it would have been $X = x'$ otherwise. For instance, an economic model might predict liquidity problems of a bank, and, in response, customers withdraw their assets, causing the very liquidity problems that the model predicted. A more pernicious case concerns the use of risk assessment tools to predict the risk of criminal defendants to reoffend in the future, which can become self-fulfilling when high-risk individuals are imprisoned and imprisonment *itself* increases their risk to reoffend (Khosrowi and van Basshuysen 2024). Conversely, predictions might also be self-undermining when a model predicts $X = x$ and, as a result of the prediction, 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 choose to be more cautious, avoiding travel and contacts, thus decreasing infections and deaths relative to the 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. When recognizing performativity as a phenomenon, 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).

PDRS raise acute concerns about performativity, too, 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; Khosrowi et al. 2025). Specifically, predicting discoveries can plausibly causally affect the research pursuits that human researchers will undertake. Two general types of behavioral response by researchers to PDRS predictions and recommendations are to follow the prediction and invest in projects that are deemed likely or promising, or strategically eschew the prediction 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 and their mixtures makes clear that it is largely uncertain how, exactly, PDRS may impact the discovery trajectories ultimately pursued.

What is clear, however, is that PDRS are unlikely to be merely *predictive* tools: if they work as intended, they must be able to inform actual discovery pursuits, and if they do, it should be recognized that PDRS do not merely predict, but can (and perhaps should) causally affect discovery trajectories. Such 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 scientists' imagination or creativity, making the identification of relevant questions, hypotheses and experiments more efficient, their predictions and recommendations must have *some* causal bearing on eventual discovery trajectories, if only by accelerating discoveries that would have been made regardless. 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). Neglecting performative aspects, however, leaves unclear what PDRS' performative powers in shaping research trajectories are and should be, and how PDRS that have the potential to steer discovery efforts in specific directions should be designed and deployed.

To better understand PDRS' potential performative impacts, we can envision an epistemic landscape with a set of in principle possible discoveries D, as shown in Fig. 1.

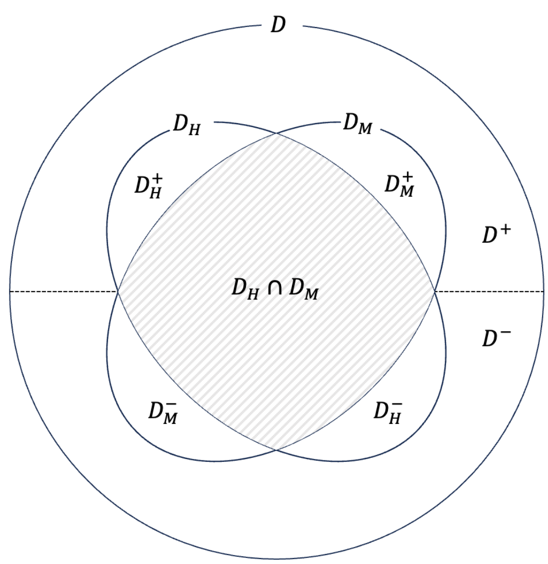

Figure 1. The space of possible discoveries D, and the discovery envelopes D_M and D_H, with and without machines.

The circle describing 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. 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. With these partitions in place, we can now imagine the two ellipsoid discovery envelopes: D_M (for 'machines') which contains discoveries that can be achieved in the presence of PDRS and D_H (for 'humans') that can be achieved by human investigators without PDRS. These envelopes may overlap significantly ($D_M \cap D_H$), but also come apart importantly. Specifically, there are two areas of interest: $D_M^+ \cup D_M^-$, the set of all discovery trajectories that may be taken with PDRS but couldn't be taken without them and $D_H^+ \cup D_H^-$, the set of all discovery trajectories that may be taken without PDRS but couldn't be taken with PDRS; each irrespective of how pursuitworthy the discoveries at issue truly are. In short, there will likely be some nonempty set of discovery trajectories that, perhaps for now, are inaccessible to PDRS-based discovery but may be accessible to human investigators, e.g. because PDRS do not have the creative or imaginative capabilities to conceive of them. Equally, if PDRS work as intended, there will also be a nonempty set of trajectories that humans would fail to consider at all, or fail to consider as sufficiently promising, without PDRS.

With this map in view, we can now more clearly imagine that performative PDRS may steer discovery trajectories towards a distribution at the intersection $D_M \cap D_H$, and at the union $D_M^+ \cup D_M^-$ respectively. Where, exactly, PDRS may steer discovery efforts within these spaces remains unclear, however, as there seem to be at least two different goals that PDRS are meant to promote: 1) supporting or accelerating the pursuit of discovery trajectories that would be pursued anyway, i.e. $D_M \cap D_H$, and 2) promoting the pursuit of (promising) discovery trajectories that would not be pursued without PDRS, i.e. D_M^+. Such efforts, however, come at the risk of steering discovery efforts towards D_M^-. What these possibilities suggest is that how, exactly, PDRS impact the distribution of discovery trajectories that are pursued remains an open, empirical question and that the answer to this question is importantly determined by how PDRS are designed, how human researchers respond to their predictions and recommendations, and by larger visions and understandings of what PDRS' proper roles in supporting/automating agenda-setting should be.

