

Extrapolation and its limits

Stephen Yablo

Once again, it would be nice if I could explain the topic with examples, but as it is we're going to have to make do with anecdotes. The first concerns a conversation Einstein is supposed to have had with a puzzled citizen, or maybe it's just a Jewish folk tale:

"How does the telegraph system work? I don't understand how they can make a message go down a wire."

"Simple, imagine a giant dog, with his head in Moscow and his tail in Leningrad. Pull the tail in Leningrad and the head barks in Moscow."

"Yes, but what about the wireless telegraph? How does that work?"

"The same way, but without the dog."

Another comes from the 1980 presidential debates between Ronald Reagan and Walter Mondale. Reagan had been showing colossal ignorance about world affairs; someone asked him about Valery Giscard D'Estaing, the president of France, and he said, "I don't believe I have heard that name." The moderator asks Mondale if it bothers him that there's so much Reagan doesn't know. "No," Mondale said, "it's not what he doesn't know that bothers me, it's what he knows for sure that just isn't true." (Borrowed from Will Rogers apparently.)

In both these stories is that you've got a hypothesis A that implies another one B – pulling the dog's tail to get its head to bark implies there's a dog there, and knowing that food stamps are being used by layabouts to buy vodka implies they are being used in that way – and the pair of these is supposed to determine a weaker hypothesis that is, as we might put it, $A-B$, or A stripped of its implication that B . Mondale for instance alleges that Reagan quasi-knows various things where to quasi-know that p is, roughly, knowing that p , stripped of its implication that p .

This kind of implication-stripping, or scaling a hypothesis back so that it no longer implies what it used to, presents something of a challenge to analytic philosophy's traditional self-image. Frege, Russell, and Moore sought where possible to identify contents "from below" by showing how they could be built up out of weaker contents. They never, as far as I know, tried the opposite approach, approaching a content "from above" by first overshooting the target and then stripping away unwanted extras. One could see this as just an oversight on their part, but the fact is that logical addition is pretty well understood --- it's just conjunction --- while logical subtraction is a bit of a

mystery. I suspect it's no accident that Wittgenstein, as he began tearing himself free of the analytic paradigm, also began to wonder about logical subtraction, asking what is left if we subtract a man's arm going up from his raising that arm, and what is left if we subtract from the fact that *It hurts!* the fact that the sufferer is me.

Logical subtraction is a mystery, but that's not to say philosophers don't sometimes give it a try. Goodman considered a statement lawlike iff it was a law, except it might not be true. Parfit explains quasi-memory in something like the way we explained quasi-knowledge on behalf of Will Rogers. The scare quotes sense of a moral term is (something like) the regular sense minus any implication that the act is thereby commendable. But although philosophers do sometimes engage in the act of subtraction, they're also nervous about it, perhaps without always knowing why.

So that's one reason for looking further at logical subtraction; it's common in philosophy, at the same time as philosophers have doubts about it. But the immediate reason for looking further at this is that we are forced to, to complete the story begun the other day about content-parts.

Here's the account we gave Monday: B is part of A iff the inference from A to B was, first, truth-preserving, and second, aboutness-preserving. This was then further explained in terms of what made A and B true or false. Suppose we call X a *decider* for hypothesis S if it's either a truthmaker for S or a falsemaker for S . Then

$B \leq A$ iff (i) A implies B
 (ii) each B -decider is implied by an A -decider

(understood to mean that each truthmaker is implied by a truthmaker and each falsemaker is implied by a falsemaker). From this definition it's pretty easy to see that content-part has the most elementary properties expected of a part/whole relation:

reflexivity $A \leq A$,
antisymmetry if $B < A$ then not $A < B$, and
transitivity if $A \leq B$ and $B \leq C$, then $A \leq C$.

But these properties just ensure content-part is a partial order; and that's not quite enough to warrant use of the term "part." After all, earlier in the alphabet than is a partial order on letters and that doesn't make 'a' in any intuitive sense part of 'z.' Higher-than is a partial order on musical notes and that doesn't make middle C part of middle D.

