Making Enough of a Difference: Graded Difference-Making and Levels of Explanation

Harjit Bhogal

October 29, 2018

Abstract

Here is a puzzle: Given we live in a physical world why are there are many phenomena that we typically do not explain in physical terms? For many phenomena it looks like the correct 'level of explanation' is not the fundamental physical level.

In the paper I give an account of what makes for the right level (or levels) of explanation. I do this by developing an account of *explanatory goodness*.

The account is based on the idea that explanations should cite the things that make a difference to the explanandum, but it avoids the problems that, I argue, face existing difference-making approaches to the levels question.

We live, or so I will assume, in a physical world. The basic constituents of the world are physical. Every phenomenon in the world is, ultimately, a physical phenomenon. However, there are some phenomena that we typically explain without explicit reference to physics. Instead we explain them in sociological, or economic, or biological terms and not in physical terms.

On the face of it, this is puzzling. If every phenomenon is ultimately physical, why do we explain many phenomena in non-physical terms?

One very natural thought is that pragmatic considerations provide the answer here. We just are not capable of enquiring into, for example, the economic systems that we are interested in by using physical vocabulary. The economic phenomena are so complicated when understood in physical terms that epistemic and computational limitations prevent us from, for example, doing the physics of exchange-rates.

There is clearly something to this. But it can't be the whole story – certain high-level explanations seem preferable to even the physical explanations we would give if the epistemic and computational limitations were relaxed. Consider, for example, the discovery of eighteenth century doctor John Arbuthnot that more males had been born than females in London in each of the last 82 years. There is a possible microphysical explanation of this fact, one which lays out separately all of the physical details which led to each male or female birth. But this does not seem like a good explanation.

Kitcher (2001, pp. 70-72) argues that 'even if we had this "explanation" to hand, and could assimilate the details, it would still not advance our understanding'. On the other hand, an evolutionary biological explanation does much better. The argument that there is an evolutionary tendency to produce a 1:1 sex ratio at the time of sexual maturity along with the fact that males are more prone to infant mortality than females, explains why there is evolutionary pressure for more males to be born than females. This evolutionary explanation is pitched at the right level, the microphysical explanation is not.

Another example: Consider an economic explanation. We can explain why the dollar declined in value by noting that the central bank cut interest rates thus reducing the incentive for investors to put their money in US accounts and so reducing demand for the dollar. This economic explanation gives us understanding in a way that no physical explanation could – explanations that talk about the movement of all the particles that make up the economic system seems, in a sense, unenlightening. The economic level seems to be the right one for explaining this phenomenon — the physical level is not the right level.

So, sometimes it seems that the best explanation for a phenomenon uses the tools and vocabulary of the higher-level sciences. That is to say, sometimes, the right 'level of explanation' for a particular explanandum is not the physical level. Why this is so has puzzled many philosophers — Franklin-

Hall (2016, p. 3) says that making sense of why the correct level of explanation is sometimes non-physical 'has been a kind of Holy Grail in the philosophy of science, long sought but never found'.

There is a difficult question, then, about why the right level of explanation is sometimes not physical. But this is an instance of a more general question: why is right level of explanation what it is? Why is the right level of explanation for the phenomenon of economy class being uncomfortable the economic level, rather than the physical level or the chemical level or the anthropological level?

This question of what makes for the right level of explanation is the one that I will take on in this paper. The way I'm going to do this is by focusing on one previous approach to this question about levels – the *difference-making approach*. This is a extremely attractive and intuitive approach but, at least as it has been developed so far, it fails to pick out the right levels of explanation. However, I'm going to build an account that retains the spirit of the difference-making approach while avoiding the problems faced by previous versions. To do this I will introduce two dimensions of *explanatory goodness* – two ways in which explanations can be better or worse – and argue that the trade off between these two dimensions picks out the right level of explanation.

1 The Difference-Making Approach

So let's start by considering the difference-making approach. It is, perhaps the currently dominant approach to the question of levels of explanation(see, in particular, Strevens (2008) and Woodward (2010). In this section I'm going to describe the approach and then note that it faces a significant problem.

The central idea is that explanations should cite all and only the things that make a difference to the truth of the explanandum. So, for example, the microphysical explanation of Arbuthnot's regularity —the fact that more males than females were born every year in London —, the one that lays out separately all of the physical details which led to each male or female birth, contains *too much information*, in particular, it contains information that does not make a difference to the explanandum.

There are different conceptions of exactly what makes something a difference-maker, but the intuitive thought is that a fact A makes a difference to B when A is relevant to whether to not B holds. One natural way to develop this thought is by saying that for two facts, A and B, A makes a difference for B if were not-A to hold then not-B would hold. Arbuthnot's regularity, counterfactually depends upon the fact that there are higher rates of infant mortality in males and so, on this conception, that fact counts as a difference-maker. The precise microphysical details of any particular birth, on the other hand, do not count as a difference-maker since if the details of any particular birth were different there would still be more males born than females. So we can start to see how the difference-making approach could favor higher-level explanations.

We shouldn't stick too closely to the conception of difference-making as counterfactual dependence — it is ultimately untenable for reasons closely related to the problems that face accounts of *causation* as counterfactual dependence. But it gives us the flavor of a difference-making approach.

The difference-making approach is, to my eyes, and the eyes of many other philosophers, extremely attractive. The idea that explanations should cite the things that are relevant to the whether the explanandum holds is very intuitive — after all not just any information is explanatory, rather the information needs to be targeted in a certain way at the explanandum. And it gives the start of a plausible answer to why higher-level explanations are sometimes to be preferred over lower-level explanations.

