

# Proof and Psychology

Sharon Berry

April 16, 2010

[add Frege quote about ‘Let us sharply distinguish logic from psychology’]

## 1 Introduction

“Are all mathematical truths provable?” This is a classic question in philosophy of mathematics - which I will not try to answer in this essay. However, even starting to think about this question forces us to consider another question, which I will try to answer.

It only makes sense to ask ‘What statements of kind X are provable?’ to the extent that we have some notion of what counts as a proof. There are well known issues about informal proofs, but even if we just consider explicit formal proofs we face a big question about which derivations-in-a-formal-system qualify as genuine proofs. If just being provable in *some* formal system were a sufficient condition for being provable, then every sentence S (true or false) would be provable. For, just consider the formal system of first order logic plus S as an axiom. Or, consider the axiom-less formal system which has all the inference rules for first order logic, plus the rule that you can infer S from empty premises.

Thus, if we want to even start to think about what’s provable, we are forced to ask ourselves a deeper question: what un-argued premises -if any- can occur in a genuine proof, and what inferences? (Note that here that I mean ‘proof’ in the sense of “proof that P”, not in the sense of “proof that P follows from Q”, since clearly any premises can occur in a ‘proof’ of the latter kind, and all statements are ‘provable’ in this sense, since all statements follow from a contradiction.) Another way of putting this same question is: what formal systems F have the property that all proofs-in-F are also proofs simpliciter?

In this paper I will consider three initially plausible strategies for answering this question, but I will argue that only the last one is really promising. The strategies focus on, respectively: truth, indubitability of propositions, and idealization of human psychology. If I am right we get a relationship between logic and psychology which would have shocked Frege (hence the quote above): to the extent that there is any coherent notion of proof and provability simpliciter (as opposed to proof-in-a-formal-system, and provability-in-a-formal-system) this reflects idealized but still rather arbitrary and contingent facts about human psychology.

## 2 Mere Truth

The first strategy says simply that the proofs in a formal system can be considered proofs simpliciter, iff all the premises allowed by that formal system are true, and all the inferences allowed by it are truth-preserving. This idea has the clear appeal of tying the epistemic notion of proof closely to truth and reliability (which, intuitively, are extremely relevant to justification). For, note that the belief forming method of believing the results of proofs in a formal system  $S$  will be extremely (in fact, perfectly) reliable, when  $S$  is a formal system of the kind just specified. This is because, it's impossible to get to a false conclusion by starting with only true premises, and proceeding by only truth-preserving inference steps. Hence, all possible worlds in which someone applies the method (believe the results of a proof in  $s$ ) are ones in which they thereby arrive at a true belief. Another advantage of this theory is that it doesn't appeal to any potentially puzzling facts about primitive epistemic normativity, all it uses are facts about which propositions are true (are all the premises true? are there any combinations of sentences satisfying the inference procedure, such that the premise sentences are all true and the conclusion sentence is false?).

Sadly, however, this proposal cannot hope to capture anything like the intuitive notion of proof. For one thing, according to this account we still get the result that every true mathematical sentence  $S$  is provable, trivially, by the one line proof " $S$ ". But this conflicts dramatically with the intuitive notion of proof. Just think about the classic philosophical question we started with, 'can there be unprovable philosophical truths?' If by 'provable' we just meant 'derivable from true premises via truth-preserving inferences' the answer to this question would be an obvious yes.

Also, here's another way in which the simple definition proposed above fails to fit the intuitive notion of provability. Most people think that  $P \neq NP$  is true, but we have not yet found a proof of it. But on this theory, if  $P \neq NP$  is true, the obvious one-line derivation " $P \neq NP$  therefore  $P \neq NP$ " would count as a proof of it. Also, consider the fact that philosophers and others have often wondered whether there are unprovable mathematical truths. On this view of proof, the answer would be an immediate 'no'.

On to the next strategy then.

## 3 Indubitability

The next strategy is to invoke the notion of indubitability. Maybe a proof has to have premises that are not just true (as per the above) but indubitably true, and inference procedures that are not just truth-preserving but indubitably truth preserving. This is a classic idea, and might be attributed to thinkers as diverse as Descartes Frege and Kant.

But, in the past century or so, the notion of indubitability has come under serious attack. Let's assume, for now, that indubitability isn't just supposed to be the psychological property of being a proposition that most normal humans

(for some choice of ‘most’ and ‘normal’) accept. Saying that the premises and inferences in a proof need to be indubitable in this sense would be an example of what I am calling the psychological strategy, (which will be discussed in the next section).

