Logical Nihilism: could there be *no* logic?

Gillian Russell

This is the final draft of a paper that is forthcoming in a special issue of *Philosphical Issues* on Philosophy of Logic and Inferential Reasoning, edited by Cory Juhl and Josh Schechter.

Logical monists and pluralists disagree about how many correct logics there are; the monists say there is just one, the pluralists that there are more. Could it turn out that both are wrong, and that there is no logic at all? Such a view might with justice be called *logical nihilism* and here I'll assume a particular gloss on what that means: nihilism is the view that there are no *laws* of logic, so that all candidates—e.g. the law of excluded middle, modus ponens, disjunctive syllogism *et. al.*—fail.[1,2]

Nihilism might sound absurd, but the view has come up in recent discussions of logical pluralism.[3] Some pluralists have claimed that different logics are correct for different kinds of case, e.g. classical logic for consistent cases and paraconsistent logics for dialethic ones. Monists have responded by appealing to a *principle of generality* for logic: a law of logic must hold for *absolutely all* cases, so that it is only those principles that feature in all of the pluralist's systems that count as genuine laws of logic.[4] The pluralist replies that the monist's insistence on generality collapses monism into nihilism, because, they maintain, every logical law fails in some cases.[5]

[1] (Cotnoir, forthcoming) distinguishes two kinds of logical nihilism, the first of which is the view that *natural* languages have no correct logic, and the second of which is that the consequence relation is empty. The view I have in mind here is the second.

[2] Monism has been the default view for many centuries, but it has recently been explicitly defended by Priest (2006). Pluralists include Carnap (1937), Varzi (2002), Beall and Restall (2006), Russell (2008) and Field (2009). There is not much recent mainstream literature on this kind of logical nihilism. Cotnoir (forthcoming) discusses a related view by the same name. Two close relatives of my version of nihilism are defended by Mortensen (1989), one based on the idea that nothing is necessary, the other on the idea that nothing is true in all mathematical models. Mortensen, on my reading, is a nihilist about logical truth, though I am unsure whether he would want to generalise this to logical consequence. Nihilism is also discussed in terms of models in Estrada-González (2015).

[3] Beall and Restall (2006); Priest (2006); Bueno and Shalkowski (2009); Russell (2013); Cotnoir (forthcoming)

[4] For example "The obvious reply to this argument is that it is only truth-preservation over *all* situations that is, strictly speaking, validity. One of the points about deductive logic is that it will work come what may: we do not have to worry about anything except the premises." (Priest, 2006:202)

[5] "...we see no place to *stop* the process of generalisation and broadening of accounts of

From this interchange we can extract a sketch of an argument:

$$\frac{\begin{array}{l}\text{To be a law of logic, a principle must hold in complete generality.}\\ \text{No principles hold in complete generality.}\end{array}}{\text{There are no laws of logic.}}$$

Pluralists do not intend this as an argument for logical nihilism, of course. They think that nihilism is absurd, and since many who work on non-classical logics find the second premise plausible, they intend the argument above as a reductio on the first premise: the monist's assumption that logic must be completely general. Still, that premise has both intuitive appeal and support from historical writers on logic, and—as I will explain in the first section of this paper—nihilism is not, in the end, absurd. These two things—the plausibility of premise 1 and the non-absurdity of the conclusion—turn this sketch of a reductio on premise 1 into a sketch of a direct argument for nihilism.

In the first part of this paper I clarify what logical nihilism amounts to. In the second, I look the above argument for logical nihilism in more detail. Then in the third and final section, I suggest a sensible response, inspired by Lakatos' influential *Proofs and Refutations*. I argue that the method of *lemma incorporation* outlined in that book can be applied—with beneficial results—in logic, and in particular, that it is appropriate in response to the nihilist.

1 Logical nihilism

Logical nihilism is the view that there are no laws of logic, but what this amounts to will depend on what a law of logic *is*. Textbooks usually stipulate a definition (perhaps giving readers a sense that this matter is uncontroversial) but there are several distinct views in the philosophy of logic literature. I will approach the question in 2 stages, looking first at what syntactic form a law of logic should take, and then at the interpretation of principles of that form, where this will determine what it would be for a logical law to be true. The overall goal is to get clear on what it would mean for *none* of them to be true.