However, the difference-making approach faces a major problem. Put simply, special science explanations typically ignore lots of difference-makers — once we have given the special science explanation there are still additional physical facts that make a difference to the occurrence of the explanandum. So, the difference-making approach rejects these special science explanation as missing out on difference-makers.

The best way to see this is by focusing on an influential version of the difference-making approach given by Strevens (2008). (Though Woodward (2010) faces a very similar problem.)

Strevens' account of explanation starts with a set of low-level physical facts, including physical laws, that entail the phenomenon under investigation.[1] Roughly, the idea is to remove information from the premises piece by piece while still retaining the entailment of the phenomenon. For example, if the original premises included the facts that a particular object had mass of 10g and was moving at 2m/s then we might replace these facts with one stating that the object had momentum of over 15gm/s, if that fact is enough for the entailment to still go though. When the entailment has been abstracted as far as it can go then the premises that remain explain the phenomenon.

However, the demand that the explanans entails the explanandum means that the account has problems recommending explanations that are sufficiently high-level. Lots of higher-level explanations are such that the explanans does not entail the explanandum — there are ways that the explanans could be true that do not lead to the explanandum.[2]

Imagine, for example, I have some peas and I explain why they are smooth and not wrinkled by noting that they were bred from parent plants that were homozygous in the allele associated with smoothness. This is a good classical genetics explanation. But the explanans does not necessitate, and so does not entail, the explanandum — there are cases, like those where mutations occur, where the peas are bred from parents that are homozygous in the relevant allele but the peas end up wrinkled and not smooth. So Strevens's account does not count this as an explanation. The putative explanation misses out on difference-makers, like the physical facts that determine that a mutation did not occur in this case, and so doesn't count as a genuine explanation.

Imagine, for example, I have some peas and I explain why they are smooth and not wrinkled by noting that they were bred from parent plants that were homozygous in the allele associated with smoothness. This is a good classical genetics explanation. But the explanans does not necessitate, and so does not entail, the explanandum — there are cases, like those where mutations occur, where the peas are bred from parents that are homozygous in the relevant allele but the peas end up wrinkled and

[1] There has to be a restriction on the type of entailment here in order to avoid derivations – like that of the height of the flagpole from the length of the shadow – that are obviously not relevant to explanation.

[2] Notice that many other accounts of explanation — most obviously simple causal accounts, like that of Lewis (1986) — do not require such an entailment. See Woodward (2003, Ch. 4) for a detailed discussion of whether accounts of explanation should require entailment.

not smooth. So Strevens's account does not count this as an explanation. The putative explanation misses out on difference-makers, like the physical facts that determine that a mutation did not occur in this case, and so doesn't count as a genuine explanation.

Similarly, Strevens's account rules out many intuitively good high-level explanations. Consider, for example, the explanation of a particular person getting the job from the fact that they performed best at the interview. Or the explanation of the formation of a volcanic arc from the fact that one tectonic plate subducts under another leading to cracks in the crust of the upper plate. Or the explanation that dollar declined in value because the central bank decided to cut interest rates (thus leading to fewer investors putting their money in US bank accounts). Strevens' account won't allow us to give any of these explanations.[3]

Rather, in all these cases Strevens' account says that we should add in further physical information to the explanans — enough so that the explanans entails the explanandum. But this is just in conflict with scientific practice — the explanations that we actually see in the special sciences are, like the explanations mentioned above, not ones where the explanans entails the explanandum.

This isn't just a problem for Strevens — this type of issue affects difference-making accounts more generally. Take any explanation where the explanans doesn't physically necessitate the explanandum. Then there are physical facts that make a difference to whether we are in one of the cases where the explanans does lead to the explanandum, or one of the cases where it does not. And so, those physical facts make a difference to the explanandum.

Any explanation, then, where the explanans doesn't physically necessitate the explanandum leaves out physical facts that are difference-makers. And so, any such explanation is ruled out by the difference-making approach. But again, almost every explanation we actually see in the special sciences does not have the explanans physically necessitate the explanandum. And so the difference-making account is in conflict with scientific practice.[4]

[3] Strevens does go on to add further nuances to his account, allowing that in some situations explanations need not entail the explanandum – for example, he develops a theory of probabilistic explanation. But none of these seem to help in the cases under consideration.

[4] We would have to change the form of this argument very slightly to target Woodward's (2010) account of *proportion-*

However, the core idea of the difference-making approach still seems compelling – we want explanations that leave out that which is irrelevant to the explanandum, and it does seem like microphysical explanations of higher-level facts include things that don't make a difference. Is there any way that we can save the spirit of the approach?

Here is one attempt that seems very natural. Can't we avoid the problem for the difference-making account by relaxing the condition that explanations need to cite *all* difference-makers? That condition is the one that leads to the problem for the difference-making account, since higher-level explanations typically leave out some difference-makers.

But it's not enough to stop there. If we don't need to cite all difference-makers which ones do we need to cite? The most obvious answer is that we should cite the facts that make *enough* of a difference. So the defender of the difference-making approach might say 'Sure, low-level facts make a difference, low-level physical facts do make a difference to phenomena like the peas being smooth, for example, but they don't make *much* of difference, so it is reasonable to ignore them in giving our explanations.' Or alternatively, they might say, 'Sure, low-level facts *sometimes* make a difference, but they don't make a difference *most of the time*, so it is reasonable to ignore them in giving our explanations.'

Such thoughts seem intuitive, but on reflection is hard to know what to make of them. Let's start with the second thought. It is not true that the physical facts only sometimes make a difference to the smoothness of the peas. The facts that determine whether there is a mutation always make a difference — sometimes they lead to the peas being smooth and sometimes they lead to the peas being wrinkled, but, either way, it makes a difference that the physical facts are one way rather than another.