So, if indubitability isn’t to be understood psychologically, what is it? The classic idea is that indubitable propositions are ones that it’s literally impossible to doubt, for the following reason: These propositions are so central to thought, that someone who tried to doubt them wouldn’t count as thinking at all (so a fortiori they wouldn’t count as doubting either). Note that this is not just a matter of how these propositions relate to contingent human psychology, but rather arises from necessary truths about what kinds of possible psychological states would count as thinking. Such propositions could be called laws of thought, in the sense that no one could count as thinking who didn’t accept them.

And a similar proposal can be made for inference procedures: indubitable inference procedures would be patterns of inferences such that anyone who wasn’t willing to make these inferences wouldn’t count as thinking at all. Actually, there are two ways we could go here. We could either say that something like modus ponens is an indubitable inference iff no one who didn’t *accept the principle* could count as a thinker, or iff no one who wasn’t *disposed to accept every instance* of the procedure could count as a thinker.

The main problem with this, I claim, is that there simply *don’t seem to be* any premises or inference procedures which are indubitable in the relevant sense. With regard to indubitable beliefs, Timothy Williamson has emphasized that even propositions, like “All vixens are vixens” don’t have the property of being literally impossible to doubt. A philosopher who has an unconventional view of how universal generalization works, and accepts a conspiracy theory whereby there aren’t really any foxes might reject this sentence because they think that to accept “All Xs are F” commits one to “There are some Xs”. Surely such a person would count as thinking. And plausibly they would count as meaning the same things as we do by key words like “all” and “vixen”. Thinking about the issue more generally, what’s going on is this. Plausibly, a certain amount of accuracy (or at least agreement with us) is required for someone to count as thinking. If someone made assertions completely at random, saying first “snow is white” and then “snow is not white” and then “cats grow on trees, therefore snow is white” we would not count them as thinking at all. And this motivates the idea that there may be certain particular propositions (the indubitable ones) which anyone who could count as thinking would have to accept. But, when we consider our intuitions about particular cases, it turns out that what’s required to attribute someone a thought involving vixens, or universal generalization, is not that they have any *particular* belief. Instead require (roughly) that they have enough of the right particular beliefs and methods of reasoning. As a result, we don’t get particular propositions which count as indubitable in the sense required by this classic definition. For any particular obvious truth, we can always imagine someone who denies it but still counts as a thinker because they accept enough of the right things about everything else. And the same

story applies with regard to principles of reasoning. It may seem *prima facie* like someone who rejected modus ponens (either by suspending judgement as to whether this inference form was invalid, or by refusing to make certain inferences of this kind) couldn't count as thinking. But, in the first case, just imagine someone who is disposed to accept every instance of modus ponens, but just isn't sure that there aren't some odd edge cases where this intuitively plausible principle fails. Or, in the second case, imagine someone who actively is unsure as to whether it is truth preserving when applied to some special domain (like reasoning about the infinite, or our own language). These restrictions might seem odd or psychologically implausible, but actually they are quiet analogous to restrictions we actually do place on another intuitively attractive principle - Tarski's T schema. That is, we accept that all normal sentences with same form as "Snow is white' is true iff snow is white", but we don't accept a general principle of reasoning that lets you infer any instance of this schema, on pain of running into contradiction when we consider the liar sentence "This sentence is false". For if (\*) "The sentence next to the (\*) is false" is true iff the sentence next to the 1 is false, then if the starred sentence is false, it's true, and if it's true, then it's false. Thus, to sum up, we can't say that what distinguishes the kind of premises and inferences that can permissibly occur in a proof, is that they are literally indubitable, in the sense that nothing would count as doubting them. For, it seems, there are no sentences or inference patterns which actually have this property.

## 4 Psychological accounts

This brings us to the third strategy, which I will advocate.

On this view, a sentence's provability or un-provability depends on its relation to idealized but still somewhat arbitrary facts of human psychology in the following way. The point of a proof is to relate some truth of interest to the kind of true mathematical principles and truth preserving mathematical inferences which human beings are inclined to accept - to beat a path which ultimately connects basic true statements and valid inferences which people find obvious together in such a way that they ultimately lead to some surprising new mathematical consequence. And the special class of truths which can occur as premises in such a chain (and valid inference procedures which can form links in this chain) is ultimately distinguished by nothing more than its relationship to contingent human psychology.

However, this mention of psychology raises two questions, which need to be answered in order to make this kind of view plausible.

