Richer than reduction

David Danks

Departments of Philosophy & Psychology

Carnegie Mellon University

Abstract

There are numerous routes for scientific discovery, many of which involve the use of information from other scientific theories. In particular, searching for possible reductions is widely recognized as one guiding principle for scientific discovery or innovation. However, reduction is only one kind of intertheoretic relation; scientific theories, claims, and proposals can be related in more, and more complex, ways. This chapter proposes that much scientific discovery proceeds through the use of *constraints* implied by those intertheoretic relationships. The resulting framework is significantly more general than the common reduction-centric focus. As a result, it can explain more prosaic, everyday cases of scientific discovery, as well as scientists' opportunistic use of many different kinds of scientific information. I illustrate the framework using three case studies from cognitive science, and conclude by exploring the potential limits of analyses of scientific discovery via constraints.

1. Routes to discovery

The diverse paths and techniques for scientific discovery, invention, and construction form perhaps the most heterogeneous part of science. There are many ways and methods, whether structured or intuitive, to develop a novel scientific theory or concept. In fact, people have sometimes thought that scientific discovery does not—perhaps, could not—exhibit any systematic

patterns at all. While this latter pessimism is arguably unwarranted, the skeptics are correct that there is great diversity in routes to scientific discovery. At one extreme, a relatively minimal type of discovery occurs when the scientist starts with an existing theory, and then adjusts its parameters in light of new data. For example, a novel experiment might reveal the importance of a previously unconsidered causal factor. A more speculative type of scientific discovery depends on analogical reasoning, as that can lead the scientist to consider entirely new classes or types of theories. Alternately, various abductive or inductive strategies can point towards scientific theories, models, or concepts that have not previously been considered. And of course, there might be no explicit or conscious "method" at all in a case of scientific discovery; it might, from the perspective of the scientist herself, be the result of unexplainable inspiration.

This paper explores a particular set of methods for scientific discovery—those that use constraints from other scientific theories. Scientific innovation and discovery is often driven by consideration of *other* (folk and scientific) theories and models, where the resulting constraints can be both structural and substantive. Past discussions of this constraint-based scientific discovery have almost always centered on reductionism or reductionist commitments as a discovery strategy (Bechtel & Richardson, 2000; Schouten & de Jong, 2012; Wimsatt, 1980). More specifically, there are two principal ways to use reductionism as a method for scientific discovery and innovation. First, suppose that one has a theory T_H that captures the higher-level (in some relevant sense) phenomena or structure. Reductionism, as an overarching meta-scientific commitment, implies that there must be some lower-level theory T_L—in fact, potentially many such theories if there are many lower levels—such that T_H reduces to T_L. (For the moment, I leave aside the question of the meaning of 'reduces to'.) Scientific discovery at the T_L-level can thus be guided by our knowledge of T_H: the higher-level theory can provide substantial information about features of T_L (e.g., the relevant inputs and outputs), and thereby significantly reduce the possibility space. For example, the search for

underlying causal mechanisms is frequently guided in exactly this way by a higher-level theory about the structure of the system or the functional roles of various components (e.g., Darden, 2002; Darden & Craver, 2002). Of course, this use of reductionism does not eliminate the need for discovery; even though T_H might reduce the space of possible T_L's, it will rarely uniquely determine one particular T_L. Thus, we will still need to use one or more strategies from the previous paragraph, such as adjustment in light of novel data. Nonetheless, we can use reductionist commitments as a substantive "downward guide" to scientific discovery, and thereby greatly simplify the task.

A second way to use reductionism as a discovery strategy starts with a lower-level T_L that specifies particular components of the system (perhaps mechanisms in a strong sense, perhaps not). We can then seek to discover a T_H that captures the functional roles or higher-level regularities and relations of the system, and that reduces to T_L. For example, we might have a robust scientific theory about some set of regulatory mechanisms within a cell, and then aim to find a higher-level theory that captures the patterns that result from interactions of these mechanisms in particular environments. More generally, T_H will typically incorporate elements of T_L as particular realizations or implementation specifications of the T_H-components. This lower-level information significantly constrains the possible functional, computational, or causal roles for elements of T_H, precisely because we require that T_H reduce to T_L. Although T_L might sometimes uniquely determine T_H at some levels (e.g., if T_H is the asymptotic behavior of dynamical model T_L), the discovery situation will typically be more complex: the proper T_H may depend on our explanatory goals, or specific initial conditions, or aspects of the background context. This second use of reductionism and reductionist commitments does not obviate the need for scientific discovery, but nonetheless provides guiding "upward constraints" that can significantly speed or improve that discovery.