What more might be needed? The problem with earlier in the alphabet than is that there isn't anything you can point to as 'a's other parts, the one or ones whereby it exceeds 'z.' Likewise there isn't anything you can point to as the extra bit or bits that are found in middle D but are absent from middle C. Evidently a relation does not count for us as a kind of parthood unless it meets the further condition that when X is a proper "part" of Y, there is something left over: Y has other proper "parts" that are disjoint from X (have no parts in common with it).

So, a putative part/whole relation is properly so called only if, in addition to the three conditions above, it satisfies

leftover a thing can't have just one proper part; there's always a second disjoint from the first.

I call it the leftover principle in honor of Wittgenstein's question: what is left over if we subtract from the fact that I raise my arm the fact that my arm goes up? A proper content-part of *A* disjoint from *B* will be called a (logical) leftover. Should it be that there's a distinguished leftover *R* that in some intuitive sense makes up the difference between *A* and *B* – at the very least, *A* holds in the same worlds as *B&R* – it will be a candidate for the role of remainder, written *A-B*.

Taking stock, the fate of content-parts, it seems, is very much tied up with the existence of logical leftovers and remainders. Are there such things? When *A* implies *B*, is there always something, or are there always some things, that you can point to as what *A* adds to *B*? Philosophers are confident types and so their first inclination is to look for a recipe that delivers a remainder in every case and by a uniform method. I am going to suggest something like that myself, but what proposals have been made so far?

By far the most usual idea here is that *B-A* is $B \supset A$. An argument in support of this was given by J.L. Hudson in "Logical Subtraction" (*Analysis* 35, 1975) Greatly simplified it goes like this. Think first of numerical subtraction. $a - b$ is the number c which added to b gives us back a . Logical subtraction should work as far as possible like that. *A-B* should be the *C* such that *B&C* is equivalent to *A*. Problem, however: there is no such *C*, or rather there are too many choices of *C* that make that equation true. Solution: look then for a canonical *C*, presumably the strongest statement which combines with *B* to yield (something equivalent to) *A*, or the weakest such statement. The strongest is *A&B*, that is, since *A* implies *B*, *A* That's *too* strong. We want a *C* that picks up where *B* leaves off, not one that by repeating *B* makes it irrelevant. The weakest statement satisfying the equation is $B \supset A$. Lacking an alternative let's go with that.

To see how it works, let's apply the Hudson hypothesis to the simplest possible case: $p \& q - p$. The Hudson theory says $p \& q - p$ is $p \supset p \& q$. This might already strike you as funny because it seems as clear as anything what is left over when you subtract one of two independent conjuncts is the *other* conjunct, in this case q . And q is a strictly stronger hypothesis than $p \supset p \& q$, because it is not implied by $\sim p$ as the material conditional is. That's just an intuitive argument but we can say something more principled too. By definition, when B is part of A , $A-B$ is supposed to be a different part of A . At the very least then $A-B$ should be part of A .

Clearly p is part of $p \& q$. So the principle tells us that $p \& q - p$ should be part of $p \& q$. Certainly if $p \& q - p = q$, then that condition is met; q too is part of $p \& q$. But what if $p \& q - p$ is $p \supset p \& q$? Is that part of $p \& q$? It is only if each of the conditional's truth-makers is implied by a truthmaker for the conjunction. But this is very far from being the case. Indeed one truthmaker for the conditional, namely p 's negation, is incompatible with all truthmakers for the $p \& q$. That completes the proof that $A-B$ can't be $B \supset A$; it can't because $B \supset A$ isn't part of A and $A-B$ is so defined that it must in these cases be part of A .

Of course some other recipe for the remainder might do better; I will be arguing that one does do better. It has to be admitted, though, the prospects look poor, because there look to be clear counterexamples to the idea that a remainder $A-B$ always exists. What does this shirt is scarlet add to this shirt is red? What does we danced badly add to we danced? ("Badly"?) What does, to give a more contemporary example, Samantha knows water is wet add to water is wet or indeed water exists? The fact is that A sometimes seems to have no parts whatever that make sense apart from B , and certainly none substantial enough to make up the difference. Some logical parts are inextricable from their wholes.

