SOMETIMES PROOFS ARE NOT EXPLANATIONS

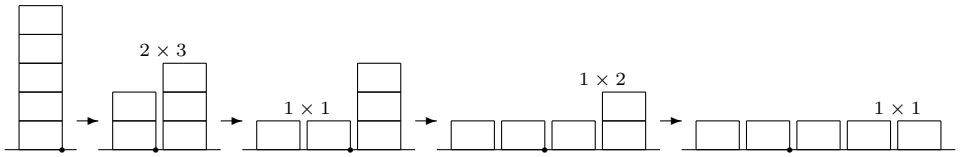
CHRISTOS KAPOUTSIS

Abstract. To wonder why a theorem is true before reading a proof of it is normal. To wonder why a theorem is true after we have read somebody else's proof of it is a little strange. But to wonder why a theorem is true after having actually proved it on our own, that is very weird. Still, it happens. We present a simple example of this phenomenon and attempt an analysis. Although our first steps lie within the theoretical framework of why-questions, we quickly need to pay attention to the extra-logical features that a mathematical proof necessarily has as an entity upon which a reader's mind computes.

1. Introduction

1.1. **A game.** On the top of a table lies a stack of b boxes. You want to break this stack into b stacks of 1 box each via a sequence of legal moves. A legal move consists of picking one of the stacks currently on the table and breaking it into two smaller stacks. Every such move is worth a number of points, which is equal to the product of the sizes of the two stacks that it produces. Your final score in this game is the total number of points that your moves are worth.

For example, starting with $b = 5$ boxes, one of the many ways you can reach 5 stacks of height 1, is to proceed as shown in the following sketch:



Your first move creates a new stack out of the 3 topmost boxes of the original stack, a move that is worth $2 \times 3 = 6$ points. The three following moves are worth 1, 2, and 1 points, respectively, for a final score of $6 + 1 + 2 + 1 = 10$.

How should you play this game in order to achieve the highest possible final score? What is the best strategy? Please spend a minute trying to answer this question before reading on.

1.2. **A sum.** What is the sum of the first n natural numbers? A nice closed-form formula for this sum is well-known and easy to prove. We state and prove this little theorem now, before using it in later sections.

Theorem 1. *For any $n \geq 1$, we have $\sum_{i=0}^{n-1} i = \frac{n(n-1)}{2}$.*

Proof. By induction on n. For the base case, we have $\sum_{i=0}^{1-1} i = 0 = \frac{1(1-1)}{2}$, as required. For the inductive step, we assume that the claim holds for some

Date: Spring 2005.

$n \geq 1$ and try to prove it for $n+1$. We calculate:

$$\sum_{i=0}^{(n+1)-1} i \;=\; \sum_{i=0}^{n} i \;=\; \Big(\sum_{i=0}^{n-1} i\Big)+n \;=\; \tfrac{n(n-1)}{2}+\tfrac{2n}{2} \;=\; \tfrac{(n+1)n}{2},$$

where the third equality holds because of the inductive hypothesis. Hence, the claim is true for $n+1$, as well, which completes the induction. □

1.3. **A basket of apples.** Consider the following setting:

Bob is eating apples from a basket that originally contains 10 of them.

Now consider the following problem with respect to this setting:

When the basket gets empty, how many apples will Bob have eaten?

Please solve this problem before moving on.

You are probably surprised I call this question a problem. Maybe you have read the setting and the question twice, to make sure you have not misread anything. Or you may think there is a catch. No. This is an honest problem, with an easy solution: by the time the basket is empty, Bob will have eaten ten apples. So then, what is the point?

This problem is extremely easy. You would probably say it is a *no-brainer*. I agree, but I think I have in mind something much closer to the literal meaning of the phrase than you do. Not, of course, that somebody without a brain could solve this problem. No, because one would need her brain to understand the question, in the first place. And that is the point.

Here is a plausible, I think, way to describe how easy this problem is.

First, we read the description of the setting. Then, we *understand* it. This certainly involves some computation in our minds, and one could speculate on what this computation consists of. Maybe we identify in our long-term memory objects that look like "basket", "apple", "Bob", and so on, and fetch them into our short-term memory for further processing. We probably also identify the process of "some original quantity being consumed", and bring this into our short-term memory for further processing, as well. Note that we do not need to have seen a basket of apples before and a Bob eating from it. It is enough that we know that people have names and that "Bob" can be one. And it is enough that we have experienced ourselves eat from a bowl spoonful after spoonful, walk home one step at a time, spend our stipend dollar after dollar, live through the day hour after hour. Overall, analogies are made between the objects and processes in the setting and similar objects and processes stored in our long-term memory. These are then brought into our short-term memory and computed upon in order for the meaning of the setting to emerge.

I suggest that, during this computation that assembles the meaning of the setting, the answer to the question "When the basket gets empty..." emerges, too, as an unavoidable by-product. In other words, it is hard for us to understand the setting without also producing somewhere in our short-term memory the prediction that eventually Bob will eat all ten apples and the basket will get empty.

Now, having understood the setting, we read the question; and understand it, by assembling its meaning in a similar way. When this is over, we attempt to find the answer. But the answer is already there. It has been lying in our

short-term memory all this time. Finding it is more like a look-up operation rather than a cumbersome search. In particular, no further analogies are necessary and our long-term memory does not need to get involved. So, the problem is a no-brainer because we have spent our brains understanding it rather than searching for its solution.

Of course, the details of the speculation above are almost certainly wrong, if they mean anything at all. They are there just to fill in the gaps of a story that I had to say to adequately describe the much weaker suggestion: That, in attacking a problem, we first spent some effort trying to understand the setting that gives rise to it and then some effort trying to find the solution. That, in the first stage, we unavoidably make some steps toward the search for the solution, by involuntarily producing facts which are not strictly necessary for understanding the setting. And that one reason why a problem may be very easy is that these unavoidable steps immediately consume all there is to the problem.