Regardless of which strategy we pursue, the exact constraints will depend on both the details of the scientific case, and also the particular account of 'reduction' that one employs. For example,

syntactic theories of 'reduction' (e.g., Dizadji-Bahmani, Frigg, & Hartmann, 2010; Nagel, 1961) will emphasize discovery through manipulations of the symbolic representations of the theories. In contrast, causal theories of 'reduction' (e.g., Churchland, 1985; Hooker, 1981a, 1981b) will focus on discovery of similar causal roles or capacities across the theories. However, *all* theories of 'reduction' agree that the relevant relation involves a very close connection between the two theories. Thus, scientific discovery via reductionism inevitably results in a new scientific theory that is tightly coupled with the pre-existing theory—either discovery of a T_H that reduces to the existing T_L, or discovery of a T_L to which the existing T_H can reduce. This tight connection between old and new theories provides much of the power of reductionism as a discovery strategy (when it is successful). For a given T_H, there will often be a relatively small class of lower-level realizations or implementations that actually exhibit the precise higher-level phenomena. For a given T_L, information about the relevant initial or background conditions will often almost determine the higher-level T_H. And we gain enormous further benefits if we can discover a suitable $<T_H, T_L>$ pair, as we can use each to refine the other, combine them into integrated multi-level models, and thereby establish cross-level, cross-theory, and cross-disciplinary connections.

However, although there can be significant benefits from requiring a reductionist connection between T_H and T_L (regardless of direction of discovery), such a connection comes with a significant cost: the required tight couplings are usually very difficult to establish. First, all extant theories of 'reduction' require that both T_H and T_L be full scientific theories, even though scientists frequently work with vaguer or more uncertain not-quite-theories (e.g., observation of a correlation between two factors, or knowledge that some manipulation produces a probabilistic change in a target variable). Second, reductionist discovery must involve levels that are an appropriate distance from one another, as reductions are very hard to establish across large "level gaps." Third, the requirements for a full reduction are often quite stringent, and so we might not be able to establish

the appropriate connections between T_H and T_L (though searching for those connections could potentially be useful for discovery purposes). Fourth, for a given T_L, there might simply not be an appropriate T_H at our desired level, as we might not be able to abstract away or modularize the implementation details in T_L. Fifth, for a given T_H, the relevant distinctions or objects might not be respected in T_L (e.g., cognitive symbols might not be directly found in neural models), and so T_H could actually be a misleading guide for scientific discovery.

Reductionism and reductionist commitments are very powerful guides for scientific discovery, but also very limited. If we look more broadly, we can find many cases in which information from other scientific theories has been used for scientific discovery, but where those uses simply cannot be understood in terms of the search for reductions. Reduction is, however, only one intertheoretic relation of many, and so we might suspect that scientific discovery via reductionist commitments is only one way to employ information other scientific theories. Perhaps we can have a more general, more useful model of scientific discover by considering alternative intertheoretic relations. This chapter aims to provide such an account via the use of intertheoretic constraints generated by those relations; reductive constraints are simply one special case. To that end, Section 2 provides a more general account of the notion of 'intertheoretic constraint', with a particular eye towards their use in discovery. Section 3 then uses that account to explicate several cases of scientific discovery in the cognitive sciences. Those case studies might well seem banal and ordinary, but that is part of the point: everyday scientific discovery is largely a matter of trying to fit together disparate puzzle pieces, and scientists employ many different constraints and relations—not just reduction—to find the next piece of the puzzle.

2. Discovery via constraints

There are many different intertheoretic relations, involving multiple theoretical virtues. Reduction is one salient relation, and it holds when there is a tight coupling—syntactic, semantic, causal, functional, or other—between two theories. Autonomy is a different intertheoretic relation that obtains when features of theory T_A are essentially independent of T_B. For example, macroeconomics is thought to be explanatorily autonomous from quantum mechanics. More controversially, psychology has been claimed to be ontologically autonomous from neuroscience (Fodor, 1974, 1997). These two relations of reduction and autonomy clearly fall at the extremes; theories can be related to one another in more subtle and fine-grained ways, as we will see in Section 3. Importantly, these intertheoretic relations imply intertheoretic constraints (thlough perhaps an empty set of constraints, as in the case of autonomy). For example, if T_H reduces to T_L, then if T_L is true, then T_H must also be true.[1] Moreover, this constraint (or its contrapositive: given a reduction, if T_H is false, then T_L must be false) does much of the work when reductionism is used as a guide for scientific discovery, which suggests that perhaps much scientific discovery proceeds through the use of intertheoretic constraints of all sorts, not just those grounded in reductions.

A general account of intertheoretic constraints should include reduction and autonomy as special case intertheoretic relations, but should also apply more generally, though that requires some complications.[2] At its most abstract, a scientific theory S (or model, or claim, or…) *constrains* another theory T relative to some theoretical virtue V just when the extent to which S has V is relevant in some way to the extent to which T has V. That is, an intertheoretic constraint exists between S and T if S's status with respect to V (e.g., truth, simplicity, predictive accuracy, explanatory power, etc.) matters in some way for T's status with respect to the same V. For example, the existence of a reduction relation between T_H and T_L yields (at least, on most accounts of 'reduction') the constraint

[1] Readers who are skeptical about notions of 'truth' with regards to scientific theories should instead substitute 'accurate' or 'approximately true' or whatever notion they prefer.
[2] For space reasons, I only summarize my account of intertheoretic constraints here. More details and discussion can be found in chapter 2 of Danks, 2014, or Danks, 2013.

that T_L's truth implies T_H's truth. That is, the truth of T_L is directly relevant to whether T_H is true. Crucially, though, a reduction implies this tight constraint only for some theoretical virtues (e.g., truth). The existence of a reduction relation does not, for example, necessarily imply any constraint with respect to explanatory power, as T_H could reduce to T_L but provide explanations with different scope and generalizability. Moreover, although I have been using the word 'theory' in this paragraph, this account of 'constraint' does not actually require *S* and *T* to be full-blown theories. Relevance can arise between scientific claims, data descriptions, partially specified models, and other kinds of not-quite-theories, and thus constraints based in those intertheoretic relevance relations can obtain between them. Of course, the specific relation underlying particular constraints could have more stringent requirements of the relata (e.g., a reduction requires theories), but that is not intrinsic to intertheoretic constraints more generally.