Now take the first thought. This also seems misguided, at least on it's most obvious interpretation. It seems false that the physical facts that determine whether there is a mutation don't make much of a difference to the smoothness of the peas. They make the difference between the peas being smooth

ality since his view takes *variables* to be the relata of explanatory change. But this is a simple fix.

and them not being smooth, surely this is a substantial difference.

So, is there no reasonable way of developing this idea that explanations should cite things which make enough of a difference? Should we just give up on difference-making providing the answer to the question about levels? I think the answer to both questions is no. We can still develop the idea that explanations should cite things which make enough of a difference and defend the spirit of the difference-making approach.

To do this, we needs a graded-conception of difference-making — developing this, as we have just seen, is tricky — and some sense of what makes for *enough* of a difference. In the rest of the paper I'm going to develop a view in the spirit of graded-difference-making and show how it can answer the question about levels of explanation.

But before I actually give my account, there is a little bit of groundwork to lay.

2 Explanatory Goodness and Explanatory Correctness

Let's start by distinguishing between *explanatory goodness* and *explanatory correctness*. An account of explanatory correctness is just an account of when something counts as an explanation and when it is not. And account of explanatory goodness is an account of when explanations are better or worse.

A difference-making component to explanation could be built into an account of explanatory correctness, as Strevens (2008) does, or added as a additional virtue or good-making feature of an explanation, as Woodward (2010) does. Since I'm am developing a sense in which difference-making can be graded, it's natural for me to take the latter strategy, since goodness can be naturally graded while correctness cannot

So, I'm going to develop my account as an account of dimensions of explanatory goodness — ways in which explanations can be better or worse. We will see how these dimensions of explanatory goodness capture difference-making intuitions and answer the question about levels of explanation.

But in developing the account we are going to need the notion of explanatory correctness too. We won't need any one specific account, just the generic notion of explanatory correctness will be enough. But for definiteness, and ease of presentation, I'm going to fix on a simple causal account of explanation, given by Lewis (1986). The account says that A explains E's occurrence if and only if A gives information about the network of causal relations that leads to E.

We should stop here to note one important constraint that was perhaps implicit in Lewis's account but it will be useful to make explicit: Very unnatural or disjunctive properties cannot be part of the explanans of correct explanations. For example, imagine we tried to explain the decline in the value of the dollar by appealing to the deterministic fundamental laws of nature, and the fact *the world was in A_1 or A_2...or A_n at it's initial time*, where A_1 through A_n are all and only the possible initial conditions that would, along with the deterministic laws, entail the decline in the dollar. This, I take it, is not acceptable as an explanation, because of it's disjunctive character. Neither would an explanation that introduces a single property, F, that replicates the disjunction by applying to all only cases where A_1 through A_n hold.

Note that ruling out these unnatural or disjunctive properties is something that everyone who wants to understand high-level explanations has to do. Some properties, the natural or projectable ones, can be used in other explanations, while the unnatural, disjunctive, or gerrymandered properties cannot. I'm not going to give an account of naturalness here though, that's a big project that's I take on elsewhere.

Whether or not you like this causal account of explanatory correctness doesn't really matter, as we will see more clearly later, the account of explanatory goodness is not committed to it. (It is committed to the claim that unnatural properties can't explain though.) But there is a dialectal reason for using this account: It's a very minimal account — it sets a very low bar for something to count as an explanation. Consequently, using this account of correctness makes very clear the all the interesting work for identifying the level of explanation is done by the account of explanatory goodness.

3 Robustness

Again, the aim is to develop an account driven by the thought that explanations should cite the things that make *enough* of a difference to the explanandum, and then to use this to answer the question about levels of explanation. And to do that we need to start by developing a graded conception of difference-making. As we saw in section 1, the most obvious ways to do this run into problems.

But there is, I think, a way to make sense of the the idea that some facts can make more of a difference and others less. To see this, consider a different example. Imagine that you drop an ice cube in warm water and the ice cube melts. You can explain the ice melting by citing the fact that it was dropped into warm water. Alternatively, you could (at least in principle) explain the melting of the ice cube by citing the precise microstate, call it M, of the ice-cube water system and showing how the physical laws lead from that to the melting. Something seems problematic about this second explanation though — it appears to be pitched at the wrong level, just as, for example, the microphysical explanation of Arbuthnot's regularity is pitched at the wrong level.

When such cases are discussed in the literature on statistical mechanics it is heavily stressed that most, in fact, nearly all, of the possible microstates of the ice-cube water system would lead to the ice melting when it is dropped in the water. Though not *all* microstates — there are some possible microstates that would lead to the ice cube *growing* when it is dropped in warm water, but those microstates are incredibly rare and unlikely.

The fact that the overwhelming majority of the microstates lead to the melting, it is often thought, provides a reason why the it is reasonable to appeal to the higher-level explanation that just cites the ice being dropped in the water rather than having to appeal to the precise microstate (see, for example, Meacham (2010, p. 1117), Albert (2000, pp. 150-151), (Maudlin, 2011, pp. 309-318)). I think this is right, the fact that most of the microstates lead to the melting tells us about why it is bad to give the microphysical explanation that cites M.

This is because there is a sense in which the particular microstate M which holds doesn't make much of a difference to the melting of the ice cube, because *if almost any other microstate held the ice cube* *would still have melted* after being dropped in the water. To put it another way, M holding is not *required* or *close to required* for the melting of the ice cube. If M had not held then the melting would almost certainly have still occurred.

We can generalize these ideas into a measure of graded-difference making. In particular, I propose a dimension of explanatory goodness called ROBUSTNESS. The central idea of ROBUSTNESS is that an explanation of B from A is better if more (in the sense of a higher proportion) of the ways that B could hold are such that, in that situation, A explains B.