2. Dissecting the Game

Let us return to the game of Section 1.1 and to our question there:

What is the best strategy for playing this game?

If you have tried it, you may have observed a quite strange phenomenon: several different strategies yield the same final score! This is no coincidence.

Theorem 2. *The final score is independent of strategy.*

Proof 2A. I will prove a stronger claim, that for b boxes in the initial stack:

no matter how you play on b boxes, your final score will be $\frac{b(b-1)}{2}$.

The proof is by (strong) induction on b. The base case $b = 1$ is simple: the game ends before you make any move, so your final score is 0, or $1(1-1)/2$.

For the inductive step, I start by assuming that, for any $a = 1, 2, \ldots, b-1$, any strategy for playing the game on a stack of a boxes will bring you $a(a-1)/2$ points. Then, I consider the game on a stack of b boxes: No matter what your strategy is, you will start by breaking this initial stack into two stacks of sizes x and $b-x$, respectively, for some $1 \leq x \leq b-1$; you will then proceed to break the two new stacks until every box on the table is in its own stack. Your first move will be worth

$$x(b-x)$$

points, while breaking the two new stacks will bring you, respectively,

$$\frac{x(x-1)}{2} \qquad \text{and} \qquad \frac{(b-x)(b-x-1)}{2}$$

points (by the inductive hypothesis). Therefore, your final score will be

$$\begin{aligned} & x(b-x) + \tfrac{x(x-1)}{2} + \tfrac{(b-x)(b-x-1)}{2} \\ & \qquad = xb - x^2 + \tfrac{x^2}{2} - \tfrac{x}{2} + \tfrac{b^2}{2} - \tfrac{bx}{2} - \tfrac{b}{2} - \tfrac{xb}{2} + \tfrac{x^2}{2} + \tfrac{x}{2} \\ & \qquad = \tfrac{b^2}{2} - \tfrac{b}{2} = \tfrac{b(b-1)}{2}, \end{aligned} \tag{1}$$

as required. This concludes the induction, and the proof of the claim. □

Now, is this a good proof? There is no doubt that it meets all reasonable standards of a mathematically *rigorous* proof. But is it a *natural* one, in the sense that after you have read it you are left with no questions whatsoever about it? Put another way, suppose you try to reproduce this proof, to tell yourself the same explanatory story as to why all strategies achieve the same score. As you follow the reasoning of the proof, is every step going to look like the most natural next step you could make at the particular point? Or is it the case that for some steps you will not be able to immediately justify why you are making them? Try this experiment.

2.1. **Where did *that* come from?** The complaint that you probably have is that the number $b(b-1)/2$ is brought into the discussion with absolutely no justification. Prior to its introduction, there is nothing to suggest that the invariant value of the final score should be this particular one. It looks like somebody who knew more has tipped me to it. As Polya [3] would say, the value enters the proof as a "deus ex machina".

Question 1. *Where did the value* $\frac{b(b-1)}{2}$ *come from?*

Answer. To answer this question, I retrace my thoughts in the construction of the proof. First, the suspicion that all strategies yield the same final score, naturally led me to the assumption that knowing this invariant score as a function of b would help me prove the theorem. Then, to find this function, say $f(b)$, I knew it was enough to apply *any* strategy and calculate what final score it yields. But since any strategy would do, I naturally picked the *simplest* one: "as long as the original stack has 2 or more boxes, remove its topmost box". Under this strategy, f clearly satisfies the recurrence

$$f(b) = (b-1) + f(b-1),$$

since you first earn $1 \times (b-1) = b-1$ points by removing the topmost box from the b-tall stack and then another $f(b-1)$ points for bringing down the remaining $(b-1)$-tall stack. Applying this recurrence repeatedly, I got

$$\begin{aligned} f(b) &= (b-1) + f(b-1) \\ &= (b-1) + (b-2) + f(b-2) \\ &\cdots \\ &= (b-1) + (b-2) + \cdots + 1 + f(1) \\ &= (b-1) + (b-2) + \cdots + 1 + 0, \end{aligned}$$

which is exactly $b(b-1)/2$, according to Theorem 1. □

So, the mystery behind the discovery of the value of the final score is resolved. Hopefully, no other mysteries remain as to how I found this proof: Experimentation led me to the suspicion that all strategies produce the same score. This suspicion led me to the discovery of an expression for this score in terms of b. Then, I could restate the theorem as a claim about the natural number b. At that point, induction was a natural technique to apply. And applying it could not have been easier.

Overall, there is now no difference between you and me. Previously, I was the person that had built the proof and you were a person that had simply read it. Naturally, there was stuff that I knew but you did not.

As a result, although you were convinced that the theorem was true, you remained (partially) puzzled as to how the proof was found. That is, you were in the position of a *reader* who is *convinced* but *puzzled*. Resolving this puzzlement required some (easy, but still) non-trivial reasoning, that you either discovered on your own or you found in the answer to Question 1.

Let us use the name Proof 2A′ for the full argument that we get by incorporating the answer to Question 1 into Proof 2A. Then, what I just said in the previous paragraph is that with respect to Proof 2A′ I have no more information than you do; that is, information-wise we are both in the position of the person that has discovered the proof. So, it is no more "you, me, and Proof 2A". Rather, it is "us and Proof 2A′". And not only are we both convinced that the theorem is true, but we are also both free of any *puzzlement as to how the proof of the theorem was found*.

At this point, it is tempting to conclude that we are also free of any *puzzlement as to why the theorem is true*. After all, what else is there to a theorem other than its proof? If we have been able to prove it, then we cannot possibly be puzzled as to why it is true, can we? Well...

2.2. **But why is it true?** Why is the theorem true? Is there a single *reason* which we can put our finger on and say "That's why!"? Or, if no single reason is behind it, is there a *succinct story* that will carefully select those aspects of the world that make it necessary for the theorem to hold?