1.1 The Form of Logical Laws

Several kinds of principles might be considered important enough to be given the honorific of 'law'. One kind is exemplified by the law of excluded middle and the law of non-contradiction, which are often written:

$$\vDash \phi \vee \neg\phi$$

$$\vDash \neg(\phi \wedge \neg\phi)$$

These consist in an initial double turnstile symbol (representing the property of logical truth) and then a single sentence schema, containing zero or more

cases. For all we know, the only inference left in the intersection of (unrestricted) *all* logics might be the *identity* inference." [6]

logical constants ($\neg$, $\wedge$, $\vee$) and metalinguistic variables (here ϕ). A traditional view—sometimes associated with Hilbert and Frege—is that this is the form of a logical law. On this view, other principles with the appropriate form include:

$$\begin{array}{lll}
\vDash & \phi \rightarrow \phi & \\
\vDash & \phi \rightarrow (\psi \rightarrow \phi) & \\
\vDash & ((\phi \rightarrow \psi) \rightarrow \phi) \rightarrow \phi & \text{(Peirce's Law)} \\
\vDash & \phi & \\
\vDash & \phi \wedge \psi &
\end{array}$$

The latter two have the appropriate form, but they would not normally be considered laws because they aren't normally considered to be *true*. (Whether any of the others are true is one of the things that is at issue in this paper.)

There are two reasons not to stick with this traditional conception of a law when characterising logical nihilism. First, we now know how to generalise logical truth to logical consequence, and there seems no reason to ignore principles with premises on the left of the turnstile—like modus ponens and disjunctive syllogism. And second, if all it took to be a logical nihilist was commitment to the view that there are no logical truths, then some logicians who would not regard themselves as nihilists—and don't seem to deserve the title—would get counted as such. For example, Strong Kleene logic is a logic on which there are no logical truths, though modus ponens and disjunctive syllogism both hold.[7] It seems wrong to classify Strong Kleene logicians as logical nihilists.

The remedy is obvious though. We permit laws of logic to feature a name for a set of premises on the left hand side of the turnstile (which now expresses the more general notion of logical consequence) as in modus ponens and disjunctive syllogism:[8]

$$\begin{array}{rrccl}
\phi \rightarrow \psi, & \phi & \vDash & \psi & \text{(MP)} \\
\phi \vee \psi, & \neg\phi & \vDash & \psi & \text{(DS)[9]}
\end{array}$$

Now the Strong Kleene logician is not misclassified as a nihilist. But why stop there? Perhaps we should consider more complex principles which state connections between different claims of logical consequence, such as:

$$\begin{array}{llllr}
\text{If} & \Gamma \vDash \psi, & \text{then} & \Gamma, \phi \vDash \psi & \text{(thinning)} \\
\text{If} & \Gamma \vDash \phi \wedge \psi, & \text{then} & \Gamma \vDash \phi & (\wedge\text{-E})
\end{array}$$

However in the present paper I reserve the words "law of logic" for simple statements of logical truth and consequence *in which the turnstile is the main predicate* and specifically exclude conditional laws of entailment and sequent

[7] See e.g. (Sider, 2010:79) for further details.

[8] Tarski (1936), Gentzen (1964) Logical principles with such a premise-conclusion form also have a much older claim to being called laws of logic, since Aristotelian syllogisms also took premise-conclusion form.

[9] Here we use the usual convention of omitting set-theoretic brackets from around the name of the set (in list notation) mentioned on the left-hand side of the turnstile. So $\phi \rightarrow \psi, \phi \vDash \psi$ is really an abbreviation of $\{\phi \rightarrow \psi, \phi\} \vDash \psi$.

rules like thinning, cut, and the sequent forms of conjunction elimination. The reason is this: a natural interpretation of the claim that there is no logic is that the extension of the relation of logical consequence is *empty*; there is no pairing of premises and conclusion such that the second is a logical consequence of the first. This would make any claim of the form $\Gamma \vDash \phi$ false, but it would not prevent there from being correct conditional principles.[10] For example, each of the following (the last of which is usually thought to be false) would all hold for the trivial reason that the antecedent of the conditional could never be true:

$$\begin{array}{llllr} \text{If} & \Gamma \vDash \psi, & \text{then} & \Gamma, \phi \vDash \psi & \text{(thinning)} \\ \text{If} & \Gamma \vDash \phi, & \text{then} & \Gamma \vDash \phi \wedge \psi & (\wedge\text{-I}) \\ \text{If} & \Gamma \vDash \psi \text{ and } \Delta, \psi \vDash \phi & \text{then} & \Gamma, \Delta \vDash \phi & \text{(cut)} \end{array}$$

An empty logical consequence relation is nihilism enough, and so for the rest of this paper I will take logical nihilism to be the view that there are no laws of logic, where a law of logic takes the form:

$$\Gamma \vDash \phi$$

1.2 The Interpretation of Logical Laws

Nihilism is the view that no such principles are true and so to evaluate that thesis we need to know what the truth of the claims would require. That might seem as if it ought to be obvious to anyone working on logic, but there are three main approaches here—essentially three different ways to understand the turnstile—which I'll call the *cases*, *interpretations* and *universalist* views.[11]