There's the dilemma, then: on the one hand, leftovers should always exist when B is properly part of A , otherwise content-part is not really a notion of part. On the other hand, sometimes when B is part of A , B seems inextricable from A – dancing is part of dancing badly but it can't be extricated – and this would seem to mean that the expected leftovers just aren't there.

Now, to go by its title this is a paper about extrapolation, and it's not clear to begin with what extrication has to do with extrapolation. I'm going to propose that it is a kind of extrapolation. But I admit there's no evident connection to begin with. What the word "extrapolate" most immediately brings to mind is Hume's puzzle about why the observed part of reality should resemble the unobserved part – why the greenness of these emeralds should confirm the

hypothesis that other emeralds are green as well. Call that the puzzle of inductive extrapolation. That's not our topic today obviously.

If I say our topic has more to do with projection than confirmation, you might think of Goodman's new riddle of induction: in what respects exactly is the unobserved part of reality supposed to resemble the unobserved part? These emeralds are as much grue as green, after all; why should it be the greenness that one expects to carry over to other emeralds, rather than the grueness? That's the puzzle of projective extrapolation – what are the inductively fruitful ways to project from the cases in hand? Projective extrapolation is for sure puzzling. But it's not our topic today either.

If I say the kind of extrapolation at issue today has more of a logical flavor – it's more to do with going on in the same way than the inductively fruitful way – then you might be reminded of Kripkenstein's rule following paradox: what is there in the speaker's head, or outside it, to determine what counts as a word's applying in the same way to new cases? The color samples that guide my application of "green" have been just as much grue as green. Why should "green" in my mouth not be true of unexamined grue (so blue) emeralds rather than unexamined green ones? Here it's truth-conditions we're trying to extrapolate so call this the puzzle of alethic extrapolation. Alethic extrapolation is even more puzzling than projective and inductive in Kripke's view:

Wittgenstein has invented a new form of skepticism. Personally I am inclined to regard it as the most radical and original skeptical problem that philosophy has seen to date (WRPL, 60).

The kind of extrapolation I want to talk about also has to do with truth- or application conditions, and it is also suggested by certain passages in Wittgenstein. But type 4 extrapolation is in some ways more puzzling even than alethic. One reason for this is that, as Kripke acknowledges, the traditional puzzles are skeptical in nature; no one really doubts that inductive, projective, and alethic extrapolation by and large work. Type 4 extrapolation, as we will see, may or may not work. Another reason is that type 4 problems remain even if, as Kripke says, we "waive [the] basic and general skeptical problem" that Wittgenstein on Rules and Private Language mainly concerns.

I say "mainly concerns" because type 4 extrapolation does perhaps make an appearance in an Appendix to the book. It's an appendix on the so-called "conceptual problem of other minds"; how do we even see what it means for another to suffer. Wittgenstein says:

If one has to imagine someone else's pain on the model of one's own, this is none too easy a thing to do: for I have to imagine pain which I *do not feel* on the model of the pain which I *do feel* (300)

He gives an analogy meant to evoke the appropriate sense of bewilderment:

[Suppose] I were to say: "You surely know what 'It is 5 o'clock here' means; so you also know what 'It is 5 o'clock on the sun' means. It means simply that it is the same time there as it is here when it is 5 o'clock (PI, 350)

Kripke comments that

... the '5 o'clock on the sun' example seems obviously intended as a case where, without the intervention of any arcane philosophical skepticism about rule-following, there really is a difficulty about extending the old concept – certain presuppositions of our application of this concept are lacking. ...Wittgenstein seems to mean that, *waiving his basic and general skeptical problem*, there is a special intuitive problem...illustrated by the 5 o'clock on the sun example (WRPL, 118-9)

This gets close to the topic today. How is it possible to extend an old concept, or content, to an area where some of its presuppositions, or more generally implications, are lacking? How is it possible to extend the content of Ouch, it hurts! to people such that when you hit their thumb with a hammer, it doesn't in truth hurt a bit? How is it possible to extend the content of Reagan knows that p to worlds where p fails? How is it possible to extend the content of Oscar believes water is wet to worlds where there is no water?