This high-level characterization of 'intertheoretic constraint' is qualitative and vague in certain key respects (e.g., what does it mean for *S*'s theoretical virtues to be "relevant" to *T*'s virtues?), but is already sufficiently precise to highlight some notable features (see also Danks, 2014). Perhaps most importantly, this account implies that constraints are objective, not subjective: the constraint obtains if *S* is actually relevant for *T*, regardless of whether any scientists realize that it is relevant. In fact, a common scientific activity is the discovery of novel intertheoretic relations and constraints that were previously unknown (but were present all along). A less-obvious implication is that constraints are, on this account, comparative in both relata: whether *S* constrains *T* with respect to *V* depends not only on *S* and *T* themselves (and their relations), but also on the alternatives to *S* and *T*. This property of constraints might be surprising, but follows immediately from the focus on relevance, as whether one theory or model is relevant to another will depend on what we take to be the serious alternatives. For example, suppose that *T* is a particular associationist model of human (psychological) causal learning that uses prediction errors in learning (e.g., Rescorla & Wagner, 1972),

and *S* is neural evidence that prediction errors are computed in the brain. Is *S* relevant for whether *T* is the correct theory? If the only alternatives to *T* are other models that use prediction errors (e.g., other standard associationist models, such as Pearce, 1987), then the answer is "no," as *S* does not provide information that distinguishes between them. However, if the alternatives to *T* include models that do not directly employ prediction errors (e.g., more rationalist models, such as Griffiths & Tenenbaum, 2005), then the answer is potentially "yes," as *S* might rule out (or make less plausible) some of these alternatives to *T*. More generally, relevance (of all different types) can depend on what else might have been the case, and so the alternatives to *S* and *T* matter.[3]

The use of these intertheoretic constraints in scientific discovery is relatively direct and immediate. Suppose that I am trying to discover a new scientific theory, model, or other account of phenomena *P* (in domain *D* and at level *L*) for purposes or goals *G*. For this discovery task, the first step is to list possibly-relevant theories and models S_1, …, S_n (and their corresponding sets of competitors $\mathbf{S}_1$, …, $\mathbf{S}_n$). These *S*'s are my scientific beliefs and knowledge that could perhaps imply constraints that are relevant to our theory of *P* (at level *L* for goal *G*). They might be about other phenomena in *D*, or characterized at a different level, or offered to fulfill a different function, but still potentially related. I almost presumably have some ideas about what kind of theory or model is desired, even if only a vague sense. That is, we can assume that I have some set **T** of possible "targets," where **T** will frequently be infinite, or involve a number of unknown parameters, or otherwise be very broad.

Given these components, the use of intertheoretic constraints for scientific discovery is straightforward, at least in the abstract: (a) for each $S_i/\mathbf{S}_i$, we determine the *G*-constraints (i.e., the constraints that are relevant for the scientific goal) that they imply for **T**; (b) aggregate the *G*-

[3] As an aside, notice that this alternative-dependence implies that the particular constraints that scientists entertain can depend on contingent historical facts about the science (that influence the set of alternatives considered), even though the existence and nature of those constraints are not history-dependent.

constraints together, perhaps deriving further implied constraints; and (c) compute the resulting impacts on **T** (e.g., ruling out certain possibilities, or making others more likely). In some special cases, this process will result in only one T_j at the end, in which case the constraints were fully sufficient for our discovery problem. More typically, this process will reduce the possibility space, but not fully determine *T*. We can then look for additional *S*'s (since it will not always be obvious *a priori* which scientific beliefs are potentially relevant), or try to discover additional constraints implied by the current *S*'s (since we cannot use a constraint if we do not know about it), or turn to one of the other types of discovery strategy outlined at the beginning of this chapter (e.g., collecting novel data to further refine or specify the scientific theory or claim).

Scientific discovery via reductionism can easily be understood in terms of this schema. Consider the "bottom-up" strategy in which we know T_L and are trying to discover T_H. In this case, T_L largely sets the domain and phenomena, and other factors (perhaps extra-scientific) determine the level and set of possible target T_H's. This discovery problem is truth-centric,[4] and so we are concerned with truth-constraints: given that T_L is true, how does this constrain the possible truth of elements of $\mathbf{T_H}$? The notion of a truth-constraint is significantly more complicated than one might initially suspect (see Danks, 2014 for details), but it is relatively simple if we require that the target T_H be reducible to T_L: any candidate T_H that is inconsistent with T_L in the relevant domain can be eliminated. That is, we get exactly the constraint that is used by reductionists in scientific discovery. And a similar analysis can be given for "top-down" reductionist discovery in which we know T_H and are trying to discover T_L. Thus, the much-discussed reductionist strategies are simply special cases of this more general account of the use of intertheoretic constraints for scientific discovery.

