Abstraction and its Limits: Finding Space For Novel Explanation

Eleanor Knox
King's College London
eleanor.knox@kcl.ac.uk

October 4, 2012

Abstract

Several modern accounts of explanation acknowledge the importance of abstraction and idealization for our explanatory practice. However, once we allow a role for abstraction, questions remain. I ask whether the relation between explanations at different theoretical levels should be thought of wholly in terms of abstraction, and argue that changes of variable between theories can lead to novel explanations that are not merely abstractions of some more detailed picture. I use the example of phase transitions as described by statistical mechanics and thermodynamics to illustrate this, and to demonstrate some details of the relationship between abstraction, idealization, and novel explanation.

Introduction

> ... if you are not 'hipped' on the idea that *the* explanation must be at the level of the ultimate constituents... there is a very simple explanation here. The explanation is that the board is rigid, the peg is rigid, and as a matter of geometrical fact, the round hole is smaller than the peg, the square hole is bigger than the cross-section of the peg. (Hilary Putnam [22, p.94])

The idea that higher level theories can be more explanatory of some phenomena than lower level ones is a appealing one. In some ways, it is obviously right;

to describe, say, a biological phenomenon in terms of molecular dynamics (or, God forbid, quantum field theory) is to lose all hope of understanding in a morass of detail. And yet, until quite recently, the idea was not properly fleshed out in the literature on scientific explanation; causal, deductive-nomological and unificationist models all appeared to imply that more detail, or a more fundamental theory would always lead to better explanations.

The last few years have seen more attention paid to the importance of abstraction and idealization to explanation. Michael Strevens' book, *Depth* [25] suggests a central role for abstraction in causal explanations. In the philosophy of physics literature, Robert Batterman [3, 2, 4, 5, 7, 8] has drawn attention to a host of cases in which different theories are connected by limiting relations, and argued that related phenomena cannot be explained using only the resources of the more fundamental theory. In doing this, Batterman goes further than merely arguing for a role for abstraction, and argues that there is a sense in which fundamental theories may be regarded as explanatorily inadequate.

In this paper, I also want to go a step beyond merely noting the importance of abstraction for explanation (although my account will differ significantly from Batterman's). I will examine the possibility of *novel explanation*. Putnam's famous example, which considers the explanation of a square peg's failure to fit into a round hole, seems to be a case in which more than one explanation of a phenomenon is possible. On the one hand, we have a common sense explanation in terms of rigidity and geometry, and on the other we have (or rather, don't have, but are aware of in theory), an explanation in terms of the microscopic constituents of the peg and hole. Suppose we grant that, even if we were in possession of the microscopic account, the macroscopic account would provide a better explanation. And suppose we also grant that abstraction is an important part of the explanatory story, that the throwing out of detail often leads to better explanations. We might then ask whether abstraction is the whole story: can we account for the explanatory utility of the higher level explanation purely by noting that it is a distant abstraction of the more fundamental explanation? Should we model all explanations as situated along a sliding scale of abstraction, with higher level explanations deriving their considerable explanatory power from their ability to concisely summarize relevant information from some fundamental picture?

I will argue here that abstraction is not the *whole* story. When we change from one theory to another, the change of variables involved can induce novel explanation. However, abstraction is an important part of the story here; it is precisely because the abstraction that we deem appropriate is highly sensitive to changes in variables that the explanations given by one theory may be deemed novel with

respect to another theory. When we properly understand the role of abstraction, we appreciate that *explanatory value* may be irreducible, even where theoretical reduction is possible. Particular kinds of complex variable change make the abstraction techniques of the higher level theory opaque from the perspective of the lower level theory.

Although the account I give here has the potential to be applied quite widely, I will work through a detailed example that has been much discussed in the philosophy of physics literature, that of phase transitions. Phase transitions require an appeal to an asymptotic limit for their explanation within statistical mechanics (in this case the thermodynamic limit, which involves taking the particle number parameter to infinity). Along with other cases of asymptotic inter-theoretic relations, phase transitions have sometimes been held to be *emergent* phenomena, where emergence can mean something quite weak; a common theme is that emergence involves novelty or autonomy. But emergence is a weasel word, and novelty and autonomy are often left undefined. This paper gives an account in which the explanatory value of one theoretical description may not always be reducible to the explanatory value of another, and argues that variable changes involving asymptotic limits can play a role in this irreducibility of explanatory value. It may therefore be construed as giving an account of emergence in terms of explanatory novelty, and as explaining why asymptotic limits can lead to cases of emergence. However, if this is an account of emergence, it is a very weak one. The examples I will discuss are all cases in which reduction is possible in quite a strong sense. As such, I add my voice to a number of physicists [1] and philosophers of physics [10, 11] who believe that theories can be novel despite being successfully reduced to another theory. But the reader who is allergic to the term emergence, or who believes it to be by definition incompatible with reduction, may pass over any emergence talk in this paper. What is really of interest here is the sense in which higher level explanations may be novel.

I begin in section one by giving a toy example that illustrates the interplay between abstraction and changes of variable in a very simple context. The example illustrates the way in which changes of variable can lead to changes in explanatory abstraction, and thus lead to judgments of explanatory value at one level that cut across those at another. However, this example is too simple to illustrate the case convincingly, first because the kind of change of variable involved is too simple, and second because the variable change fails to lead to an interesting theory. This motivates the need to wade into the deeper waters of real physics.

In section two, I introduce a more realistic and contentious example, that of phase transitions. I argue that the puzzles surrounding phase transitions can be

solved. However, they provide a realistic and interesting example of the complexities involved in real variable changes, which typically a subtle interplay of idealization and abstraction. Phase transitions are also of interest because they've been held by some authors to involve emergence. I argue that, when the puzzles surrounding phase transitions are solved, no obvious novelty remains. However, examination of the phase transition case indicates the existence of a particularly irreversible change of variables between statistical mechanics and thermodynamics.

Section three follows on from the discussion phase transition by taking a broader look at the issues of idealization and abstraction raised.