I hope you see some connection between this broadly Wittgensteinian issue of content-extrapolation – of how to extrapolate contents beyond the region of logical space where they strictly apply – and the earlier, quasi-logical, issue of content-subtraction – how to subtract from *A* one of its implications *B*? Because the proposal is going to be that they are one and the same operation. Subtracting an implication *B* from *A* (or abstracting away from that implication, or bracketing that implication – I won't distinguish these) just is extrapolating *A* beyond the *B*-region of logical space, outside of which it does not strictly speaking apply. I suspect it's not a coincidence that Wittgenstein, author of the best-known content-extrapolation problem (5 o'clock on the sun), is also responsible for the best-known subtraction problem:

When I raise my arm, my arm goes up. And now the problem arises: what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm? (PI, 621)

It seems to me that the two styles of problem are at bottom the same. Instead of asking how we extrapolate "It's 5 o'clock here, in Boston" to "It's 5 o'clock there, on the sun," he might have asked how it is possible to subtract from "It's 5 o'clock here in Boston" the implication or presupposition that "here" is a certain distance from the sun. Instead of asking what is left over when we subtract my arm going up from the fact that I raise my arm he might have asked how to extrapolate the condition of my raising my arm to other parts of logical space, where my arm doesn't go up.

I mentioned earlier a dilemma about leftovers and remainders: on the one hand they better exist, for content-parts to be parts, on the other hand they don't always exist, because of examples like red minus colored, or swimming three laps minus swimming. One could try a moderate approach on which remainders sometimes exist and sometimes don't. But philosophers being what they are they tend to take rather extreme positions. You find one camp – I'll call them the mysterians – insisting remainders never exist – and others – the more pollyannish types – insisting they always exist. Here's the

mysterian thesis: one feels there *must* be a remainder, but there is really no must about it. the very notion of a logical remainder is unclear, perhaps irremediably so.

This is what Robert Jaeger maintains in a paper called "Action and Subtraction" (*Phil Review* 82, 1973):

The question "What is left over?"...presupposesthat there is exactly one statement with certain logical properties (A&S, 321). [But] whereas there is exactly one number x such that $x+2=5$, it is not the case that there is exactly one statement R such that " R & my arm goes up" is logically equivalent to "I raise my arm" ("Action & Subtraction," 328)

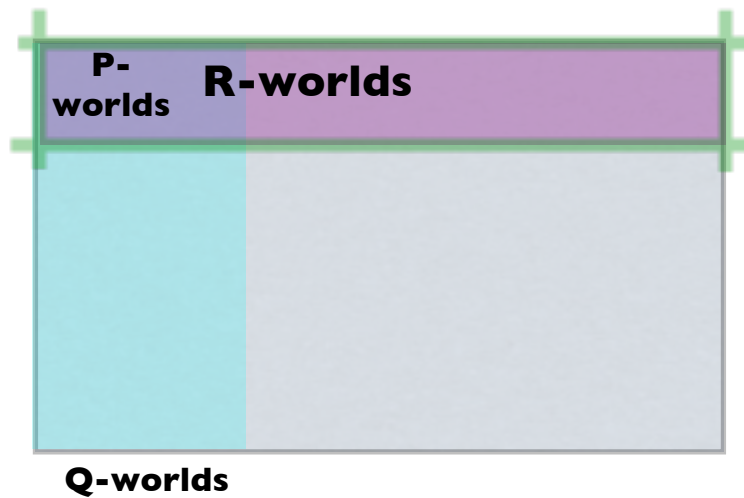
Hudson responds as I have already said that

if there are several different propositions whose conjunction with Q is P , then ...the weakest....of these shall be considered *the* difference between P and Q ("Logical Subtraction," 131)

This gives us a

pollyannish antithesis: the feeling that there *must* be a remainder is quite correct. the remainder $P-Q$ is the "best" (weakest?) R such that $Q \& R$ is equivalent to P .