This picture of "scientific discovery via intertheoretic constraints" is similar to, but (at least) generalizes and extends, the constraint-inclusion model of scientific problem-solving (Nickles, 1978,

[4] Again, readers should freely substitute their preferred term for 'truth', such as 'accuracy' or 'approximate truth'.

1981). The constraint-inclusion framework for scientific discovery (and confirmation) contends that scientific problems, not theories, are the relevant units of inquiry, and that the goal of inquiry is a satisfactory answer, not truth (Laudan, 1981; Nickles, 1988). "Constraints" then provide critical information about what would count as a satisfactory answer to a problem: any such answer must satisfy the relevant constraints (Nickles 1978). Scientific problems are not *defined* by constraints, but they form a major part of the characterization of problems, and provide one way to gain understanding about the structure of a scientific problem. As Nickles (1981) puts it: "The more constraints on the problem solution we know, and the more sharply they are formulated, the more sharply and completely we can formulate the problem, and the better we understand it." (p. 88)

There are several shared features of the constraint-inclusion model and the framework proposed in this section: (a) problems, questions, and goals are central, not simple truth; (b) constraints play an important role in the scientific discovery process; and (c) the existence of constraints does not depend on contingent scientific history or human psychology, though our awareness of them might depend on these factors. At the same time, though, we employ somewhat different understandings of 'constraint'. Most notably, the constraint-inclusion model focuses on relatively "hard" or quasi-logical constraints, where these are derived for a particular problem. For example, a "reductive" constraint *C* on problem solutions specifies that the solution (whatever it may be) must be re-representable as specified by *C* (Nickles, 1978); this type of constraint thus focuses on the mathematical relations between syntactically characterized scientific theories. Of course, there are many kinds of constraints in the constraint-inclusion model, but in general, "every single constraint, by definition of 'constraint', rules out *some* conceivable solution as inadmissible." (Nickles, 1981, p. 109; emphasis in original) In contrast, my constraints need only influence plausibility without definitively ruling anything in or out. A constraint can be useful even if nothing is *inadmissible* as a result of it. In addition, my account does not identify constraints with problems, but rather provides

an account of how they arise from intertheoretic relations. The "discovery via constraints" model proposed here thus generalizes the constraint-inclusion model by allowing for "soft" constraints, and also provides an account of the source of problem-specific constraints in terms of the potentially relevant theories (and their intertheoretic relations).

Of course, one might object that my account is too high-level and abstract to be useful, precisely because it attempts to cover a wide range of cases and constraints. In general, there is only a limited amount that can be said if we restrict ourselves to talking in terms of letters—*D*'s, *G*'s, *S*'s, and so forth—rather than specific domains, phenomena, and discovery problems. For example, we need to know the relevant goal(s), as the very same **S** might truth-constrain **T**, but not explanation-constrain **T**. Thus, the trajectory of scientific discovery for one goal can be quite different than for another goal.[5] The details make a critical difference, and it is hard to evaluate this account without considering its applicability to particular cases of scientific discovery, and so we now examine some particular instances of scientific discovery.

3. Case studies of constraint-driven discovery

This section considers three case studies from cognitive science, each of which shows an instance of constraint-based scientific discovery, and that collectively show how scientific discovery can be an iterative process in which the outputs of one episode can be the inputs or constraints of the next. Although all three examples are drawn from cognitive science, I suggest that the lessons apply across many scientific disciplines. I focus on these examples only because I know them best, not because there is anything special or distinctive about them (at least, with respect to the use of intertheoretic constraints for discovery). In fact, as noted earlier, these case studies should hopefully

[5] This goal-dependence does not necessarily imply some sort of goal-dependent pragmatism or perspectivism (though I do also endorse that; see, e.g., Danks, 2015). Rather, this dependence is just a generalization of the old reductionist observation that two theories could stand in a reduction relation without thereby constraining one another's explanations in any interesting or informative way (e.g., Putnam, 1975).

seem somewhat anodyne, as one claim of this chapter is that "discovery via constraints" is a completely normal and regular scientific activity.

3.1. Representations of causal knowledge

People have a great deal of causal knowledge about the world: we know which switches cause the lights to turn on; we know ways to alleviate pain; we might understand the causes of the functioning of a car engine; and so on. Causal knowledge is arguably one of the key guides throughout our cognition (Sloman, 2005), and the first case study focuses on this phenomenon *P* of causal knowledge, within the domain *D* and level *L* of cognitive psychology/science. In particular, consider the discovery problem of finding a theory (or not-quite-theory) *T* that describes the structure of these cognitive representations. There are many different plausible candidate theories, as our causal knowledge might be structured as: lists of pairwise associations (e.g., Shanks, 1995); stimulus → response or environment → action mappings (e.g., Timberlake, 2001); causal graphical models (e.g., Danks, 2014; Griffiths & Tenenbaum, 2005); or in some other way.[6]

