



The Problem-Ladeness of Theory

Daniel Levenstein^{1,2} · Aniello De Santo³ · Saskia Heijnen⁴ · Manjari Narayan⁵ · Freek J. W. Oude Maatman^{6,7} · Jonathan Rawski^{8,9} · Cory Wright¹⁰

Accepted: 1 September 2024
© Society for Mathematical Psychology 2024

Abstract

The cognitive sciences are facing questions of how to select from competing theories or develop those that suit their current needs. However, traditional accounts of theoretical virtues have not yet proven informative to theory development in these fields. We advance a pragmatic account by which theoretical virtues are heuristics we use to estimate a theory's contribution to a field's body of knowledge and the degree to which it increases that knowledge's ability to solve problems in the field's domain or problem space. From this perspective, properties that are traditionally considered epistemic virtues, such as a theory's fit to data or internal coherence, can be couched in terms of problem space coverage, and additional virtues come to light that reflect a theory's alignment with problem-solving agents and the context in a societally embedded scientific system. This approach helps us understand why the needs of different fields result in different kinds of theories and allows us to formulate the challenges facing cognitive science in terms that we hope will facilitate their resolution through further theoretical development.

Keywords Pragmatism · Problems · Scientific theories · Cognitive science

Introduction

Cognitive science and its adjacent fields are facing a significant need for theory development. The field of Psychology is experiencing simultaneous crises in replicability (Anvari & Lakens, 2018; Makel et al., 2012; Nosek et al., 2022), poor theory (Eronen & Bringmann, 2021; Muthukrishna & Henrich, 2019), and practical relevance (Giner-Sorolla, 2019). An influx of new experimental techniques in the field of neuroscience has motivated the emergence of alternative theoretical frameworks (Richards et al., 2019; Barack & Krakauer, 2021; Begley et al. 2012), which have yet to significantly impact clinical treatments of neuropsychiatric conditions despite their aim of doing so (Jorgenson et al., 2015). The field of Linguistics faces a disconnect between engineering advances (Mitchell & Krakauer, 2023) and the field's theoretical development, which has stagnated amid an increased focus on behavioral experiments and data analysis.

While these challenges may seem disparate, they each reflect a failure of existing theories to address the needs of the aforementioned fields (Devezer et al., 2021; Goldrick, 2022; Nurse, 2021; van Rooij & Baggio, 2021) and prompt questions of how to evaluate the relative qualities of developing theories. In the philosophy of science, these

✉ Daniel Levenstein
daniel.levenstein@mcgill.ca

¹ Montreal Neurological Institute, McGill University, Montreal, QC, Canada

² Mila, Montreal, QC, Canada

³ Department of Linguistics, University of Utah, Salt Lake City, UT, USA

⁴ Cognitive Psychology Unit, Institute of Psychology, Leiden University, Leiden, The Netherlands

⁵ Dyno Therapeutics, Watertown, MA, USA

⁶ Department of Theoretical Philosophy, Faculty of Philosophy, University of Groningen, Groningen, The Netherlands

⁷ Department of Philosophy of Behavioural Science, Faculty of Social Science, Radboud University, Nijmegen, The Netherlands

⁸ Department of Linguistics & Language Development, San Jose State University, San Jose, USA

⁹ Department of Linguistics and Philosophy, MIT, Cambridge, USA

¹⁰ Department of Philosophy, California State University Long Beach, Long Beach, CA, USA

are known as “theoretical virtues”: the properties by which we judge, or should judge, scientific theories (Kuhn, 1977, Laudan, 1984). Theoretical virtues include properties such as empirical accuracy, internal consistency, and simplicity (Keas, 2018; Schindler, 2018) and are sometimes construed as properties that provide epistemic justification, or support for belief, in a theory (Douglas, 2014a, 2014b; Schindler, 2018). While philosophers of science have debated how to balance competing virtues (Matthewson & Weisberg, 2009; McMullin, 2009) and assess their relative importance (Elliott & McKaughan, 2014; Mackonis, 2013; Rosales & Morton, 2021), this discourse has not yet had a large influence in the cognitive sciences (Roedelein, 1997; Mizrahi, 2022).

We contend that a pragmatic approach, which construes science as a social institution that progresses through solving certain kinds of problems (Doppelt, 1981; Kitcher, 2013; Laudan, 1977), might be useful for informing scientific methodology in this regard. Growing out of the classical projects of Dewey (1938), Peirce (1878), and James (1907), the pragmatist program has had a recent resurgence with the work of, e.g., Chang (2022), Kitcher (2013), and others (Misak, 2007). The core of this approach is “the pragmatic maxim”: that “to attain clearness of our thoughts on an object, we need only to consider what conceivable effects of a practical kind the object may involve” (James, 1907, p 29). While this approach may seem of limited scope beyond so-called practical matters, pragmatism concerns itself broadly with practices (of which science is one, Brigandt, 2013; Longino, 1990), which it sees as systems of activities defined by their aims (Chang, 2022). For example, clarifying a concept’s meaning is just as much an aim as treating an illness. A pragmatic approach is thus particularly well suited to analyzing theoretical virtues, because in asking what properties make a theory good, we must inevitably ask what a theory is good for. That is, it prompts us to acknowledge our aims with regard to scientific theories and identify those properties that are virtuous *by virtue of their ability to further those aims* (Kuhn, 1983).

It is our intent to (1) communicate a pragmatic perspective of science to researchers in the cognitive sciences, which we believe will be valuable for making progress in those fields, and (2) advance a pragmatist account of theoretical virtues, using the cognitive sciences as a case study. Cognitive science is an interdisciplinary field at the intersection of multiple topics (e.g., psychology, neuroscience, linguistics, and artificial intelligence, Miller et al., 2003), which are each a field of study in their own right. We refer to these fields, along with cognitive science, as “the cognitive sciences” (Sobel, & Li, 2013), to emphasize a diverse set of research communities with motivations and problems of interest that overlap with, but are not identical to, those of cognitive science “proper.” Each of the cognitive sciences is currently experiencing an increasing adjacency to emerging

technology (e.g., neurotechnology, social media, and large language models). This in turn puts them in increasing adjacency to “nonscientific” issues such as misinformation, privacy, and engineering practices. Thus, the cognitive sciences provide a unique opportunity to study how the different aims of different research communities, and their relationship to societal aims, can shape scientific research. We hope that the development of such a perspective and its communication to practicing researchers will be valuable to theory development in their respective fields.

The central premise of our account is that scientific theories are “problem-laden”—they are developed with the aim of being effective in solving nonscientific problems, the primary method of their development is through the solution of scientific problems, and they are judged based on their problem-solving ability. Each of these sources of influence shapes the content and form of a theory, from which it cannot be separated. We begin by developing an account of scientific problems that centers on the collective aims of a research community while admitting a wide range of individual motivations for scientific research. We consider research communities to act as societal agents, whose component members are joined by a collective aim to develop a body of effective knowledge in a specific domain. We then argue that theoretical virtues are heuristics researchers use to estimate a theory’s impact on a field’s ability to solve problems in its domain or “problem space.” From this perspective, virtues that are often construed as reasons for belief in a theory can be seen as heuristics for a theory’s ability to cover a field’s problem space, and additional theoretical virtues emerge that reflect a theory’s ability to facilitate its use by problem-solving agents and its context in a societally embedded scientific system. These “agential” virtues can involve operations a theory performs on a field’s problem space, such as its ability to relate previously disparate problems, open new problems for investigation or move a field’s problem space closer to societal interests. Throughout, we provide examples from the cognitive sciences that illustrate the concepts introduced, which we hope will help guide theory development in these fields.

What’s a Scientific Problem?

Science as a Problem-Solving Institution

Although accounts of science that emphasize problem-solving appear as early as Aristotle (Quarantotto, 2020), the work of Laudan (1977) is epicentral to their modern iterations. In contrast to an earlier focus on the logic of theory change and justification (e.g., Popper, 1959), Laudan proposed that science is “essentially a problem-solving activity” and that theories matter only “insofar as they provide adequate solutions

to problems” (Laudan, 1977). This approach developed ideas from Kuhn (1962) and Shapere (1969), which placed heavy emphasis on historical and sociological analysis and the “puzzle-solving” practices of “normal science.”

The central role of problem-solving aligns with methodological guidance from scientists, funding bodies, and editors at all stages of the research process. When developing a research project, “constructing and formulating research questions is... perhaps the most critical aspect of all research” (Alvesson & Sandberg, 2013, p. 1). When applying to fund that project, the primary piece is a specific aims page that “demonstrates a problem [and] proposes aims that work toward a defended solution” (Monte & Libby, 2018, p. 1042). While conducting research, a scientific problem provides a direction for long-term and day-to-day decisions (Beveridge, 1950), and when communicating the results of that work, the abstract and introduction must first “communicate what is missing in the literature (i.e., the specific problem)” (Mensch & Kording, 2017, p. 4) and “convince your readers that you have identified an important, open scientific question that they should care about” (Plaxco, 2010, p. 2263).

Thus, there seems to be an agreement among scientists that (1) there are things called “scientific problems,” which (2) are the primary motivation for doing research. However, scientific guidance gives little treatment to what a problem actually is or what makes one scientific. In other words, “choosing good problems is essential for being a good scientist. But what is a good problem, and how do you choose one? The subject is not usually discussed explicitly within our profession.” (Alon, 2009, p. 726). As a result, scientists frequently disagree whether something is “actually” a problem (e.g., Seth, 2016) or whether a problem’s solution will benefit the field. While these kinds of debates are part of healthy and productive scientific discourse, we argue that they would benefit from a common ground as to what is actually being debated.

A Problem is a State of Affairs in Which an Agent’s Aims are Unmet and is Defined by the Constraints Under Which it Would be Solved

While Laudan extensively discussed the role that problems play in the evaluation of theories, he did not—like scientists themselves—provide extensive guidance as to what a problem actually is or what makes one scientific (Nickles, 1981). Instead, the modern treatment of problem-solving is generally thought to have originated with Newell and Simon (1972), who defined a “problem” as a constrained search in possible configurations, or states, of a domain (see also Holyoak, 1995). For example, the game of checkers is a problem in which players search to find a configuration in which all of their opponent’s pieces are captured or cannot

move, under the constraints of movement. According to this account, a problem consists of an initial state, a goal state, and the allowable moves in the domain, and a solution is simply a sequence of operations that conforms to path constraints and leads to the goal state.

However, Newell and Simon’s definition is only concerned with well-defined problems: those with an explicit state and operations by which it can be transformed toward a goal. In contrast, it is difficult to know what corresponds to the state of a scientific problem, let alone the “allowable moves” in science (Feyerabend, 1975; Nickles, 1980). To address this concern, Nickles (1981) and Haig (1987) developed an account that emphasizes constraints on the solution itself. According to their account, a problem consists of (1) a set of constraints, or criteria, on what counts as a solution and (2) a demand that an object satisfying the constraints (i.e., a solution) be found. This account captures cases in which, rather than a prespecified goal state, we have conditions that, if met, would constitute a solution to the problem. It further accounts for problems that allow partial solutions, in which more constraints (of potentially varying degrees of importance) are satisfied to varying degrees.

Further, by requiring a “demand” for a solution, Nickles’ and Haig’s account also recognizes that not all constraints are necessarily problems—there must be some motivation or impetus to find a solution. This idea is further elaborated in agent-based accounts, according to which problems are only defined with respect to the circumstances and aims of invested agents (Elliott, 2021). For example, a boy’s ruptured appendix is a problem for the parents interested in his well-being, while the treatment with antibiotics is a problem for the invading bacteria. In each case, the same situation presents a different problem for each agent by virtue of their different aims. Agency is itself an active area of interdisciplinary research (Dennett, 1989; Kauffman, 2002; Mitchell, 2023; Nguyen, 2020), which goes beyond the scope of this work. However, as a working definition, we consider an “agent” to be an entity in an environment which has one or more aims (due to internal or environmental states with higher or lower value) and abilities with which it can act to pursue those aims (Mitchell, 2023). Agents can occur at various levels of complexity and organization, from single-celled organisms whose aims include the ingestion of nutrients until the next cell division, to a human whose aims include having more leisure time, and even collective and/or societal agents such as an ant colony whose aims include keeping the colony alive and the queen reproducing or a corporation whose aims include maximizing value for shareholders (List & Pettit, 2011).

Elliott (2021, p. 1014) provided a concise and encompassing summary of these various accounts, stating that “a problem is a state of affairs in which something valued is harmed or is obstructed from reaching an end both valued

and assigned to it.” According to this account, specifying a problem requires (1) a set of propositions that describe a *situation*, including obstructed aims, (2) a set of propositions listing the *agents* who have assigned value or desired ends to items in the state of affairs, and (3) a set of propositions that describe *constraints* on the problem’s solution. By spelling out the propositions necessary to claim “X is a problem,” this “general propositional model” captures key elements that readily capture the colloquial sense in which we refer to problems in everyday life, as well as the technical problems encountered in professional contexts: problems have problem-havers (they are defined with respect to agents for whom the situation is a problem-for), problems are context-specific (they involve a specific state of affairs in which the agents’ aims are unmet), and problems are solved when the constraints entailed by these aims are satisfied.

Scientific Problems are Problems for a Research Community

While the propositional model provides a comprehensive and general account of what constitutes a problem, it is not immediately apparent how it might apply to the problems encountered in scientific research. For example, the stated motivation of a recent study (Sun et al., 2023) is to solve the problem that “it’s unclear why systems consolidation only applies to a subset of hippocampal memories.” Who, in this case, are the relevant *agents*, what is the *situation* such that their *aims* are unmet, and what are the *constraints* on a successful solution?

Nominally, scientific problems are those of professional interest to scientists. Thus, it might seem that the problem-solving agents for scientific problems are the researchers themselves. However, a cursory analysis suggests that Sun et al. do not have a direct stake in the problems they solve (aside from curiosity and professional interest). Instead, scientific papers and grant applications are often framed to address an unmet need of society or a subset of its members, and the institution of science is often “justified” (e.g., by funding agencies or universities) by its potential for societal benefit (Douglas, 2009). However, this emphasis on societal problems does not readily mesh with the textbook view that the aim of scientific research is knowledge for its own sake (Pâslaru, 2023). Instead, scientific problems generally involve a gap in knowledge, a disconnect between existing theories or a methodological challenge. While such problems *may* be motivated by potential action possibilities (e.g., curing a disease or building a new technology), these concerns are rarely the immediate aim of scientific problems, and critically, achieving these goals is rarely seen as a criterion for their solution.