It probably sounds strange that I am asking this question right after we have finished a proof of the theorem. Isn't this proof already the story that I am after? Not really. To see what I mean, try removing from this proof all unnecessary mathematical formalities and focus on the heart of its argument: what does it say? The reason why our strategy does not matter when we start with b boxes on the table is because

> it does not matter on the 1-box stacks that we will eventually create; and, because of this, it does not matter on the 2-box stacks that we will create, either; and because of these two facts, it does not matter on the 3-box stacks that we may create; and so on, up to the b-box stack that we start with.

Are you happy with this story? I am not. Although I do understand that the irrelevance of strategy is a property that propagates from shorter to taller stacks, I still do not see how each step in this propagation happens.

My discomfort is not surprising, of course, as this little story has completely ignored the content of the inductive step. We cannot possibly achieve understanding without incorporating that argument. Let us try. What does the argument say? It says that, no matter how we break a stack of b boxes, the $x(b-x)$ points from our move will nicely combine with the scores $x(x-1)/2$ and $(b-x)(b-x-1)/2$ that we will earn from the two new stacks, so as to build a total that is exactly as claimed.

But how does *this* happen? I can see that the math works out in Equation (1), but why does it? Isn't it strange that these three quantities 'conspire' the way they do? In particular, notice how conveniently the several occurrences of x in (1) end up canceling each other, leaving the final score independent of the way the first stack was split. What kind of a coincidence is this? Can you explain it? I cannot. To me, it is a mystery.

Hopefully, you agree with me that there *is* something here to discuss. And you also feel the discomfort that, somehow, the reasons why the theorem is true have been screened from us. That, natural as it may be, Proof 2A′ is not telling us what is really happening. I will assume that you indeed feel this way, and I will continue discussing 'our' puzzlement about the situation. If this is not the case, I can only suggest that you read on: most probably, the point will be clearer at the end of the next section. Either way, it is important to mention that I am not inventing these issues as part of some strange philosophical scheme. This is the sincere puzzlement that students have actually had after being exposed to Theorem 2.

So, here is a summary of our situation:

- We have a *rigorous* proof. We are convinced the theorem is true.
- We even have a *natural* proof. One that, apart from establishing the truth of the theorem, also resolves all questions as to how itself was discovered. We (could) have proved the theorem by ourselves.
- We still do not *see* why the theorem is true. We remain puzzled.

Overall, we are in the position of a *prover* who is *convinced* but *puzzled*. We encode this strange situation in the following question.

Question 2. *But why is the theorem true?*

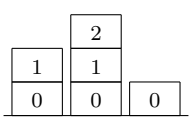
Notice that our puzzlement is, in fact, as serious as it would have been if, before finding or reading Proof 2A′, we were told by some oracle that we trust that the theorem is indeed true.

2.3. **That's why!** I am about to give a different proof of Theorem 2. But I believe this proof is already in your mind, waiting for the correct questions to put it together. So let me first ask you these questions.

- What is the *height* of a box? That is, if you freeze the game at any particular point, what number would you assign to each box in order to appropriately describe how high on the table that box is?
- What is the *total height* of the b boxes on the table?
- What is the total height when the game starts? What is the total height at the end of the game?
- How does the total height change in each move? Does it increase or does it decrease? By how much?

Try to answer these questions before moving on. The hope is that the proof will then pop out. If not, you should just read the next paragraph.

Proof 2B. Consider a configuration of the boxes on the table. For any given box, define its *height* to be the number of boxes below it in its stack. Define the *total height* on the table to be the sum of the heights of all boxes. For example, in the configuration



the labels indicate the heights of the boxes; the total height is 4. (Can you now complete the proof without reading on? Try it!)

Now consider a single move in the game, where the x topmost boxes of a stack of size m are removed to form a new stack. What change does this cause in the total height? Clearly, only the height of the x boxes that are moved is modified. For each of them, it decreases by an amount equal to the number of boxes that remained in the stack; that is, by $m - x$. Hence, the total height decreases by

$$x(m - x).$$

Which is exactly the number of points that the move is worth!

Therefore, in each move the player increases his score by exactly the amount by which he decreases the total height. Hence, in the end his final score will equal the overall decrease in total height. But this overall decrease is independent of strategy: no matter how the player plays, the original and final configurations of the game are fixed. □

So, here is a succinct story that I believe adequately answers Question 2. The reason why our strategy in this game will not matter is because

> no matter how we play, in each move we increase our score by as much as we decrease the total height of the configuration; so, our final score will necessarily be the difference between the initial and the final total height.

Are you satisfied with this story? I am, and I hope so are you. The next section tries to explain what makes this story a satisfactory one. Before that, let us finish with a couple of remarks.

First, as already mentioned in the little story above, the new proof can immediately give us the value of the invariant score.

Corollary. *The common score of all strategies is* $\frac{b(b-1)}{2}$.

Proof. In the starting configuration, all b boxes are on the same stack, so their heights cover all numbers in $0, 1, 2, \ldots, b-1$, and hence the total height is $b(b-1)/2$, by Theorem 1. In the final configuration, every box is on its own, so that all boxes have height 0, and thus the total height is 0. Therefore, the overall decrease in total height in the course of the game is $b(b-1)/2$. □

The second remark addresses the concern that the introduction of the concepts "height of a box" and "total height on the table" is as unnatural as the introduction of $b(b-1)/2$ in Proof 2A.

Question 3. *Where did the concept of "height" come from?*

Answer. Again, I retrace my thoughts in building Proof 2B. After Proof 2A, I already knew that the final score was always $b(b-1)/2$. Searching for a 'better' proof, I tried to identify this quantity on the table as the value of some 'natural' magnitude there:

where on the table can I find the value $b(b-1)/2$*?*

In this search, it was natural to think of the well known fact of Theorem 1: if I could identify the numbers $0, 1, 2, \ldots, b-1$ somewhere on the table, I would also have the required value, as their sum. So, the question became:

where on the table can I find the numbers $0, 1, 2, \ldots, b-1$*?*

But then, since there were b boxes on the table, it was natural to try to identify each of these numbers as the value of some property of one of the boxes. The concept of "height of a box" was the first idea. □

2.4. **The second proof is better.** What makes Proof 2B so much better is of course the introduction of the concept of 'height'.