The first step in scientific discovery via constraints is to provide the different $S_i/\mathbf{S_i}$—that is, the potentially relevant scientific claims or theories, as well as their relevant, plausible alternatives. For the particular phenomenon of causal knowledge, there are an enormous number of potentially relevant scientific claims; for simplicity, we consider only two. First, there is substantial empirical evidence that people understand (perhaps implicitly) many of their actions as having a relatively uncaused (i.e., self-generated) component (Hagmayer & Sloman, 2009), and this emerges at a very young age (Rovee & Rovee, 1969). This understanding is arguably part of the reason that we have experiences of free will: we see our actions as not caused solely by the environment around us, but rather attribute some of the causation to ourselves. There are many alternatives to this claim S_1 (i.e.,

[6] As a matter of historical interest, these were the three main types of theories of causal structure representation being proposed in cognitive science in the early 2000's.

other members of $\mathbf{S}_1$); for example, people might understand their choices as entirely determined by environmental conditions, or by their own prior cognitive or emotional state.

A second, potentially relevant scientific claim S_2 is that people's decisions are appropriately responsive to indirect information about the state of the world. Obviously, we adjust our decisions so that they are appropriately tuned to the world. The relevant feature of human decision-making here is that we can use information that is not immediately relevant in order to make inferences about those factors that are directly relevant. For example, I might not bother to flip a light switch if I see my neighbor's lights are off, as I might thereby infer that the power is out in my neighborhood. That is, we employ disparate pieces of information to shape our decisions in order to maximize our chances of achieving a desired outcome. Alternative scientific possibilities to this S_2 are that people's decisions might depend only on local or immediate factors, or even be truly random in important ways.

Now consider the scientific discovery problem of the nature of our cognitive representations of causal structure. If S_1 holds (rather than some other possibility in $\mathbf{S}_1$), then our representations must enable us to derive predictions of the outcomes of exogenous actions. In particular, the representations should honor the basic asymmetry of intervention for causal relations (Hausman, 1998): exogenous actions to change the cause C (probabilistically) change the effect E, but exogenous changes in E do not lead to changes in C. Thus, our cognitive representations cannot be composed solely of lists of associations, as those are symmetric in nature. On the other side, if S_2 holds, then our representations of causal structure must be relatively integrated or unified, since we can use disparate pieces of information to shape or constrain our choices. Thus, they cannot consist solely in environment → action mappings, as those are not "tunable" in the appropriate way.[7] If we

[7] One might object that they could be tunable, if we understood "environment" in an appropriately broad and rich way. The problem is that this move then makes the mappings essentially unlearnable, as every experience now involves a unique, never-before-seen environment.

think back to our original set **T** of possibilities, we find that only causal graphical models can satisfy the constraints implied by both S_1 and S_2. And as a matter of historical fact, causal graphical models are currently the dominant theory of cognitive representations of causal structure knowledge, in large part because they are the only representations that can explain diverse reasoning, inference, and decision-making abilities such as S_1 and S_2 (Danks, 2014).

In this case study, scientific discovery occurred partly through understanding how our prior scientific commitments and beliefs constrained the possibilities. We did not need to perform a new experiment, or engage in analogical reasoning. More importantly for this chapter, the process looks nothing like "discovery via reduction." There is simply no possibility of a reduction relation between "causal representations are structured as causal graphical models" and either S_1 or S_2. None of these claims rise to the level of a full-fledged theory (as required for a reduction). More importantly, these not-quite-theories are not accounts of the same phenomena at different levels, and so a reduction would simply be inappropriate. In order to make sense of this example, we need to see scientific discovery about *P* (= the structure of our causal knowledge) as shaped and informed by other scientific claims that are relevant because they impose constraints, not because they are involved in a reduction.

3.2. Concepts based on causal structure

For the second case study, we turn to the phenomenon of conceptual representation in our cognitive psychology: that is, we plausibly want to discover the structure of our everyday concepts, such as DOG, STUDENT, or APPLE, though with the recognition that there might be multiple types of concepts depending on the particular domain or even individual (e.g., Barsalou, 2008; Machery, 2009). Many different theories of conceptual structure have been proposed over the years (Murphy, 2004), and so we have a rich set **T** of theoretical possibilities, and a correspondingly difficult

scientific discovery problem. One natural S_1 is the empirical finding that people frequently (and often spontaneously) group together different individuals on the basis of their shared or similar causal structure (Carey, 1985; Keil, 1989). The contrast class $\mathbf{S}_1$ here includes, for example, the claim that perceptual similarity always determines grouping. And given this S_1, we can sensibly include S_2 about causal structure: namely, people's representations of causal knowledge are structured like causal graphical models. These S_1 and S_2 constrain the space of theories of conceptual representations: at least some of our concepts are (likely) structured as causal graphical models (Rehder, 2003a, 2003b). Moreover, S_1 provides us with a relatively precise characterization of when our concepts will have that structure.