We can translate this thought into the language of possible worlds to give the official formulation of ROBUSTNESS: An explanation of B from A is better if that explanation holds in more, that is, a higher-proportion, of the physically possible worlds where B holds.[5]

So, we can see that the microphysical explanation of the ice melting scores badly on ROBUSTNESS since there are many physically possible worlds where the ice melts but in very few of them is the melting explained by the holding of the particular microstate M. In fact, in only very few of the physically possible worlds where the ice melts does M hold. In the overwhelming majority of physically possible worlds where the ice melts this microphysical explanation does not apply; for most (in fact nearly all) of the ways that the melting could hold, M does not explain the melting.

The higher-level explanation which just cites the fact that the ice cube was dropped in warm water to explain the melting scores much better on ROBUSTNESS — lots of the worlds where the ice melts are such that the explanation is that the ice was dropped in warm water.

Similarly, take the microphysical explanation of the decline in the dollar — one that cites the movement of all the particles that make up the economic system. This explanation scores badly on ROBUSTNESS because the explanans of that microphysical explanation holds in very few worlds, so the explanation must also hold in few worlds. In the overwhelming majority of physically possible worlds where the dollar declines in value this microphysical explanation does not apply. The economic explanation does much better on ROBUSTNESS — lots of the worlds where the dollar declines in value

[5] This dimension is similar to the account of 'explanatory depth' given by Weslake (2010).

are such that the explanation is that interest rates declined. More generally, ROBUSTNESS will tend to disfavor explanations that are very specific, because those explanations can only hold in few worlds.

Again, we can see the intuitive force of ROBUSTNESS by seeing how it captures a graded sense of difference-making. Consider an explanation that scores well on ROBUSTNESS. If A explains B and this explanation scores very highly on ROBUSTNESS then A is close to required for B. This is because there are few situations where B occurs without A — in most of the physically possible worlds where B occurs it is explained by A, so there are only few physically possible worlds where B occurs and A doesn't.

To put it another way, if A does not hold then it will be rare that B holds since nearly all of the cases where B holds it is explained by A. A, then, makes a substantial difference to whether B occurs.

On the other hand, if A explains B and this scores badly on ROBUSTNESS then A is not close to required for the occurrence of B. If the explanation scores badly on ROBUSTNESS, typically, this is going to mean that A holds in few of the worlds where the B holds — there are many physically possible worlds where B occurs without A.[6] It is rare for A to be part of the cause that leads to B, most of the time when B occurs, it is caused by something else. In this sense, A does not make much of a difference to B. The microphysical explanation of the ice melting is a good example of this — it is rare for the specific microstate M to be part of the cause of the melting. M is not required, to close to required, for the melting — in this sense it doesn't make much of a difference to the melting.

Accepting ROBUSTNESS as a dimension of explanatory goodness is a way of cashing out the difference making intuition — it implies that explanations are better, at least along one dimension, if they cite facts that make more of a difference.

So we have a graded notion of difference-making. This is part of what we needed to fix up the difference-making account and so to answer the question of levels of explanation. But this isn't all

[6] This is only true 'typically' because it is possible that the explanans holds in a similar (or wider) range of worlds as the explanandum, but only explains the explanandum in a few of these worlds. In this case, the explanation would score low on ROBUSTNESS. But given the extremely minimal account of explanatory correctness we are working with, these cases will be very rare so, for the most part, I will ignore them in what's to come.

we needed. Our aim was to develop the idea that explanations should cite the things that make enough of a difference. So, now we need some grip on what is ‘enough’.

4 Precision

One option is to just have a threshold for ROBUSTNESS — we demand that explanations should score highly enough on ROBUSTNESS and so make enough of a difference. But lots of clearly bad explanations meet this criteria.

As we noted, ROBUSTNESS tends to favor less specific explanations over more specific ones. Consequently, lots of explanations can score well on ROBUSTNESS in virtue of being very *unspecific*. If A is part of the causal nexus that leads to B and holds in many of the worlds where B holds then the explanation of A from B will score well on ROBUSTNESS. Consider such an explanation of Arbuthnot’s regularity — the fact that more males than females were born in London every year for 82 years in the eighteenth century. It is part of the causal nexus that led up to this fact that there were humans in London at the start of the eighteenth century. And in most of the possible cases where Arbuthnot’s regularity holds there were humans in London at the start of the eighteenth century. So the explanation of Arbuthnot’s regularity that just says that there were humans in London at the start of the eighteenth century scores well on ROBUSTNESS.

In one sense this is the right result — the fact that there were humans in London at the relevant time clearly makes a difference to the truth of Arbuthnot’s regularity. But on it’s own this isn’t a good explanation — it is too unspecific; it misses out on almost everything of importance.

What exactly is wrong with very unspecific explanations like these? Intuitively, the issue is that just saying that there were humans in London at the start of the eighteenth century doesn’t do much to lead us to expect that more males would be born than females each year. Although it is part of the causal history of the births we can infer almost nothing about whether those births will be male or female from the fact that there were humans in London.

The idea that explanations should make the phenomena in question expectable is a classic idea in the philosophy of explanation, notably expressed by Hempel (1965) and Salmon (1989). Hempel's account, and many subsequent accounts, like for example, Strevens' account discussed in section ?? — take this idea of expectability to the extreme by saying that an explanans must *necessitate* the explanandum, so the explanandum is *guaranteed* by the explanans.[7] And it does seem like a virtue of an explanation if the explanans guarantees the explanandum. But even if an explanation does not have this feature, it is good if it approximates it — it is good, that is, if the explanans makes the explanandum more expectable. Very unspecific explanations do badly on this count.