The new concept becomes the center of the entire network of concepts that the game gives rise to: a configuration of the boxes is now only a number, the total height that it defines; a move by the player is a mere number, too, the decrease in total height; this same number is also the player's gain. By understanding what the 'height' of a box is and how it behaves throughout the game, one almost immediately has a grasp of all aspects of the game that are relevant to the question why the final score is independent of strategy. In effect, the new proof *tightly reorganizes the discussion* around a well-defined center and in a way that distinctly reveals the connections among the different relevant concepts.

However, it is as important that the new center of the discussion is very *intuitive*, as you may have already verified on your own by successfully answering the questions in the beginning of Section 2.3.

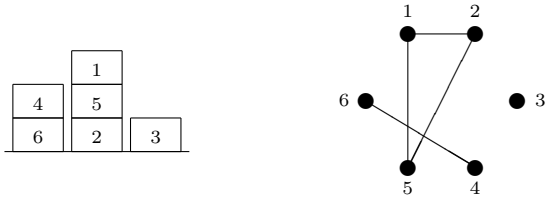
Understanding what the height of a box is involves no thinking at all. It is as simple as 'how high the box is'. When Proof 2B defines it, it does not really describe a property of the boxes that we did not know of; it simply makes us aware of its importance in the discussion. The property has always been there and, after we are reminded of it, it is almost impossible to keep thinking of the boxes without also having their height in mind.

Moreover, the behavior of the property is equally familiar. That splitting a stack will somehow decrease the height of some of its boxes and that at the end of the game the total height will be 0 are facts that we know before we think about them. Equally familiar is the interaction of this behavior with the behavior of our score: some initial total height is being gradually transferred from the table into our score until it is completely consumed. Rings a bell? Put ten apples and Bob's stomach in the place of the initial total height and our score and you will see a process that in Section 1.3 we agreed we can reason about almost without using our brains.

These two attributes of Proof 2B, the *tight reorganization* of the discussion around a very *intuitive* new center, seem (quite vaguely, sure enough) to explain why we consider it a significantly better proof. We will clarify and elaborate in later sections. Before that, another interesting point is due.

2.5. **No, *that*'s why!** Coming up with the notion of 'height of a box' is only one way to start an explanatory proof of Theorem 2. Here is another.

Proof 2C. Assume an arbitrary (but fixed throughout the game) numbering of the boxes from 1 to b and consider a configuration of them on the table. We define the *adjacency graph* for this configuration to be a graph with vertices $\{1, 2, \dots, b\}$ and an edge between vertices i and j if and only if boxes i, j belong to the same stack. For example, for the numbering indicated by the labels on the boxes, the configuration on the left has the adjacency graph shown on the right:



Note that, at the start of the game all boxes are in the same stack, so the adjacency graph contains all possible edges. In contrast, when the game is over, each box is in a stack of its own, so the graph contains no edges at all.

Now, consider a move in the game, where the x topmost boxes of a stack of size m are removed to form a new stack. How does this change the adjacency graph? Clearly, no edges are added, but some are removed. In particular, an existing edge between vertices i and j is removed if and only if boxes i and j are separated by the split; that is, if and only if one of them is among the x topmost boxes of the stack being split and the other one is among the remaining $m - x$. Hence, the number of edges being removed equals the number of couples of boxes being separated; that is, equal to

$$x(m - x).$$

Which is exactly the number of points that the move is worth!

Hence, each move increases the player's score and decreases the number of edges in the adjacency graph by the same amount. So, the final score of the player has to equal the overall decrease in the number of edges. But this overall decrease is determined by the starting and final graphs only, and these do not depend on how the player plays. □

How does this new proof compare to the previous two? Clearly, it is as rigorous and establishes the truth of the theorem as strongly. But how good is it in terms of explanatory power? You hopefully agree with me that it is clearly better than Proof 2A (for the same reasons that made Proof 2B better as well), but not as satisfactory as Proof 2B. We will discuss this latter comparison in the next section. Before that, let's finish with two remarks similar to those that followed Proof 2B.

First, the new proof also gives us the invariant value of the final score.

Corollary. *The common score of all strategies is* $\frac{b(b-1)}{2}$.

Proof. In the initial graph, all possible edges among the b nodes are drawn. To count them, iterate over all nodes and count the number of edges leaving out of each one. You will get a total of $b(b-1)$, because on each of the b nodes you will see all edges that can depart from it toward the other $b-1$ nodes. But you will also have counted each edge exactly twice, once when you were on its one end and once more when you were on its other end. So, you should divide your count by 2. The initial graph has $b(b-1)/2$ edges.

In contrast, the final graph has 0 edges. Therefore, the overall decrease in the number of edges in the course of the game is $b(b-1)/2$. □

Second, we should again address the "where did this come from" question.

Question 4. *Where did the "adjacency graph" come from?*

Answer. As in the answer to Question 3, I tried to identify $b(b-1)/2$ as the value of some natural quantity on the table. This time I recalled that (as shown in the proof of the last corollary) this number is exactly the number of edges in a complete graph with b vertices, which is a well-known fact. It was natural then to associate the configuration on the table with such a graph. It was clear that the vertices of the graph should correspond to boxes on the table, so it remained to interpret what the presence of an edge would mean. It had to mean that some relation between the corresponding boxes should hold. Being in the same stack was the first such relation to come to mind. □

2.6. **The second proof is still better.** Like Proof 2B, the new proof also has the two important attributes that it *tightly reorganizes* the discussion around an *intuitive* new center, the adjacency graph.