In section four, I return to the relationship between thermodynamics and statistical mechanics, and argue that the value of certain kinds of thermodynamic explanation is not explicable from the perspective of statistical mechanics. The abstractions used in thermodynamic explanation are not readily available in statistical mechanics, precisely because of the complex nature of the variable changes involved in moving between the two theories. The kinds of considerations involved in the case of phase transitions therefore do introduce novelty of a sort, but in a more indirect way than is sometimes suggested.

This paper touches on vast topics, and there is much that will go undiscussed. My account of phase transitions will be only a sketch, in part because I think the topic has been well-discussed elsewhere, and in part because I hope that the results here are of interest to a general audience that a technical discussion of statistical mechanics might exclude.

Moreover, although this is a paper on explanation, I will not recommend a model of explanation, nor will I give a full-bodied account of the circumstances under which some abbreviation of a detailed description counts as an explanatory abstraction. I will however assume that abstraction should be part of our account of explanation itself, and not merely of our account of the pragmatics of explanation. In assuming this, I deny the picture of explanation painted by Peter Railton [23], in which our theory of explanation (deductive-nomological, causal, or whatever) defines the *ideal explanatory text*, which we then select from based on context. Rather, I hold there to be cases in which abstraction leads to better explanation in a robust sense, not merely because of an audience's limited capacity for absorbing detail, but because some details are of much greater explanatory relevance than others, or because the feature to be explained is the robustness of some feature under changes of detail.[1] This is the kind of account articulated by

[1]The contrast between a view on which the value of abstraction is a matter of pragmatics and

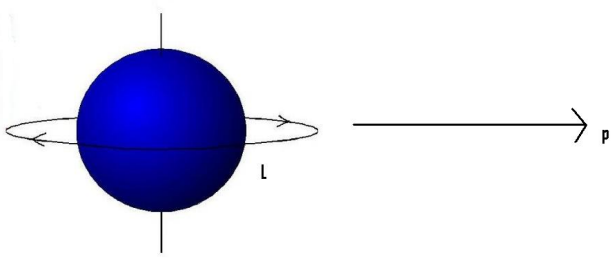

Figure 1: The spinning ball

Michael Strevens for causal explanation, and it is my hope that something like this can be developed for those cases of explanation that are not covered by the causal account.

1 Abstraction and changes of variable: A toy example

The main aim of this paper is to demonstrate that the kinds of variable changes involved in moving from one theory to another can lead to explanatory novelty. Real examples of this will, however, be complex enough that the phenomenon at issue is hard to spot. It's therefore helpful to begin with a very simple example of a change of variables, and consider what happens in this toy model.

Consider a Newtonian model of a spinning ball, moving with constant velocity in the x direction, as shown in Figure 1. If no forces act on the ball, the equations governing its motion are very straightforward; if we point out that both angular

the one assumed here might be thought of as a contrast between epistemic and metaphysical senses of explanatory value. In the kind of account of abstraction I assume here, abstraction is valuable not only beacuse of our limited capacity for detail, but because it also allows us to track real relevance relations in the world. I take it to be a matter of metaphysics that certain coarse-grained variables are well-correlated with other coarse-grained variables. This more-than-epistemic view of abstraction will feed into a more-than-epistemic view of explanatory novelty. I'll have more to say about this in the concluding section of this paper.

and linear momentum are conserved, there is not much more to be said:

$$L = l \tag{1}$$

$$p = k \tag{2}$$

where L and p are angular and linear momentum respectively, and l and k are constants. If at time $t_0 = 0$ the ball begins at position $x = 0$ and angle of rotation $\theta = 0$, for constant momentum the equations of motion are:

$$\frac{L}{I}t = \theta \tag{3}$$

$$\frac{p}{m}t = x \tag{4}$$

where I is moment of inertia and m is mass. If we are now asked to explain the position of the ball at a given time, it's obvious that angular momentum is an irrelevant variable. A good explanation should only appeal to linear momentum and equation (4). We have here a simple case of explanatory abstraction; a full description of the set-up is less explanatory than a simpler account that includes only the relevant detail.

However, consider an equally accurate mathematical description of the set-up in terms of changed variables:

$$A = pm + L \tag{5}$$

$$B = pm - L \tag{6}$$

The new equations relevant to explaining the position of the ball are:

$$A = a \tag{7}$$

$$B = b \tag{8}$$

$$\frac{A + B}{2m^2}t = x \tag{9}$$

Equation (9) is dimensionally correct, and even appeals to the same 'observable' quantities, (m, x and t) as equation (4). But in our transformed variables we can no longer eliminate a variable and perform the abstraction. Any explanation of the position of the ball will appeal to two different variables with the dimension of angular momentum.

What can this toy example teach us? It demonstrates that one common kind of abstraction, variable reduction, is highly sensitive to a change of variables. Although equations (4) and (9) are equivalent, (9) is not, by the lights of its variables, an explanatory abstraction, whereas (4) is. Part of what is valuable about the explanation is lost in the variable change.

However, the example is too simple to make the point completely. After all, faced with equation (9), a reverse change of variables almost jumps off the page; the simpler explanation in terms of one variable is only a small algebraic step away. But now suppose we are dealing with the kinds of variable change involved in real inter-theoretic relations; these kinds of changes will be much more complex; the backwards move is unlikely to be obvious. In fact, equations governing variable change will rarely involve anything as simple as an algebraic relationship. Typically, they will include 'irreversible' mathematical processes like summation; information will be lost, and a change back to the old variables will not be possible without prior knowledge of the previous theory. In general, once we have kicked away the ladder that lead to our variable change, the way back to the original theory and its abstraction techniques will not be reconstructible.