In order to deal with this issue, I propose that we should accept another dimension of explanatory goodness: PRECISION. PRECISION will tell us what is bad with these very unspecific explanations. But also, our main idea was that explanations should cite things that make enough of a difference. Difference-making is measured by ROBUSTNESS; what makes for enough of a difference is given by PRECISION — explanations should score well enough on ROBUSTNESS whilst still scoring reasonably well on PRECISION. (As we will see, this isn't so easy, because in many cases there is a trade-off between PRECISION and ROBUSTNESS.)

The formulation of PRECISION is very similar to that of ROBUSTNESS, but the guiding idea is very different. The central idea of PRECISION is that an explanation of B from A is better if more (in the sense of a higher proportion) of the ways that A could hold are such that, in that situation, A explains B.

We can translate this thought into the language of possible worlds to give the official formulation of PRECISION: An explanation of B from A is better if that explanation holds in more, that is, a higher-proportion, of the physically possible worlds where A holds. (To compare, ROBUSTNESS said that an explanation of B from A is better if that explanation holds in more, that is, a higher-proportion, of the physically possible worlds where B holds.)

The evolutionary explanation of Arbuthnot's regularity which cites the greater infant morality of

[7] Or at least, the simple version of Strevens' account does this, ignoring the nuances mentioned in footnote 3.

males and the pressure towards a one-one sex ratio at the time of sexual maturity scores well on PRECISION because most of the physically possible worlds where there is greater infant morality of males are such that Arbuthnot's regularity holds — that is, more males are born than females each year — and this is explained by the greater infant morality of males.

On the other hand, the explanation of Arbuthnot's regularity which just cites the fact that there were humans in London at the start of the eighteenth century does worse on PRECISION. This is because there are many physically possible worlds where the explanans holds, that is, where there were humans in London at the start of the eighteenth century but where Arbuthnot's regularity does not hold, and so where the explanans doesn't explain the explanandum. For example, there are many physically possible worlds where females have greater infant mortality, and so in those cases Arbuthnot's regularity does not hold.

Or consider the ice cube case again: The explanation of the melting of the ice cube that just cites the fact that that you dropped it in warm water scores very highly (but not maximally) on PRECISION because in nearly every possible world where you do drop the ice cube in the water the dropping leads to it melting. But the explanation that cites the precise microstate M and the deterministic laws will score maximally on PRECISION[8]. Notice that PRECISION tends to favor more specific explanations over less specific ones.

One useful way to think about PRECISION is as implying a kind of graded necessitation. If A explains B and that scores highly on PRECISION that means that in most of the physically possible worlds where A holds, A explains B. And that implies that in most of the physically possible worlds where A holds, B holds. So, when the PRECISION of this explanation is high then A 'comes close' to necessitating B within the range of physically possible worlds.

And so now we can see how PRECISION captures the idea that explanations should lead us to expect the explanandum. If the PRECISION of an explanation is high then the explanans comes close to necessitating the explanandum, in the way described above, and so the explanans should lead us to

[8] Assuming determinism, which I will be throughout the paper

expect the explanandum.[9]

So, I claim, PRECISION is a second dimension of explanatory goodness. Explanation, I claim, should cite things that make as much of a difference as possible while still making the explanandum expectable. That is, they should score as well as possible on ROBUSTNESS while still scoring well on PRECISION.

These are the two dimensions of explanatory goodness. But how do they work together to get the plausible results regarding levels of explanation?

5 Specificity of Explanations

The first step in answering this question is to draw out the implications that my view has for the degree of specificity, or in other terminology – the degree of *abstraction* – that an explanation should have. That is the focus of this section. The second step, I'm going to consider the implications for the right level of explanation — that is, what tools and vocabulary we should use to explain which facts. I'll discuss that in the next section.

In order to see the implications of my view we need to appeal to the combination of PRECISION and ROBUSTNESS. I'm not, though, going to claim that there is a single unique way to weight these dimensions against each other. It is plausible that different scientific fields have different preferences for PRECISION and ROBUSTNESS. The typical physicist, I suspect, values PRECISION highly; the typical sociologist values ROBUSTNESS more. The typical biologist perhaps somewhere in between. I don't think any of these preferences are mistaken.

Relatedly, identifying a unique balance of PRECISION and ROBUSTNESS would lead us to identify a single correct level of explanation for each explanandum — one explanation that 'best balances' the

[9] In fact, PRECISION is a particularly attractive measure of this sense of expectability – more attractive, for example, than the probability of the explanandum conditional on the explanans. The expectability intuition is that the *explanation* should render the explanandum expectable. If B has a high probability conditional on A this might be true for reasons that are totally independent of any explanation of A from B. But, if the PRECISION of an explanation of B from A is high then this means that given A we should expect B *in virtue* of the explanatory connection between A and B.

two dimensions. But this would be the wrong result. An account of levels of explanation should not, for example, say that the psychological level is the right level at which to explain the phenomenon of loss aversion, for example, and the neuroscientific level is not. Scientific practice is pluralist about the level of explanation here, and so we should be too.