I want to argue that there is considerable truth to both sides, and that the extrapolation model helps us see how this can be. Graphically the model is this:



The question is: what requirements on R are symbolized by this diagram? It seems to me there are three. For R to extend P beyond Q – for it to go on in the same way, as it were – R should meet these three conditions:

1. *within* Q , R is true/false in a world iff P is
2. *within* Q , R is true/false in a world \underline{w} for the same reasons that \underline{w} is $P \& Q$ rather than $\neg P \& Q$ (the reverse)
3. *outside* Q , R is true/false for the same reasons as within.

1. says that R is equivalent with P within Q , so call it equivalence. 2. Speaks to the reasons why R is true/false within Q , so call it reasons. 3. takes a bit more explanation. R 's acquiring truthmakers (or falsemakers) as you crossed the Q -border would mean that R for one kind of reason in Q -worlds and another in non- Q worlds – it "changed direction" as it crossed the Q border. 3. then is a kind of orthogonality condition and so I call it orthogonality. For our purposes today we can ignore reasons and work just from equivalence and orthogonality.

Back to the main claim: suppose that the remainder R has got to agree with P within Q , and agree with itself across the Q -border; then, I claim, there is truth in both the mysterian and the pollyannish position.

The truth in mysterianism. The lesser truth is that $Q \supset P$ is a terrible candidate for the role of P - Q . (Already said it isn't part of P but this is a different objection.) $Q \supset P$ agrees with P within Q , no problem there. But it's very far from agreeing with itself across Q . It is true for quite different reasons when Q is false than when Q is true, for the fact that Q is false is itself a reason for $Q \supset P$ to be true.

The greater truth is that it is dangerous to assume that there must be an intuitively satisfying remainder, even in cases where there are clear leftovers, as raising my arm clearly adds to my arm's going up that I in some sense meant my arm to go up. What would the remainder be in this case?

I will my arm to go up is too **weak**. It doesn't follow from my willing my arm to go up (R) and its going up (Q) that I raised my arm (P), since my arm might have gone up for other reasons. This is a violation of equivalence. *I will my arm up* isn't equivalent to *I raise my arm* even with the region where my arm goes up.

I effectively will my arm to go up is too **strong**. One reason it is false in *arm-stays-down* worlds is precisely that my arm stays down, making the act of will *ineffective*. This is obviously not a falsity-maker that obtains also in *arm-goes-up* worlds. *I effectively will my arm to go up* thus takes on new falsity-makers as we pass out of the region where my arm goes up. This is a violation of orthogonality. R should be false for the same sorts of reasons when my arm doesn't go up as when it does.

The truth in pollyannaism The remainder R is supposed to be a proposition that agrees with P within Q and, if you like, agrees with itself across the Q -boundary. In a minute I will argue in pollyannaish fashion that a proposition like that always exists.

But hold on: don't we know that in some cases it cannot exist? Haven't we already agreed with Jaeger that subtraction is not always well defined? There is a subtle ambiguity here. It is one thing to say that subtraction is well-defined as a logical operation on propositions. That means there is always such a proposition as $P-Q$. It is another thing to say the proposition itself is well-defined. That means that go to any world you like, the proposition is true or false there. This is in fact the synthesis: when Q is intuitively inextricable from P , a proposition $P-Q$ still exists, just don't try evaluating it at (too many) worlds where Q fails.