To navigate this challenge, we consider a research community to be a collective societal “agent”: a group of

researchers whose communal aim is the development of a body of knowledge that can be used to solve others’ and future problems involving phenomena in a specific domain (Casadevall & Fang, 2015; Frankel, 1980). This body of knowledge is stored and communicated through the publication, training, or institutional events like conferences and can include, e.g., descriptive and explanatory accounts of phenomena, models and conceptual frameworks by which they can be understood or methodological know-how by which they can be observed and manipulated. To meet the aim that their body of knowledge be useful (or “operationally coherent,” Chang, 2017, 2022), members of a research community can take collective “actions” including, e.g., additions or changes to its body of knowledge or demonstrations of its problem-solving utility.¹ Scientific problems can then be seen as problems for a research community, or “field,” which result from an unmet aim of a useful body of knowledge,² as well as problems for researchers in that field by virtue of their adoption of its communal aim.

The “situation” of a scientific problem thus consists of the current state of a field’s knowledge and a proposition that changing some aspect of that knowledge could increase its problem-solving efficacy. The constraints on that problem’s solution arise from the specific way(s) in which the field’s knowledge could better satisfy its aims. Together, the situation and constraints are often presented as the “background” of a research project, and the most important constraint is often designated by a “research question”: a concise interrogative sentence used to convey a problem and serve as a shorthand for its central

¹ The degree to which a loosely organized group, such as a research community, can be said to take actions is an active area of debate (see, e.g., List and Pettit, 2011, Porello et al., 2014, Tollefsen 2015). Indeed, further work is needed to characterize the distributed manner in which a research community’s body of knowledge is modified, and in what ways, if any, it can be said to be “agential” action of the research community. However, a concerned reader can take this as “as if” agency, in which the combined actions of individual group members together acting as if they were an agent, working toward shared aims.

² We note that the claim that a research community’s primary aim is knowledge with problem-utility (or “operational coherence” Chang 2017, 2022) does not imply this need be the motivation of its individual members, who may themselves be driven to develop knowledge purely for curiosity, and does not negate the various other aims of science such as understanding (Khalifa 2020), which can coexist with, or even be explained by, a primary aim of problem-utility. Further, the aim of problem-utility applies to a research community’s body of knowledge as a whole and not necessarily to any one piece of that knowledge. Just as the research community is an emergent societal-level agent, problem-utility is a societal-level goal that emerges from the social structures that propagate and maintain scientific institutions and their embeddedness in society. Historians of science have noted the key role research communities and their social structures and norms played in the development of science (Wootton, 2015, Shapin 1994).

constraint (a solution should adequately answer the question). In Box 1, we present an analysis of the situation and constraints for the problem solved by Sun et al. (2023) presented at the beginning of this section.

However, the statement of a research question is insufficient to fully specify a problem's solution criteria (Nickles, 1978). For example, when a developmental psychologist asks "How does visual acuity develop?," this question poses a different problem than the one posed by a physiologist with the same question. Where the psychologist may be looking for solutions in terms of experiences during critical stages of development, the physiologist is likely looking for solutions in terms of the response properties of neurons in the visual system. These "tacit constraints" (Polanyi, 1966) are conditions on the solution which are not fully articulated, but whose presence is indicated by the judgments and actions of competent practitioners of a discipline.

The prevalence of tacit constraints suggests that scientific problems are generally ill-defined (Bechtel & Richardson, 2010; Reitman, 1964; Simon, 1973). In addition to unspoken constraints, it is often impossible to know all of the phenomena that might be relevant to a given problem, what knowledge will prove useful to unforeseen extrascientific problems or how it might contribute to the development of future knowledge that does so. Further, scientific problems are generally ill-posed: rarely does a unique solution exist that satisfies the specified constraints. These features are indicative of so-called wicked problems (Schickore, 2020): those that have no unambiguous solution because of incomplete, contradictory or changing requirements.

The wickedness of some scientific problems brings up a potentially disconcerting question: are scientific problems ever really "solved"? We maintain that they are, in two different ways. First, scientific problems are solvable to the extent that researchers explicitly define their solution criteria. Once specified, a problem can be definitively solved by an object that satisfies the constraints (Nickles, 1978). Highly theoretical problems often involve constraints on mathematical structures for which definitive answers can be found, while applied problems can be well defined with respect to a quantifiable goal of a successful application. For example, "what are the possible energy states of a double well potential under the Schrodinger Equation?" is a well-defined problem in physics with a demonstrable solution, and "how can we treat Alzheimer's disease?" has well-defined constraints on its solution involving the reduction of symptoms in Alzheimer's patients and, ideally, a reduction of the number of patients needing care. While the problem itself does not specify *how* the disease should be treated, by attachment to a quantifiable extrascientific target, it is possible to definitively say the problem was solved or at least ameliorated to a quantifiable degree.

Box 1 Problem—"Why does systems consolidation only apply to a subset of hippocampal memories?" (Sun et al., 2023)

Agent: Research community—Systems Neuroscience

Situation:

Systems consolidation theory (SCT, Squire & Alvarez, 1995) and complementary learning systems theory (CLS, McClelland et al., 1995) are prominent theories in the field of systems neuroscience. Roughly, CLS claims that the mammalian brain has two complementary learning systems: one, located in the neocortex, is the basis for the gradual acquisition of structured knowledge about an animal's environment, while the other, located in the hippocampal formation, supports rapid learning of individual experiences. SCT maintains that the offline replay of hippocampally stored memories supports the transfer of new information from those experiences into broadly distributed circuits across the neocortex.

These theories can account for a wide range of experimental phenomena. For example, they are used to explain the effects of anterograde amnesia following hippocampal lesions (Scoville & Milner, 1957; Squire, 1992) compared to the apparent lack of memory effects following localized cortical lesions (Lashley, 1950), the presence and spatiotemporal coupling of replay and reactivation in hippocampal and cortical circuits during sleep (Girardeau & Lopes-dos-Santos, 2021; Ji & Wilson, 2007; Wilson & McNaughton, 1994), the effects of targeted perturbations during sleep and learning via electrical or optical methods (Girardeau et al., 2009; Maingret et al., 2016), and the spatial localization and behavioral dependence on memory-associated "engram" cells after learning (Kitamura et al., 2017). Further, CLS/SCT has been informative for a range of applied and extrascientific problems—for example, they have inspired the development of artificial neural network architectures (Mnih et al., 2015) and explain patterns of memory deficits in epilepsy (Gelinas et al., 2016) and Alzheimer's disease (Zhen et al., 2021), which has led to potential therapeutic targets (Lee et al., 2020).

However, there are a number of experimental observations which are not accounted for by the theories. For example, it has been observed that the ability to recall information learned during some experiences remains dependent on the hippocampus for the entire lifetime of an animal (Gilboa & Moscovitch, 2021). This is a shortcoming of the systems neuroscience's body of knowledge, as it indicates that the theory will not be able to effectively inform, e.g., further experiments (how to design an experiment in a new memory paradigm when the predominant theory cannot predict whether the memory will be hippocampal or cortical-dependent?) or potential external agents who wish to use the theory (how to develop treatments to alleviate epilepsy-related memory symptoms when your theory on which your treatment is based is not reliable?).

Constraints:

The primary constraint is that a solution should answer the research question, by providing an explanation for the observations that systems consolidation only applies to a subset of hippocampal memories. The question is framed as a "why?" question, for which a frequent approach in the field is to provide a normative explanation which demonstrates that some state of affairs is optimal for the solution of some task under neurally plausible conditions (Levenstein et al., 2023). A solution should thus consist of a statement of the "task" being performed by systems consolidation and a demonstration that the observations (selective memory consolidation) are beneficial for its solution. The task, as well as the conditions under which it is optimized (which are also frequently called "constraints"), should be supportable by existing knowledge in the field. In addition, a solution would ideally not disrupt the ability of the field's body of knowledge to account for other, already solved, problems using SCT/CLS.

Proposed Solution:

The authors propose a modification of systems consolidation theory ("generalization-optimized systems consolidation"), according to which hippocampal memories are only consolidated when it aids generalization. In support of this solution, the authors introduce a new neural network formalization of systems consolidation which reveals an overlooked tension in SCT: unregulated neocortical memory transfer can cause overfitting and harm generalization in unpredictable environments. Thus, the observation of selective consolidation can be explained by a postulate that memories which remain hippocampal-dependent are not generalizable.

However, the majority of scientific problems are somewhere between these two extremes. For example, the problem of selective memory consolidation (Box 1) does not include explicit constraints on its solution, and one could imagine solutions that appeal to different kinds of explanation (e.g., mechanistic rather than normative, Brigandt, 2013; Levenstein et al., 2023) or a different operationalization of, e.g., “memories.” This necessitates the second sense in which scientific problems are “solved”: provisional solutions through community-based methods of evaluation and acceptance. This includes an oftentimes messy process of (pre- and post-publication) peer review and consensus among a research community as to whether a proposed solution is adequate. This form of solution is always “provisional” because what is seen as an adequate solution in one socio-historical context may not be in another, as additional constraints (e.g., new data) become available, standards change or alternative solutions are presented. The degree to which science depends on community-based assessment and its susceptibility to subjective opinion has led to a potentially relativistic view of scientific progress (Laudan, 1990). However, such a view downplays (1) historical evidence of progress in the ability of scientific knowledge to address applied concerns with increasing accuracy and scope of prediction or manipulation of phenomena (Silver, 2000, i.e., increasing operational coherence, Chang, 2017), (2) the grounding of scientific knowledge in effective action and experimentation (Strevens, 2020, Hacking, 1983; Chang, 2004), and (3) the social processes used by research communities to achieve objectivity between subjective individuals (Douglas, 2004, 2009; Longino, 1990).

Thus, scientific problems are not different in *kind* from other problems (they are situations in which an agent’s aim(s) are unmet, with a set of constraints on what would count as a solution), but they are distinguished by the *identity* of the problem-having agent: a research community, whose body of knowledge is not meeting its aim of being usable for nonscientific problems in its domain. Further, they are uniquely wicked due to the unspecified nature of how that knowledge might be useful to solve extrascientific problems or the constraints on what changes would constitute an improvement.

Research Communities Have a Shared Problem Space by Which They Judge Theories

A Research Community Has a Shared Problem Space: The Set of Problems in its Domain

Centering scientific problems on a research community leads to a critical question: what is in a research community’s

domain? First and foremost, a research community’s domain contains *the set of empirical phenomena that are the subject matter of its research* (Frankel, 1980). For example, the field of neuroscience’s domain contains phenomena relating to neurons and the nervous system, while the field of linguistics’ domain contains the phenomena of language. The phenomena in a field’s domain are generally considered to be its primary distinguishing characteristic (Casadevall & Fang, 2015; Darden, 1978). However, it’s important to note that research communities and their domains are not mutually exclusive—research communities can have overlapping domains, and the domain of one can even be completely subsumed by another (Casadevall & Fang, 2015). For example, many of the phenomena in linguistics’ domain, as well as many of its members, also belong to the field of cognitive science, and the field of hippocampal electrophysiology is almost entirely within the domain of neuroscience.

In addition to empirical phenomena, a research community’s domain contains *any problems that agents external to the field, such as non-scientists or researchers in other fields, might have related to those phenomena* (Frankel, 1980, Fig. 1A). For example, the problems of “How to treat patients with language disorders?” and “How to effectively teach a second language?” involve language and are thus in the domain of linguistics. While these problems are not themselves scientific (they are problems for speech-language pathologists and educators, respectively, and are neither directly nor exclusively solved by scientific research), they are in the research community’s domain by virtue of their relationship to phenomena in that domain (Love, 2008), and its aim to develop knowledge that can be used to solve them. Such problems are “external” to the research community, and it can be argued that research communities form as a means for developing knowledge that facilitates their solution (Frankel, 1980).