The cases view is prevalent in recent work on pluralism. It says that $\Gamma \vDash \phi$ is true if and only if in any *case* in which all the members of Γ are true, ϕ is true as well. If we identified cases with possible worlds we would get the formulation—familiar to many philosophers—on which $\Gamma \vDash \phi$ is true iff every possible world in which all the premises are true is one in which the conclusion is true. The more general definition quantifies over cases because formal work deals with *models* rather than possible worlds (and there are lots of reasons not to identify models and possible worlds) and this allows it to remain neutral between conceptions of logic that take consequence to be defined using impossible cases, incomplete cases, and other kinds of formal model. The cases approach allows us to say more about what logical nihilism amounts to: it is the view that for any set of premises Γ and conclusion ϕ whatsoever, there is a case in which every member of Γ is true, but ϕ is not.

[10] A note about vocabulary: arguments are often said to be neither true nor false, but rather valid or invalid. This is correct as far as it goes, but a principle containing a turnstile as its main predicate can be regarded as a sentence making claim about the relevant argument. Such a claim will be true if the argument is valid, false if it is not. Hence the nihilist can be said to believe that there are no *true* atomic claims attributing logical consequence.

[11] The first two are distinguished in Etchemendy (1999). The third is found in (Williamson, 2013:94–95) and Williamson (2017).

An alternative approach to logical consequence can be found in Tarski (1936) and Quine (1986) and in some ways fits even more naturally with model-theoretic semantics. On the *interpretations* view $\Gamma \vDash \phi$ is true iff whatever (syntactically appropriate) interpretation is given to the non-logical expressions in Γ and ϕ, if every member of Γ is true, then so is ϕ. For example, if our argument is $Pa, a = b \vDash Pb$, then the interpretations approach says that the argument is valid iff there is no interpretation of P, a and b (assuming we are treating $=$ as logical) such that Pa and $a = b$ are true, but Pb is not. Models are understood as offering us different interpretations of the non-logical expressions, and hence if we find a model in which Pa and $a = b$ is true but Pb is not, the principle is not true. On the interpretations conception then, logical nihilism is the view that for every argument, $\Gamma \vDash \phi$, there are interpretations of the non-logical expressions in Γ and ϕ which would make every member of Γ true, but ϕ not true.

Finally, the universalist approach to consequence is newer (Williamson, 2013; 2017) but clearly continuous with tradition. Its exposition is easiest if we begin with logical truth. Call the result of replacing every non-logical expression in a sentence or sentence schema ϕ, with a syntactically appropriate variable ϕ's *shell* (distinct expressions with distinct variables, predicate expressions with predicate variables etc.) Then $\vDash \phi$ is true iff the universal closure of ϕ's shell is true. For example, $\vDash \psi \vee \neg\psi$ is true iff $\forall X(X \vee \neg X)$ is true, and $\vDash \forall x Fx \rightarrow Fa$ is true iff $\forall X \forall y(\forall x Xx \rightarrow Xy)$ is true. To proceed further, we would need to generalise this to a definition of logical consequence.[12] Call the result of replacing all the non-logical expressions throughout an argument with syntactically appropriate variables the argument's shell. Then $\Gamma \vDash \phi$ is true just in case the argument's shell is such that *on every assignment* on which all the premises of the argument are true, the conclusion of the argument is true. An assignment is a function taking each non-logical expression to a syntactically appropriate value (perhaps objects for term-variables, properties or sets for predicate-variables, and propositions for sentential variables.) On this approach, logical nihilism would be the view that for any argument, there is an assignment which makes all the premises true without making the conclusion true.

For the sake of making progress in a reasonable number of words, I'm going to assume the interpretations approach for the purposes of laying out an argument for logical nihilism. But it is an important and non-trivial question whether different approaches to logical consequence result in weaker or stronger logics, and especially, whether some approaches admit of a more robust rejection of the weakest logic of all.