Firstly, if facts about human psychology are contingent, than doesn't this kind of theory have to say that it's contingent what propositions are provable, and what counts as a proof? But, surely it would be very odd to say that if there weren't any people, or if people's psychology changed the standard proof that there are infinitely many primes would cease to count as a proof. So this looks like a problem. Here, I simply want to specify that my proposal says the

the notion of proof *rigidly designates* the class of sentence strings which start with true premises and truth-preserving premises that are obvious to actual humans (existing on earth now). Hence, just as water is still H<sub>2</sub>O at possible worlds where different stuff runs in the rivers and lakes of earth's counterpart, the usual proof that there are infinitely many primes still counts as a proof at possible worlds in which our counterparts resolutely reject some of the premises or inferences in that proof.

It's interesting to note that our counterparts on such a world will have generally have their own notion, twin-proof, which they call "proof" and use to convince each other of mathematical statements, just as we do. (Just as denizens of Putnam's Twin Earth would have a word "water" which refers to twin-water, the substance that flows in their rivers and streams, and plays a similar role in their non-scientific lives to the role water plays in ours). The functions of proofs in building physical theories and awarding mathematics prizes will be much the same, though -since they accept fewer immediate statements/inferences -there will be a number of truths which we have access to but the denizens of these worlds do not. Theories in mathematical physics which are extremely tractable for us might be much less useful to them (e.g. they might propose Newtonian Mechanics, but not accept certain inferences which the needed to evaluate various integrals which we do know how to evaluate, and this might lead them to build worse bridges and cannons). Or, on some fortunate worlds, the difference will be reversed, and our counterparts will be inclined to make immediate and correct inferences about the solutions to certain differential equations which are quite untractable with respect to all methods that we accept.<sup>1</sup>

The second big concern is that we haven't specified the relevant notion of idealization, and to address the possibility that actual human psychology may not allow for any suitable idealization. Clearly, these issues are related.

Now, my aim in this paper is to argue for the psychological strategy in general: to argue that attempts to generate a notion of provability out of pure mathematical facts about what is true and/or truth preserving fail, as do philosophical appeals to a supposed class of propositions such that nothing would count as doubting them, and any notion of provability we have reflects contingent facts of human psychology. So, ultimately, I won't argue for any one particular way of idealizing what goes on in human brains (indeed, future neuroscientific and psychological study of what actually happens in the brain, may well suggest fruitful and physically natural ways of idealizing what happens in the brain, which we would never have guessed a priori). And, I want to suggest that (absent some, as yet undiscovered, new anti-psychological notion of proof, unconsidered in this paper) if there ultimately proves to be no coherent way of

---

<sup>1</sup>Indeed, it's an interesting further question just how much extra immediate mathematical knowledge (or, at least true belief), could be built into creatures of roughly our size and shape, occupying worlds with similar physics. It seems like a disposition to accept any particular mathematical truth - that can be stated sufficiently succinctly- could be added. But could there also be creatures whose brains performed Malament-Hogarth machine like super-tasks (thanks to Peter Koellner and Micheal Potter for independently raising this possibility), such that the set of true matheamtical sentences which they were inclined to immediately accept wasn't even recursively enumerable?.

idealizing what goes on in the brain to generate a definite class of assumptions and inferences which are "the kind of things people will accept in proofs" our notion of proof is itself correspondingly indeterminate.

## 5 Example Psychology-based accounts

However, I think it will help clarify the proposal, and make the latter claim possible, to consider some candidate dimensions of idealization, and the accounts or proof they generate.

In effect, we have already seen that idealizing us merely as "finite beings" or "thinkers" doesn't generate a non-trivial notion of proof or provability. Every stateable proposition is 'provable' in some finite axiom system. And, not only will different thinkers will accept different mathematical and logical claims without argument, but there don't seem to be any statements or inferences which have the property of being shared by literally all creatures that would count as thinkers. So, the relevant idealization will have to be a bit more hands-on than that.

One natural candidate for idealization is suggested from Putnam's notion that certain mathematical statements are either "obvious or reachable from the obvious via obvious steps". Here, the idea is that there's a certain fairly well defined class of true statements and valid inferences which nearly all adult humans are inclined to find obvious, once exposed to typical modern mathematical training. Not only this, but all the extra statements which many *but not* all nearly people find obvious, are statements which could be proved or refuted from statements and inferences which everyone does find obvious.

This kind of idealization fits well with normal mathematical experience. Indeed, it's one of the most distinctive and attractive features of mathematics that even the most abstruse and daunting mathematical arguments are accessible to everyone in this way. Certainly there's a wide variation in the mathematical steps and inferences which different people find obvious. An inference that takes one line in a journal article can take a whole page when written out in enough detail to seem compelling to a hobbyist with a college math major like myself. So there's a vast difference in what propositions seem obvious.