That the new organization of the discussion is as successful as the one in Proof 2B cannot be denied. The two proofs follow the same pattern: a quantity associated with the central concept (the number of edges in the adjacency graph; the total height on the table) decreases in every move, and the decrease always equals the increase in the player's score, so that the final score is the total decrease of this quantity. The same apples-and-Bob argument, but with a different kind of apples in each case.

What makes Proof 2B more satisfactory is therefore its being more successful in the selection of the central concept: the total height on the table is more intuitive than the adjacency graph. In a sense, the heights of the boxes are *on the table*, their presence and behavior being within our immediate perceptive faculties. In contrast, the adjacency graph seems to be a concept *on the side*, whose behavior requires some (straightforward, but still) non-trivial extra step in order to understand and work with.

3. Requesting an Explanation

Questions 1 through 4 are typical of the kind of questions that students ask after they have been exposed to a proof and while they are still trying to "understand". As already seen, Question 2 is of a sharply distinct nature.

All other questions request some kind of explanation as to how a certain concept used in the proof was arrived at. They are questions *about the proof*; about its discovery; about the history of the prover's thoughts during and behind the construction of the proof. Naturally, the prover never asks questions of this kind. It is a (convinced, but) puzzled reader that poses them.

In contrast, Question 2 is *about the theorem*, about the fact being proved. It is a request for a 'deeper' reason why the theorem actually holds. And it is characterized by the surprising property that it can be posed even by the person that has just finished constructing the proof.

In this study we focus on questions of the latter kind. Questions of the former type are not less interesting; discussions of them can be found in Polya's [3] and in the related analysis by Sandborg [4].

3.1. **Why-questions.** The connective that introduces Question 2 serves only as a reminder of the context in which the question has been asked. If

we remove it, Question 2 is simply

(2) *Why is the theorem true?*

or, using the statement of Theorem 2 explicitly,

Why is the final score independent of strategy?

A framework for analyzing questions of this type, called why-questions, in the context of scientific explanation has been developed by van Fraassen [6].

3.1.1. *Motivation and definitions.* The main motivation there has been the resolution of two important problems faced by the philosophy of science in previous attempts to describe scientific explanation: *rejection* and *asymmetry*. The first refers to the situation where a scientific theory rejects a request for explanation as illegitimate, although the request lies clearly within the theory's domain. The second refers to the situation where, between two equivalent propositions, a theory uses one as an explanation of why the other holds, but not conversely. An example (from [6]) that illustrates both problems is the following: According to atomic physics, each chemical element has a characteristic atomic structure and a characteristic spectrum (of light emitted upon excitation). So, the proposition that an element exhibits a particular atomic structure is equivalent to the proposition that this element exhibits the corresponding particular spectrum. Now, if we ask why an element has a particular spectrum, the explanation is given in terms of its atomic structure. But if we ask why the element has a particular atomic structure, a response in terms of its spectrum is not considered an explanation. The theory actually rejects this latter question.

The direction followed by van Fraassen bases on the crucial realization that "scientific explanation is not (pure) science but an application of science" (p. 156). Hence: "An explanation is not the same as a proposition, or an argument, or list of propositions: it is an *answer*. (Analogously, a son is not the same as a man, even if all sons are men, and every man is a son.) An explanation is an answer to a why-question. So, a theory of explanation must be a theory of why-questions." (p. 134). And since any such theory must take into account the *context* in which a why-interrogative is asked, so must any theory of explanation.

According to van Fraassen, a why-question expressed by an interrogative in a given context is a triple $Q = \langle P, X, R\rangle$, where (concrete examples of these definitions are given in the next section):

- P is the *topic*: the proposition which is the subject of the interrogative; the interrogative asks for an explanation why P is true.
- X is the *contrast-class*: a set of propositions that contains P and all alternatives to it; the interrogative asks for an explanation why it is the case that P as opposed to some other member of X.
- R is the *relevance relation*: a relation between topic/contrast-class pairs and propositions; among the possibly many reasons why P and not some other member of X, the interrogative asks for one that is R-related to $\langle P, X\rangle$.

Then, a *direct answer* to Q is any proposition B which is true exactly when $\langle P$; and, for all $P' \in X - \{P\}$, not P'; and $A\rangle$, for A some proposition that

R-relates to $\langle P, X\rangle$. This A is the *reason* offered by B. For simplicity, we also refer to B as "Because A" and to A as an *answer*.

In addition, to ask Q presupposes that there exists a direct answer to it which is true. So, the *presupposition* of Q is that only P is true in X and that there exists a true proposition in R-relation to $\langle P, X\rangle$. If this presupposition is false according to the body K of accepted background theory and factual information that is associated with the current context, then Q does not *arise* at all in this context and can therefore be rejected.

3.1.2. *An example.* Suppose Q is the question expressed by the interrogative $I =$"why does the sun exert a force on earth?". Then the topic P is the proposition expressed by the sentence "the sun exerts a force on earth".

In the rather strange context where we would be asking why the sun exerts a force on *earth* as opposed to some other planet, the contrast-class X' would consist of all propositions that can be expressed by the sentence scheme "the sun exerts a force on planet x", as x ranges over all planets. Note how the definitions above capture our discomfort in this question: With this topic P and contrast-class X', the question presupposes that the sun does not exert a force on any planet other than the earth. In the ordinary context, the associated factual information makes this presupposition false. Therefore, this question does not arise and, if asked, it would be rejected.

Now consider the more natural context, in which the interrogative I would be just asking why there *is* a force as opposed to not there being a force. Then, the contrast-class would simply be $X = \{P, \text{not } P\}$.

For these P and X, at least three relevance relations could emerge.

For a natural context, suppose I is one of the questions in a high school exam. Then the relevance relation R relates to $\langle P, X\rangle$ propositions that lie within Newton's theory. One of them is, of course, the proposition A expressed by the sentence "every mass exerts a force on every other mass". Hence, "Because A" is a possible direct answer to $Q = \langle P, X, R\rangle$.