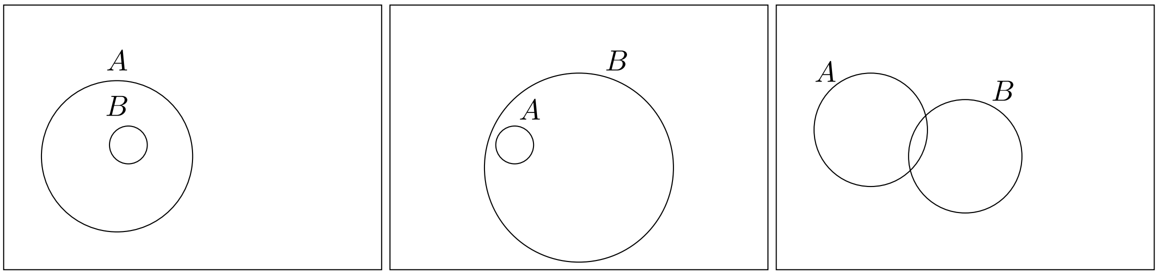
Rather, only very weak assumptions about how PRECISION and ROBUSTNESS weight against each other are needed in order for the account to get the right results about levels of explanation. All I need to do is rule out very extreme preferences for PRECISION over ROBUSTNESS, and vice versa. These extreme preferences exhibit themselves in cases where people would accept a big loss in PRECISION to get a tiny gain in ROBUSTNESS or vice versa. Of course, it's vague what counts as an 'extreme' preference. But hopefully our discussion going forward with only appeal to relatively clear cases.[10]

(There's a question about what this 'ruling out' comes to. Am I saying that those extreme preferences are irrational? I am not. Rather, the idea is that people typically do not have such extreme preferences, so in attempting to make sense of what levels of explanation we typically favor in what cases it is legitimate to ignore those preferences. If someone has a very extreme preference for PRECISION, say, and subsequently favored much more detailed explanations than we normally do, I would not claim that they are mistaken, rather they are just not in keeping with the explanatory standards of the community.)

So now we are in a position to see how PRECISION and ROBUSTNESS tell us how specific an explanation should be. The first step is to notice that the ideal case, where PRECISION and ROBUSTNESS are maximal, is one where the explanans and the explanandum 'overlap' in the range of physically possible worlds – that is, they hold in exactly the same range of physically possible worlds. To see this look at Figure 1.

In the diagrams the boxes represent the space of physically possible worlds and the circles represent

[10] A brief methodological aside: Ultimately, we are trying to understand and account for a feature of scientific practice — we are trying to make sense of the way certain levels of explanation are taken in scientific practice to be acceptable for a particular explanandum and some are not. This is not a particularly sharp phenomenon — there are no bright lines to be seen in the practice between explanations that clearly acceptable and those that are not. Disagreement and vagueness abound. So, we should not expect, or perhaps even desire, our philosophical account to be totally free of vagueness and to draw bright lines.



(a) The explanation will score low on PRECISION.

(b) The explanation will score low on ROBUSTNESS.

(c) The explanation will score low on both PRECISION and ROBUSTNESS.

Figure 1: PRECISION and ROBUSTNESS together imply proportionality

the range of worlds where the propositions A and B hold. In 1(a) A holds in a much wider range of worlds than B, and so an explanation of B from A is guaranteed to score low on PRECISION because in most of the worlds where A holds, A does not explain B.

1(b) is the opposite case. B holds in a much wider range of worlds than A and so an explanation of B from A is guaranteed to score low on ROBUSTNESS because in most of the worlds where B holds it is not explained by A. The microphysical explanation of the decline in the dollar is a case of this kind.

When PRECISION and ROBUSTNESS are both maximal A and B perfectly overlap. Deviations from this overlap reduce explanatory goodness. This, on it's own, tells us a lot about the level of specificity an explanation should have – the specificity of the explanans should match the specificity of the explanandum.

(This idea of 'matching' or 'proportionality' means that the account is in the spirit of Yablo's (1992) discussion of causation. However, my account doesn't put in the claim about matching by hand – rather it follows from independently motivated thoughts about explanations making the explanandum expectable, and citing things that make a substantial difference. And, more importantly, PRECISION and ROBUSTNESS allow us a measure of how things can be more or less proportional, thus allowing the account to be a graded difference-making account.)

In most realistic cases, though, we won't have perfect matching between explanans and explanandum. Let's consider a more realistic case then – take the pea case mentioned earlier: I have some peas, and I can explain why they are smooth and not wrinkled by noting that they were bred from parent plants that were homozygous in the allele R that is associated with smoothness. This is a good explanation, but it does not score maximally on PRECISION – there are cases where the peas were bred from parents that were homozygous in the allele associated with smoothness but where the peas do not end up smooth.

For example, there is always the possibility of mutations leading to the peas being wrinkled. Similarly, disease or dehydration, or a variety of other factors could lead to the wrinkling. Nevertheless PRECISION is high because such cases are rare – in most of the physically possible worlds where the peas were bred from parents that were homozygous in the allele associated with smoothness that fact explains the peas' smoothness.

We could give an alternative lower-level explanation that is designed to do better on PRECISION. For example we could give an explanation that cites the specific molecular details of the inheritance. This molecular explanation would have an advantage on PRECISION over the genetic explanation because it would rule out the possibility of certain mutations. But, since the genetic explanation already scores well on PRECISION this advantage will be relatively small. The molecular explanation, on the other hand, would score *far* worse on ROBUSTNESS than the genetic explanation – the huge amount of added detail needed to move from a classical genetic explanation to a full molecular account leads to a substantial loss in ROBUSTNESS. Unless we have very extreme preferences for PRECISION over ROBUSTNESS, then, my account is going to favor the genetic explanation over the molecular explanation.

Or, to put it another way, the molecular explanation includes lots of facts that don't make much of a difference to smoothness of the peas — the peas could easily still be smooth even if the precise molecular details were different — and these facts don't add enough to the expectability of the explanandum to be worth including in a good explanation.

Another example: Consider the case of the dollar declining in value that we discussed above. The microphysical explanation of the decline in the dollar – the one that that cites the movement of all the particles that make up the economic system does a little better than the economic explanation on PRECISION but far worse on ROBUSTNESS. That is to say, the explanation cites facts that don't make much of a difference, and this isn't compensated for by a big gain in PRECISION.