We should note though, that if consequence is defined in any of these familiar ways, then purported logical laws make claims that are extremely *general* and so, in one sense, very strong. It is because of this that logical nihilism is not as

[12] This is not entirely straightforward, because we don't normally quantify into arguments, but rather into sentences. The method given here was chosen for retaining the spirit of the universalist approach while allowing for the treatment of non-compact logics, as well as logics in which the validity of an argument can't be defined in terms of the logical truth of a related conditional.

absurd as one might have thought. Consider what some dialetheists say about the classical law called *explosion*.

$$\phi, \neg\phi \quad \vDash \quad \psi \quad \text{(explosion)}$$

Where ϕ is a dialetheia—that is, a sentence ϕ such that both ϕ and $\neg\phi$ can be true at once (perhaps the Liar sentence is like this)—the premises may both be true, while ψ is false. Hence (explosion) is not a law of logic. Still, a dialetheist can hold that dialetheias are rare—perhaps we only get them with certain kinds of metalinguistic self-referential sentences—and where ϕ is *not* a dialethia, that explosion will be truth-perserving. This kind of dialetheist will say that explosion is not a logical law, but that there are many contexts where it is ok to use it—both in everyday life and in mathematics—because it is truth-preserving in those contexts. Intuitionists may say the same thing about the law of excluded middle and more generally proponents of weaker logics may engage in the process of classical recapture in which they explain the appeal and limited scope of various classical principles which they (in full generality) reject as laws of logic.

A nihilist can take the same approach to *every* putative logical law that these friends of weak logics take to the classical laws they have abandoned: they aren't *really* laws, perhaps because of quite rare and esoteric counterexamples. But the principles will do just fine in many contexts. Perhaps all the standard instances of conjunction introduction are truth-preserving. The nihilist can allow this, so long as she thinks there are some—perhaps very specialised—cases in which it fails.[13]

2 Towards logical nihilism

Henceforth I'll assume the interpretations approach to logical consequence, on which logical nihilism is the view that for every principle of the form $\Gamma \vDash \phi$ there is an interpretation of the non-logical expressions in Γ and ϕ such that every member of Γ comes out true but ϕ does not. Such an interpretation would be a counterexample to the principle. If it turns out that there are no such counterexamples, and that on *every* interpretation of those non-logical expressions on which each member of Γ is true, ϕ is also true, then the principle will be a logical law, and nihilism will be false.

So now observe that for the interpretations account of logical consequence to have any hope of being satisfactory, we need a sufficiently rich library of interpretations; if our only available interpretation for ϕ is the value *true* then there will be no interpretation which makes the conclusion untrue, and a fortiori no counterexample. This would be so even for obviously invalid arguments like affirming the consequent:

[13] This also explains how it is possible to give an *argument* for logical nihilism, without the view itself undermining that argument; each of the steps in the argument might be truth-preserving without it being the case that *all* arguments of that form are truth-preserving.

$$P \to Q, Q \models P \tag{AC}$$

The way to avoid (AC) being classed as a law is to enrich our library of interpretations for atomic sentences to include the value *false*. (This is well-motivated, since there are lots of sentences which *are* false.) Then we can consider the interpretation I such that $I(P) = F$ and $I(Q) = T$. Since it makes both premises true and the conclusion false, this interpretation is a counterexample to (AC).

If our interpretation values for atomic sentences are *true* and *false*, then we might think—as classical logicians do—that the law of excluded middle is a logical law:

ϕ	ϕ	$\vee$	$\neg\phi$
T	T	T	T
F	F	T	T

But it is natural to wonder whether this principle—like (AC)—is only masquerading as a logical law because we do not have a library of interpretations rich enough to give a counterexample. Many philosophers think that some sentences are *neither true nor false*, perhaps for reasons of reference failure, or vagueness, or because they concern future contingents. Yet these sentences can still feature in arguments that we want to assess for validity. Suppose we add N (for *neither*) to our set of interpretations for atomic sentences, and, for the sake of argument, assume the Strong Kleene interpretation of the connectives, and that N is not a designated value (so not a way of being true), then we get a counterexample to the (LEM) as well, e.g. an interpretation on which it is not true.

ϕ	ϕ	$\vee$	$\neg\phi$
T	T	T	T
F	F	T	T
N	N	N	N

Strong Kleene logicians are not nihilists and Strong Kleene logic has some logical laws—such as (MP), (DS), and (explosion)—but again we might wonder whether this is only an artifact of the poverty of our library of interpretations. Some think that when we enrich our language enough to include metalinguistic constructions and the truth-predicate—enough to construct paradoxical sentences like the Liar—we get sentences that can take both the truth-values *true and false* (call this status *Both*). Assuming that the truth-conditions for the connectives function in the new cases as in FDE, this give us counterexamples to (DS) and (MP) and, famously, (explosion) (in each case the conclusion sentence gets *false*, while both premises get *Both*.)