The toy theory is also misleading in another way; there is an obvious sense in which the transformed variables of (7)-(9) are the wrong variables in which to describe basic mechanics. The example would be more telling, and more interesting, if our new variables themselves lead to successful abstraction techniques. Although that's not the case here, this will be the case when a change of variables takes us from some more basic theory to a higher level theory, that is, when the change of variables expresses a reductive relationship. Typically, the higher level phenomenological theory is developed independently of, and often prior to, the more fundamental theory. Its variables have been designed to lead to successful abstraction by its own lights, and it will thus have its own standards of explanatory value.

Given all of this, it is not at all clear why we should expect judgements of explanatory value to map neatly from one theoretical level to another. What counts as a good explanation in a given theory is often a matter of finding an explanation in as few relevant variables as possible. However, this kind of variable elimination itself depends on the choice of variables. Even when neat reductive relations are possible, the move between theories will involve a move to new variables and new judgements of explanatory value. An abstraction that leads to a good explanation in the higher level theory may 'cut across' the division between variables naturally made within the more fundamental theory. Even though the higher level description may be reduced to, and understood in terms of, the lower level description, the

lower level will not agree with the higher level as to whether the description constitutes a good explanation; from the more fundamental perspective, the selection of information may look hopelessly arbitrary.

It is in this sense that the explanations offered by higher level theories may be novel with respect to some more fundamental theory. Changes of variable induce changes in explanatory abstraction, and these lead to explanations that are not merely abstractions from the description provided by the underlying theory. The case for this will be made stronger by a real example, where two different theories both have considerable explanatory pedigree, and where the relationship between them is particularly complex.

2 Phase Transitions

To get a better handle on how real changes of variable work, I'll look at the relationship between statistical mechanics and thermodynamics. This is far from a simple topic; the degree to which thermodynamics is reducible to statistical mechanics is a matter of some debate, and individual instances of reduction are generally controversial. Nonetheless, the reduction of thermodynamics (henceforth TD) to statistical mechanics (henceforth SM) is more complete than most; we have reasonable statistical mechanical definitions of a number of thermodynamic variables.

To get at one such change of variable, and the complexities involved in it, I'll spend some time discussing phase transitions. Phase transitions include changes of state, such as water boiling or carbon dioxide sublimating, as well as some less familiar phenomena, like sudden changes of magnetization. These all have in common that they involve sudden changes to the large scale physical properties of a system ('sudden' here means corresponding to some very small change in some control variable). Thermodynamics proves adequate to describe and categorise a wide variety of these effects via a discontinuity in a variable of the thermodynamic theory.

I will here describe only the first-order phase transition associated with a change of state of some system held at a stable temperature away from critical temperature (at or near certain temperatures, the behaviour of certain systems becomes particularly odd), but much of what is said here applies to other phase transitions as well. Even in our simple case, we will see that a puzzle arises: thermodynamics characterises phase transitions as involving a discontinuity in the free energy of a system. However, the *statistical mechanical* function corresponding to

the free energy of the system can only possess discontinuities if we take the limit of the function as the number of particles becomes infinite. It therefore seems, at first blush, as if statistical mechanics predicts that phase transitions can only occur in infinite systems.

Before describing the problem, and how to begin to solve it, in a little more detail, it's worth defending my choice of example. There is an ongoing and lively debate about the status of phase transitions; why choose such a contested example of inter-theory relations with which to make a general point about explanation? Oddly, this is actually one of the more *straightforward* examples available within the range of SM/TD boundary problems.[2] The problems here have been well treated and (in my view) solved within the existing literature. This solution can even be stated with relative brevity, making my job a good deal easier. Phase transitions involve asymptotic limits, and have therefore been held by several different philosophers to be cases in which interesting novelty arises. Discussing phase transitions will give me the opportunity to analyse the importance of asymptotic limits.[3] My eventual claim will be that asymptotic limits lead to particularly robust and interesting changes of variable, and thus to novel explanations.

2.1 First order phase transitions at constant temperature: a sketch

For a simple example of a phase transition,[4] let us take a system at constant temperature, whose volume is changed slowly enough that the system stays close to equilibrium at all times. Thermodynamics predicts that such a system will undergo a phase transition when there is a non-analyticity in the free energy of the system; that is, when the free energy of the system or one of its derivatives undergoes a discontinuous change. Figure 1 shows the predicted thermodynamic behaviour of a system in terms of its Helmholtz free energy (A, here given as A_{TD} to make it clear that this is a thermodynamic quantity) and pressure (P) plot-

[2]This will raise eyebrows among readers familiar with the philosophy of physics literature. There is a great deal more to be said about phase transitions than will be touched on here. For example, I will not discuss the strangely universal behaviour shared by diverse systems when at or near critical temperature, nor the renormalization group techniques used to explain this behaviour. However, the issues here can be divorced from these interesting questions for our purposes. For recent more physics oriented treatments see [20],[12] or [21].

[3]In [11] Jeremy Butterfield discusses a range of such cases.

[4]Even for this simple example, what I give here is only the barest of sketches. For a detailed treatment see e.g. [24].

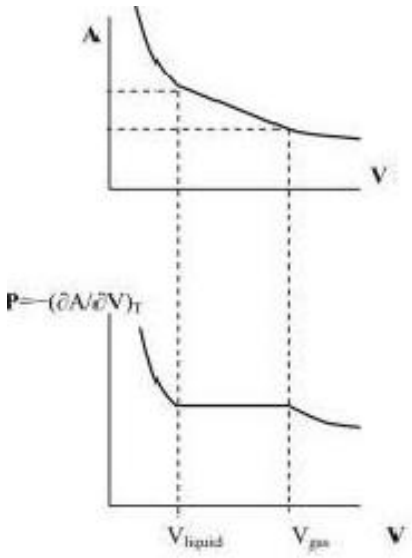

Figure 2: Helmholtz free energy A at a constant (less than critical) temperature. From [24].

ted against volume. The Helmholtz free energy here is defined in terms of the thermodynamic variables U (internal energy), T (temperature) and S (entropy):

$$A_{TD} = U - TS \tag{10}$$

When we move to a statistical mechanical description of the same phenomenon, we give the Helmholtz free energy in terms of SM variables:

$$A_{SM} = -k_B T \ln Z \tag{11}$$

Here, k_B is the Boltzmann constant, T is temperature, and Z is the *partition function*, given by:

$$Z = \sum_i \exp(-\frac{E_i}{k_B T}) \tag{12}$$

Z is obtained by summing over all microstates with some particular energy E_r, and will depend on the number of particles in the system.