On the other hand, suppose that I happens to be the n-th successive interrogative of the form "Daddy, why…?" asked by the 7-year-old while he is observing the night sky through the telescope and after his father has quickly summarized to him the essentials of Newton's theory. Then the interrogative is really a concrete way to express the deeper question why forces between masses exist in the first place, and the relevance relation R' relates $\langle P, X\rangle$ only to propositions that are in this spirit. In particular, A is not R'-related to $\langle P, X\rangle$, and "Because A" is not an answer to $Q' = \langle P, X, R'\rangle$. In fact, in the associated background theory of Newton, no true proposition is in R'-relation to $\langle P, X\rangle$. Hence, the presupposition of Q' fails, and Q' does not arise. The poor father will have to reject the question.

Finally, suppose I is one of the rhetorical questions asked by a philosopher as he describes how man invents concepts —the phlogiston, the force, the electron— that help him organize his experience into networks of knowledge that are as unified and as tight as possible. Then I is really asking why Newton's story of the world needs the sun to exert a force on earth, and the relevance relation R'' follows this unusual spirit. For example, the philosopher may go on to answer that this is because the earth revolves around the sun; that is, Newton needs the sun to exert a force on earth so that

his story both stays concise and accounts for the observation that the earth revolves around the sun. So, the proposition P' expressed by the sentence "the earth revolves around the sun" is R''-related to $\langle P, X\rangle$. Notice how the twisted context allows P' to be offered as an explanation for P, when in the ordinary context it is P that is actually offered as an explanation for P'.

3.1.3. *Evaluation of explanations.* Because of what problems it set out to solve, the framework of why-questions focuses on deciding whether some proposition A may be offered as an answer to a particular question or not. Although this alone is an important problem, it is also important to be able to evaluate answers.

So, if A may indeed be offered as an answer to a question $Q = \langle P, X, R\rangle$, how good an answer is it? Van Fraassen mentions three ways in which this evaluation can be done in a context with background knowledge K (p. 146):

> The first concerns the evaluation of A itself, as acceptable or as likely to be true. The second concerns the extent to which A *favours* the topic P as against the other members of the contrast-class. [..] The third concerns the comparison of *Because* A with other possible answers to the same question; and this has three aspects. The first is whether A is more probable (in view of K); the second whether it favours the topic to a greater extend; and the third, whether it is made wholly or partially irrelevant by other answers that could be given. (To this third aspect, Salmon's considerations about *screening off* apply.)

Recall that, in the Reichenbach-Salmon sense, an answer A' screens off another answer A from the topic P if and only if the probability of P given both A and A' is the same as the probability of P given A' alone.

3.2. **Our explanation request as a why-question.** We now model Question 2 as a why-question. Namely, we build the why-question expressed by the interrogative of (2) when this is asked in the context of Section 2.2.

The topic P is the proposition that the theorem is true, i.e., that the final score is independent of strategy. The contrast-class is just $X = \{P, \text{not } P\}$, since the only alternative that we considered was that the theorem is false.

To determine the relevance relation we first note that the interrogative in (2) can be asked even in a situation where one has read the statement of the theorem but has seen no proof of it yet. Then, the why-question would have (the same contrast-class X, and) a relevance relation R_m that would allow any mathematically rigorous argument as an answer. However, in the context of Section 2.2, an answer must not only be mathematically rigorous but also satisfy our demand for 'deeper' reasons, for a story that will 'unveil the conspiracy'. Hence, the implied relevance relation R_d is clearly a restriction of R_m. In particular, it renders Proof 2A irrelevant while it allows Proofs 2B and 2C.

Overall, when modeled as a why-question, Question 2 is the triple

$$Q_2 = \langle P, X, R_d\rangle$$

for the P, X, and R_d just described. One can quickly see that this modeling completely avoids the issues that we want to address. In particular, we still have no analysis of

- the difference between Proof 2A on the one hand and Proofs 2B and 2C on the other, and
- the difference between Proofs 2B and 2C,

which is what we are really after. However, the modeling does give us a nice vocabulary to describe these issues and thus make our conclusions more general. In this vocabulary, the first of the two differences above is the difference between R_m and R_d, while the second one asks for the evaluation of answers that have passed the R_d test.

3.3. **R_m versus R_d.** So, how do R_m and R_d differ? What makes one argument succeed in satisfying our request for 'deeper' reasons and another one fail? What is the difference between an argument that manages to trigger in our mind this feeling of 'understanding' and one that doesn't? What is this feeling of 'understanding', anyway?

We certainly cannot give precise descriptions of the relations R_m and R_d. But we have agreed that, although each one of the Proofs 2A, 2B, and 2C passes the R_m test, only Proofs 2B and 2C make it through R_d. Sections 2.4 and 2.6 have already discussed what promotes the two proofs, by saying that each of them manages to *tightly reorganize* the argument around a very *intuitive* new center. To say anything more than just this, we need to allow ourselves to speculate. So, although the rest of the section is written in the form of a claim, it is really only a reasonable suggestion.

To understand is *to create an analogy with the physical world.* To map the new experience to past experience about the physical objects that surround us. To find between the new phenomenon and an already experienced physical one an isomorphism which preserves the behavior of the participating concepts. To invent a metaphor through which discussing the new phenomenon is like discussing an old physical one. To reduce the new stuff to stuff about our everyday interaction with the physical environment. (Here, the meaning of the words analogy, mapping, isomorphism, metaphor, reduction is that of *metaphor* as described in [1] and [2].)

This is exactly what Proofs 2B and 2C did. The first one reduced the discussion to a discussion about how high the several boxes collectively stay on the table. The second one mapped the discussion down to a discussion about how many lines remain drawn between several dots on a page. Both described the evolution of the game in terms of an apples-and-Bob process. Overall, analogies re-described the setting of the theorem in terms of everyday experience that exists in almost everyone's memory.