So we can start to see, then, how my account can pick out the right degree of specificity. The ideal case is one where the specificity of the explanans matches that of the explanandum. But in most cases we don't get this perfect matching and we have to weigh PRECISION and ROBUSTNESS against each other.

6 Levels of Explanation

We have now seen the full account of explanatory goodness; we have seen how it flows from a kind of graded-difference making view of explanation; and we have seen the implications it has for how specific explanations should be. So the main work of the paper has been done.

I want to end by doing something a bit more ambitious. Again, the previous section all about how my account picks out the right degree of specificity for explanations. But I want to take a step further and see what my account says about the right *level of explanation* – that is, the tools and vocabulary we should use in explaining particular facts. That's what I'll do here.

Consider a realistic special science explanation. Take, again, the economic explanation of the US dollar declining in value that cites the decrease in US interest rates, noting that this leads to a decreased preference of investors for putting their money in US bank accounts. We already saw that my account favors this explanation over an alternative physical explanation that cites the movement of every particle that made up the economic system. But, I'm going to argue that my account favors this economic explanation over *any* physical explanation, because there is no explanation of the decline in the dollar given at the physical level that scores similarly well on PRECISION and ROBUSTNESS. And so the economic level is preferred over the physical level for this phenomenon. More generally, this shows us how my account can favor one level of explanation over another.

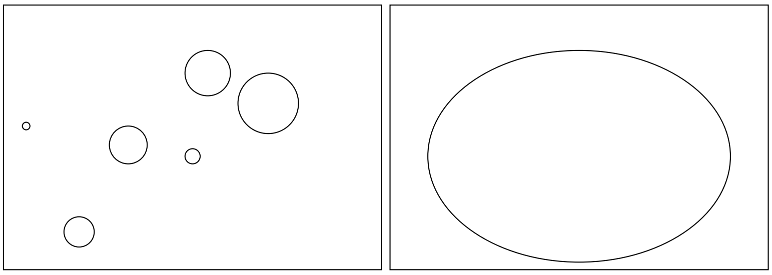
To start, notice that the fact under investigation – the decline in the value of the dollar – is radically multiple realizable. It could be realized by extremely different lower-level systems. For example, the decline in the value of the dollar in an economy similar to that of 19th century America would be realized by a few men changing their dispositions to trade currency. In an economy similar to our current, vastly more technologically complicated, economic system it is realized very differently. For example, part of the realizer of the decline in the dollar consists in the state of computer systems in foreign exchange markets in London, Frankfurt and Tokyo.

The type of multiple realizability pointed at here is not simply that there are certain facts which are consistent with more than one low-level realizer. That is a weak sense of multiple realizability in which everything but the most specific fundamental physical facts count as multiply realizable. Here I'm considering a stronger sense of multiple realizability – a fact is multiply realizable in this stronger sense if the possible low-level realizers are 'scattered' throughout the space of possible worlds.

The low-level realizer of the decline in the dollar in an economic system that resembles the modern US economy is extremely different from the realizer in a system similar to that of 19th century America. But further, these realizers don't share any (non-disjunctive) low-level feature that can be used to pick out all and only the cases where the dollar does fall in value. It's not the case that all the worlds where the dollar declines in value have total energy within a certain range, or anything like that. This is what I mean by saying that the realizers are scattered – they have no (non-disjunctive) low level feature in common.

Figure 2 illustrates this. Figure 2(a) depicts the possible realizers of the decline in the dollar as scattered throughout the space of physically possible worlds – there is no non-disjunctive low-level feature that is shared by just these disparate realizers. Figure 2(b) shows a fact that is not multiply realizable in this way. The fact holds in a huge number of physically possible worlds, so there are realizers that differ substantially from each other, but they are not scattered.

We can now see why any explanation of the economic fact in physical terms would do badly. For an explanation to score well on PRECISION and ROBUSTNESS the explanans must hold in many of the



(a) The realizers of the decline in the dollar are scattered.

(b) Some facts hold in many possible worlds but without their realizers being scattered.

Figure 2

cases where the explanandum holds but few of the cases where the explanandum does not. That is, there must a reasonable degree of overlap between explanans and explanandum. But given the scatteredness of the economic explanandum — that is, given the way that the decline in the dollar holds in a scattered range of worlds — any physical explanans would have to be extremely disjunctive in order to score reasonably well on PRECISION and ROBUSTNESS. And remember, as we noted in section 2 these very disjunctive or unnatural explanantia are not part of genuine explanations, so no genuine physical explanation will score well on PRECISION and ROBUSTNESS.

Figure 3 represents this. The red circles here represent possible low-level physical explanantia and the black circles again represent the scattered economic fact. Genuine, non-disjunctive, low-level explanations would be like one of the two cases represented here. Either the low-level explanation is like 3(a) and the explanans would hold in a much narrower range of cases than the explanandum, and so would score poorly on ROBUSTNESS. Or the explanation is like 3(b) where the explanans holds in many cases where the explanandum does not, and so scores poorly on PRECISION.