One might object that this example is not really a case of scientific discovery, but rather is "simple" scientific reasoning. However, this characterization is overly simplistic and dismissive. As an historical matter, the story that I provided in the previous paragraph largely captures the scientific history: causal graphical models were only proposed as a possible model of some concepts once people combined the information in S_1 and S_2. The causal model theory of concepts was "discovered" largely by thinking through the implications of these constraints. More generally, this objection assumes a sharp distinction between scientific reasoning and scientific discovery, but part of the point of these case studies is precisely that there is no bright line to be drawn. Scientific practice partly consists in trying to put various pieces together into a relatively integrated account. That integration can involve both discovery (e.g., proposing an entirely new theory of conceptual representation in terms of causal graphical models) and reasoning (e.g., showing the relevance of empirical findings that are not directly about the nature of conceptual representations).

This case study also demonstrates the dynamic nature of these processes in two different ways. First, notice that S_2 here is T from the previous case study. The product of some scientific discovery will itself usually imply constraints on other **T**'s, though those might not immediately be recognized

by the scientists. These connections provide one way in which a single empirical finding can have wide-ranging "ripple effects": the impact of an empirical finding is not necessarily limited to the immediately relevant scientific question or problem, as the answer to that question can imply constraints that help answer a second question, which can thereby imply constraints for a third question, and so on.[8] Second, this "discovery via constraints" is dynamic in nature because it leads to a new theory with novel empirical predictions that can subsequently be tested and explored (e.g., Hadjichristidis, Sloman, Stevenson, & Over, 2004; Rehder, 2009; Rehder & Kim, 2010). And those experiments and observations provide additional, novel S_i claims that further constrain our theory of conceptual representations, either in detailed structure or in the scope of a particular theory. Scientific discovery and reasoning do not proceed in a discrete, staged manner, but rather involve a complex dynamic between using constraints to develop new theoretical ideas, and using ideas to find novel constraints.

3.3. Goals and learning

The third case study looks a bit more like a case of "traditional" scientific discovery than the prior two. Consider the general question of the role of goals—more generally, beliefs about future tasks—on what and how we learn from the environment. Arguably, almost all major theories of (high-level) learning in cognitive psychology assume that goal or future task information only influence the domain from which we learn, but do not further influence the method or dynamics of learning. For example, essentially all theories of concept learning assume that I have domain knowledge about which features are potentially relevant to the new concept, but that the goal and (beliefs about) future tasks do not otherwise influence my concept learning. That is, given the same

[8] A framework for characterizing and modeling this dynamics represents another extension of the constrain-inclusion model of Laudan, Nickles, and others.

domain and same stimuli, learning is (on all of these theories) predicted to have the same dynamics. However, this dominant assumption has been called into question, and a new theory was discovered or proposed, in large measure by considering constraints implied by other scientific commitments.

The first theoretical claim S_1 that is potentially relevant to this problem is that much of our learning depends partly on attention. If we do not attend to a factor, then we typically learn less about it (e.g., Desimone & Duncan, 1995; Huang & Pashler, 2007), though some learning can occur even when we do not consciously attend to the items (DeSchepper & Treisman, 1996). We do not need to make any particularly strong theoretical commitments about the nature of attention here. Rather, S_1 simply expresses the fact that attention and learning are sometimes closely connected. The relevant contrast class $\mathbf{S}_1$ here includes the possibilities that attention does not directly modulate learning, or that attention is merely a necessary condition for learning (i.e., a "gate" on learning) rather than influencing it in a more nuanced fashion.

The second theoretical claim S_2 is that attention allocation depends partly on one's current task or goal. That is, my current task influences the particular way that I allocate my attention across my perceptual or cognitive field. For example, the current task or goal helps to determine which dimensions of objects are salient, and so which dimensions or objects are subsequently ignored as I perform that task (Maruff, Danckert, Camplin, & Currie, 1999; Tipper, Weaver, & Houghton, 1994). More colloquially, people pay much less attention to things that do not matter for their tasks, though they do not necessarily completely ignore those features of the stimuli or environment. As with S_1, this claim is likely not particularly surprising or controversial, though the human mind could have functioned differently (e.g., selection of task-relevant factors might have involved only cognitive mechanisms, rather than lower-level attentional processes).

Both S_1 and S_2 are widely (though not universally) endorsed in cognitive psychology, and both imply constraints on whether goals might influence the dynamics of learning. For concreteness,

consider only two theoretical claims in **T**: (a) "goals only determine domain of learning input," labeled $T_{current}$ since it is the assumption of most current learning theories; and (b) "goals influence learning dynamics," labeled T_{new} since it is a novel theory (in this domain). Now consider the constraints implied by S_1 and S_2 for the two possible, though different, tasks of "learning for goal *A*" or "learning for goal *B*" (e.g., "learning to *predict* a system's behavior" vs. "learning to *control* a system's behavior"). By S_2, we should expect differential attention allocation; by S_1, we should expect this differential attention to translate into differential learning. That is, S_1 and S_2 jointly raise the plausibility of T_{new} and decrease the plausibility of $T_{current}$, even though none of these theoretical claims stands in any particular reductive relation with one another. Their intertheoretic relationships are more complicated, but equally able to support scientific discovery. In fact, this case study was historically a true case of discovery: although the analysis here evaluates T_{new} and $T_{current}$ as contemporaneous competitors, T_{new} was originally invented and proposed (in Danks, 2014) only after recognizing that the constraints implied by S_1 and S_2 were in significant tension with $T_{current}$, and so a new theory was needed. Moreover, subsequent experimental results spoke strongly in favor of T_{new} (Wellen & Danks, 2014; see also Hagmayer, *et al.*, 2010).