But, and this is the remarkable thing, we always assume that the mathematician can -with enough time and patients- break down the statements that he finds obvious until they can be proved from premises and inferences that *we* find obvious as well! This extra-ordinary capacity to produce agreement, and connect individual mathematical judgements back to premises and inferences which nearly everyone finds obvious and compelling contrasts starkly with pretty much every other field of human endeavor. Can people with unusual moral intuitions often hope to find a proof of the claim that they find intuitively obvious from simpler claims that are obvious to everyone? Can I hope to back up my immediate literary intuition that the last few pages of *Lolita* are obviously not supposed to be a hallucination, using only premises and inferences which X and Y (the proponents of that theory) would find obvious. And, even in the sciences

don't we often find acrimonious disputes between different scientists about how to interpret some agreed on body of data, that cannot be resolved in this way? Who would hope to find a proof of their preferred interpretation of Quantum Mechanics, from premises and inferences which everyone involved found immediately obvious? The closest thing to this kind of disagreement in mathematics would be the conflict between intuitionists and classicists, or between those who advocate and those who reject classical logic, but these are marked exceptions, and the proportion of people who hold the dissenting views are relatively quite small.

This profound amount of agreement *suggests* that there might be a fairly sharply delineated collection of statements and (clearly, syntactically specified) inferences which a) nearly everyone is disposed to find obvious b) allow one to prove all other mathematical truths that feel obvious to a sufficient number of people. If this is true, it would be plausible to say that the intuitive notion of provability applied to the property of being provable from true sentences and truth-preserving inference procedures which have this kind of general obviousness.

But, on the other hand, we can't be certain of this a priori. There might turn out to be some new claim about number theory which feels as immediately obvious as the least number principle or  $2+2=4$  to 50

Another kind of idealization is much higher level. It thinks not only about the statements which everyone is disposed to find obvious when exposed to a normal (contemporary) mathematical education, but of historical factors which might have influenced this. This view involves a kind of stepping back from mathematical practice. Rather than thinking of our access to mathematical knowledge in terms of things like assume PA, the ZF axioms etc, apply first order logic and other inferences, it thinks about the processes which lead us adopt the kind of math education which leads people to find these basic statements like axioms, obvious.

At this level, mathematics looks more like science. Clearly mathematicians have experience based hunches about which mathematical conjectures are likely to turn out to be correct. They don't just choose what statements they are trying to prove out of a hat! Instead, experience with related questions gives them intuitions about which lines of inquiry are likely to lead to a proof, what the answer to certain open questions are likely to be, and this guides where they direct their research energies.

This kind of mathematical intuition about what proofs people will and won't find can seem mysterious, if you just think of it as intuitive access to the mathematical realm, without having any kind of proof. It sounds like mathematicians are dimly seeing the structures which they will eventually prove things about. But this needs not be true in anything more than a metaphorical sense. For, note that these kinds of mathematical intuitions are constantly checked and corrected by experience. You might start out something is a great approach to a given type of problem, and then find no proof after years of research, while some other avenue that initially felt less likely to allow a proof yields success very quickly. You might have an intuition that the answer to a certain open question

will be yes, and then learn about someone proving that the answer is no. Unless intuitions about how to look for proofs are radically unlike the intuitions developed by skilled practitioners of other human activities like pottery or hunting, it would be surprising if such experiences didn't effect one's intuitions. And, given this kind of constant experience and correction, it doesn't seem implausible that mathematicians intuitions wind up being better than chance.

Now, somewhat more controversially, you might think that yesterday's experience honed hunches effect today's mathematics education in such a way that someone trained today would find new statements obvious (and allow them as premises in proofs). In particular, mathematicians whose experience lead to have the intuition that some (actually true) statement  $T$  was true, but who did not accept as premises/unargued inferences in a proof anything from which it could be proved. But then they might teach the next generation of mathematicians in such a way that the next generation did accept premises/inferences from which  $T$  could be proved.

Obviously there's much more to say about whether zoomed-out view of human mathematical reasoning is plausible, but my only point in mentioning it here is to note that, if correct, it would allow for a quite different way of idealizing the processes that lead us to form mathematical beliefs.