(There is a slight complexity here. We might worry that we have overestimated how scattered the realizers of the explanandum are because what we are trying to explain is a *particular* decline in the dollar. And we might think that any particular decline in the dollar that is in fact realized by the modern US economy could not be realized by an economy that looks like that of 19th century America – the way we individuate the event does not allow this. Rather, it has to be realized by

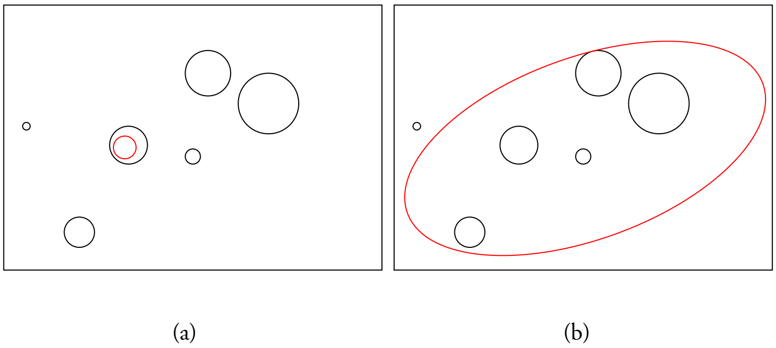


Figure 3

something at least roughly similar to the modern US economy. But even this more finely individuated event is still scattered across a wide range of worlds. The realizer of the decline in the dollar could partially consist of trades made predominantly in China, or in Germany, or other financial centers. And it could partially consist in decisions make by various different investment banks, with various different decision-making procedures. Even if these different possible ways that this more finely individuated fact could hold aren't as strikingly different as the modern economy and the economy that looks like 19th century America, they are clearly still different enough that an explanation of the fact in fundamental physical terms would have to be extremely disjunctive to score well on PRECISION and ROBUSTNESS.)

So, any physical level explanation of the decline in the dollar will score badly on explanatory goodness. The economic explanation that cites the drop in interest rates will do much better. This is because the economic explanans is itself scattered, and in particular, it is scattered in a way which, to a large extent, *matches the scatteredness of the economic explanandum*. Just like the decline in the value of the dollar, the drop in interest rates can be realized by economies which look like 19th century America and economies which resemble the modern US economy. So, the economic explanation can score much better than *any* physical explanation on PRECISION and ROBUSTNESS.

My account correctly favors the economic over the physical level of explanation here. And more generally, when we have good special science explanations of 'scattered' facts my account will favor those explanations over physical explanations for exactly the same reasons – genuine physical explanations of such facts will score badly on either PRECISION or ROBUSTNESS.

In the cases we have considered so far the explanandum was a higher-level fact. If the fact we want to explain is a low-level physical fact then it is clear that explanations in terms of higher-level facts will score extremely badly on PRECISION. A fact stated in macroscopic terms will not be able to give an explanation of a fact about a subatomic particle that scores acceptably on PRECISION. If we don't have extreme preference for ROBUSTNESS, then, we will prefer the explanation of facts about subatomic particles to come at a low-level.

These cases show how, after ruling out extreme preferences, my account can favor one level of explanation over another. However, my account will not, in general, pick out a unique right level of explanation; sometimes there are explanations at multiple different levels which will all score reasonable well on explanatory goodness. For example, there is nothing in my account to suggest that the neuroscientific explanation of an agent taking a certain bet must be preferred over the economic one. The neuroscientific explanation may do a better on PRECISION and the economic explanation better on ROBUSTNESS but there is no reason to think that one is acceptable and the other is not. But, as we noted, this type of pluralism looks like the right result.

7 Conclusion

One final thing to note: Now that we have the full account in view, we can see that it is not committed to the specific causal account of explanatory correctness that we were working with. Both PRECISION and ROBUSTNESS were defined in terms of the generic notion of explanatory correctness, not the particular causal account.

We can conjoin the account of explanatory goodness with a variety of different accounts of explanatory correctness and it will have the same implications for levels of explanation as when we appealed to the minimal causal account. In particular, whatever account of explanatory correctness we use it will have the implication that 'A explains B' entails both A and B. This implication means that if PRECISION and ROBUSTNESS are maximal then A and B fully overlap in physically possible worlds. Further it means that if B does not hold in many worlds where A holds then PRECISION will be low; and if A does not hold in many worlds where B holds then ROBUSTNESS will be low. It was these claims that did the work in our reasoning about levels of explanation. So the results of the paper are not tied to the minimal causal account of explanation.

We started out with a puzzle: Even though the world is ultimately physical, the correct level of explanation is not always the physical level. The difference-making approach to this question is very attractive, but the way it has been developed so far has been unsatisfactory – it leads us to reject many good high-level scientific explanations.

But we don't need to give up the idea that difference-making is the key to understanding levels of explanation. We can develop the idea of graded-difference making and the intuition that explanations should cite facts that make enough of a difference. I have done this by introducing two dimension of explanatory goodness, one of which is a graded measure of difference making, and arguing that the trade-off between these dimensions has implications for what level we should pitch explanations. We are not left with a unique best level of explanation for each phenomenon, but we are left with an account that makes sense of the which explanatory approaches we find in science and in ordinary life and which we do not.

References

Albert, D. Z. (2000). *Time and Chance*. Harvard University Press.

Franklin-Hall, L. R. (2016). High-level explanation and the interventionist's 'variables problem. *British Journal for the Philosophy of Science 67*(2), 553–577.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press.

Kitcher, P. (2001). *Science, Truth, and Democracy*. Oxford University Press.

Lewis, D. (1986). Causal explanation. In D. Lewis (Ed.), *Philosophical Papers Vol. Ii*, pp. 214–240. Oxford University Press.

Maudlin, T. (2011). Three Roads to Objective Probabilty. In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*. OUP.

Meacham, C. (2010). Contemporary Approaches to Statistical Mechanical Probabilities: A Critical Commentary–Part I: The Indifference Approach. *Philosophy Compass*.

Salmon, W. C. (1989). 4 decades of scientific explanation. *Minnesota Studies in the Philosophy of Science 13*, 3–219.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press.

Weslake, B. (2010). Explanatory depth. *Philosophy of Science 77*(2), 273–294.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy 25*(3), 287–318.

Yablo, S. (1992, April). Mental Causation. *The Philosophical Review 101*(2), 245–280.