Now, it's clear that by having both a thermodynamic and a statistical mechanical expression for the free energy, we ought thereby to have an expression for

the thermodynamic free energy in statistical mechanical terms. That is, we should have a bridge law that reflects the change in variables involved in moving from SM to TD. However, things are not this simple: Problems arise when we note that Z is a sum of analytic functions, and is hence itself an analytic function. This also makes A analytic (an analytic function of an analytic function is analytic). And an analytic function is infinitely differentiable; it cannot have the kind of discontinuities in its derivatives shown in figure 1.

However, we can find a way round the problem if we take the thermodynamic limit, that is, if we examine the behaviour of Z as $N \to \infty$, where N is the number of particles. In this limit Z can indeed have appropriate singularities. So in order to recreate the predicted thermodynamic behaviour in statistical mechanics, we must treat the system as being infinite in both particle number and volume (leaving $\frac{V}{N}$ constant).

2.2 Locating and dissolving the mystery

Real systems are finite. It is therefore puzzling that we should have to appeal to an infinite limit in order to account for phase transitions.[5]

In [13], Craig Callender states the puzzle as arising due to a tension between four propositions:

> 1. Real systems have finite N.
> 2. Real systems display phase transitions.
> 3. Phase transitions occur when the partition function has a singularity.
> 4. Phase transitions are governed/described by classical or quantum statistical mechanics (through Z).
>
> [13, p.549]

At least one of these propositions must be given up, and all four options have been explored in the literature. (Philosophical opinions, like gases, tend to expand to fill all corners of a given space.) But of particular interest here is the possibility of denying 4, and declaring phase transitions irreducible in one sense or another. This is the approach Batterman takes when he writes:

[5]The very abbreviated account below roughly follows that given in Paul Mainwood's work [17, 18]. For a detailed account in the same vein see Jeremy Butterfield and Nazim Bouatta [12].

> My contention is that thermodynamics is correct to characterize phase transitions as real physical discontinuities and it is correct to represent them mathematically as singularities. Further, without the thermodynamic limit, statistical mechanics would completely fail to capture a genuine feature of the world. Without the thermodynamic limit, in fact, statistical mechanics is incapable even of establishing the existence of distinct phases of systems. [5, p.12][6]

Given the statistical mechanical model, it is not clear why we should insist that phase transitions must be 'real physical discontinuities'. Certainly, no amount of data will establish a discontinuity, and it's hardly obvious that we witness discontinuous behaviour when we see water boil. Of course, data can't establish the lack of a singularity either. Batterman claims that to assume that there is no physical discontinuity is to beg the question - to assume the completeness of statistical mechanics when that is precisely what is at issue. However, unless we have bought wholesale into Batterman's anti-foundationalism, it seems very odd to think that we are not allowed to privilege the mathematical form of a more fundamental theory over one that is approximate and phenomenological.

Callender's suggestion is that we deny 3, and insist that phase transitions do occur even when there isn't a singularity. Instead, we can simply insist that:

3*. Phase transitions occur when the partition function is such that there would be a singularity in the thermodynamic limit.

With this definition, the tension between Callender's four propositions vanishes.

This solution seems roughly right, but all our work is not quite done. It is important to get clear on the role that thermodynamics is playing on our puzzle. Mainwood argues that the puzzle can be generated *without any reference to thermodynamics whatsoever*; it is in fact a problem internal to statistical mechanics.

It turns out that even if we consider the problem of modeling phase transitions within statistical mechanics, in complete ignorance of the thermodynamic treatment, the infinite limit is unavoidable. For one thing, the finite N case is computationally intractable. When we move to the infinite case, we render the problem of predicting phase transitions easier in part because we abstract away from boundary conditions. By ignoring the complex interaction of water molecules with the sides of the kettle, and eliminating details such as the shape of the kettle itself, we allow for generalised predictions of the phenomenon of water boiling.

[6] Chiang Liu [15, 16] also shares something like this view.

It is an assumption of this paper that such abstraction methods are not a mere matter of mathematical convenience, that abstracting from detail can lead to an increase in explanatory power; we have here an example that illustrates this. In this case, the increase in explanatory power is quite pronounced, for it is only when we take the thermodynamic limit that mathematical structures emerge that allow us to distinguish between various phase transitions of different types. This categorization and characterization of different phase transitions is a key success of statistical mechanics. That the $N \to \infty$ limit is crucial to this success makes it apparent that the pressure to understand the theory's reference to the infinite case comes from within statistical mechanics itself.

With this awareness, and our replacement proposition 3* in hand, the puzzle begins to look more benign. What we appear to have is a case of a rather extreme idealization[7] required for the modeling and explanation of phase transitions in statistical mechanics. This idealization allows for a successful abstraction technique, which enhances the explanatory power of the theory.

As Mainwood points out, the problem is now one of justification: how do we know that by taking the thermodynamic limit, we abstract in such as way as to reveal structural features of the actual (non-infinite) system? To some degree, this idealization is justified by its own predictive and explanatory success, but more justification is desirable. In particular, it would be nice to have better mathematical models which would allow us to more fully understand the relationship between the finite N case and the thermodynamic limit. But this is not a deep conceptual worry of the kind that the problem is sometimes taken to imply. Moreover, we do at least seem to have the beginnings of such models for simple systems. Butterfield and Bouatta [11, 12] have argued that we can see asymptotic behaviour as emerging *before* the limit in various cases. In the case of simple non-critical phase transitions, a model called the Ising model is relevant; this demonstrates simple phase transitions for finite two-dimensional magnetic systems. Of course, the systems we have been discussing are neither magnetic nor two-dimensional, but there is reason to think that similar results apply to the phase transition under discussion here. If we see models like the Ising model not as providing adequate alternative predictions, but rather as justifying the use of the infinite limit, then they have considerable explanatory power. They lend plausibility to the idea that taking the

[7]John Norton argues in [21] that taking the thermodynamics limit involves approximation rather than idealization, but I'll continue to refer to the phenomenon under issue here as idealization. Although the difference between idealization and approximation is conceptually important, nothing in *my discussion here* turns on which of these categories the use of the thermodynamic limit falls under.

thermodynamic limit picks out features of the actual finite partition function.