Let's now work towards a recipe for remainders. One objection to the $Q \supset P$ theory is that $Q \supset P$ is never part of P . ($Q \supset P$ has truthmakers I_i not implied by truthmakers for P .) Another, related, objection was that $Q \supset P$ isn't orthogonal to Q since it has different truthmakers when Q is false. A third much more direct objection it this the theory makes it way too hard for $P-Q$ to be false. The only way it can be false is for P to be false and Q true. But to go back to the simplest case, $a \& b - a$ is surely false if a and b are both false. But $a \supset a \& b$ is true when a and b are both false. But if

(1) $P-Q$ is false iff P is false and Q true,

how can we weaken it? P false, Q true is the limiting case of a phenomenon we can call P “adding falsity” to Q , a phenomenon that can obtain even when Q is false. The proposal is that

(2) P - Q is false iff P **adds falsity** to Q ,

where adding falsity, or being additionally false, or being false not just because Q is false, is the relation $a\&b$ bears to a when a is false.

What relation is that? Well, a is false for a reason that doesn’t depend on b , witness the fact that it could still obtain even if b were true. Interestingly this is a relation that makes sense even if P doesn’t imply Q ($a\&b$ adds falsity to $b\&c$ when a is false) so the following definitions are meant *not* to assume that X implies Y :

(3) X adds falsity to Y in \underline{w} iff $X\&Y$ has a Y -compatible falsity-maker in \underline{w} .

(4) X adds truth to Y in \underline{w} iff $\neg X\&Y$ has a Y -compatible falsity-maker in \underline{w} .

And now here are truth-conditions of P - Q :

(5) P - Q is

- i. false in \underline{w} iff P adds falsity to Q in \underline{w}
- ii. true in \underline{w} iff P adds truth and no falsity to Q in \underline{w}
- iii. undefined iff neither true nor false, that is, neither P nor its negation adds falsity to Q .

Two things here deserve comment.

Why in ii. do we say “adds truth *and no falsity*”? This because there is nothing to prevent a sentence adding *both* truth and falsity. Suppose I as a matter of fact have two dogs, a boy dog Sparky and a girl dog Barky. Let Y be *I have exactly one dog*, and let X be *I have only girl dogs*. $X\&Y =$ *My one dog is female* is false for the Y -compatible reason that I have a boy dog. $\neg X\&Y =$ *My one dog is male* is false for the Y -compatible reason that I have a girl dog. So *I have only girl dogs* adds both truth and falsity to *I have only one dog*—truth in that it gets one of my dogs right and falsity in that it gets the other wrong.

It says in iii. that the remainder is undefined iff neither P nor its negation adds falsity to Q . An example to show how that happen: $Q =$ *France has a king*, $P =$ *The king of France is bald*. (I am assuming P presupposes Q in the old-fashioned sense that it and its negation both imply Q .) Does P add falsity? $P\&Q =$ *France has a bald king* is false, I would say, because France lacks a king—not because

France lacks a bald king. This by proportionality, since the “bald” is an irrelevant extra. That falsity-maker is not Q -compatible, though, since it’s Q ’s negation. So *The king of France is bald* adds no falsity to *France has a king*. Neither, for similar reasons, does *The king of France is not bald*, at least not on a narrow scope reading. ($\neg P \& Q = \text{France has a non-bald king}$. This is false because Q is false.)

That neither *The king of France is bald* nor its negation adds falsity seems like one possible explanation of the fact that, as Strawson said, *The king of France is bald* seems unevaluable—at any rate much less evaluable than *I had breakfast with the king of France this morning*. The difference would be explained by saying that *I had breakfast...* does add falsity; it is false because I ate alone, which is fully compatible with France having a king off to the side.

- (6) Hypothesis: Let S presuppose a falsehood π . S suffers from catastrophic presupposition failure (it’s intuitively unevaluable) iff neither S nor its negation adds falsity, or equivalently $S \neg \pi$ is undefined.

This of course suggests that S will strike us as evaluable if $S \neg \pi$ is defined, and so we might as well make that explicit:

- (7) Hypothesis: Let S presuppose a falsehood π . S suffers from *non-catastrophic* presupposition failure (it’s intuitively evaluable) iff $S \neg \pi$ is undefined; and it strikes us as true/false iff $S \neg \pi$ is true/false.