The aim to develop, maintain, and communicate a body of effective domain-specific knowledge leads to new problems—scientific problems that are “internal” to the research community (Fig. 1A). These include what Laudan referred to as “first order” *empirical problems*: “anything about the observable world which strikes us as odd or otherwise in need of explanation” (directly contradicting the field’s aim of a usefully comprehensive body of knowledge about the phenomena in its domain), as well as “higher order” *conceptual problems*: problems about the soundness of higher order structures the field develops to deal with “first order” empirical problems (Laudan, 1977). For example, whether non-context-sensitive dependencies are possible in a language is an empirical problem for the field of linguistics, while whether such a dependency exists over a tree versus a string representation is a conceptual problem, as it reflects an inconsistency between two theories (Graf, 2022). A field’s internal problems also include “toy problems,” which are used

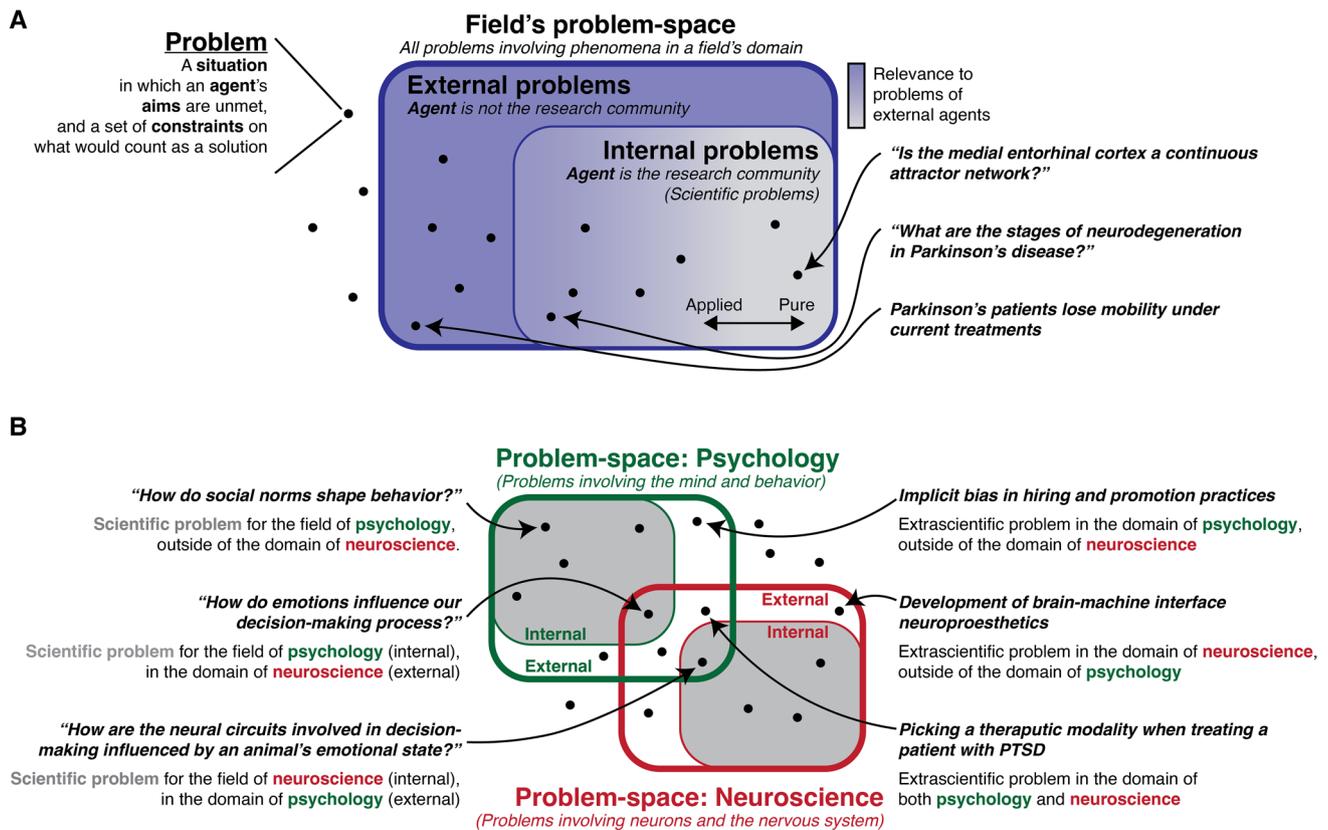


Fig. 1 Problem space. **A** A field's problem space is the set of problems relating to phenomena in its domain, including problems for extrascientific agents or other research communities (external problems) and problems for the research community due to a shortcoming of its body of knowledge (internal problems). Internal problems can

have a range of relevance to external agents, which can be considered a spectrum from pure to applied problems. **B** Research communities can have overlapping problem spaces, as many problems involve phenomena in the domains of multiple fields

as test cases or exploratory grounds for theory development.³ For example, the double well potential in quantum mechanics (Holstein, 1988) did not correspond to anything in the known world (let alone of societal relevance) and was not an internal inconsistency; it was instead a tractable problem taken on by physicists to develop quantum theory—to demonstrate its feasibility, to develop methods by which it could be used, and to understand its implications.

³ Rather than exclusive categories, these various types of problems should be seen as descriptive attributes that indicate something about, e.g., the identity of problem-having agents, the phenomena of relevance to the problem, or the role the problem plays in the research community and its body of knowledge. Indeed, in many cases, problems cannot be cleanly discriminated to a single type. For example, Laudan noted a “continuous shading between straightforward empirical and conceptual problems,” similar notes have been made about pure and applied problems (Douglas 2014b), and toy problems can become pedagogical as they become well-studied or empirical if situations are discovered that they apply to (both of which happened to the double well potential, Jelic & Marsiglio, 2012).

A field's internal problems can range from what is traditionally described as “pure” to “applied” concerns, based on the degree to which they are connected to external problems (Roll-Hansen, 2017; Yaghmaie, 2017; Fig. 1A). For example, an internal problem in the field of neurobiology could be directly relevant to an external problem (e.g., the problem of “why do people develop Parkinson's disease?” is directly related to the problem of “how do we treat people with Parkinson's disease?”) or the possible use of the knowledge could be unknown (e.g., “how is sleep regulated in the fruit fly?”). However, the distinction between pure and applied problems is rarely clear and may not even be an accurate or beneficial characterization of scientific practice (Douglas, 2014b). For example, the solution of “purely” scientific problems often leads to the ability to tackle “applied” concerns (e.g., “how is sleep regulated or dysregulated, in humans?”), and the ability to successfully inform unforeseen applications is some of the best supporting evidence for the epistemic validity of scientific knowledge (even that which was developed under the auspices of pure science).

Together, the problems in a research community's domain comprise its *problem space* (Fig. 1): the set of problems of professional relevance to a research community. Problem space is as important for determining a field as the phenomena in its domain (Love, 2008, Fig. 1B), and one cannot get a full understanding of the actions of a research community without considering its problem space as it is the primary driver of scientists' research efforts.

A Theory is Judged by Virtue of its Impact on a Field's Ability to Solve Problems in its Domain

Among the components of a field's body of knowledge, scientific theories have gotten the most attention from the philosophy of science (Suppe, 1977). Past work in the field has treated theories as refutable explanatory frameworks for a set of observed facts (Popper, 1959), foundations of scientific paradigms (Kuhn, 1962) or research programs (Lakatos, 1970), instruments for predicting phenomena (van Fraassen, 1980), and well-tested proposed truths about the world (Psillos, 1999), among others (Godfrey-Smith, 2003). Whereas traditional conceptions have considered theories to be necessarily composed of logical sentences (the syntactic view) or mathematical models (the semantic view, Lutz, 2017), the pragmatic view has emphasized a more heterogeneous composition (Love, 2013; Winther, 2021)—especially for theories at earlier stages of development than is often considered in philosophy of science (Rohrlich & Hardin, 1983). For example, it has been noted that theories can have a variety of structural components (Craver, 2002a) with varying degrees of formalism, depending on, e.g., their historical trajectory or as needed to suit their function (Love, 2013).⁴ Whatever form they take, scientific theories are generally considered to be higher-order components of a field's body of knowledge which have a domain (a set of

phenomena they pertain to), are selective accounts that omit features of those phenomena (abstraction, Jones, 2005), and contain deliberate falsehoods of the remaining features (idealization, Potochnik, 2017). As a result, different research communities can have qualitatively different theories that suit the particularities of their respective domains, even when comparing the theories of fields as similar to the cognitive sciences (Box 2). However, we note that our problem-centric account of theoretical virtues (1) is agnostic to theories' composition and requires only that they are parts of a research community's body of knowledge which are developed to increase its problem-solving efficacy and (2) is not limited to scientific theories; it can readily be applied to other scientific products such as models.

Box 2 Multi-level theories in the cognitive sciences

In fields that study complex systems (like the cognitive sciences), it is common to consider phenomena at distinct levels (Anderson, 1972; Love, 2012; Oppenheim & Putnam, 1958). For example, a well-known approach in cognitive science is to separate computational, algorithmic, and implementational levels of analysis (Marr, 1982; Marr & Poggio, 1976), and it is common in neuroscience to distinguish levels of organization (e.g., cellular, circuit, and systems) (Churchland & Sejnowski, 1988, Churchland/Sejnowski, 1994) and levels of mechanistic (Bechtel, 1994) or causal (Craver, 2007) explanation. However, it has been difficult to rigorously identify a single set of distinct levels based on phenomena themselves that correspond to the ways this approach is used in practice (Love, 2012, Potochnik, 2021, but see Machta et al., 2013). Rather than statements about discretization in the world, these level-based approaches can be seen as a useful problem-solving strategy (Levenstein et al., 2023), which balances the need to focus on a subset of aspects of phenomena (in order to effectively solve problems in a causally complex world), with the need for a limited number of shared abstractions (rather than a large number of problem-specific abstractions), and is especially useful in the study of biological systems or those studied under a framework of computation (for which the abstraction of processes and functions is a critical strategy, Colburn & Shute, 2007; Wouters, 2003).

In addition to distinct levels, multiple lines of work have emphasized the need for accounts that unify abstractions at different levels into multi-level theories (Craver, 2002b, Bernston & Norman, 2021, O'Malley et al., 2014). In neuroscience, this often involves a division of labor into descriptive explanations (which idealize a phenomenon at a given level of abstraction), mechanistic explanations (which explain how idealizations at one level emerge from those at lower levels), and normative explanations (which explain idealizations at one level by appealing to their ability to perform a function at a higher level) (Levenstein et al., 2023). For example, the solution to the problem in Box 1 is a normative explanation in which a phenomenon (selective memory consolidation) is explained by appealing to its optimality to perform some task (memory performance in generalizable environments, as modeled at a higher level of abstraction). In addition, CLS/SCT contains descriptive explanations that idealize, e.g., hippocampal and neocortical processes, memories, and other components of the theory at, e.g., behavioral, circuit, and computational levels of abstraction, which are connected by mechanistic explanations for *how* memories are initially formed (Nadel et al., 2012) and consolidated (Klinzing et al., 2019) and normative explanations for *why* memories should be separated into complementary learning systems (Roxin & Fusi, 2013). This multi-level theory surrounds a "core" (Lakatos, 1970) idea: that the mammalian brain has two

⁴ We note three senses in which one might consider a theory to have a function: (1) theories are (to some extent) intentionally designed entities, and thus, there is a purpose they were designed for by virtue of the goals of their designer agents, (2) theories are components of a research community's body of knowledge and thus have a function by virtue of their contribution to furthering the research community's goal of problem-solving utility, and finally (3) theories themselves may be subject to memetic selection (Hull, 1990; Shrader, 1980)—some are "replicated" through their propagation in scientific and extrascientific practice, while others are not. One can consider a theory's "function" to be the properties/features/contribution for which it is selected in the same way a protein's function can be considered the contribution it plays toward furthering its gene's selective fitness. In each case, the answer is "problem-solving utility"—a theory is designed to solve some problem(s), it furthers a research community's goal by solving some problem(s), and it is selected/maintained because it continues to be used to solve some problem(s).

complementary learning systems and the offline replay of hippocampally stored memories supports their integration into cortical circuits. However, without the multi-level “belt,” the emperor has no proverbial clothes—it cannot make any predictions, and even the terms in the core are meaningless.

The Marrian level scheme further specifies analysis at computational, algorithmic, and implementational levels. In this case, specific cognitive phenomena can be characterized in terms of the computations they enable (goals/task), the algorithms by which they do those computations (without referring to the specific implementation), and the ways in which they are implemented in neural (or non-neural!) substrates. This level scheme is especially useful for theories in the field of cognitive science (which seeks to build a useful body of knowledge about cognitive phenomena) in which the explanatory target is a specific computation/cognitive phenomenon, due to multiple realizability (Marr, 1982)—that is, the same computation could be performed by many different algorithms, and the same algorithm could be implemented in different substrates.

In contrast, for theories in which the target is a neuroscientific phenomenon (i.e., in the field of neuroscience), the descriptive/mechanistic/normative division allows researchers to (1) characterize neural phenomena at a variety of levels of abstraction; (2) understand how they emerge and how they operate in ways that facilitate understanding, prediction, and possibly informed/effective manipulation; and (3) understand the functions that they serve for the operation of the brain and behavior and how those functions may be enhanced or disrupted (e.g., in disease). This flexibility is necessary because, in addition to multiple realizability, neuroscience faces an issue of multi-functionality—the same neural phenomenon can play a role in multiple functions (e.g., it does not make sense to think about the singular function of inhibitory neurons, but rather their role in many different functions, from the computation of contextual modulation, Keller et al., 2020, to maintaining a stable balanced state in neural populations, Sadeh & Clopath, 2021), many phenomena are subject of study long before a connection to specific functions are established, and there is disagreement as to what the relevant functions even are (Buzsáki, 2020, Poppel & Adolfs, 2020) or if they can all be well explained as computation (Brette, 2018; Marder & Goaillard, 2006; Richards, 2018).

Where Laudan considered a theory itself to be the solution to a problem, we consider instead that they are *used as part* of a solution. This accounts for their use in solving nonscientific problems (e.g., Newtonian mechanics does not solve the problem of “how do we send a lander to the moon?”; it is used by engineers as part of the problem’s solution) as well as scientific problems (e.g., Newtonian mechanics is used as part of a proposed solution to the problem of planetary motion, in which researchers make a model that represents the position of the sun, planets, etc. and compare calculations made using the model to observational data). Further, a problem’s solution often requires the use of multiple theories (e.g., solving the problem of planetary motion requires appealing to additional theories of optics used to collect the data and one of an unobserved 8th planet in the case of the orbit of Uranus). Just as an experiment can only ever test a constellation of theories (Harding, 1976), a problem is solved by a constellation of theories and the way in which they are combined to form an object that satisfies the problem’s constraints.

When faced with a technical problem to solve, extrascientific actors (e.g., policymakers and engineers) must decide to use one scientific theory over another and thus must make a judgment as to which theory is best for their problem at hand. From the perspective of these “theory users,” the question of which theory is best is superficially trivial: for a medical professional, the best theory is one that can inform effective drug design or successful treatment of a patient; for a machine learning engineer, the best theory is one that informs increased performance of their neural network on a specific computational task. That is, the best theory is the one that meets the needs of the problem at hand, as specified by the constraints on its solution.