This brief discussion raises a number of issues concerning the nature of abstraction and its relation to idealization; these will be the topic of the next section. However, before moving on, it's worth returning to the issue of *novelty*. One reason for our interest in phase transitions was that they were held to involve *novel behaviour*. Bob Batterman thinks phase transitions involve novelty in rather a strong sense; they involve physical singularities that are 'not there' in the finite statistical mechanical description. However, if we want to resist this (and the above discussion seems to suggest that resistance is far from futile), one might wonder whether all novelty is thereby dissolved. Jeremy Butterfield doesn't think so:

> I take emergence as behaviour that is novel and robust relative to some comparison class. I take reduction as, essentially, deduction. The main idea of my first rebuttal will be to perform the deduction after taking a limit of some parameter. Thus ... we can deduce a novel and robust behaviour, by taking the limit $N \to \infty$ of a parameter N.
>
> But on the other hand, this does not show that the $N \to \infty$ limit is physically real, as some authors have alleged. For my second main claim is that in these same examples, there is a weaker, yet still vivid, novel and robust behaviour that occurs before we get to the limit, i.e. for finite N. And it is this weaker behaviour which is physically real. [11, p.1065]

However, now the problem is that, once one has demonstrated that a given behaviour is to be expected even in the finite case, the obvious sense of novelty proposed by Batterman is no longer available. What then, do we mean when we say that phase transitions exhibit novelty? My answer here (which will be explained in detail in section 4) is that, qua part of thermodynamics, phase transitions exhibit explanatory novelty, where explanatory novelty is to be cashed out in the terms suggested in section one. I make no claim that this captures Butterfield's meaning in the above, nor that this explanatory novelty is the only kind of novelty to be found in phase transitions, but it at least provides one way of thinking about novelty in this reductive context.

3 A tangled web: abstraction, idealization and robustness

I am interested here in the change of variable that takes place when we express the thermodynamic free energy in terms of the statistical mechanical partition function. Considering phase transitions demonstrates that this change of variables is far from simple, and thus introduces a number of complications that were brushed under the carpet by the toy example of section 1. It's therefore worth making a few distinctions, and exploring the relationship between abstraction and explanation, and other concepts like idealization and robustness. One distinction, between idealization (or approximation) on the one hand, and abstraction on the other, is particularly relevant here:

- *Abstraction*: The elimination or omission of detail (often the omission of variables).
- *Idealization (and approximation)*: The introduction of strictly false statements (either via the introduction of a model (idealization), or by making false assumptions about the target system (approximation)).[8]

This distinction mirrors one made by Martin Jones [14], as well as by Ernest McMullin [19].[9] Because the above often go hand in hand, it is easy to lose sight of the distinction, but a little attention to the toy example of section one will demonstrate that not all cases of abstraction need involve idealization: in that case angular momentum was genuinely irrelevant to the linear trajectory of the ball, and thus one could abstract away from a full, detailed description of the system without any need for false assumptions. Likewise, taking, say, the mean kinetic energy of a system involves abstracting away from a great many details without necessarily introducing any idealization.

However, my interest here is as much in the relationship between the two concepts above as in their distinction. In the phase transition case, we have an idealization involving the introduction of a false statement: that the particle number is infinite. This idealization allows us to abstract away from various details

[8] Again, the difference between idealization and approximation can be of great importance, but I'll here use the term 'idealization' as a proxy for both.

[9] The terminology here is as tangled as the web itself, and I don't pretend that the terms I use are defined uniformly throughout the literature. McMullin discusses the difference between *Galilean idealization* and *Aristotelian idealization*, but the distinction is similar to the one above.

including the boundary conditions. Thus, in this case, idealization *facilitates* an explanatory abstraction.

An oft-used example can illustrate this neatly. Consider any inclined plane calculation that neglects friction, by modeling the phenomenon on a frictionless plane. Merely setting the frictional force to zero isn't abstraction; the frictionless plane may have zero friction, but the equations for friction can still be expressed, however trivially. Abstraction comes after the idealization, when we omit friction and its associated equations from our explanation of the motion of the body down the inclined plane. Idealization here acts as a precursor to abstraction. In real-life examples of explanatory abstraction idealization will often have a role to play in facilitating the relevant abstractions.

Let us turn now to this question of justification. In order to facilitate explanatory abstraction, idealizations must be justified. That is, it must be shown that the false assumptions made do not significantly impact on the features of the target system that are of interest. This is where a new concept, *robustness*, enters the picture. If a feature of a system is robust, this means that it is stable under perturbations of some other variables. In the phase transition example, in order to justify our use of $N \to \infty$, we need to show that the features revealed in this limit reflect features of the finite N system. We demonstrate this by demonstrating that the features revealed in the thermodynamic limit are robust: they are invariant under changes of N and under changes of the boundary conditions(as long as N is large enough). The $N \to \infty$ idealisation is justified just if the number of particles and shape of a kettle are irrelevant to phase transition behaviour. Robustness demonstrations are of course often difficult and very involved, even for well-understood cases; a great deal of work in both physics and philosophy can be seen as attempting to provide such demonstrations.[10]

Thus the picture we have in the case of phase transitions is one of an idealization that allows us to abstract away from detail. The idealization, and hence the abstraction, require justification via some kind of robustness demonstration; we possess some tools, such as the Ising model, for providing such a demonstration. What of explanation? In the case above, the abstraction is required within statistical mechanics for the explanation of phase transitions; even if we could solve the many-body problem for some particular system required to predict a phase transition, we wouldn't have revealed the kinds of broad structural features gen-

[10]Butterfield's well worked out examples in [11] can be seen as offering robustness demonstrations (among other arguments). Several of Batterman's examples also offer this kind of demonstration: see in particular his account of the rainbow in [5].

erally associated with phase transitions. I take it this scenario is quite typical: the full story within a given theoretical framework will often involve a combination of idealization and abstraction. The elimination of a variable will often only be possible if an effect is assumed to be small, or if a system is thought to be well idealized by some simpler system. As we eliminate detail in this way, we will often acquire better explanations; structural features of the theory once obscured by detail will become clear, and problems become more tractable.