ϕ	ψ	ϕ	$\neg\phi$	ψ
T	T	T	F	T
T	F	T	F	F
T	N	T	F	N
T	B	T	F	B
F	T	F	T	T
F	F	F	T	F
F	N	F	T	N
F	B	F	T	B
N	T	N	N	T
N	F	N	N	F
N	N	N	N	N
N	B	N	N	B
B	T	B	B	T
B	F	B	B	F
B	N	B	B	N
B	B	B	B	B

It is plausible that this weakens the logic about as far as (FDE)—the logic of first degree entailment. I don't intend from here to go through each FDE law and show that *some* counterexample exists. Instead I'll take two of the most plausible FDE laws—identity and conjunction introduction—and argue that a rich enough conception of an interpretation gives us counterexamples to each. What I hope you'll find plausible is that if this is the case for laws as basic as identity and conjunction introduction, then nothing is safe: logical nihilism could well be true.

So first, note that in addition to vagueness and the ability to self-refer, another phenomenon of natural language that can complicate our logic is *context-sensitivity*. Context-sensitive expressions can shift their meanings from context to context. Logicians usually handle this in one of three ways. 1) they ignore the phenomenon (much as one might ignore self-reference to focus on simpler languages) and deal only with languages which do not contain context-sensitive expressions, or 2) they stipulate that the context is not allowed to change over the course of an argument, or 3) logicians who are especially interested in context-sensitivity (e.g. Kaplan (1989)) may turn such expressions into logical constants, thus removing them from the scope of the interpretation function (and into the definition of satisfaction for the language.) But context-sensitivity is extremely widespread[14] and we don't—at least for practical reasons—want to make every expression into a logical constant. So let us suppose that such expressions within the domain of the interpretation function.

We might still pursue 2) and stipulate that the context never changes over the course of an argument, but this looks less acceptable if we think that there is a special kind of context-sensitive expression whose interpretation is sensitive to *linguisitic context*. Consider, for example, the little-known atomic sentence

[14] One might even think that every sentence is context-sensitive, on the grounds that every sentence contains a tensed verb.

'SOLO'. 'SOLO' is true when it appears alone, as an atomic sentence, but false any time it embedded in a larger construction. Hence the atomic sentence 'SOLO' is always true, but the conjunction 'SOLO $\wedge$ SNOW IS WHITE' is always false. So now consider the argument form conjunction introduction:

$$\phi, \psi \vDash \phi \wedge \psi \tag{$\wedge$I}$$

When ϕ is SOLO and ψ is *snow is white*, all the premises are true but the conclusion is false:

$$\begin{array}{l} \text{SOLO} \\ \text{Snow is white.} \\ \hline \text{Snow is white} \wedge \text{SOLO.} \end{array}$$

We can model this with a formal interpretation function by allowing the function to take a second argument, corresponding to the linguistic context—embeddedness or solo—of the sentence atom.[15]

And now it is relatively straightforward to see how we might get a counterexample to identity, assuming our language is rich enough: let 'PREM' be an atomic sentence whose value is *true* when it features in the premises of an argument and *false* when it features in the conclusion (or in any other linguistic context). Then we have a counterexample to identity, for the following instance has a true premise, and a non-true conclusion:[16]

$$\frac{\text{PREM}}{\text{PREM}}$$

3 Resisting nihilism with lemma incorporation

One way to respond to the counterexamples above is to embrace them wholeheartedly. Logician nihilism is true, there is no logic. This might be pleasingly dramatic, but it is an uncomfortable thought for many logicians.

A less uncomfortable response would be to argue that the purported counterexamples are not genuine. Essentially one would argue that they fail to provide genuine *interpretations* for the argument form, of the kind we meant to quantify over when we defined logical consequence in terms of truth-preservation

[15] See Russell (2017) for a formal presentation of how this could work.

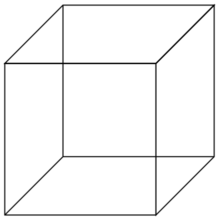
[16] Small print: does this really get us all the way to nihilism? Strictly speaking, that depends on what logical constants are in the language. In particular, if you have $\top$ as a 0-place truth-functor that is always interpreted as *true*, then you will have some arguments with $\top$ and things like $\top \vee \top$ as a conclusion. These will be valid regardless of the interpretation, so with these logical constants you won't get all the way to nihilism. I don't highlight that very much here because it seems to me that a logic with only these laws may as well be logical nihilism, in that everything that seems bad about the one seems bad about the other. For example, neither will be useful for doing metatheory. Or assessing proofs in arithmetic. Elsewhere I've called the view on which there are *hardly any* valid arguments *logical minimalism* and suggested that it is just as bad as nihilism. (Russell, 2017)

across *all* interpretations. This is a substantial philosophical commitment: after all, why not count such context-sensitive interpretations?[17] And it seems in danger of disallowing interpretations *because* they provide counterexamples to a theorem we wish to defend, as if we're refining ‘interpretation’ in the face of these problems to save our familiar theorems. Call this—anticipating what is to come—the *monster-barring* response.