In practice, identifying which theory fits the bill is rarely straightforward. Indeed, training in fields informed by scientific research generally consists of learning about the kinds of problems encountered in their respective practices, scientific theories that tend to be useful for solving them, and the methods and strategies by which they can be effectively applied. Theory selection then involves an assessment of which available theories contain objects that can be mapped to the phenomena in the problem’s situation or solution criteria (the problem’s “phenomena of relevance”) and if those objects can be made to correspond to those phenomena in the context relevant to the problem. This second requirement is called empirical adequacy (van Fraassen, 1980) or evidential accuracy (Keas, 2018), of which some degree is thought to be a baseline requirement for scientific theories (Douglas, 2009). While it may be challenging to say just how much accuracy is “adequate,” or necessary to accept a theory from a scientific perspective, the degree to which a theory user cares that a theory is accurate and the phenomena they care that it is accurate about are strictly determined by solution criteria of their problem at hand. This can be considered a principle of problem sufficiency: a theory user only needs to consider if a theory’s correspondence with phenomena is *sufficient* to meet the constraints of a given problem. While further accuracy is unlikely to be detrimental, tradeoffs often occur (Heijnen et al., 2024). For example, more accurate theories, especially about complex systems like those encountered in the cognitive sciences, often require accounting for more aspects of the phenomenon, which may or may not be known, easily measurable or easily calculable. A problem’s solution criteria thus provide a critical guide as to how much a theory user should weigh empirical accuracy and about what, relative to other considerations.

Like extrascientific agents, scientists make decisions about which theories to use in the course of research. For example, in Box 1, Sun et al. appeal to the neuroconnectionist theory that artificial neural networks are a good model for their biological counterparts (Box 3). As with their nonscientific counterparts, the best theory for a scientific problem is the one that best meets the needs of its

solution criteria, which determine, e.g., the desired degree of adequacy with respect to specific phenomena (Love, 2008). Unlike their nonscientific counterparts, scientific problems are solved for the purpose of developing scientific theories. That is, scientific researchers are not only theory users but are also, and predominantly, theory developers.⁵ In that capacity, decisions are not only made about which theories to *use*, but also about, e.g., which theory to *develop* (or “pursue,” Laudan, 1977), which parts of it to change or which problems to attempt to solve to demonstrate its problem-solving ability. As a result, where a theory user judges theories based on problem sufficiency, the theory developer judges them based on a principle of *problem coverage*: its contribution to the research community’s body of knowledge as a whole or, more specifically, the *set of problems in the field’s problem space it facilitates the solution of* (Fig. 2; Laudan, 1977; Love, 2008).

A Problem-Centric Account of Theoretical Virtues

The Evidential and Coherential Virtues are Heuristics for Pluralistic Problem Coverage

In a perfect world, we would simply measure a theory’s problem coverage (by, e.g., comparing the set of problems solvable by a field’s body of knowledge with and without the theory, Fig. 2) and keep those that covered more problems than their competitors. Indeed, Laudan (1977) imagined such a calculus as the way theories are ultimately compared. However, we cannot actually assess which problems a theory could be used to solve—first, because we cannot actually count all possible problems in a field’s problem space and, second, because many have not yet been solved (they are rendered solvable, but not solved, by the theory). In place of the ability to make the actual calculation, scientists must estimate a theory’s problem coverage using more easily accessible properties. These properties can then be used as heuristics or rules of thumb for rendering a judgment or making a decision in situations of insufficient time or incomplete information (Bechtel & Richardson, 2010). For example, a common heuristic for a good move in chess is how well it controls the center squares, which a player uses because they cannot calculate a full tree search of the move’s

possible implications. Similarly, a common heuristic for a good theory is how well its objects correspond to experimental observations of phenomena, as it will likely be usable to solve problems for which those phenomena are relevant.

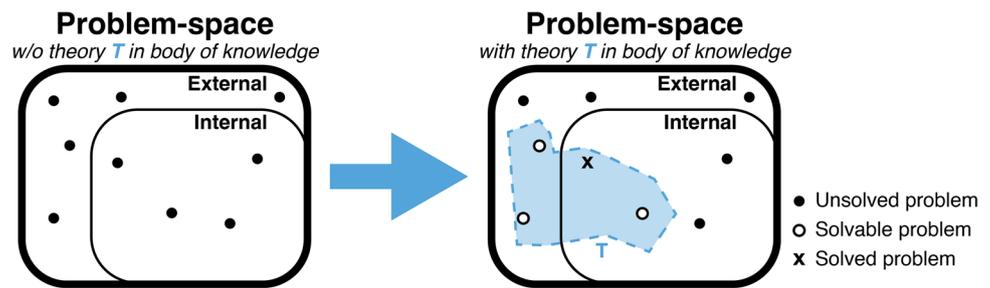
The idea that we judge the quality of scientific theories based on a set of specific properties, or theoretical virtues, is generally attributed to Kuhn (1977), in response to claims that his paradigm-defining work left theory choice a matter of “mob psychology” which “cannot be based on good reasons of any kind” (p. 356). Numerous virtues have been proposed, including testability, empirical accuracy, simplicity, unification, consistency, coherence, and fertility (Schindler, 2018). Keas (2018) has proposed a systematic organization of these virtues, by which they can be divided into four distinct kinds (Table 1): those about how well theoretical components correspond to events and regularities in the world (evidential virtues), those that pertain to how well theoretical components fit together (coherential virtues), those that possess an aesthetic shape that is qualitatively different from the logical-conceptual fit of the coherential virtues (aesthetic virtues), and those that can only be instantiated as a theory is cultivated after its origin (diachronic virtues). While there may be virtues that do not obviously fit in this classification (e.g., testability or falsifiability) and other classifications have been proposed (e.g., Douglas, 2014a, 2014b; Wojtowicz & DeDeo, 2020), we refer to Keas’ classification because the distinction of evidential and coherential virtues corresponds nicely to heuristics based on theory-phenomenon relationships and theory-theory relationships, respectively, and as we will see in the following section, the aesthetic virtues can be subsumed into a new category based on theory-agent relationships.

The evidential virtues are widely regarded to be the most important desiderata for a theory.⁶ While this is often attributed to a claim that theory-observation correspondence is the strongest evidence for the truth of a theory, a pragmatic perspective suggests that these virtues are instead proxies for a theory’s problem coverage, which is based on an underlying assumption that correspondence between a theory’s objects and observations of a specific phenomenon indicates an ability to solve problems involving that phenomenon (Fig. 3A). Consider, for example, evidential accuracy, i.e., the virtue of fitting empirical evidence well. A theory which has no fit to empirical evidence will not be useful for any problems that require accurate accounting for any observable phenomena. As the degree of correspondence between the theory and specific phenomena increases, the theory will be able to cover progressively more problems: those with constraints

⁵ While theory development is sometimes construed as a separate stage from theory-testing, performed by a separate group of “theorists,” theory development is a field-wide, collective endeavor. Even strictly “experimental” researchers develop a research community’s body of knowledge (its theories), in that they are developing empirical descriptions of phenomena, and even applied research develops a theory by demonstrating the ways in which it can be used to solve various external problems.

⁶ “First, a theory should be accurate within its domain. That is, consequences deductible from a theory should be in demonstrated agreement with the results of existing experiments and observations” (Kuhn, 1977).

Fig. 2 Problem coverage



that require accurately accounting for those phenomena to a degree of accuracy less than or equal to that provided by the theory (per the principle of theory sufficiency). Thus, the degree to which a theory matches experimental observations (its evidential accuracy) can be couched, or “cashed out,” in terms of the theory’s ability to be used to solve problems that involve phenomena for which that correspondence has been demonstrated.

Because coverage is calculated with respect to a specific problem space, the phenomena for which a theory’s accuracy is judged and how accurate depends on the problem space of the field for which it is being developed (Love, 2008, Fig. 3A). One might think that the phenomena in the field’s domain delineate these boundaries. However, it is impossible to know ahead of time what scale or phenomena may turn out to be relevant for problems involving the phenomenon in its domain. Instead, the phenomena in a field’s domain determine the problems in its domain, which in turn determine the phenomena represented by objects in its theories. For example, the molecular details of cellular translation or the calculation of information-theoretic measures may not seem necessarily in the domain of neuroscience, but their utility to solve neuroscientific problems brings them within the purview of its theories. In contrast, even though neurons are made of quarks and the details of government influence behavior, these objects are rarely, if ever, in neuroscientific theories. The precision of a theory and the phenomena it accurately depicts are generally only developed with consideration of the level of accuracy that is sufficient to solve problems in a field’s problem space.

In addition to evidential accuracy, Keas (2018) identifies two evidential virtues with progressive expansion of scope (Table 1): causal adequacy, which reflects “the degree to which a theory’s causal factors plausibly produce the effects in need of explanation,” and explanatory depth, which reflects “the degree to which a theory excels in causal history depth or other depth measures such as the range of counterfactual questions that its law-like generalizations answer regarding the item being explained.” We suggest that the progressive scope of these virtues reflects heuristics for a theory’s problem coverage based on how well a theory’s correspondence with observations is able to generalize beyond the specific conditions or phenomena

for which it has been demonstrated (Fig. 3B). For example, theories with a high degree of causal adequacy generally contain objects that represent “underlying causes”—parts and interactions that together produce observable phenomena (Craver, 2007). Because one can manipulate these theoretical objects in ways that mimic conditions in the world, such causal or mechanistic models can be used to extrapolate beyond the bounds of their originally observed data, including unobserved conditions or the effects of perturbations (Ellner & Guckenheimer, 2006; Pearl & Mackenzie, 2018). Thus, causal adequacy indicates that a theory will be able to effectively contribute to the solution of problems requiring accuracy about the phenomena it has been tested for, but in circumstances in which that correspondence may not have been explicitly demonstrated.

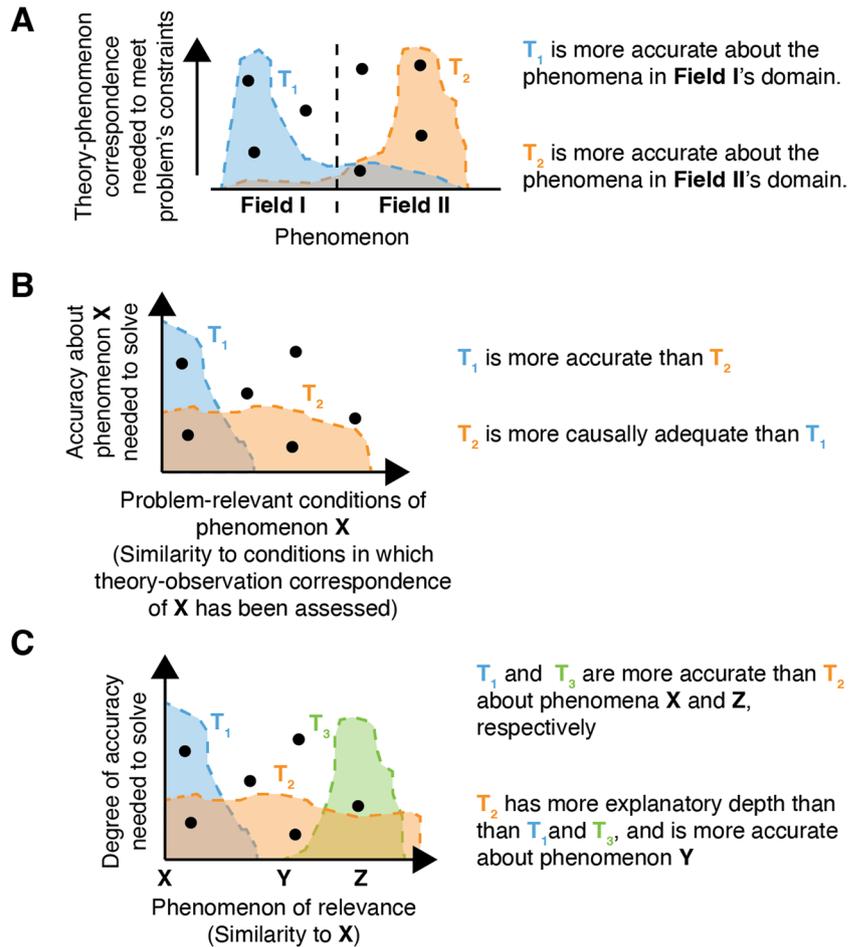
Explanatory depth further expands the scope of domain expansion indicated by causal adequacy. Theories with a high degree of explanatory depth provide extensive causal history in the case of their ability to explain events or contain law-like statements which are able to handle a large range of counterfactual (what-if-things-had-been-different) questions about the phenomena in their domain (Fig. 3C). For example, the theories of plate tectonics, Darwinian evolution, and Newtonian mechanics all have high degrees of explanatory depth, due to the extensive explanations they provide about geological observations, the diversity of species, and the motion of objects, respectively. A theory with explanatory depth does not just account for observed phenomena in unobserved conditions but can be used to uncover previously unobserved phenomena as well. Explanatory depth is thus a heuristic for a theory’s ability to cover problems involving a broader domain of phenomena beyond those the theory was originally developed for or has been specifically applied. This reflects a significant expansion of the theory’s domain to entire classes of phenomena and thus dramatic expansion of its coverage of problem space.

Theories with a higher degree of causal adequacy or explanatory depth need not have a higher degree of evidential accuracy. “Abstract models” in the cognitive sciences often sacrifice a high degree of evidential accuracy for causal adequacy or explanatory depth (O’Leary et al., 2015). These abstract models are often able to account for many different phenomena, but each to a low degree of accuracy.