However, nothing I've said here speaks to explanatory *novelty*. The above seems simply to be a case of abstraction. If this were all there was to our story about explanation, the answer to my question - *Can we account for the explanatory utility of the higher level explanation purely by noting that it is a distant abstraction of the more fundamental explanation?* - would be a straightforward *yes*. But, as the toy example of section one suggests, things are not so simple once a change of variable is involved. So far our discussion has stayed within the realms of statistical mechanics and explanatory novelty has not yet raised its head. I propose that explanatory novelty does emerge once we look at the move to thermodynamic variables.

4 Novel explanation in thermodynamics

Let us return to the relationship between thermodynamics and statistical mechanics. The last section told a complex story about the explanation of phase transitions within statistical mechanics. I'll argue here that the right view of this story leads to a picture of the relationship between statistical mechanics and thermodynamics that allows for explanatory novelty. One of the lessons we learn from considering phase transitions is that at least one of the variable changes involved in the bridge laws between the two theories is of a particularly complex and irreversible form. This change of variables itself involves abstraction. Novelty enters the picture when we consider an additional set of abstractions: those required when we explain things at the thermodynamic level. I'll concentrate here on the importance, within thermodynamics, of abstractions based on the work/heat distinction, and argue that they demonstrate explanatory novelty in the sense under discussion here: the abstractions used cut across the distinctions naturally made in statistical mechanics.

4.1 Changing variables

Our discussion above emphasised the description of phase transitions *within* statistical mechanics; they could be defined entirely via considering Z, the partition function. If this analysis is right, where does thermodynamics enter the picture? The answer is, rather late, only at the point at which we actually make the theoretical identification:

$$A_{TD} = -k_B T \ln Z_\infty \tag{13}$$

Here the subscript in Z_∞ is intended to indicate that this does not connect the thermodynamic free energy A_{TD} directly to the realistic, finite SM partition function, but rather expresses the free energy as a function of the idealized Z in the thermodynamic limit. I take it that this is at least a potential bridge law[11] linking statistical mechanics and thermodynamics, and, as such, represents a change from SM to TD variables. It's the bridge law we need if A_{TD} (which does have non-analyticities) is to be strictly equal to some function of statistical mechanical variables.

Nonetheless, there are problems with this identification, at least if it is supposed to hold for all systems: many systems don't possess a well-defined thermodynamic limit! There are several approaches one might take to this. One option is to deny that (13) ever gives the correct bridge law, and perhaps insist that the real bridge law is only an approximate one:

$$A_{TD} \approx A_{SM} = -k_B T \ln Z \tag{14}$$

Filling in this notion of approximate equivalence will, however, be difficult, and some might question whether a law like (14) has any place in a successful reduction. Another option is to insist that (13) only applies to systems with a well-defined thermodynamic limit, and hope that an alternative expression can be found for systems without one. The full bridge law would then be disjunctive, but I'll assume here that, if such an approach were successful, this very benign (and explicable!) disjunction needn't undermine the reductive project. However, I won't try to solve this problem here. The fact that the thermodynamic limit isn't well defined for many systems is an ongoing research problem in statistical mechanics; after all, without the thermodynamic limit, we have great difficulty

[11] In as much as I discuss reduction here, I have in mind a loosely Nagelian model of reduction as the deduction of one theory from another with the aid of (potentially quite substantive) bridge laws.

in describing phase transitions. Moreover, I don't intend to argue for the reduction of thermodynamics to statistical mechanics, merely to investigate a kind of explanatory novelty that might hold even if such a reduction succeeded.

With that in mind, let's consider what kind of change of variable (13) is, and what it might tell us about the relationship between statistical mechanics and thermodynamics if it were a correct bridge law. As a result of the appeal to Z_∞, this bridge law involves a *mathematically irreversible* operation; there is no way to derive the actual statistical mechanical partition function from the infinite idealization. This is a simple consequence of the fact that taking the thermodynamic limit involves the elimination of detail. Wherever a change of variables itself involves abstraction, a backwards move to our old variables will be impossible. As such, the kind of variable change expressed by (13) is very different from that expressed by (5) and (6), our momentum variable changes in the toy example.

What if (13) turns out not to be the correct bridge law, but some other is available? It seems fair to say that any variable change that can capture the fact that A_{TD} genuinely does have non-analyticities will also share this feature of irreversibility; any way of fleshing out the 'approximately equals' relationship will have to take into account the fact that A_{SM} is a more fine-grained quantity than A_{TD}, because it depends on features of the partition function washed out by the thermodynamic description. So even if (13) proves not to be the correct bridge law, we can expect a bridge law that is mathematically irreversible.

4.2 Work and heat in thermodynamics

Do our new variables play a role in explanatory abstractions whose value can't be understood from the perspective of statistical mechanics? The answer seems to be yes. Consider the following rather mundane question, and its answer within thermodynamics: *Why do diesel engines, unlike petrol engines, not need spark plugs?*

We answer this by considering the adiabatic compression of the gaseous air/fuel mixture in the combustion chamber. As we compress the gas the temperature rises according to the relationship

$$T_f = T_i \left(\frac{V_i}{V_f} \right)^{\gamma - 1}, \tag{15}$$

where $T_{i/f}$ is the initial/final temperature, $V_{i/f}$ is the initial/final temperature, and γ is the adiabatic index or heat capacity ratio of the gas, which measures the ratio

of its heat capacity at constant pressure to its heat capacity at constant temperature. We then note that the autoignition temperature of diesel is lower than that of petrol (210°C for diesel vs. 246-280°C for petrol). As a result, it possible to reach diesel's, but not petrol's, autoignition temperature with realistic engine compression ratios.