In sum, these three case studies provide three different ways in which intertheoretic constraints can be used to suggest or "discover" new scientific ideas. In contrast with "discovery via reduction," this account in terms of "discovery via constraints" can explain how disparate theories, as well as claims and other not-quite-theories, can inform and guide our scientific investigations. One might be concerned by the relatively prosaic and banal nature of these case studies, as we often think about scientific discovery as something grand or transformative. However, scientific discovery is also an everyday phenomenon, as particular scientists discover novel ways to synthesize or unify disparate scientific pieces. This type of everyday scientific thinking requires explanation and clarification just as much as the grand discoveries of Newton, Einstein, or others. And while reduction might

sometimes be the basis of everyday scientific discovery, the more typical case is to use multidimensional intertheoretic constraints in order to add new pieces to our scientific puzzles.

4. Constraints all the way up?

The previous section focused on small-scale cases of scientific discovery via constraints, largely to provide enough detail to help demonstrate the mechanics of the framework. These smaller cases also clearly demonstrate that constraints are doing the relevant work, rather than full-blooded reductions. At the same time, one might wonder whether the "discovery via constraints" approach might be useful for understanding much larger-scale scientific discovery.[9] So, I close with some speculative thoughts about whether even scientific "revolutions" (using the term in a very broad way) could be understood in terms of discovery via constraints. At first glance, it is not obvious how constraints might be playing a role, particularly given the many stories about the crucial role of creativity in inventing or discovering scientific theories with wide scope. These stories highlight the role of **T** as a "free parameter" in the present account: I provided no explanation or account about how or why particular theoretical possibilities are included in **T**, even though discovery is (in a certain sense) limited to the elements of that set, and creativity or intuitive insight might be one way to generate elements of **T**. On the "discovery via constraints" view, creativity in theoretical innovation can thus have a very large impact, even though it is not directly modeled or explained.[10] We can only consider the impact of various constraints on a theoretical idea if we recognize the idea as possible or worth considering, and imagination or creativity might help explain why some possibility is included in **T**.

[9] Thanks to Donald Gillies for encouraging me to consider this possibility, even after I had initially dismissed it.
[10] That being said, creativity could perhaps be modeled as discovery via constraints in the following way: suppose creativity results, as some have suggested (e.g., Simonton, 1999), from profligate, unguided idea generation, followed by thoughtful pruning of the outputs. This pruning process could potentially be based on the use of constraints, and so we have the beginnings of a picture in which *all* discovery is based on constraints. Of course, this proposal does not explain the "idea generator," and much more work would need to be done before we have a full story in terms of constraints. Nonetheless, it is suggestive that even the "singular creative act" might be captured by this framework.

In many cases of scientific revolutions, this creative innovation is an important part of the overall story, but a single creative act is almost never the full story of any scientific revolution. Constraints arguably play a large role in the dynamics of scientific change that can result *after* the initial innovation. In many scientific "revolutions," there is a significant initial shift in approach or "paradigm" (again, using the terms broadly) that is followed by significant work to put the empirical and theoretical pieces back together inside the new framework. The initial creative idea alone typically predicts and explains many fewer phenomena than were captured using the prior scientific theory and paradigm. Completion of the scientific revolution thus depends on finding auxiliary theories, background conditions, special cases, and other additional theories and not-quite-theories that generate explanations and predictions. Discovery via constraints will frequently play a significant role in these discoveries: these additional scientific elements can be discovered by trying to integrate constraints from the initial innovation, as well as constraints from prior empirical data and other posits that "survive" the revolution. For example, the Copernican revolution that shifted astronomy to a heliocentric view of the solar system started with a creative innovation, but then required substantial work to determine the appropriate constants, parameters, structures, and so forth. A dichotomy is sometimes drawn between periods of "revolutionary" and "normal" science, but a scientific revolution typically requires many steps of normal science along the way, and those can all (I argue) be fruitfully understood in terms of discovery via constraints. Discovery via constraints might (or might not) help us understand creativity or true innovation, but much of the rest of the process of large-scale scientific change can potentially be helpfully modeled as discovery via constraints.

In general, the process of scientific discovery often employs constraints from other scientific ideas, claims, theories, and not-quite-theories. These constraints result from the complex, multidimensional intertheoretic relationships that obtain between these pieces and the to-be-

discovered scientific claim. Reduction is one salient intertheoretic relationship, and a source of particularly powerful constraints when it obtains. The corresponding "discovery via reduction" can thus also be particularly powerful, but only when the reduction relation obtains. In actual scientific practice, and particularly in everyday science, reductions are rarely forthcoming. Instead, scientific discovery proceeds through the opportunistic use of less powerful, but more widespread, constraints grounded in weaker intertheoretic relationships. "Scientific discovery via intertheoretic constraints" includes "discovery via reduction" as a special case. More importantly, it provides us with a richer, more nuanced understanding of some ways in which scientists develop novel ideas and theories.