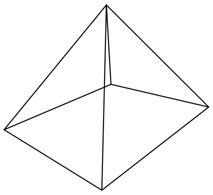
What I want to show in this final section of the paper is that these are not our only options. And moreover neither is our best option. I will do this by developing a useful analogy between the dialectical situation here and one depicted in Lakatos (1976). Lakatos' text is a dialogue between a teacher and his students. It explores proposed counterexamples to Euler's formula, a geometrical conjecture about the relationship between the number of faces (F), edges (E) and vertices (V) of a polyhedron:

$$V - E + F = 2$$

The equation works with some familiar examples, such as cubes and pyramids:



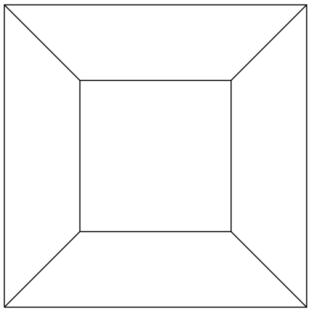
$8 - 12 + 6 = 2$



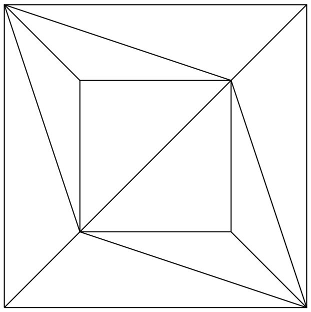
$5 - 8 + 5 = 2$

Moreover the teacher offers an ingenious proof that it works for all polyhedra. We are asked to imagine that an arbitrary polyhedron is made of a rubber sheet (so that it is hollow inside) and then to imagine cutting away one of the faces, before stretching what remains onto on a flat plane. For example, with the cube above we might cut out the top face, and then stretch the rest onto a flat plane so that it looks like this:

[17] It might be substantial in that it argues that this is how we *should* define ‘interpretation’ in the definition of logical consequence, or it might be historically and psychologically substantial in that it argues that this is what logicians *have always meant* when they used the word ‘interpretation. Or both.



Assuming V-E+F=2 for the original polyhedron, V-E+F=1 for this new flat network (because it has the same number of vertices and edges, but one less face.) Then if a face is not already a triangle, we carve it into triangles, perhaps like this:



Finally we remove these triangles one by one. When we do so we either remove an edge—in which case one face and one edge disappears, so $V-E+F=1$ iff it did before—or we remove 2 edges and a vertex—in which case one face, 2 edges and a vertex disappear, and again $V-E+F=1$ iff it did before. Eventually we have a single triangle, for which clearly $V-E+F=1$, since $3-3+1=1$. So the original network satisfied $V-E+F=1$ and the original polyhedron satisfied Euler's formula: $V-E+F=2$. So far so good. We have a proof of the geometrical conjecture and as student Delta suggests:

> "You should now call it a *theorem.* There is nothing conjectural about it anymore."[18]

And then something interesting happens. Some of the students propose odd counterexamples to the so-called theorem: what about a polyhedron which is formed by pressing the faces of two differently-sized cubes together?[19] Or one

[18] (Lakatos, 1976:9)

[19] Lakatos holds that proofs are defeasible and so allows a non-standard use of the word *theorem* that explicitly allows it to be applied to claims that are proven but *false*. This terminology is quite useful in discussing his work, so I'll follow him here and from now on I'll leave out the *so-called* before the word *theorem*.

formed by cutting a cube-shaped hole out of the middle of a larger cube (nested cubes)? Or a polyhedron shaped like a picture frame?

In Lakatos' dialogue, the group is conflicted about the appropriate response to the counterexamples. Student Gamma thinks them devastating to the theorem:

> "Sir, your composure baffles me. A single counterexample refutes a conjecture as effectively as ten. The conjecture and its proof have completely misfired. Hands up!"[20]

But student Delta is appalled that anyone would take the counterexamples seriously:

> "But why accept the counterexample? We proved our conjecture—now it is a theorem. I admit that it clashes with this so-called 'counterexample'. One of them has to give way. But why should the theorem give way, when it has been proved? It is the 'criticism' that should retreat. It is fake criticism. This pair of nested cubes is not a polyhedron at all. It is a *monster*, a pathological case, not a counterexample."[21]

In response to the series of counterexamples Delta proposes a series of increasingly technical definitions of 'polyhedron' (designed to exclude the new shapes from counting) and doughtily maintains that each was the intuitive content of 'polyhedron' as it was used in the conjecture all along. Lakatos calls Delta's approach 'monster-barring'.