Proof 2A tried to move toward the same direction. It first described the invariability of score as a property of the stacks that can possibly appear on the table throughout the game, and then used induction to describe how this property climbs up from shorter to taller stacks backward in time. But after that point, the analogy broke. In describing the elementary steps in this climb, the proof switched platforms, to transfer us to the world of symbolic manipulation, in line (1).

The problem with this switch is *not* that by jumping to symbolic manipulation we lifted ourselves into some abstract world that is not in any way mapped to our physical experience. It is exactly by analogy to the physical world that we can actually understand symbols and manipulate them. The variables of an arithmetic expression are like fresh objects on the page. We can move them around, regroup them, take them away in pairs, as if they were pebbles of different colors, or "guys" (a metaphor quite common among students of mathematics) belonging to different teams. So, throughout the calculations in line (1), we are still in some correspondence to the physical world. We could not do otherwise.

The problem is that this correspondence does not compose well with the analogy that the argument of Proof 2A had already established via its inductive structure. What does the occurrence of $\frac{b}{2}$ in line (1) mean exactly in terms of how high the boxes are on the table and what move I have made? And why does it come with a negative sign? What about the first occurrence of $\frac{x^2}{2}$? What about the second one? What does it mean, in terms of my first move, that these two occurences together cancel out with the negative occurrence of x^2? And how do you make sense of the fact that, while we start and end the manipulation with 'apples', in between there are both 'apples' and 'oranges'? (The three original terms and the final term are all squares of numbers of boxes, but in between we also see terms that are just numbers of boxes.)

Are there answers to these questions? There have to be! In fact, some of them are easy. But some are harder. Can you answer them? Maybe you can, but how long will it take you? In any case, I am confident that you think about these issues only now that I am asking and, in particular, long after the proof has convinced you about the truth of the theorem.

In general, at the beginning of every symbolic manipulation, the variables of the arithmetic expression are fresh objects that map to concepts of the main context, the one that gives rise to the expression. When we start the manipulation, this mapping is forgotten, the fresh objects enter the algebraic battle field semantically uncharged, and fight their way through each line. Those that finally make it across the last equal sign put on again their original meaning, and that is the only thing that we keep as we return to the main context. This tactic of forgetting, manipulating, then remembering again is exactly the power of algebra, what makes it so widely applicable. It is by severing the links to the main context that we manage to perform all these calculations. If we had to maintain these links all the way through, to constantly have in mind the meaning of each variable, some easy calculations would immediately become hard or impossible.

Overall, in order to achieve efficiency, we switch our correspondence with the physical world: from the one implied by the main context, we move to the one implied by the context of symbolic manipulation, and then back to that of the main context. When the two correspondences do not compose well, the price that we pay is in explanatory power. What intervenes between the start and the end of the symbolic manipulation stays behind an impenetrable brick wall. Our only option is to simply walk around it, by switching to the appropriate correspondence, and verify the correctness of what lies behind.

This is where Proof 2A fails. It uses more than one analogies with the physical world and these do not compose well. As a result, it does not provide us with one analogy that can take us throughout the entire argument. Under these conditions, we can verify but we can not understand. In the general terms of R_m and R_d, the suggestion is that *a mathematically rigorous argument can make it through R_d only if it provides a single analogy to the physical world that can be used throughout the entire proof.*

3.4. **Inside R_d.** So far, the modeling of Section 3.2 and the description of R_d in the previous section have sketched the gross outline of a systematic reasoning that can weed Proof 2A out of the possible answers to Question 2, exactly as our intuition has already dictated clearly but unsystematically.

The next issue that we want to address is of course the comparison between Proofs 2B and 2C. According to our reasoning, both should be declared possible answers to Question 2, which is in par with our original intuition. But this intuition also tells us that 2B is better than 2C. Does this show up in our modeling?

We should first check whether the three criteria suggested by van Fraassen in Section 3.1.3 for the evaluation of explanations can tell this difference.

Clearly, each of the proofs is true with probability 1 and supports the topic P of the question against its alternative in the strongest possible sense. Hence, both proofs score maximally in the first two criteria, so that the first two aspects of the third criterion cannot distinguish between them.

The last aspect of the last criterion cannot help, either. If we interpret it as implying the Reichenbach-Salmon criterion for an answer being "screened off" by another (which is the only interpretation suggested by van Fraassen), then all relevant probabilities are 1, rendering the criterion completely blind. Of course, some other interpretation of "made wholly or partially irrelevant" might do better. But it is hard to imagine one that will decide Proof 2C is made irrelevant by Proof 2B without at the same time also deciding that Proof 2B is made irrelevant by Proof 2C, as the concepts involved in the two proofs are tightly entwined.

It is easy to describe this entwinement precisely. Fix a configuration of the boxes on the table and consider one of the stacks in this configuration. On the one hand, each box in this stack has a height and the sum of these heights is a number at most the total height of the configuration. Call this number the *total height* of the particular stack. On the other hand, each box in the stack corresponds to a node in the adjacency graph for this configuration and thus the stack corresponds to a subgraph of the adjacency graph. Call this subgraph the *subgraph* of the particular stack. Now, it is easy to see that all nodes in this subgraph connect to each other and to no nodes outside the subgraph. Hence, if the stack has size m and therefore total height $m(m-1)/2$, its subgraph has m nodes and $m(m-1)/2$ edges. Overall, a configuration consisting of s stacks corresponds to an adjacency graph consisting of s complete graphs as connected components, each the subgraph of a stack. Splitting one of the stacks to create two new stacks is the same as removing the appropriate edges from the corresponding subgraph to create the subgraphs of the new stacks.