Table 1 The theoretical virtues, as classified by Keas (2018), and their pragmatic/problem-centric construal

Evidential virtues	About how well theoretical components correspond to events and regularities in the world. Pragmatic construal: heuristics for a theory's problem coverage based on the relationship between the theory and phenomena, Pragmatic construal (Keas et al., 2018)
Evidential accuracy	A theory (T) fits the empirical evidence well (regardless of causal claims) Ability to cover problems that require empirical accuracy to a certain degree about specific observed phenomena
Causal adequacy	T's causal factors plausibly produce the effects (evidence) in need of explanation Ability to solve problems about the originally observed phenomena that require manipulation or prediction with extrapolation to unseen circumstances
Explanatory depth	T excels in causal history depth or in other depth measures such as the range of counterfactual questions that its law-like generalizations answer regarding the item being explained Ability to be applied to problems involving other, related phenomena
Coherential virtues	Pertain to how well theoretical components fit together Pragmatic construal: heuristics for a theory's problem coverage based on internal structure and relationship to other theories
Internal consistency	T's components are not contradictory Pragmatic construal Without—may give conflicting solutions. Unable to get a clear answer from the theory itself
Internal coherence	T's components are coordinated into an intuitively plausible whole; T lacks ad hoc hypotheses—theoretical components merely tacked on to solve isolated problems Suggests T will be able to be readily applied to other problems without further problem-specific modification
Universal coherence	T sits well with (or is not obviously contrary to) other warranted beliefs (e.g., other theories) Without—conflicting solutions with other theories: which theory to choose?
Aesthetic virtues	Possess an aesthetic shape (fittingness) that is qualitatively different from the logical-conceptual fit of the coherential virtues Pragmatic construal: subsumed by agential virtues, based on the relationship between the theory and various agents
Beauty	T evokes aesthetic pleasure in properly functioning and sufficiently informed persons Pragmatic construal A theory aligns with the aesthetic preferences of theory-using agents (agent appropriateness)
Simplicity	T explains the same facts as rivals but with less theoretical content A theory plays to the usability and understandability constraints of theory-using agents (agent appropriateness)
Unification	T explains more kinds of facts than rivals with the same amount of theoretical content -> Communal facilitation
Diachronic virtues	Can only be instantiated as a theory is cultivated after its origin
Durability	T has survived testing by successful prediction or plausible accommodation of new data Pragmatic construal Examples of a theory's problem-solving accomplishments
Fruitfulness	T has generated an additional discovery by means such as successful novel prediction
Applicability	T has guided strategic action or control, such as in science-based technology

Fig. 3 Problem coverage and the evidential virtues. Schematics of the different forms of problem coverage estimated by the evidential virtues. Problems are designated by points which, when shaded, are rendered solvable by the presence of a given theory in a field's body of knowledge



For example, the continuous-rate units used in many artificial neural networks (Yang & Wang, 2020) are a significant abstraction of the complex geometry and electrical properties of neurons and as a result sacrifice a large degree of correspondence to their observed electrical activity. However, they can be used to model the activity of individual neurons across diverse brain regions and cell types (Doerig et al., 2023) and even the activity of entire populations (Wilson & Cowan, 1972), their latent factors (DePasquale et al., 2023), neural subcompartments (Jones & Kording, 2021) or abstract representation spaces (Rumelhart et al., 1987). This is similar to the case in physics, where the most general laws (those with the most explanatory depth) almost never correspond to experimental or applied uses. Instead, these laws are generally supplemented with phenomenological laws and correction factors which are needed to make them apply to what is actually observed (Cartwright, 1983).

Indeed, a research community will often maintain more abstract theories along with those with a higher degree of evidential accuracy (Box 3). For example, there are multiple alternative accounts of the electrical activity of single neurons (Gerstner, 2014), each of which involves different abstractions and idealizations and has different degrees of

(in)accuracy under different conditions or assumptions. Where it might seem illogical under a “theories as proposed truths” framework to maintain multiple theories with overlapping domains (e.g., as they cannot all be, strictly, true), a framework of problem coverage explains why a field might maintain a population of overlapping theories: they collectively cover the problems in a field’s domain (Fig. 4). This population naturally includes theories about different phenomena (and thus are potentially applicable to problems for

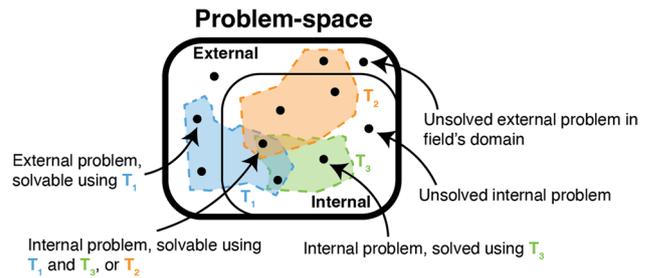


Fig. 4 Pluralistic problem coverage of a field's problem space. Fields have multiple theories that together aim to cover all problems in their problem space

which those phenomena are relevant) but also includes the proliferation of different theories about the same phenomenon (e.g., at different levels of abstraction, Love, 2012). This is because (1) the world is causally complex, and thus, problem-solving requires selective attention to the more relevant aspects of a subset of phenomena (Khalifa, 2020; Potochnik, 2017) and (2) the needs and competencies of different agents are highly diverse, requiring different theories that might meet these needs. As a result, theories pay for their success on some problems with failure on other problems. We consider this to imply a “no free lunch” principle of scientific theories: that no theory can cover all problems, even about the same phenomenon, which is analogous to similar principles in optimization and induction (Wolpert, 2023). This suggests that a kind of theoretical pluralism should be maintained by a research community, with a population of theories that make different idealizations to collectively cover the various constraints of the problems in its problem space (Brigandt, 2013).

Pluralism has become a popular view among philosophers of biological sciences, as it has repeatedly proven difficult to square strict reductionism with the actions of scientists in those fields (Dupre, 1993, Mitchell, 2003). Pluralism is not without its detractors, however. For example, it has also been argued that pluralism is only a temporary state, en route to the one true theory at the lowest (most fundamental) level of abstraction (Oppenheim & Putnam, 1958), and one might worry that extreme disunity of scientific theories will break apart the communication, social enterprise, and debate-to-consensus process on which a research community relies (Strevens, 2020) or that an “anything goes” mentality will result in a loss of scientific standards (Reisch, 1998). Where the evidential virtues can be seen as heuristics for a theory’s problem coverage based on theory-phenomenon relationships, the coherential virtues can be seen as heuristics for problem coverage based on a theory’s internal structure or relationship to other theories, which counterbalance the risks of unchecked pluralism.

To understand the relationship between the coherential virtues and theoretical pluralism, we consider the implications if the most expansive one, universal coherence, is lacking. Universal coherence states that a theory should not be contrary to other warranted beliefs, such as other well-supported theories. If a research community maintains a piecemeal collection of problem-specific theories, it can be difficult for a theory user to know a priori which to use when problems arise that involve phenomena in the domain of multiple theories. If the theories are in general agreement but differ in, e.g., their degree of evidential accuracy, then a theory user can simply choose the theory with sufficient accuracy for their problem at hand or according to some other criteria. However, if the theories disagree (e.g., are inconsistent), the theory user

is left with a conundrum of which to use. Further, solving some problems requires at least the partial integration of concepts and explanations from different theories (Brigandt, 2010), which is not possible if the theories are incoherent. Thus, incoherence has the effect of decreasing a theory’s problem coverage—the field’s body of knowledge is less able to solve some problems (those in the domain of, or requiring the combined use of, incoherent theories) with the theory than without it. Indeed, identifying and rectifying incoherent theories with overlapping problem spaces are often a strong motivator of research. However, if the theories are never used to solve the same problems, or if they are compatible (overlapping but without explicit disagreement), there is rarely a strong drive to unify, as theories about the same phenomenon at different levels of abstraction are generally useful for different problems. Thus, universal coherence can be seen as a virtue to the extent the two theories have overlapping problem spaces (Brigandt, 2010), rather than overlapping domains of phenomena.

The argument above holds for the other coherential virtues, though each with a less expansive scope. If a theory includes ad hoc components tacked on to solve isolated problems (internal incoherence), then a theory user cannot rely when facing additional problems that the theory would not need further ad hoc modifications. If the theory is internally inconsistent, then it cannot be relied on to give a consistent solution at the time of its use. In these cases, theory users could end up with different possible solutions using the same theory, given the variation in the ad hoc modifications they employ. This would result in a kind of “pluralism” that is vacuous; it is not theoretically based, but rather based on confusion in the theory.

Box 3 Theoretical pluralism in neuroscience

The field of neuroscience is experiencing a rapid development of new neurotechnologies, which enable large-scale recordings of up to thousands of neurons simultaneously, possibly from multiple brain regions, with cell type specificity, during complex behavior, and the ability to selectively manipulate neurons in the population. While this has had significant benefits to the field, resulting in a wealth of new data and the ability to perform targeted manipulations of neural systems, it has revealed that many of the field’s former theories are unable to account for the new data or inform the reliable use of those manipulations. These former theories were, for the most part, developed to account for the activity of a small number of neurons, with selective attention to interpretable neurons, and population-level theories which involved idealized collections of these interpretable units, or highly abstract models/descriptions, have been unable to account for the complexity observed with the new methods. As a result, the field has seen the development of a number of candidate theories:

Deep learning framework: Explanations of the neural computations underlying cognition should focus on objective functions, learning rules, and architectures (Richards et al., 2019)

Hopfieldian view: Cognition can be well explained by transformations

between or movement within representational spaces that are implemented by neural populations. To be contrasted with a “Sherringtonian view,” that cognition can be well explained by point-to-point communication between neurons organized into circuits (Barack & Krakauer, 2021)

Neuroconnectionism: Artificial neural networks (ANNs) are a highly suitable computational language to model the brain computations underlying cognition: sufficiently abstract to be computationally tractable and reproduce cognitive functions, while still being close enough to biology to relate to, implement, and test neuroscientific hypotheses (Doerig et al., 2023)

Phylogenetic refinement: Biologically plausible theories of behavior (including cognition) can be constructed by following a method of “phylogenetic refinement,” whereby they are progressively elaborated from simple to complex according to phylogenetic data on the sequence of changes that occurred over the course of evolution (Cisek, 2019)

Inside-out view: Rather than appealing to cognitive terms, theories of neural phenomena should be framed in terms of intrinsic patterns which are selected and grounded by action and prediction (Buzsáki, 2019)

A problem-centric view would suggest considering the coverage of each theory—what problems does each cover, and where is the overlap? It is interesting to note that these are all very general theories (“views” or conceptual “frameworks”), which specify how explanations should be framed but have little to say about specific phenomena (Levenstein et al., 2023). Thus, they have the potential for large and overlapping problem coverage (all refer to neural phenomena of “cognition”), with a high degree of explanatory depth but a low degree of empirical accuracy. In order to solve problems requiring a notable degree of accuracy, these theories would need to be combined with other, more specific, theories about the phenomena of relevance. For example, in addition to the standard systems consolidation theory (SCT, Box 1), there have been multiple developments and alternative theories that all pertain to phenomena of memory, sleep, and hippocampal-cortical interactions:

Multiple trace transformation theory: Memory traces are formed in the hippocampus and neocortex at the time of encoding. Hippocampally dependent, episodic or context-specific memories transform into semantic or gist-like versions that are represented in extra-hippocampal structures. To the extent that episodic memories are retained, they will continue to require the hippocampus, but the hippocampus is not needed for the retrieval of semantic memories (Winocur et al., 2010)

Memory indexing theory: The role of the hippocampus is to form and retain an index of neocortical areas activated by experiential events. The reactivation of the stored hippocampal memory index (e.g., during sleep) will reactivate the associated array of neocortical areas, resulting in a memorial experience and, through cortico-cortical plasticity, establishing a cortically based memory trace (Teyler & DiScenna, 1986)

Synaptic homeostasis hypothesis: During sleep, spontaneous activity renormalizes synaptic strength across neuronal populations. Memories are consolidated through comprehensive sampling of statistical regularities incorporated in neuronal circuits over a lifetime (Tononi & Cirelli, 2014)

This illustrates multiple kinds of pluralism relevant in neuroscience—the field maintains theories with different degrees of explanatory depth, in which broader theories have a wider domain but require more domain-specific theories, as well as theories with a similar degree of coverage of problems in highly overlapping domains.

The Agential Virtues Reflect a Theory’s Ability to Facilitate its Use by Agents and its Context in a Societally Embedded Scientific System

Scientists use theories for reasons that are not just about a problem at hand, and they develop them for reasons other than their ability to directly solve specific problems. For example, a scientist might try to develop one theory over another because the first is more relevant to a specific audience or the subject of ongoing debate, and the “aesthetic” virtues (Table 1) of beauty and simplicity are often used to justify theory selection. Initially, this might seem like a problem for a problem-based view of theoretical virtues: why would these considerations, which seem unrelated to a theory’s problem-solving ability, influence which theories are used and developed? We next propose a set of theoretical virtues—the agential virtues—which reflect a theory’s ability to facilitate its use by agents and its context in a societally embedded scientific system. Where the evidential and coherent virtues are heuristics for a theory’s problem coverage based on theory-phenomenon and theory-theory relationships, respectively, the agential virtues are those based on the relationship between a theory and specific agents. Like the other sets of theoretical virtues, we identify three agential virtues with progressive expansion of scope (Table 2), by considering the needs of theory users, research communities, and society.

The least expansive agential virtue, agent appropriateness, refers to the degree to which a theory aligns with the capacities of its intended users. These may be other researchers in the same field, researchers in another field or non-scientific agents. For example, in a field where linear algebra is not part of the standard curriculum, theories that do not require its use might be preferred over those that do, and a theory with implications for mental health may be preferable if it refers to phenomena that are easily measured or manipulated by clinical practitioners. Agent appropriateness acknowledges the fact that most of the problems rendered solvable by a theory are not immediately solved by its presence, and they must then be solved (either by other researchers in the case of scientific problems or by external agents in the case of extrascientific problems). By aligning a theory with the abilities of those theory users, agent appropriateness renders more problems more readily solvable and thus indicates a higher degree of problem coverage for a theory.

Agent appropriateness does not simply refer to agents’ technical abilities. We maintain that the virtues of beauty and simplicity can both be accounted for under the umbrella of agent appropriateness, in that both are defined with respect to a specific agent. Beauty is simply an alignment with a specific agent’s intuitive and aesthetic preferences.

Table 2 The agential virtues

Agential virtues	Reflect a theory's ability to facilitate its use by agents and their context in a societally embedded scientific system. Heuristics for a theory's problem coverage based on the relationship between the theory and potential problem-solving agents
Definition	
Agent appropriateness	T fits the capacities of its intended theory-using agents
Communal facilitation	T supports the health and efficacy of its research community
External alignment	T aligns with societal benefit

This alignment might be expected to make theories more easily transmissible to agents with those preferences (Boyer, 1998) and thus facilitate their use.⁷ Similarly, simplicity can be seen as an alignment with an agent's information-processing capacities. Where one might define an agent-independent metric of simplicity using information-theoretic methods (Sterkenburg, 2016), e.g., by quantifying the number of bits needed to express the theory or the number and order of terms in a mathematical model, its calculation depends on the choice of formal structure in which a theory is expressed, and there is information in the model's "construal" (how those terms are interpreted to correspond to phenomena, Weisberg, 2013) which cannot be quantified in the same way. Further, such a calculation rests on an assumed ideal decoder, which does not necessarily reflect the information-processing abilities of a theory user. We suggest that the theoretical virtue of simplicity can instead be understood as alignment with what is cognitively simple for problem-solving agents, which is a heuristic for the theory's problem coverage by way of its usability—what does a theory user find simple to use—and understandability—what does a theory user find simple to understand. While this does in many cases align with formal simplicity—e.g., fewer theoretical objects, it is defined with respect to the capacities of a specific theory-using agent and is thus under the purview of agential alignment.