I hope the analogy to the logical case is clear: in both cases we have an initial conjecture, and then the conjecture is given a proof and thereafter called a theorem. The theorem is simple and elegant. Some counterexamples are proposed and the counterexamples are creative, interesting, but also a bit odd and non-standard. To use Student Delta's word, they are "monsters." Student Gamma is the analogue of our logical nihilist; he is willing to scrap the entire project and all the class' hard work on the basis of a weird counterexample. Delta is the monster-barrer, determined to protect the initial theorem by refining and redefining the terms it employs if necessary.[22] It is the teacher's view (and so, presumably, Lakatos') that there is a third strategy which is better than either of the others—*lemma incorporation*.[24]

Lemma incorporation involves giving a detailed proof of the original conjecture, and then seeking out the assumption—or lemma—that fails in the case of the monstrous counterexample. This assumption is then incorporated into a new statement of the conjecture. In the case of Euler's theorem and the nested

[20] (Lakatos, 1976:13)

[21] (Lakatos, 1976:14)

[22] "I turn in disgust from your lamentable 'polyhedra', for which Euler's beautiful theorem doesn't hold. I look for order and harmony in mathematics, but you only propagate anarchy and chaos."[23]

[24] (Lakatos, 1976:23)

cube counterexample, the guilty assumption was that a polyhedron can always be stretched flat after one of the faces has been removed—the hole in the middle of the nested cubes prevents this. The teacher incorporates this lemma by calling polyhedra which satisfy the assumption of stretchability 'simple,' and amending the theorem to say that for *simple* polyhedra, $V - E + F = 2$, or even more suggestively, that the Euler characteristic of a simple polyhedron is 2.[25]

Although these amendments happen in response to the monsters, the monsters are *not* simply listed as ad hoc exceptions to the theorem. Lemma incorporation requires that the original proof is scrutinized to find the assumption which fails in the case of the monster, and then incorporates *that* assumption into the new theorem.

Lemma incorporation is superior to nihilism in geometry (i.e. to Gamma's suggestion that we abandon the theorem outright) because it preserves and builds on the progress we have made in learning about polyhedra, rather than throwing everything away. Lemma incorporation is also superior to Delta's monster-barring, because it uses the interesting counterexamples to drive the study of polyhedra forwards, because it does not require us to have already been using a definition of *polyhedron* that somehow perfectly excluded the monsters (even though we didn't yet know that we needed to worry about them) and because it tends to lead to more fruitful mathematics.

Suppose we try it with logic. Classical logic doesn't just tell us that the (LEM) is a theorem; it gives us a model theoretic proof of that theorem. So our first step will be to elaborate that proof. Here is one way to do it:

> Either ϕ is true in a model M, or it is false. In the first case, $\phi \vee \neg\phi$ is true in M because of the truth-clauses for $\vee$. In the second case, $\neg\phi$ is true in M because of the truth-clause for negation, and so again $\phi \vee \neg\phi$ is true in M. So either way it is true in the model, and—since M was arbitrary—it is true in all models. So $\phi \vee \neg\phi$ is a logical truth.

Now we need monster. Our Strong-Kleene counterexample to (LEM) in section 2 used sentences with the value *neither*.[26] So we examine our simple proof and realise that our assumption that the sentence could only be *true* or *false* is violated by the monster. Hence our culprit is the assumption that sentences can only be true and false. Still, perhaps there *are* some sentences which can only be *true* or *false*—sentences in the language of arithmetic might be like—and our result would hold for these. Our new theorem reads: *for any ϕ which can only be true or false*, $\phi \vee \neg\phi$ is a logical truth. Just as the geometry teacher dubs polyhedra which satisfy the stretchability lemma *simple*, so we could give a name to sentences which meet our assumption. Perhaps *bivalent* would be suitable. Then we can retain the proof above as a proof of:

[25] (Lakatos, 1976:33–34)

[26] Probably this does not seem the *most* monstrous of our counterexamples, but I am using it to illustrate the method because the proof of (LEM) is slightly more substantial than that of, say, (ID). Proofs that are too obvious or too short sometimes generate more confusion than is ideal for illustrative purposes.

(1) For all bivalent ϕ, $\vDash \phi \vee \neg\phi$.

We have incorporated the violated lemma into the logical law, and the counterexample to the old law is not a counterexample to the new version.