This example might be chosen almost at random from some vast list of thermodynamic explanations, for it displays a feature that's exceedingly common: it assumes that the process in question is *adiabatic*. Adiabatic processes are those that involve no heat transfer between the system and its environment; all energy transferred is transferred as work. If a process is fast or particularly well-insulated this is a reasonable assumption, and it's very widely used, especially in explanations of engine function.

Thus the value of this explanation, and others like it, rests on an abstraction - all details relating to heat transfer are excluded from the explanation. This abstraction results in a *better* explanation: the details pertaining to heat transfer would greatly complicate the equations, and fail to give additional insight into the phenomenon above. The abstraction depends on a distinction that is absolutely basic to thermodynamics; that between work and heat.

How does this explanation relate to a statistical mechanical description? There aren't any obvious problems with reduction here (beyond perhaps some general problems inherited from the incompleteness of the reduction of TD to SM); in fact, there exists a strangely simple reductive account of γ that relates it to the number of degrees of freedom in a given gas. Likewise, we do have something like definitions of heat and work in terms of statistical mechanical variables; as it happens, one way of getting a handle on the distinction in SM is via A, the Helmholtz free energy.[12] Recall our original thermodynamic equation for this:

$$A_{TD} = U - TS. \tag{16}$$

A here represents the amount of energy available for work, U is the internal energy of the system, and TS (temperature $\times$ entropy) represents something like the 'heat energy' of the system (only something *like* the heat energy, because heat only really makes sense as form of transferred energy). So our bridge law (13) (or whatever other bridge law replaces it) gives an account of a quantity related to work in statistical mechanical terms. Given that we also have expressions for

[12]The most direct way to think about the distinction between heat and work would be to take a closer look at the first law of thermodynamics. However, our expression for the free energy is intimately related to this, and we have already examined the form of the variable change involved.

entropy in statistical mechanical terms, the prospects for reduction look good.

Let us also assume that there isn't any problem *in principle* with describing the ignition of a diesel engine in more fundamental terms, via either classical mechanics or quantum mechanics, and applying statistical mechanical techniques. Doubtless such a description would be extremely complex, and very difficult to achieve in practice! However, if we were to possess such a description, I'd be happy enough to call it an explanation of the phenomenon. But it would certainly not be the *best* explanation of diesel engine behaviour: our thermodynamic explanation above seems to do a much better job.

Can we understand the value of this thermodynamic explanation from the perspective of statistical mechanics, even if we have both a reduction and a statistical mechanical explanation? I think not. The simplest way in which we might understand the explanatory value of the thermodynamic account from within statistical mechanics would be if we could understand the thermodynamic description as being a straightforward abstraction from the description proposed in more fundamental terms. However, this is exactly what the change to thermodynamic variables seems to block. By the lights of statistical mechanics, why should it lead to better explanation to perform abstractions based on quantities related to the logarithm of the partition function, or, worse, on the logarithm of the partition function in the infinite limit? When we rewrite the explanation above in terms of statistical mechanical variables, we can no longer see why we've thrown out some details, and kept in others.

One way of thinking about the unnaturalness of the work/heat distinction from the perspective of statistical mechanics is to think simply in terms of the kinetic theory of gases. Here, the distinction between work and heat must be something like the distinction between the kinetic energy associated with random, chaotic motion of molecules, and the kinetic energy associated with aggregate motions of molecules. Even if we can find some way of carving these up, it's hard to see how this could be a *natural* distinction; after all, from the perspective of the kinetic theory, it's all just kinetic energy.[13]

For this reason, it seems reasonable to say that the explanation offered by thermodynamics of diesel engine phenomena and the like is *novel*, in that its status as a *good* explanation cannot be understood from the perspective of the description given by the lower level theory. The abstraction involved in explanations involving adiabatic assumptions cuts across abstractions that might be natural from the perspective of statistical mechanical variables. Even if we have a theoretical re-

[13]I owe this point to Wayne Myrvold (by way of his lecture notes).

duction of the phenomenon, we can say that the thermodynamic explanation has irreducible, and thus novel, explanatory value.

How does this shed light on the importance of asymptotic relationships? The form of the bridge laws is highly relevant to the above conclusion. If the law involved a simple relationship between statistical mechanical variables, then abstractions based on the new variable might well look like helpful explanatory abstractions by the light of statistical mechanics: only a little back-translation would be required. It is precisely the complexity and mathematical irreversibility of a bridge law like (13) that leads to the conclusion that the explanatory value of the TD description is inexplicable from an statistical mechanical perspective. The fact that (13) involves a very strong idealization and a resulting high degree of abstraction[14] strengthens this point, and it's precisely the use of the asymptotic limit that establishes this idealization. Nonetheless, there's no suggestion here that this kind of explanatory novelty can only occur when an asymptotic limit occurs, nor that it's the only kind of novelty to be found in cases involving the asymptotic limit. Batterman, for example, claims that the asymptotic analysis itself is explanatorily novel.[15] Again, I would be very happy to see multiple clear-cut accounts of novelty in higher level theories, but the problem seems to me to be that clear accounts of novelty are hard to find. The argument here, then, at least puts one such account on the table.

5 Conclusions

I have presented two possibilities for accounting for the explanatory utility of higher level descriptions:

1. Higher level descriptions acquire explanatory utility only because they successfully abstract from a detailed, fundamental picture.

2. Higher level descriptions can sometimes be said to give novel explanations in a way in which mere abstractions from an underlying picture cannot.