The second agential virtue, communal facilitation, expands the scope of the first to consider the ways in which a theory facilitates the work of the research community. Often, this reflects a theory's ability to solve specific problems of critical value to the field (Fig. 5A). Where the evidential virtues are agnostic to the identity of problems covered by a theory, some

problems may be more important for a field due to their relationship to other problems and the implications their solution would have for the rest of its problem space. For example, solving a methodological problem often opens the door for others to use the method in a different context and thus facilitates the solution of previously unsolvable problems. Similarly, a theory might fill an open niche in problem space by covering unsolved problems or even by including unaccounted-for phenomena in a field's domain (thus covering many unsolved problems for which that phenomenon is relevant). Scientists often pay extreme attention to developing theories that can cover these critical or unsolved problems, even at the expense of theories that can cover more, but already solved, problems. Community facilitation accounts for the preference for or attention to developing theories that solve some problems over others.

The third agential virtue, external alignment, reflects a theory's contribution to external problems in a research community's domain and their alignment with social benefit. External alignment adds additional importance to the coverage of societally relevant problems (Fig. 5A). For example, a theory which can predict the spread and mutation of respiratory viruses may be developed for its social alignment, and theories about the effects of social isolation on cognitive development become more socially aligned during a pandemic. This is due to its support of the research community's aim—that its body of knowledge is potentially useful to solve external problems.

Together, the agential virtues account for properties a theory might be selected for, which are not about its degree of evidence or coherence, but about the ways it facilitates problem-solving by individual agents or a research community, in the interest of society. One might note that the agential virtues appear to overlap with the "diachronic virtues": those can only be instantiated as a theory is cultivated after its origin (Keas, 2018, Table 1). These include durability (a theory has survived testing), fruitfulness (a theory has generated additional discovery), and applicability (a theory has guided strategic action or control). However, where the diachronic virtues reflect accomplishments of a theory which are only observable over time, the agential virtues are properties of the agent-embedded

⁷ Further, these preferences themselves have been shaped by cultural and biological evolution, which may have selected for a preference for cognitive properties that facilitate effective problem-solving (Wojtowicz & DeDeo, 2020). Of course, determining post hoc the factors that led to a given trait is notoriously difficult, and evolutionary processes are not a guarantee of optimality (Gould & Lewontin, 1979), especially when considering cultural products and cognitive properties (Boyer, 1998; Fracchia & Lewontin, 1999).

context of a theory which can be evaluated at the time of a theory's "origin."⁸

One might wonder why we should value properties that play to the idiosyncracies of human cognition and society. Indeed, these virtues are decidedly not epistemic—on their own, they have no bearing on the justification of a theory's truth value. However, they play a significant role in the overall success of a theory, a research community, and its problem-solving efficacy and thus indirectly in its epistemic success. As has been pointed out previously, science is a social process that progresses through societal mechanisms (Kuhn, 1962; Longino, 1990), and views that ignore this are neither an accurate portrayal of science nor helpful for scientific practice. Further, as limited beings in a causally complex world (Potochnik, 2017; Wimsatt, 2007), our heuristics for judging problem-solving tools are inevitably specific to the agents who wish to solve them (Bechtel & Richardson, 2010; Wimsatt, 2007). Even if they may seem to be a negative or irrational aspect of human nature, it is important that we acknowledge and weigh these agent-specific properties in our scientific judgments and decision-making, as it is a more effective and, indeed, rational way to do science than pretending they do not exist.

Such a consideration necessitates considering the relative importance and prioritization of different kinds of virtues. Following Keas (2018) and matching the broad consensus of both philosophers and scientists (Schindler, 2022), we would advocate weighing evidential over coherent over agential virtues. For example, a highly inaccurate theory should not be used over an accurate competitor simply because it uses more familiar methods. This maintains the trustworthiness of scientific knowledge, by grounding in effective action and external problems which require a certain degree of reliability. However, even if the agential virtues are less important than, e.g., empirical adequacy, they are critical considerations in theory selection and development. For example, the cost of training in new methodology and transitions to unfamiliar research directions can be high, and given similar degrees of evidential accuracy, it may be preferable to use the easier theory. Further, theoretical pluralism allows us to maintain lower accuracy but easier-to-use theories alongside their less agent-appropriate counterparts, with knowledge of their shortcomings and the circumstances in which they are appropriate (or not) to use.

⁸ Though note that theories rarely have a singular origin, rather than a protracted period of development by a community of researchers.

Theories are Active Players in a Dynamic Problem Space

Theories Can Perform Virtuous Operations in Problem Space

Problem space is highly context-specific. It is determined by the current availability of data and methods and the current state of the field—its members, their interests, and their maturity. Some problems may be seen as more important than others at a given time or completely meaningless at another. Thus, a field's problem space changes, e.g., when new methods are developed or the needs of extrascientific agents change. Furthermore, a research community's problem space is not just a list of disconnected problems. It has a complex organization in which some problems are more closely connected, e.g., because the ability to solve one depends on the solution of another, they have overlapping phenomena of relevance, involve similar experimental techniques or are problems for the same external agents. In addition to simply solving existing problems, a theory may change a field's problem space in a way that renders it more amenable to solutions by future theories or increases the number of problems solved by other existing theories. This "indirect" coverage of a theory is readily accounted for by the agential virtues.

For example, a theory might expand a field's problem space, by identifying novel problems or by bringing new problems under the purview of a field's theories. These "problem-finding" operations are a critical and often underappreciated operation in science (Adolfi et al., 2023; Getzels, 1979). For example, the change to CLS in Box 1 opens new problems for the field of neuroscience: "How does the hippocampus know which memories are generalizable?" or "How are the predictable (and thus generalizable) elements of a memory separated from its non-generalizable (episodic) components?" Even if the proposed theory itself cannot solve these new problems, it is possible that they are solvable by other theories in the field or will lead to further theory development and their solution and thus lead to an increase in the field's problem coverage. This is a beneficial effect of the theory beyond the ability to solve existing problems, which can be captured under the virtue of community facilitation.

Alternatively, a theory might contract problem space by showing that what were previously considered to be disparate problems are closely related. Where expansion of problem space can spur progress by directing efforts toward new problems, contraction of problem space makes it more easily coverable by fewer and future theories, as problems that are closer to problem space are more easily

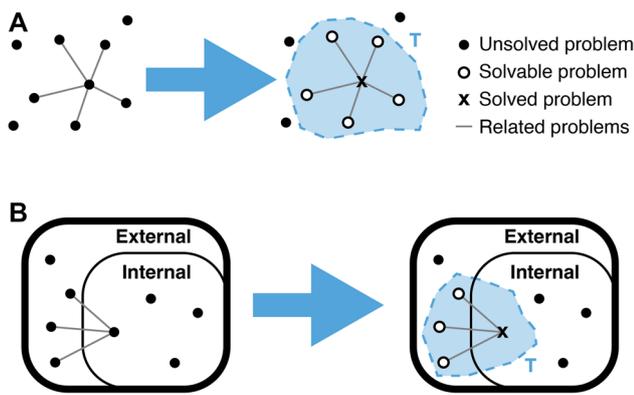


Fig. 5 Problem coverage and the agential virtues. **A** Theories can achieve community facilitation by solving critical problems that allow the solution of other, related problems in the field’s problem space. **B** Theories that facilitate the solution of societally relevant external problems are externally aligned

covered by the same theory. For example, the change to CLS in Box 1 brings problems related to the distinction between episodic and semantic memories closer to problems related to systems consolidation. In addition to rectifying incoherence (“The Evidential and Coherential Virtues Are Heuristics for Pluralistic Problem Coverage” section), unification can be seen as a beneficial contraction of problem space via coverage of more problems by fewer theories (cf. Box 4). While unification is traditionally described as a single theory explaining more facts than its competitors (Table 1), it can also account for the selection of theories that connect areas of problem space which were previously solved with disparate theories or developing connections between existing theories which previously covered different areas of problem space. Even if a theory that performs these operations can solve fewer problems than its domain-specific competitors, it may be selected or developed for its community facilitation or for its external alignment if it moves a field’s problem space closer to societally relevant problems (e.g., because it contains objects that correspond to readily manipulable or societally important phenomena).

Box 4 The replication crisis as a disconnected problem space

It has been argued that the field of psychology is experiencing a replication crisis (Anvari & Lakens, 2018; Makel et al., 2012; Nosek et al., 2022) or an inability to replicate many previously published results. While this may stem from improper use of statistical tools (e.g., p-hacking and low-powered studies, Stanley et al., 2018), it has been said that this replication crisis is more likely a symptom of a “theory crisis”—a focus on specific “effects” rather than the development of general theories (van Rooij, 2019) and a lack of theoretical formalization or mathematization (Borsboom et al., 2021; Fried, 2020; Robinaugh et al., 2021) of theories, which simultaneously are argued to lack proper conceptualization and coordination (Eronen & Romeijn, 2020) and ontological commitment (Oude Maatman, 2021).

A problem-centric view would frame this as the result of an expansive, but disconnected, problem space (each effect reflects an isolated problem) and the corresponding development of problem-specific theories which are inherently brittle due to their limited coverage of problem space. This aligns with recommendations that the field focuses on developing theories with attention to their causal adequacy and explanatory depth (Guest & Martin, 2021), perhaps even at the expense of their evidential accuracy, as such theories can cover more of the problem space. Further, the crisis can be seen to reflect a communal problem which problem space itself is fragmented; that is, each experimental effect corresponds to its own unique problem, rather than experiments that have the potential to inform multiple problems and problems that encompass observations from different experimental modalities. Together, the arguments presented in this paper suggest that theories that are able to unify problem space should be selected and developed due to their communal facilitation (cf. finding a foundational theory; Muthukrishna & Henrich, 2019).

Theories Can Solve, or Create, Emergent Problems from Collective Needs

When a research community’s needs are considered, new kinds of problems emerge that concern its health, productivity, and societal embedding. While these “communal” problems may not seem traditionally scientific, we argue that they are, insofar as they are related to the practice of science, their solution facilitates scientific progress, and scientific theories can (at least in part) contribute to their solutions. In addition to solving critical problems or performing beneficial operations in problem space, the agential virtues can reflect a theory’s ability to ameliorate these problems.

The first kind of problem is related to the content of a research community’s theories. For example, researchers might disagree on what the phenomena even are or how they should be described. At first glance, this is a traditionally “scientific” problem—a debate about the relative merits of competing theories—which might be solved by considering their respective degree of, e.g., evidential accuracy. Indeed, science is a procedure for solving this problem, of which disagreement is a healthy part of the process (Strevens, 2020). However, there is an additional communal problem which may arise: disagreement between members of the research community can result in an inability to communicate and thus hinder progress on the traditionally scientific aspects of the problem. Theories that help solve the communal problem (e.g., by developing shared models or terminology) can be virtuous via community facilitation, even if they do not solve the traditional problem itself.

Indeed, many of these communal problems can be seen as sociological problems, related to the internal structure of a research community. An effective research community thrives on a diverse population of members (Muldoon, 2013) with a complex and modular structure with groups of researchers focused on different problems or approaches (Weisberg & Muldoon, 2009). This includes, for example,

members with diverse background knowledge, as well as a diversity of approaches and distances one can be from experimental data. While this diversity is an effective tool for scientific discovery (Devezer et al., 2019), it can also result in a problematic inability for a field's members to understand or be aware of each others' work and thus an inability to apply the full range of a community's theories to their respective problems of interest. To contribute to the solution of this problem, an "introductory" theory might make the ideas or approaches of one group accessible to those of another without background knowledge or a theory might connect the theoretical objects used by different groups of researchers. In addition to the coherent virtue for problem coverage, such theories have an agential virtue for community facilitation.

Problems that run counter to societal alignment can also emerge because of the relationship between the community's problem space and extrascientific considerations. For example, in addition to supporting the amelioration of societal problems of energy production, nuclear theory introduced new problems of weapons technology and nuclear waste. This can be seen as a misalignment between the field's problem space and social/ethical concerns—where "how do we split an atom?" is a critical problem in the field's problem space, its solution introduces new external problems. This in itself introduces a communal problem to be solved (the misalignment), which might be solved, e.g., with theories that enable the use of materials in nuclear research which produce less waste. Similar issues are coming up in present work in artificial intelligence (AI), because the solution of problems in the field domain (and the ways in which they're solved) has societal applications that are misaligned with some social/ethical issues—e.g., AI fairness and discrimination using AI systems (Birhane, 2021), energy usage and global warming (Strubell et al., 2019), and the production of misinformation (Goldstein et al., 2023).

Box 5 Linguistics and language engineering

Linguistics was a leader of the cognitive revolution, through extensive mathematization and theoretical development that directly confronted the behaviorist and empiricist traditions dominant in psychology (Tomalin, 2006). Interestingly, the linguistics community has fractured amidst this push, with many researchers feeling that modern theory developments do not serve the interests of the field (Dockum & Green, 2023). At the same time, the advancement in computing hardware and accessible implementations of probabilistic models have advanced the engineering of language or Natural Language Processing (NLP, Min et al., 2021). Where previous work in the field was able to satisfy both engineering and scientific aims (Steedman, 2008), current engineering approaches have become extremely focused. This has led to debate about the importance of NLP to the field of linguistics and cognitive science more broadly (Mitchell & Krakauer, 2023; Rawski & Heinz, 2019), with some proponents claiming that the new technology reflects a paradigm shift for the field (Baroni, 2022; Wei et al., 2022), while others claim it is overhyped (Shanahan, 2022) or inconsequential (Veres, 2021) or even comes at the expense of scientific insight (Marcus, 2022; Rawski & Baumont, 2023). A problem-centric

view would consider this controversy to be a communal problem resulting from a rapid expansion of problem space, due to technological development and a rising adjacency to new external problems. Indeed, a tension between pure and applied research exists in many fields with scientific and engineering components. In many cases, the result is the emergence of distinct research communities doing work that draws from one another (e.g., Linzen & Baroni, 2021; Valvoda et al., 2022). Further, one might imagine that linguistics might develop theories that contribute to the solution of societal problems which have rapidly emerged as a result of NLP technology (Bender et al., 2021).