Lakatos' teacher suggested that lemma incorporation is better than either Gamma's nihilism and better than Delta's monster-barring. I hold that when it comes to logic too, lemma incorporation is a better response to monsters than nihilism and better than monster-barring. It is better than nihilism because it improves on, rather than abandons, our initial proof of the theorem. A nihilist might well protest that the new law above does not have the appropriate syntactic form to be a logical law, and that we are abandoning our commitment to complete generality. And indeed it seems to me that the approach endorsed here abandons the thesis that logical laws are absolutely general and in this way it avoids the argument for logical nihilism by denying the generality premise. But if our real interest is in understanding the subject matter of logic—where that is regularities in truth-preservation over sentences—then the nihilist ignores interesting facts simply because they do not fit his preconceived idea of what those facts ought to look like. If our best laws turn out to have restricted domains, then this is better than no laws at all. We should abandon the generality of logic before we abandon logic.

Lemma incorporation is also preferable to monster-barring. Monster-barring requires us to read implausible restrictions back into 'interpretation' in our original definition of logical consequence, and it fails to use the opportunity presented by the monsters to learn more. Lemma incorporation, by contrast, refines the original proofs in response to the counterexamples, and the results then naturally suggest new avenues of research. 'The Euler characteristic of a simple polyhedron is 2,' correct or not, raises new questions: what are the Euler characteristics of the various geometrical monsters? Can we identify any regularities in the distributions of Euler characteristics? Can we use these to find other non-Eulerian polyhedra? Similarly, 'for all bivalent ϕ, $\vDash \phi \vee \neg\phi$' (whether or not it is correct) immediately makes one wonder about non-bivalent sentences, different ways in which the bivalent laws might fail, and different ways we might restrict interpretations. The counterexamples to (ID) and $(\wedge I)$ can make one curious about regularities in truth-preservation over sentences whose truth-value *can* change in the course of an argument—an under-explored topic, and perhaps one that has been under-explored because of the fear of logical nihilism.[27]

[27] It's natural to ask whether this would result in logical pluralism, since there might be a correct logic for bivalent sentences, and a correct logic for sentences whose truth-value varies with position in space, and they needn't be the same logic. I think it's only a kind of logical pluralism in the sense of (Russell, 2008). The end of that paper contains a discussion of the appropriateness of calling the view 'pluralism'.

References

Beall, J. and Restall, G. (2000). Logical pluralism. *Australasian Journal of Philosophy*, 78:475–493.

Beall, J. and Restall, G. (2006). *Logical Pluralism*. Oxford University Press, Oxford.

Bueno, O. and Shalkowski, S. (2009). Modalism and logical pluralism. *Mind*, pages 295–321.

Carnap, R. (1937). *The Logical Syntax of Language*. Keagan Paul, London.

Cotnoir, A. (forthcoming). Logical nihilism. In Kellen, N., Pedersen, N. J. L. L., and Wyatt, J., editors, *Pluralisms in Truth and Logic*. Palgrave Macmillan.

Estrada-González, L. (2015). Models of possibilism and trivialism. *Logic and Logical Philosophy*, 20(4).

Etchemendy, J. (1999). *On the Concept of Logical Consequence*. CSLI, Stanford.

Field, H. (2009). Pluralism in logic. *The Review of Symbolic Logic*, 2(2):342–359.

Gentzen, G. (1964). Investigations into logical deduction. *American Philosophical Quarterly*, 1(4):288–306.

Kaplan, D. (1989). Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives. In Almog, J., Perry, J., and Wettstein, H., editors, *Themes from Kaplan*. Oxford University Press, New York.

Lakatos, I. (1976). *Proofs and Refutations*. Cambridge University Press, Cambridge.

Mortensen, C. (1989). Anything is possible. *Erkenntnis*, 30(3):319–337.

Priest, G. (2006). *Doubt Truth to be a Liar*. Oxford University Press, Oxford.

Quine, W. V. O. (1986). *Philosophy of Logic*. Harvard University Press, Cambridge, Mass.

Russell, G. (2008). One true logic? *Journal of Philosophical Logic*, 37(6):593–611.

Russell, G. (2013). Logical pluralism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. CSLI, summer 2013 edition.

Russell, G. (2017). An introduction to logical nihilism. In *Logic, Methodology and Philosophy of Science—Proceedings of the 15th International Congress*. College Publications.

Sider, T. (2010). *Logic for Philosophy*. Oxford University Press, USA.

Tarski, A. (1983/1936). On the concept of logical consequence. In Corcoran, J., editor, *Logic, Semantics and metamathematics*, pages 409–420. Hackett, Indianapolis, 2nd edition.

Varzi, A. C. (2002). On logical relativity. *Philosophical Issues*, 12:197–219.

Williamson, T. (2013). *Modal Logic as Metaphysics.* Oxford University Press, Oxford.

Williamson, T. (2017). Semantic paradoxes and abductive methodology. In Armour-Garb, B., editor, *Reflections on the Liar*, chapter 13, pages 325–346. Oxford University Press, Oxford.