So, both proofs describe the same thing but with different names. What mathematical or logical criterion could decide that one of them makes the other irrelevant without also deciding the converse? It does not seem that the last suggestion by van Fraassen can be of any help in our setting, so that all criteria are blind for the comparison that we consider. More strongly, it seems that it is the *kind* of the suggested criteria that makes them fail to distinguish between the two proofs.

But our intuition did distinguish between them, so there is a difference. In Section 2.6 we said that the better proof has been more successful in the selection of the central concept around which it organizes its argument; that, somehow, the heights of the boxes seem to be right there *on the table*, whereas the adjacency matrix is *on the side*. To make the mental shift from a configuration of the boxes to their heights is almost effortless, whereas to make the corresponding mental shift from a configuration to its adjacency graph is an easy but definitely non-immediate step. In a sense, Proof 2B does make 2C irrelevant, but not mathematically or logically; simply as in "if we can think about this right on the table, why bother moving the discussion to the side?" In other words, Proof 2B is *faster*.

We can describe this a little more precisely. First, we are comparing proofs that can be given as answers to Question 2, therefore proofs that have passed the R_d test. So, each proof establishes an analogy with the physical world, a mapping of the participating concepts to concepts related to past physical experience. Then, in view of the preceding remarks, the critical question becomes: *how easy is this mapping?* How fast can my mind perform the implied mental shift? How fast can my mind compute this reduction from new to past experience? The easier the reduction, the better the proof.

One can probably recognize here certain standard key concepts from the theory of computation [5]. However, it is important to note that the suggested criterion is clearly extra-logical, extra-mathematical. It is not one that can be decided by looking at the proof alone. The reader of the proof is as important. What reductions are easy to compute depends on who the reader is and what his past physical experience has been.

3.5. **Summing up.** So let us recap. Among all questions in Section 2, we distinguished Question 2 as a request for explaining the theorem rather than the discovery of the proof. We saw that in the theory of why-questions this special nature of Question 2 is modeled via a relevance relation that is a restriction of the one which allows all mathematically rigorous proofs.

Elaborating on the difference, we suggested that a proof is explanatory if and only if it establishes a single analogy with the physical world which alone can lead us throughout the entire argument. That is, although every proof inevitably consists of sub-arguments that base on analogies of this kind, only sometimes do these analogies compose nicely into a single one that can be followed throughout the proof. It is then that we have an explanation.

We also noticed that even between explanations our intuition may distinguish one as better than the other. We suggested that this is due to the difference in the hardness of computing the corresponding single analogy: the faster this can be done, the better the explanation.

4. Conclusion

Sometimes I have a cool idea about how to prove a theorem. I work hard for days and, happily, it works! The proof is nice, and I carefully write it down, making sure I have not missed anything. After I finally draw the little box marking the end of it, I lie back in my chair half-satisfied half-annoyed, my pencil rolling in my fingers: "Ok, it's true. But why?"

Sometimes a student wants me to go over the solution to a discrete math homework problem. I repeat the proof on the board thinking it out from scratch, so that she sees no magic is involved, only the natural next steps of an educated mind like hers. When I am done, her reaction is "Oh, ok" and then silence. I can feel the tension, so I say "But you can also think of it like this" and go on to give another, explanatory proof. Her eyes spark, "Oooh, I see", and starts re-describing the solution in her own words.

My essay has been about this "But why is it true?" or "Oh, ok—vs—Oooh, I see" phenomenon that every working mathematician, in research or in teaching, is aware of and has many stories to tell about. My main goal has been to illustrate it with a couple of simple examples that readers with no extensive mathematical background can follow —it definitely happens in more complicated math, too, but not necessarily more often. I also tried to build around this phenomenon a story that makes some sense, by using van Fraassen's theory for scientific explanation.

There is no doubt that the phenomenon is real —people bump on it every day. In addition, I believe its fit into the theory of scientific explanation is meaningful and interesting. I am also pretty confident about the general direction of the story that I have tried to tell, namely the role of fast reductions that map concepts down to past physical experience. Still, the details of my analysis are less likely to be correct, and it would be interesting to see examples where they fail.

Beyond this point, there are questions one may want to ask. For example:

What is the value of explanatory proofs in mathematics? Sure, they may make us feel that we understand, which is good. But do they also help us prove more theorems? It seems tempting to jump to a positive answer, but it is probably wiser not to unconditionally conclude so. At least in some cases, an explanatory proof seems to be little help in making further progress.

Is there always an explanatory proof? That is, can every proof be converted into an explanatory one? Again, it is tempting to answer positively. But can we? Similarly to the fact that most interesting formal systems allow for true propositions that have no proof, maybe it is also true that most interesting systems of experience allow for proofs that have no explanatory counterparts. Given our grounding of explanatory power to past physical experience, the unconditional existence of explanatory proofs would declare the mind potentially close to all mathematical problems. Both the view of the brain as the product of evolution and the evidence from mathematical practice seem to be against such a position. If this is so, what does it imply for doing mathematics, both research and teaching?

Are there general techniques to develop explanatory proofs? Even if not all proofs have explanations, many do. Are there general strategies for discovering these explanations? Namely, the same way that induction and

algebra help us find proofs for the theorems that have them, are there tools for discovering explanations for the proofs that have them?

References

[1] George Lakoff and Mark Johnson. *Metaphors we live by*. The University of Chicago Press, Chicago, 2nd edition, 2003.
[2] George Lakoff and Rafael E. Núñez. *Where mathematics comes from*. Basic Books, New York, 2001.
[3] George Polya. With, or without, motivation? *The American Mathematical Monthly*, 56(10):684–691, December 1949.
[4] David Sandborg. Mathematical explanation and the theory of why-questions. *The British Journal for the Philosophy of Science*, 49(4):603–624, December 1998.
[5] Michael Sipser. *Introduction to the theory of computation*. PWS Publishing Company, Boston, 1996.
[6] Bas C. van Fraassen. *The scientific image*. Oxford University Press, New York, 1980.