Finally, problems can emerge due to conflicts between the agential virtues and other theoretical virtues (cf. Box 5). For example, researchers might use and develop a theory because it attracts public interest or has engineering applications, in exclusion of considering its other virtues and at the expense of the development of its competitors. While this can be beneficial in terms of communal facilitation or societal alignment, e.g., by bringing people and support into the field, it can also result in misinterpretation of the field's work that could potentially cause harm to the field or its theories.

Conclusion

What are the implications of a problem-centric view of theoretical virtues? The first is that theories are inexorably problem-laden. They are developed, selected, and maintained with an eye toward the problems they can solve, and their effect on a field's problem space. Like observation, which is laden with the theories involved in its collection and motivation (Boyd & Bogen, 2009), theories are in turn laden with the problems involved in their development and selection. They retain the traces of those problems in their structural components, their formulation, and the phenomena in their domain.

Thus, the expression of theoretical virtues and their relative importance are only defined with respect to a research community's problem space. A theory is empirically adequate for (a certain problem space) or simple for (a specific user). The implications of the no-free-lunch principle are that, because no single theory can cover all problems, the virtuous properties for which we develop and select theories depend on the problems we want a theory to cover—the problem space of a research community. This results in the fields looking for different properties of their theories to suit the needs of their unique problems. As a result, fields tend to develop qualitatively different theories, which reflect their different problem spaces. This is, trivially, stating that theories are different based on the phenomena in their domain—theories from different fields need to cover problems for which different phenomena are relevant. But, less trivially, theories are different based on what people want to do with those phenomena and

the problems that were of interest during their development. The cognitive sciences have different theories, even where their domain of phenomena overlap, due to non-overlapping problem spaces.

In conclusion, we have advanced a pragmatic, problem-centric account of theoretical virtues, which draws on the cognitive sciences and theory development in those fields. This approach may be particularly useful in the cognitive sciences because it takes into account the fact that its component fields each have a distinct set of problems in their domain, resulting in distinct needs for their theories. This required adopting an account of scientific problems which is centered on a research community and the development of its body of knowledge. Such an approach has a dual benefit: (1) it distances scientific work from societal problems (the aim is to build a body of knowledge, not directly solve external problems) while (2) keeping it grounded in their solution and thus in effective action (the societal purpose of this body of knowledge is to be useful for those problems). It also acknowledges that theories can be virtuous for reasons pertaining to the role they play in improving the problem-solving efficacy of the research community or its body of knowledge as a whole, in addition to traditionally epistemic virtues which pertain to the coverage provided by the theory itself. We hope that this perspective can be helpful to both practicing scientists, as they consider the theories they use and develop, as well as philosophers of science in their considerations of a societally embedded scientific system.

Box 6 Recommendations for researchers

Problem specification. While scientists are good at spelling out the situation, it may be helpful to more explicitly spell out the constraints on the problems they are proposing to solve in their papers. There is some truth to the adage that stating the problem actually is “half the solution,” as specifying a problem’s constraints restricts inquiry to those directions which can meet them and would allow other researchers to know if any disagreements are due to different assessment if the problem has, in fact, been solved or disagreement as to what the problem actually is (i.e., different constraints)

Explicitly use theoretical virtues to guide decisions during theory development. Which theory to pursue? Which virtues does a theory already achieve to an adequate degree? Which aspects of the theory need improvement, and what can be changed or added to improve it with respect to specific virtues? While many of these assessments are already being made implicitly in the course of research, their specification can help guide researchers’ decision-making process

Characterize your field’s problem space and use it as a guide for theory development. While countless reviews are written about previous results and their interpretation, we would advise also including more thorough characterization and development of the problem spaces around specific topics

Acknowledgements The authors would like to thank the administrative team at the Lorentz Center and the organizers of the workshop: “What Makes a Good Theory? Interdisciplinary Perspectives” for facilitating this manuscript, as well as Sashank Varma and Angela Potochnik for discussions during its early development. Colin Bredenberg for comments on an early draft of the manuscript. Berna Devezer and two anonymous reviewers for helpful and constructive comments for the development of the manuscript. DL is supported by the FRQNT Strategic Clusters Program 2020-RS4-265502—Centre UNIQUE—Union Neurosciences & Artificial Intelligence—Quebec) and the Richard and Edith Strauss Postdoctoral Fellowship in Medicine.

Author Contribution All authors contributed to early discussions formulating the manuscript. The first draft of the manuscript was written by DL and all authors commented on subsequent versions of the manuscript. All authors read and approved the final manuscript.

Funding DL is supported by the FRQNT Strategic Clusters Program 2020-RS4-265502—Centre UNIQUE—Union Neurosciences & Artificial Intelligence—Quebec) and the Richard and Edith Strauss Postdoctoral Fellowship in Medicine.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

- Adolfi, F. G., van de Braak, L., & Woensdregt, M. (2023). From empirical problem-solving to theoretical problem-finding perspectives on the cognitive sciences. *PsyArXiv*.
- Alon, U. (2009). How to choose a good scientific problem. *Molecular Cell*, 35, 726–728.
- Alvesson, M., & Sandberg, J. (2013). *Constructing research questions: Doing interesting research*. SAGE Publications.
- Anderson, P. W. (1972). More is different. *Science*, 177, 393–396.
- Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, 3, 266–286.
- Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, 22, 359–371.
- Baroni, M. (2022). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic structures in natural language* (pp. 1–16). CRC Press.
- Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines*, 4, 1–25.

- Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity*. MIT Press.
- Begley, G., & Ellis, L. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531–533.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). <https://doi.org/10.1145/3442188.3445922>
- Berntson, G. G., & Norman, G. J. (2021). Multilevel analysis: Integrating multiple levels of neurobehavioral systems. *Social Neuroscience*, *16*, 18–25.
- Beveridge, W. I. B. (1950). *The art of scientific investigation*. Blackburn Press.
- Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*, 100205.
- Brette, R. (2018). What is computational neuroscience? (XXXIV) Is the brain a computer (2). <http://romainbrette.fr/what-is-computational-neuroscience-xxxiv-is-the-brain-a-computer-2/>. Accessed 13 Sept 2024.
- Brigandt, I. (2010). Beyond reduction and pluralism: Toward an epistemology of explanatory integration in biology. *Erkenntnis*, *73*, 295–311.
- Brigandt, I. (2013). Explanation in biology: Reduction, pluralism, and explanatory aims. *Science and Education*, *22*, 69–91.
- Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*, 756–766.
- Boyd, N. M., Bogen, J. (2009). Theory and observation in science, In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 edn).
- Boyer, P. (1998). Cognitive tracks of cultural inheritance: How evolved intuitive ontology governs cultural transmission. *American Anthropologist*, *100*, 876–889.
- Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press.
- Buzsáki, G. (2020). The brain–cognitive behavior problem: A retrospective. *eNeuro*, *7*, ENEURO.0069–20.2020.
- Cartwright, N. (1983). *How the laws of physics lie*. Oxford Academic.
- Casadevall, A., & Fang, F. C. (2015). Field science—The nature and utility of scientific fields. *mBio*, *6*, e01259–15.
- Chang, H. (2004). *Inventing temperature*. Oxford University Press.
- Chang, H. (2017). VI—Operational coherence as the source of truth. *Proceedings of the Aristotelian Society*, *117*, 103–122.
- Chang, H. (2022). *Realism for realistic people: A new pragmatist philosophy of science*. Oxford University Press.
- Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science*, *242*(4879), 741–745.
- Churchland, P. S., & Sejnowski, T. J. (1994). *The computational brain*. MIT Press.
- Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, and Psychophysics*, *81*, 2265–2287.
- Colburn, T., & Shute, G. (2007). Abstraction in computer science. *Minds and Machines*, *17*, 169–184.
- Craver, C. (2002a). Structures of scientific theories. In P. Machamer, & M. Silberstein (Eds.), *Blackwell guide to the philosophy of science* (pp. 55–79). Blackwell.
- Craver, C. (2002b). Interlevel experiments and multilevel mechanisms in the neuroscience of memory. *Philosophy in Science*, *69*, S83–S97.
- Craver, C. (2007). *Explaining the brain*. Oxford University Press.
- Darden, L. (1978). Discoveries and the emergence of new fields in science. In *Proceedings of the biennial meeting of the philosophy of science association* (pp. 149–160). Cambridge University Press.
- Dennett, D. C. (1989). *The intentional stance*. MIT Press.
- DePasquale, B., Sussillo, D., Abbott, L. F., & Churchland, M. M. (2023). The centrality of population-level factors to network computation is demonstrated by a versatile approach for training spiking networks. *Neuron*, *111*(5), 631–649.
- Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS ONE*, *14*, e0216125.
- Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*, 200805.
- Dewey, J. (1938). *Logic*. Holt Publishers.
- Dockum, R., & Green, C. M. (2023). Toward a big tent linguistics: Inclusion and the myth of the lone genius. *PsyArXiv*.
- Doerig, A., et al. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*, 431–450.
- Doppelt, G. (1981). Laudan’s pragmatic alternative to positivist and historicist theories of science. *Inquiry*, *24*, 253–271.
- Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese*, *138*, 453–473.
- Douglas, H. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.
- Douglas, H. (2014a). The value of cognitive values. *Philosophy of Science*, *80*, 796–806.
- Douglas, H. (2014b). Pure science and the problem of progress. *Studies in History and Philosophy of Science Part A*, *46*, 55–63.
- Dupré, J. (1993). *The disorder of things: Metaphysical foundations of the disunity of science*. Harvard University Press.
- Elliott, K., & McKaughan, D. (2014). Non-epistemic values and the multiple goals of science. *Philosophy of Science*, *81*, 1–21.
- Elliott, S. (2021). Research problems. *British Journal for the Philosophy of Science*, *72*, 1013–1037.
- Ellner, S. P., & Guckenheimer, J. (2006). *Dynamic models in biology*. Princeton University Press.
- Eronen, M. I., & Romeijn, J.-W. (2020). Philosophy of science and the formalization of psychological theory. *Theory & Psychology*, *30*(6), 786–799.
- Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*, 779–788.
- Feyerabend, P. (1975). *Against method*. Verso.
- Fracchia, J., & Lewontin, R. C. (1999). Does culture evolve? *History and Theory*, *38*, 52–78.
- Frankel, H. (1980). Problem-solving, research traditions, and the development of scientific fields. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, (Vol. 1980, No. 1, pp. 29–40). Cambridge University Press.
- Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*, 271–288.
- Gelinas, J. N., Khodagholy, D., Thesen, T., Devinsky, O., & Buzsáki, G. (2016). Interictal epileptiform discharges induce hippocampal-cortical coupling in temporal lobe epilepsy. *Nature Medicine*, *22*, 641–648.
- Gerstner, W. (2014). *Neuronal dynamics*. Cambridge University Press.
- Getzels, J. W. (1979). Problem finding: A theoretical note. *Cognitive Science*, *3*, 167–171.
- Gilboa, A., & Moscovitch, M. (2021). No consolidation without representation: Correspondence between neural and psychological representations in recent and remote memory. *Neuron*, *109*, 2239–2255.
- Giner-Sorolla, R. (2019). From crisis of evidence to a “crisis” of relevance? Incentive-based answers for social psychology’s perennial relevance worries. *European Review of Social Psychology*, *30*, 1–38.