Acknowledgments

The ideas in this paper were initially presented at the "Building Theories: Hypotheses & Heuristics in Science" conference at Sapienza University. Thanks to the audience at the conference for their comments and criticisms, particular Emiliano Ippoliti, Lindley Darden, Donald Gillies, and Margie Morrison. Thanks also to two anonymous reviewers for valuable feedback on an earlier draft.

References

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645.

Bechtel, W., & Richardson, R. C. (2000). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy, 82*, 1-22.

Danks, D. (2013). Moving from levels & reduction to dimensions & constraints. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 2124-2129). Austin, TX: Cognitive Science Society.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Cambridge, MA: The MIT Press.

Danks, D. (2015). Goal-dependence in (scientific) ontology. *Synthese, 192*, 3601-3616.

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science, 69(S3)*, S354-S365.

Darden, L., & Craver, C. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 33(1)*, 1-28.

DeSchepper, B., & Treisman, A. (1996). Visual memory for novel shapes: Implicit coding without attention. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*, 27-47.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Dizadji-Bahmani, F., Frigg, R., & Hartmann, S. (2010). Who's afraid of Nagelian reduction? *Erkenntnis, 73*, 393-412.

Fodor, J. A. (1974). Special sciences: Or the disunity of science as a working hypothesis. *Synthese, 28*, 97-115.

Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. *Nous, 31*, 149-163.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51(4)*, 334-384.

Hadjichristidis, C., Sloman, S., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science*, *28*, 45–74.

Hagmayer, Y., Meder, B., Osman, M., Mangold, S., & Lagnado, D. A. (2010). Spontaneous causal learning while controlling a dynamic system. *The Open Psychology Journal*, *3*, 145–162.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General, 138*, 22-38.

Hausman, D. M. (1998). *Causal asymmetries*. Cambridge: Cambridge University Press.

Hooker, C. A. (1981a). Towards a general theory of reduction, part I: Historical and scientific setting. *Dialogue, 20*, 38–59.

Hooker, C. A. (1981b). Towards a general theory of reduction, part II: Identity in reduction. *Dialogue, 20*, 201–236.

Huang, L., & Pashler, H. (2007). Working memory and the guidance of visual attention: Consonance-driven orienting. *Psychonomic Bulletin & Review*, *14*, 148–153.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Laudan, L. (1981). A problem solving approach to scientific progress. In I. Hacking (Ed.), *Scientific revolutions*. Oxford: Oxford University Press.

Machery, E. (2009). *Doing without concepts*. Oxford: Oxford University Press.

Maruff, P., Danckert, J., Camplin, G., & Currie, J. (1999) Behavioural goals constrain the selection of visual information. *Psychological Science, 10*, 522-525.

Murphy, G. L. (2004). *The big book of concepts*. Cambridge, MA: Bradford.

Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt.

Nickles, T. (1978). Scientific problems and constraints. In *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association* (pp. 134-148). Chicago: University of Chicago Press.

Nickles, T. (1981). What is a problem that we may solve it? *Synthese, 47*, 85-118.

Nickles, T. (1988). Questioning and problems in philosophy of science: Problem-solving versus directly truth-seeking epistemologies. In M. Meyer (Ed.), *Questions and questioning* (pp. 43-67). Berlin: Walter de Gruyter.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94(1)*, 61-73.

Putnam, H. (1975). Philosophy and our mental life. In *Mind, language, and reality: Philosophical papers* (Vol. 2, pp. 291–303). Cambridge, MA: Cambridge University Press.

Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 1141–1159.

Rehder, B. (2003b). Categorization as causal reasoning. *Cognitive Science*, *27*, 709–748.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science*, *33*, 301–344.

Rehder, B., & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *36*(5), 1171–1206.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rovee, C. K., & Rovee, D. T. (1969). Conjugate reinforcement of infant exploratory behavior. *Journal of Experimental Child Psychology, 8*, 33-39.

Schouten, M. K. D., & de Jong, H. L. (Eds.). (2012). *The matter of the mind: Philosophical essays on psychology, neuroscience and reduction*. London: Wiley-Blackwell.

Shanks, D. R. (1995). Is human learning rational? *The Quarterly Journal of Experimental Psychology*, *48A*, 257–279.

Simonton, D. K. (1999). *Origins of genius: Darwinian perspectives on creativity*. New York: Oxford University Press.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford: Oxford University Press.

Timberlake, W. (2001). Integrating niche-related and general process approaches in the study of learning. *Behavioural Processes, 54*, 79-94.

Tipper, S. P., Weaver, B., & Houghton, G. (1994) Behavioural goals determine inhibitory mechanisms of selective attention. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 47*, 809-840.

Wellen, S., & Danks, D. (2014). Learning with a purpose: The influence of goals. In P. Bello, M Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1766-1771). Austin, TX: Cognitive Science Society.

Wimsatt, W. C. (1980). Reductionistic research strategies and their biases in the units of selection controversy. In T. Nickles (Ed.), *Scientific discovery: Case studies* (pp. 213-259). D. Reidel Publishing.