If you had this view you might say that what's provable is what humans could be gotten to accept by going through this whole science-like evolutionary process of honing hunches against experience and then coming to find them obvious. Presumably there would be some possible worlds where this process lead to the adoption of false axioms, so you might want to ignore any such worlds. Also, it's not at all obvious the are sharp and general psychological facts about which transitions from experience with particulars to general hunches about how things will turn out (empirically, there seems to be a lot more disagreement about these kinds of intuitions, than about the statements and inferences which mathematicians consider acceptable premises to figure in proofs). There may well be variations in individual psychology, and arbitrary facts about the order in which one learns about various related cases may effect whether or not one forms a given hunch. If so, consider a true statement  $T$  of this kind. Maybe mathematical discoveries made in one order would lead people to eventually accept  $T$  and find it obvious enough to figure in proofs, but making the same discoveries in another order would lead people to perpetually suspend judgement as to whether  $T$ , or even to intuitively guess that  $T$  is false. Should arguments from  $T$  count as proofs? Proponents of this view would have to admit that the notion of proof is vague in this respect.

Personally I'm inclined against this version of the psychological story about truth, even if its picture of how mathematical knowledge works is correct. For, presumably we want to get a useful notion of proof, or at least one that matches with the ordinary use of the term - but this account does neither. For, if we allow that something counts as a proof iff it has true premises and truth-preserving inference procedures and *some process like the above* this could lead people to accept it's premises and inference procedures, it will be almost impossible to know that any argument which doesn't contain known falsehoods isn't a proof.

For example, right now we are intuitively confident that "no one has a proof of  $P \neq NP$ ", and it's useful and informative to say this (e.g. we tend to infer that no one has an argument for  $P \neq NP$  from premises that feel obvious to 99

Finally, one might idealize human reasoning about math in a direction that doesn't involve particular syntactically specified premises and inference rules or quasi-inductive learning. The best example of this that I can think of (still none to impressive) would be the idea that we have a "faculty of reflection" whereby we 'reflect on our own thinking' and thereby learn how to state rules that govern whichever aspects of our reasoning are formally state-able. This idea is sometimes used to argue that Godel's theorem should not apply to human mathematical knowledge, or what I have been calling the notion of provability simpliciter. For, if we had such a faculty of reflection, we would always be able to go beyond any formal system that described our mathematical reasoning, by applying the faculty of reflection to discover the formal rules for this system, and then inferring the Godelian consistency statement for this formal system. But, no formal system extending the Peano Axioms for arithmetic (as our intuitive reasoning certainly does) can include its own con statement. Hence if we have a faculty of reflection, the collection of statements which we can get to by means of it is not captureable by any formal system.

I include this possibility also just as an example, of how one might idealize particular human psychology. Prima facie, it's implausible. For, anything, it seems to be a striking fact about human psychology that we *can't* generally make this kind of jump from reasoning in a certain way, to coming up with formal rules that capture the way we reason. Philosophy would be very different if we had any reasonably active or reliable faculty of reflection that let us go from being disposed to accept certain claims about what would count as knowledge, to formally stating rules which govern our reasoning about knowledge. Everyday life would be very different if we had a faculty that let us go from being disposed to believe that someone is incompetent whenever we learn that they think well of us, to realizing that we have this trait. Literary theory and perhaps the arts would be very different if we had a faculty of reflection that would let us go from being disposed to admire art with certain features, to being able to state a rule which lists what those features are. What we seem to have, in each of these cases, is more like the kind of quasi-inductive knowledge discussed above, based on trying conjectures against our own behavior in particular cases, and developing plausible hypotheses corrected by some finite amount of data. The mere fact that the physical system producing judgements about Gettier cases, or people's character, or the merits of novels is me, doesn't (generally) seem to help me much, beyond the fact that it makes experiments cheaper.

However, to give a really compelling argument against this view, we would need a more developed story about mental faculties. Indeed, this brings us to a whole raft of interesting open questions. How ought we individuate mental faculties? What does the difference between having a faculty of doing X vs. merely having a faculty to do Y, which sometimes does X, amount to? How do descriptive facts about the workings of an actual creature's brain, relate to judgements about what faculties it has, which in turn impact our judgments

about what that creature can do “in principle”? Is the notion of mental faculties somehow conceptually confused in a sense which would impugn any attempt to idealize human reasoning? I don’t claim to have a good answer to any of these questions.

Hopefully this list of examples has made the general idea of a psychology-based account of proof more clear, and given some idea of the variety of such accounts that are possible.

## 6 Conclusion

In this paper I have argued for three things. Firstly, there’s a non-trivial question about what premises and inference rules can figure in a proof, which we need to get clear on before asking more ambitious questions about whether there can be unprovable mathematical truths. Secondly, standard answers to this question which try to cash out acceptable inferences in terms of truth and truth preserving-ness, or indubitability, fail. Thirdly (on a positive note) adopting an account of proof which appeals to contingent facts about human psychology provides a promising way of answering this question - although some versions of the psychology based account look much more plausible than others.