Currently, there is no evidence that these larger questions are considered and addressed in the PDRS literature or among philosophers of science. The performative potentials of PDRS, however, press us to confront these issues and should make us sceptical about the way PDRS are built and how their functioning is framed. Currently, following broader tendencies in ML research, many PDRS approaches transform the problem of choosing promising new discovery trajectories into a *predictive* problem. Key to demonstrating the effectiveness of such 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). But it remains unclear what we could infer from such predictive success. Specifically, existing PDRS largely predict 'off-policy', that is, they predict past distributions of discovery trajectories that could not be influenced by the predictions made. Here, predictive success indicates 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 latent in existing literature, e.g. whether a material M might exhibit property P. In essence, PDRS that can boast such predictive successes hence seem promising tools to accelerate the pursuit of *some* discovery trajectories in $D_M \cap D_H$, i.e. discoveries made anyway.

However, what we can learn from predictive success changes once PDRS are deployed in practice and predictions/recommendations get take-up by human investigators. Here, predictions are made 'on-policy', i.e. they predict discovery trajectories that are chosen, partly, *in light of* PDRS' predictions and recommendations. Here, *accurate* prediction tells a much murkier story: in an on-policy setting, predictive accuracy is not, or at least not only, a signal of whether a PDRS has successfully latched onto features of the mechanisms by which human researchers form new ideas and select questions and hypotheses to investigate, and relevant features of the world, but is rather, at the same time, a measure of the *performative power* (cf. Hardt et al. 2022) that PDRS exert on the distribution of discovery trajectories undertaken, i.e. their power to *induce* discovery trajectories that are pursued *in virtue* of being predicted or recommended (cf. Chawla 2021). Predictive success achieved in on-policy settings is hence ambiguous on whether PDRS successfully predict research that gets done anyway, or whether they merely successfully change the distribution of discovery trajectories to *whatever* is predicted. This ambiguity presses us to consider a second important concern about PDRS, which I now turn to.

4.2 What, Exactly, Is Being Predicted?

A second, major concern about PDRS is that it is often unclear what, exactly, is being predicted, and on grounds of which values recommendations are made (cf. Chawla 2021). So far, I have described PDRS as providing two kinds of outputs: *predictions* of future discoveries and *recommendations* 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. But the observed distributions of ideas, concepts, researchers, their collaborations, as well as ratings, are high-level realizations of a whole array of underlying mechanisms and norms that govern discovery processes, and discovery outcomes correlate with various underlying features, some of which are intrinsically desirable whereas others are not. 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.

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* predictive targets, i.e. the sorts of things that are difficult to measure (and hence predict) directly, but are deemed ultimately desirable when making predictions. 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 predictive targets are hoped to correlate with other desirable features, i.e. latent predictive targets 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 predictive targets 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 agents 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 confounding factors. 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. Similarly, when a community does not consider exploring the connection between A and B pursuitworthy, 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 much information about its (relative) pursuitworthiness over projects that explore other connections, such as A-C or B-D, and are less likely to occur.

This is a major problem for PDRS programmes. Developers of such systems 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 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. 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 in the evolution of discovery trajectories, it is plausible to think that 1) it must be clear by which 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, beyond their prima facie predictive targets, it remains unclear what exactly they predict and by which values they make recommendations.

What is more, even if specific predictive targets could be isolated more precisely from their potentially undesirable correlates, choosing between them may have the capacity to put research programs on significantly different tracks. 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. 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. 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 types of predictive targets may have on the distribution of discovery trajectories undertaken; and so it is unclear 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, related class of concerns about PDRS regards *novelty, creativity*, *epistemic conservatism*, and *homogenization*. In a nutshell, PDRS may be substantially restricted in how novel the connections, questions, hypotheses, and suggestions they offer can be, and might be restricted to conservative predictions on an epistemic landscape, e.g. combining familiar ideas in novel ways, but less able to make more dramatic, transformative moves on an epistemic landscape, e.g. predicting or recommending altogether new ways of approaching phenomena that are discontinuous with existing approaches (cf. Weisberg and Muldoon 2009; Ratti 2020). Relying on PDRS may hence carry the risk of conservatism, and possibly inferior epistemic outcomes. In view of such limitations, PDRS also raise concerns about the (re)homogenization of discovery (cf. Messeri and Crockett 2024; Anderson et al. 2024; Griffin et al. 2025). 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). One worry about PDRS is that they may have the capacity to 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 homogenize the distribution of discovery pursuits undertaken (e.g. owed to incentives not to depart from PDRS' predictions and recommendations).