- Girardeau, G. J., Benchenane, K., Wiener, S. I., Buzsáki, G., & Zugaro, M. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience*, *12*(1222), 1223.
- Girardeau, G., & Lopes-dos-Santos, V. (2021). Brain neural patterns and the memory function of sleep. *Science*, *374*, 560–564.
- Goldrick, M. (2022). An impoverished epistemology holds back cognitive science research. *Cognitive Science*, *46*(9), e13199.
- Goldstein, J. A. et al. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. Preprint retrieved from <https://arxiv.org/abs/2301.04246>
- Godfrey-Smith, P. (2003). *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press.
- Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *205*, 581–598.
- Graf, T. (2022). Subregular linguistics: Bridging theoretical linguistics and formal grammar. *Theoretical Linguistics*, *48*, 145–184.
- Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*, 789–802.
- Haig, B. (1987). Scientific problems and the conduct of research. *Educational Philosophy and Theory*, *19*, 22–32.
- Hacking, I. (1983). *Representing and intervening*. Cambridge University Press.
- Harding, S. (1976). Introduction. *Can theories be refuted?: essays on the Duhem-Quine thesis*. Springer.
- Heijnen, S., Sleutels, J., & de Kleijn, R. (2024). Model virtues in computational cognitive neuroscience. *Journal of Cognitive Neuroscience*, *1–12*. https://doi.org/10.1162/jocn_a_02183
- Holstein, B. R. (1988). Semiclassical treatment of the double well. *American Journal of Physics*, *68*, 430.
- Holyoak, K. (1995). Problem-solving. In E. E. Smith & D. Osherson (Eds.), *Thinking: An invitation to cognitive science* (Vol. 3, pp. 267–296). MIT Press.
- Hull, L. D. (1990). *Science as a process: An evolutionary account of the social and conceptual development of science*. University of Chicago Press.
- James, W. (1907/1975). *Pragmatism: A new name for some old ways of thinking*. Harvard University Press.
- Jelic, V., & Marsiglio, F. (2012). The double-well potential in quantum mechanics: A simple, numerically exact formulation. *European Journal of Physics*, *33*, 1651.
- Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, *10*, 100–107.
- Jones, I. S., & Kording, K. P. (2021). Might a single neuron solve interesting machine learning problems through successive computations on its dendritic tree? *Neural Computation*, *33*, 1554–1571.
- Jones, M. (2005). Idealization and abstraction: A framework. *Poznań Studies in the Philosophy of the Sciences and the Humanities*, *86*(1), 173–218.
- Jorgenson, L. A., et al. (2015). The BRAIN initiative: Developing technology to catalyse neuroscience discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*, 20140164.
- Kauffman, S. A. (2002). *Investigations*. Oxford University Press.
- Keas, M. (2018). Systematizing the theoretical virtues. *Synthese*, *195*, 2761–2793.
- Keller, A. J., et al. (2020). A disinhibitory circuit for contextual modulation in primary visual cortex. *Neuron*, *108*, 1181–1193.e8.
- Khalifa, K. (2020). Understanding, truth, and epistemic goals. *Philosophy of Science*, *87*(5), 944–956.
- Kitamura, T., et al. (2017). Engrams and circuits crucial for systems consolidation of a memory. *Science*, *356*(73), 78.
- Kitcher, P. (2013). Toward a pragmatist philosophy of science. *Theoria*, *28*(2), 185–231.
- Klinzing, J. G., Niethard, N., & Born, J. (2019). Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*, *22*, 1598–1610.
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
- Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. *Selected studies in scientific tradition and change* (pp. 320–339). University of Chicago Press.
- Kuhn, T. S. (1983). Rationality and theory choice. *The Journal of Philosophy*, *80*, 563.
- Lakatos, I. (1970). *The methodology of scientific research programmes*. Cambridge University Press.
- Lashley, K. S. (1950). In search of the engram. In Society for experimental biology. *Physiological mechanisms in animal behavior. (Society's Symposium IV.)* 454–482.
- Laudan, L. (1977). *Progress and its problems*. University of California Press.
- Laudan, L. (1984). *Science and values*. University of California Press.
- Laudan, L. (1990). *Science and relativism: Some key controversies in the philosophy of science*. University of Chicago Press.
- Lee, Y. F., Gerashchenko, D., Timofeev, I., Bacskai, B. J., & Kastanenka, K. V. (2020). Slow wave sleep is a promising intervention target for Alzheimer's disease. *Frontiers in Neuroscience*, *14*, 705.
- Levenstein, D., et al. (2023). On the role of theory and modeling in neuroscience. *Journal of Neuroscience*, *43*, 1074–1088.
- Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.
- List, C., & Philip, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.
- Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.
- Love, A. C. (2008). Explaining evolutionary innovations and novelties: Criteria of explanatory adequacy and epistemological prerequisites. *Philosophy in Science*, *75*, 874–886.
- Love, A. C. (2012). Hierarchy, causation and explanation: Ubiquity, locality and pluralism. *Interface Focus*, *2*, 115–125.
- Love, A. C. (2013). Theory is as theory does: Scientific practice and theory structure in biology. *Biological Theory*, *7*, 325–337.
- Lutz, S. (2017). What was the syntax-semantics debate in philosophy of science about? *Philosophy and Phenomenological Research*, *95*(2), 319–352.
- Machta, B. B., Chachra, R., Transtrum, M. K., & Sethna, J. P. (2013). Parameter space compression underlies emergent theories and predictive models. *Science*, *342*, 604–607.
- Mackonis, A. (2013). Inference to the best explanation, coherence, and other explanatory virtues. *Synthese*, *190*(6), 975–995.
- Maingret, N., Girardeau, G. J., Todorova, R., Goutierre, M., & Zugaro, M. (2016). Hippocampo-cortical coupling mediates memory consolidation during sleep. *Nature Neuroscience*, *19*(959), 964.
- Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.
- Marcus, G. (2022). Nonsense on stilts. *Substack*. <https://garymarcus.substack.com/p/nonsense-on-stilts>. Accessed 13 Sept 2024.
- Marder, E., & Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, *7*, 563–574.
- Marr, D. & Poggio, T. (1976). From understanding computation to understanding neural circuitry. *Massachusetts Institute of Technology*
- Marr, D. (1982). *Vision*. MIT Press.

- Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*, 170, 169–190.
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419–457.
- McMullin, E. (2009). The virtues of a good theory. In M. Curd & S. Psillos (Eds.), *Routledge companion to philosophy of science* (pp. 498–508). Routledge.
- Mensch, B., & Kording, K. (2017). Ten simple rules for structuring papers. *PLoS Computational Biology*, 13, e1005619.
- Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, 7, 141–144.
- Min, B. et al., (2021). Recent advances in natural language processing via large pre-trained language models: A survey. Preprint retrieved from <https://arxiv.org/abs/2111.01243>
- Misak, C. (2007). *New pragmatists*. Oxford University Press.
- Mitchell, S. D. (2003). *Biological complexity and integrative pluralism*. Cambridge University Press.
- Mitchell, K. J. (2023). *Free agents: How evolution gave us free will*. Princeton University Press.
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *PNAS*, 120(13), e2215907120.
- Mizrahi, M. (2022). Theoretical virtues in scientific practice: An empirical study. *British Journal for the Philosophy of Science*, 73(4), 879–902.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518, 529–533.
- Monte, A., & Libby, A. M. (2018). Introduction to the specific aims page of a grant proposal. *Academic Emergency Medicine*, 25, 1042–1047.
- Muldoon, R. (2013). Diversity and the division of cognitive labor: Diversity and the division of cognitive labor. *Philosophy Compass*, 8, 117–125.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3, 221–229.
- Nadel, L., Hupbach, A., Gomez, R., & Newman-Smith, K. (2012). Memory formation, consolidation and transformation. *Neuroscience and Biobehavioral Reviews*, 36, 1640–1645.
- Newell, A., & Simon, H. (1972). *Human problem-solving*. Prentice Hall.
- Nguyen, C. T. (2020). *Games: Agency as art*. Oxford University Press.
- Nickles, T. (1978). Scientific problems and constraints. In P. Asquith & I. Hacking (eds.), *Proceedings of the PSA* (134–148). Philosophy of Science Association.
- Nickles, T. (1980). Scientific problems: Three empiricist models. In R. Giere & P. Asquith (eds.), *Proceedings of the PSA* (3–19). Philosophy of Science Association.
- Nickles, T. (1981). What is a problem that we might solve it? *Synthese*, 41(1), 85–118.
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719–748.
- Nurse, P. (2021). Biology must generate ideas as well as data. *Nature*, 597, 305–305.
- Oude Maatman, F. (2021). Psychology's theory crisis, and why formal modelling cannot solve it. *PsyArXiv*. <https://doi.org/10.31234/osf.io/puqvs>
- O'Leary, T., Sutton, A. C., & Marder, E. (2015). Computational models in the age of large datasets. *Current Opinion in Neurobiology*, 32, 87–94.
- O'Malley, M. A., et al. (2014). Multilevel research strategies and biological systems. *Philosophy in Science*, 81, 811–828.
- Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis. *Minnesota Studies in the Philosophy of Science*, 2, 3–36.
- Păslaru, V. (2023). New textbooks for teaching philosophy of science. *Philosophy of Science*, 90(1), 200–208.
- Pearl, J. & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic books.
- Peirce, C. S. (1878). How to make our ideas clear. *Popular Science Monthly*, 12, 2860302.
- Plaxco, K. W. (2010). The art of writing science. *Protein Science*, 19, 2261–2266.
- Polanyi, M. (1966). *The tacit dimension*. University of Chicago Press.
- Poepfel, D. & Adolphi, F. (2020). Against the epistemological primacy of the hardware: The brain from inside out, turned upside down. *eNeuro*, 7, ENEURO.0215–20.2020.
- Popper, K. (1959). *The logic of scientific discovery*. Routledge.
- Porello, D., Bottazzi, E., & Ferrario, R. (2014). The ontology of group agency. In P. Garbacz, & O. Kutz (Eds.), *Proceedings: Formal ontology in information systems* (pp. 183–196).
- Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.
- Potochnik, A. (2021). Our world isn't organized into levels. In D. S. Brooks, J. DiFrisco & W. C. Wimsatt (Eds.), *Levels of organization in biology*. MIT Press.
- Psillos, S. (1999). *Scientific realism: How science tracks truth*. Routledge.
- Quarantotto, D. (2020). Aristotle on science as problem-solving. *Topoi*, 39, 857–868.
- Rawski, J., & Baumont, L. (2023). Modern language models refute nothing. *Lingbuzz Preprint*.
- Rawski, J., & Heinz, J. (2019). No free lunch in linguistics or machine learning: Response to pater. *Language*, 95(1), e125–e135.
- Reisch, G. A. (1998). Pluralism, logical empiricism, and the problem of pseudoscience. *Philosophy in Science*, 65, 333–348.
- Reitman, W. (1964). Heuristic decision procedures, open constraints, and the structure of ill-defined problems. In M. Shelly & G. Bryan (Eds.), *Human judgments and optimality* (pp. 282–315). John Wiley.
- Richards, B. A. (2018) Yes, the brain is a computer. Medium. <https://medium.com/the-spike/yes-the-brain-is-a-computer-11f630cad736>. Accessed 13 Sept 2024.
- Richards, B. A., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22, 1761–1770.
- Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, 16, 725–743.
- Roeckelein, J. (1997). Psychology among the sciences: Comparisons of numbers of theories and laws cited in textbooks. *Psychological Reports*, 80, 131–141.
- Rohrlich, F., & Hardin, L. (1983). Established theories. *Philosophy in Science*, 50, 603–617.
- Roll-Hansen, N. (2017). A historical perspective on the distinction between basic and applied science. *Journal for General Philosophy of Science*, 48(4), 535–551.

- Rosales, A., & Morton, A. (2021). Scientific explanation and trade-offs between explanatory virtues. *Foundations of Science*, 26, 1075–1087.
- Roxin, A., & Fusi, S. (2013). Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS Computational Biology*, 9, e1003146.
- Rumelhart, D. E., McClelland, J. L., & The PDP research group. (1987). *Parallel Distributed Processing*. MIT Press.
- Sadeh, S., & Clopath, C. (2021). Inhibitory stabilization and cortical computation. *Nature Reviews Neuroscience*, 22, 21–37.
- Schickore, J. (2020). Mess in science and wicked problems. *Perspectives on Science*, 28, 482–504.
- Schindler, S. (2018). *Theoretical virtues in science: Uncovering reality through theory*. Cambridge University Press.
- Schindler, S. (2022). Theoretical virtues: Do scientists think what philosophers think they ought to think? *Philosophy in Science*, 89, 542–564.
- Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, 20, 11.
- Seth, A. (2016). The real problem. *Aeon*.
- Shanahan, M. (2022). Talking about large language models. Preprint retrieved from <https://arxiv.org/abs/2212.03551>
- Shapere, D. (1969). Notes toward a post-positivistic interpretation of science. In P. Achinstein & S. Barker (Eds.), *The legacy of logical positivism* (pp. 115–160). Johns Hopkins University Press.
- Shrader, D. (1980). The evolutionary development of science. *Review of Metaphysics*, 34(2), 273–296.
- Silver, B. L. (2000). *The ascent of science*. Oxford University Press.
- Simon, H. (1973). The structure of ill-structured problems. *Artificial Intelligence*, 4, 181–201.
- Sobel, C. P., & Li, P. (2013). *The cognitive sciences: An interdisciplinary approach*. SAGE Publications.
- Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99(2), 195–231.
- Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinion in Neurobiology*, 5, 169–177.
- Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, 144, 1325–1346.
- Steedman, M. (2008). On becoming a discipline. *Computational Linguistics*, 34, 137–144.
- Sterkenburg, T. F. (2016). Solomonoff prediction and Occam's razor. *Philosophy of Science*, 83(4), 459–479.
- Strevens, M. (2020). *The knowledge machine: How irrationality created modern science*. Liveright.
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Preprint retrieved from <https://arxiv.org/abs/1906.02243>
- Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. *Nature Neuroscience*, 1–11.
- Suppe, F. (1977). *The structure of scientific theories*. University of Illinois Press.
- Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, 100, 147.
- Tollefsen, D. (2015). *Groups as agents*. Polity.
- Tononi, G., & Cirelli, C. (2014). Sleep and the price of plasticity: From synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81, 12–34.
- Tomalin, M. (2006). *Linguistics and the formal sciences: The origins of generative grammar* (vol. 110). Cambridge University Press.
- Valvoda, J., Saphra, N., Rawski, J., Williams, A., & Cotterell, R. (2022). Benchmarking compositionality with formal languages. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6007–6018).
- van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
- van Rooij I. (2019). Psychological science needs theory development before preregistration. *Psychonomic Society*.
- van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, 16, 682–697.
- Veres, C. (2021). Large language models are not models of natural language: they are corpus models. Preprint retrieved from <https://arxiv.org/abs/2112.07055>
- Wei, J. et al. (2022). Emergent abilities of large language models. Preprint retrieved from <https://arxiv.org/abs/2206.07682>
- Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy in Science*, 76, 225–252.
- Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.
- Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265, 676–679.
- Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12(1), 24.
- Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings*. Harvard University Press.
- Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia*, 48(8), 2339–2356.
- Winther, R. G. (2021). The structure of scientific theories. In N. E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.).
- Wojtowicz, Z., & DeDeo, S. (2020). From probability to consilience: How explanatory values implement Bayesian reasoning. *Trends in Cognitive Sciences*, 24, 981–993.
- Wolpert, D. H. (2023). The implications of the no-free-lunch theorems for meta-induction. *Journal for General Philosophy of Science*, 54, 421–432.
- Wouters, A. G. (2003). Four notions of biological function. *Studies in History and Philosophy of Science Part C*, 34, 633–668.
- Yaghmaie, A. (2017). How to characterise pure and applied science. *International Studies in the Philosophy of Science*, 31, 133–149.
- Yang, G. R., & Wang, X.-J. (2020). Artificial neural networks for neuroscientists: A primer. *Neuron*, 107, 1048–1070.
- Zhen, Z.-H., et al. (2021). Normal and abnormal sharp wave ripples in the hippocampal-entorhinal cortex system: Implications for memory consolidation, Alzheimer's disease, and temporal lobe epilepsy. *Frontiers in Aging Neuroscience*, 13, 683483.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.