[14] It's important to note here that abstraction has multiple roles to play in this story. The abstraction involved in taking the thermodynamic limit of the partition function count as an explanatory abstraction *within statistical mechanics*. The cross cutting thermodynamic abstraction that introduces explanatory novelty, on the other hand, is the abstraction based on the adiabatic assumption.

[15] Batterman makes this claim in his [6], when responding to criticisms from Gordon Belot [9]. However, it's not quite clear what he means by novel in this context.

An account that asserted (1) could nonetheless insist that higher level theories had considerable explanatory power, but there would be no deep difference in the *kind* of explanation they offered. Explanations given by TD, in so far as they were successful, would be valuable for the same reasons that the idealized Z_∞ is explanatorily valuable in SM. However, I have argued here that explanations offered by a theory like thermodynamics can be thought of as novel in a robust way; changes of variable induced by sufficiently complex bridge laws lead to new standards of abstraction, and thus novel explanatory strategies.

Is this conclusion a matter of metaphysics, or of epistemology? Weak accounts of emergence are often epistemic: they hold that emergence should be analysed as novelty or irreducibility relative to some given theory or state of knowledge. Such accounts are often contrasted with properly metaphysical emergence: this is the kind of account of emergence that posits novel causal powers, or irreducible properties that do not depend on our epistemic state. Where does the current account of explanatory novelty fit in this debate? One simple answer is that it fits badly. Certainly, to some ears, our discussion of explanatory novelty will have an epistemic overtone. However, in one sense, my account here is *weaker* than standard epistemic accounts of emergence; I argue here that a theory can possess novel explanatory power even when a reduction is not just possible but epistemically available! However, in another sense, my account of novelty is stronger than an epistemic account. The fact that certain choices of variable facilitate explanatory abstraction depends, I take it, on objective features of the world. The laws might have been such that abstraction was much less helpful; that micro-scale details might always have been important to phenomena. Equally, the laws might have been such that simple abstraction from detail was all that was necessary in order to capture the phenomena. But our world is not like that; it appears to be the kind of world in which there are macro-variables that do an excellent job of predicting and explaining phenomena despite being highly complex functions of micro-variables. And in some cases, some complex function of the micro-variables turns out to be irrelevant to the phenomenon at hand. Inasmuch as these features of the world are an objective matter, the novelty proposed here is metaphysical.

6 Acknowledgements

I owe particular thanks to David Wallace for comments, suggestions and discussion of a draft of this paper. Many thanks also to the YLWiP group for helpful discussion, especially to Adam Caulton, Nick Jones, and Sean Walsh. Versions

of this paper were presented in Leeds, in Bristol, to the BSPS, to King's College London, in Ghent, at the EPSA in Athens, and at Barnard College/Columbia University. Thank you to all those audiences for their suggestions.

References

[1] P.W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[2] R. W. Batterman. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction and Emergence*. OUP, 2002.

[3] R.W. Batterman. Explanatory instability. *Nous*, 26(3):325–348, 1992.

[4] R.W. Batterman. Asymptotics and the role of minimal models. *The British Journal for the Philosophy of Science*, 53(1):21, 2002.

[5] R.W. Batterman. Critical phenomena and breaking drops: Infinite idealizations in physics. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 36(2):225–244, 2005.

[6] R.W. Batterman. Response to Belot's" Whose devil? Which details?". *Philosophy of Science*, 72(1):154–163, 2005.

[7] R.W. Batterman. Idealization and modeling. *Synthese*, 169(3):427–446, 2009.

[8] R.W. Batterman. Emergence, Singularities, and Symmetry Breaking. *Foundations of Physics*, pages 1–20, 2010.

[9] G. Belot. Whose devil? which details? *Philosophy of Science*, 72:128–153,, 2005.

[10] J. Butterfield. Emergence, reduction and supervenience: A varied landscape. *Foundations of Physics*, 41:920–959, 2011.

[11] J. Butterfield. Less is different: Emergence and reduction reconciled. *Foundations of Physics*, 41:1065–1135, 2011.

[12] J. Butterfield and N. Bouatta. Emergence and reduction combined in phase transitions. *Arxiv preprint arXiv:1104.1371*, 2011.

[13] C. Callender. Taking thermodynamics too seriously. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 32(4):539–553, 2001.

[14] M.R. Jones. Idealization and abstraction: a framework. *Idealization XII: Correcting the Model. Idealization and Abstraction in the Sciences*, pages 173–217, 2005.

[15] C. Liu. Explaining the emergence of cooperative phenomena. *Philosophy of Science*, pages 92–106, 1999.

[16] C. Liu. Infinite Systems in SM Explanations: Thermodynamic Limit, Renormalization (Semi-) Groups, and Irreversibility. *Philosophy of Science*, 68(3):325–344, 2001.

[17] P. Mainwood. Phase transitions in finite systems. *http://philsci-archive.pitt.edu/8340/*, 2005.

[18] P. Mainwood. Is more different? emergent properties in physics. *http://philsci-archive.pitt.edu/8339/*, 2006.

[19] E. McMullin. Galilean idealization. *Studies in History and Philosophy of Science Part A*, 16(3):247–273, 1985.

[20] T. Menon and C. Callender. Turn and face the strange... ch-ch-changes: Philosophical questions raised by phase transitions. *http://philsci-archive.pitt.edu/8757/1/turnandfacethestrange.pdf*, 2011.

[21] J.D. Norton. Approximation and idealization: Why the difference matters. *Philosophy of Science*, 79(2):207–232, 2012.

[22] H. Putnam. Philosophy and our mental life. In *Mind, Language and Reality*, pages 291–303. Cambridge University Press, 1975.

[23] P. Railton. Probability, explanation and information. *Synthese*, 48:233–256, 1981.

[24] H.E. Stanley. *Introduction to phase transitions and critical phenomena*. Oxford University Press, 1971.

[25] M. Strevens. *Depth: An account of scientific explanation*. Harvard University Press, 2008.