This class of concerns is 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 discovery tasks have the capacity to achieve what Ratti calls *strong novelty* (Ratti 2020; see also 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 variables that best describe a new phenomenon or system without any prior 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) purport 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 system 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 between them best describe the behavior of a system.

Meanwhile, several purported demonstrations of strong novelty from ML systems 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.

There are hence reasons to be sceptical of whether *execution-level* ML systems have the capacity for strongly novel discoveries. However, so far, 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). 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 D_M overlaps sufficiently with, while also reaching significantly beyond, D_H.

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 literature have long recognized and elaborated that science is not a value-free fact-finding endeavour 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 desirable (Douglas 2009; Elliott 2017; 2022). 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). 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 evolve, 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, it seems clear that at least one important role played by human researchers is difficult to delegate to PDRS: determining by what values scientific inquiry should proceed and what goals it should cater to. Perhaps machines are better than humans at determining efficient discovery trajectories that promote finding significant truth; but determining what is significant in the first place, is up to humans on this view. This suggests that determining a suitable division of agenda-setting labor between human investigators and machines is important, but it is unclear how such labor can be effectively divided with PDRS. 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 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 above, 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 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.[3]

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. This, of course, does not imply that ML systems more broadly cannot be helpful in values discovery and negotiation, too; indeed, there may be significant scope for augmenting such processes. The point, however, is that the *determination* of goals and values central to an inquiry should heavily and authentically involve human agents, which may suggest new emphases in our conceptions of what it means to be a researcher.

4.5 What's new and what's next?

The above concerns 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 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 and scale undesirable attributes in a way that significantly amplifies concerns about them. This is not a new insight: across a range of domains, e.g. algorithmic systems used in hiring, loan approval, recidivism risk prediction, critics have argued that automation both amplify familiar concerns about, bias, fairness, and discrimination that apply to humans, too, but also raise novel issues, e.g., automation bias, the atrophying of important skills, or responsibility gaps. 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 roles in discovery may be facilitated. The discussion offered here yields five recommendations to promote this project.

1) In light of concerns about performative prediction, 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 as well as of designing them. 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 an idea) may inform recommendations of research pursuits related to the idea.

[3] 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 indeed open

2) The predictive targets of PDRS should be made more 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 predictive targets (e.g. disparities in access to funding) and how correlations with undesirable features can be disentangled and broken.

3) Building on the first two points, systematic efforts, such as through simulation studies, may 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 relevant features of human-machine interaction with PDRS and study the effects of PDRS on research trajectories regarding salient features such as speed, diversity (of ideas, approaches, etc.), originality, surprise, and others.

Together, such advances may help put PDRS research programs on track to deal with the concerns raised in this paper, and facilitate reflection and debate 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 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 augment human researchers but more fully automate central agenda-setting tasks, such as identifying novel, unrealized connections between existing knowledge items, formulating new research ideas and gauging what capacity for novelty and surprise they harbor, predicting the impact and success of a research agenda, 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 demonstrations, ongoing efforts to build PDRS have so far failed to anticipate and engage a range of important questions and concerns about what roles such systems can and should play in science. This paper has drawn out four of these 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 to steer discovery trajectories towards strongly novel outcomes, and 4) whether PDRS can account for the ineliminable role that values play in scientific inquiry.

By drawing out a series of concerns relating 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, philosophers of science and researchers in the special sciences to more fully reflect whether PDRS can, and should, assume roles distinctive of researchers. Targeting these issues through the lens of conceptual disruptions helps focus attention on the importance of basic concepts in negotiating these roles, e.g. the role-concept of 'researcher', including the qualities and abilities, such as creativity or imagination, as well as tasks and responsibilities, e.g. bringing values to bear on shaping discovery trajectories, typically associated with this role. Focusing on these issues, I suggest, helps draw attention to two key big-picture questions about the proper roles of ML in scientific discovery: 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, machine learning researchers involved in building PDRS as well as domain experts and stakeholders across the special sciences. In line with the arguments offered here, the goals and values that PDRS research pursues are far from settled, and negotiating where PDRS research should go next is an important task that we shouldn't delegate.

Declarations

The authors did not receive support from any organization for the submitted work. The authors have no relevant financial or non-financial interests to disclose.

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

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

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

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, 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

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

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

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.

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

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

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.

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

Khosrowi, D., M. Ahlers, and P. van Basshuysen. (2025). “When Predictions are More Than Predictions: Self-fulfilling Performativity and the Road Towards Morally Responsible Predictive Systems”, In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’25)*. ACM, New York, NY, USA. https://doi.org/10.1145/3715275.3732072

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, 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

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.

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.

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.

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

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

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. (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.

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

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

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.

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.