Should this hypothesis be correct then philosophical arguments that rely on intuitive judgments of truth and falsity might need to be reformulated. A certain kind of Platonist argues like this:

- i. *The number of Martian moons = the number of shoes I’m wearing* is undeniably true.
- ii. It cannot be true unless there are numbers
- iii. So there are numbers.

The challenge is to distinguish this argument from a superficial analogue:

- i. *The king of France will never own this watch* is undeniably true.
- ii. It cannot be true unless France has a king.
- iii. So France has a king.

In both cases i. can be challenged as follows. Strictly speaking these sentences are false. They strike us as true because we are a semantically forgiving tribe;

we try not to hold the falsity of their presuppositions against them. A sentence S is evaluated as though it expressed the proposition $S-\pi$. $S-\pi$ really is true in these cases. But we can't derive ontological conclusions from it because the ontological conclusion depends on π , and π has been skimmed away.

I want to return now to remainder-propositions more generally; let's assume as before that P implies Q . The truth-conditions that for all its complications, subtraction is *almost* a truth-functional operation. The truth-table looks like this:

P	Q	$P-Q$			why
t	t	t			P can't add falsity unless $P\&Q$ is false; $\sim P$ adds it because $\sim P\&Q$ is false and Q is true
t	f	X			P implies Q
f	t	f			P adds falsity because $P\&Q$ false, P true
f	f	f	t	u	the non-truth-functional case

P adds falsity P adds truth only P adds nothing

That we get this kind of remainder proposition in every case would seem to vindicates the pollyannish view, assume $P-Q$ really does extrapolate P beyond Q . It does, so pollyannaism is to that extent vindicated. Let's now try to offer something to the mysterian too.

- (8) Q is perfectly inextricable from P iff $P-Q$ is not defined in any $\sim Q$ -worlds – neither P nor its negation adds falsity to Q except when Q is true.

Notorious example of this: Tom is a tomato. $P = \text{Tom is crimson}$. $Q = \text{Tom is red}$. Perfect inextricability would mean that neither Tom is crimson nor its negation adds falsity to Tom is red , unless Tom is indeed red. So let's go to a world where Tom is not red but, say, green. P adds falsity iff $\text{Tom is a crimson-y red}$ has a Tom is red -compatible falsity-maker when Tom is green. $\sim P$ adds falsity iff $\text{Tom is a non-crimson-y red}$ has a Tom is red -compatible falsity-maker when Tom is green.

Now both $\text{Tom is a crimson-y red}$ and $\text{Tom is a non-crimson-y red}$ have falsity-makers when Tom is green. The fact that Tom is green, for example. But the question is whether there's anything about green-Tom *that Tom can keep when he's red* in virtue of which he's not a crimson-y red (or in virtue of which he's not any other kind of red, i.e. he's red if crimson).

Thinking about this I want to reach for the kind of thing Wittgenstein says in *Remarks on Color*: “There can’t be a transparent white, a luminous grey”—that kind of thing. Let’s imagine Wittgenstein has discovered there can’t be a dull crimson; crimson is inherently bold, vibrant. And let’s imagine Tom is a *dull* green. Then one could *try* to say that *Tom is crimson* is false, due not to the fact that Tom is green, but due to the lack of vitality of Tom’s color whatever it is. I can’t say for certain that no one could develop a system along these lines. But on the face of it, it seems silly; the reason it’s false that Tom is red is that Tom is green, not that Tom has some special higher order colorish property that red things and green things can share.

Crimson - red looks, then, like a case of perfect inextricability. Others that come to mind: *we danced badly - we danced, I swam 3 laps - I swam, it weighs 10 pound - you ate it weighs over five pounds, I ate almost as much cake as you - You ate cake...*It would be interesting to check these against the proposed definition of perfect inextricability. But let’s now move on to the opposite kind of perfection:

- (9) *Q* is perfectly extricable from *P* iff in each $\neg Q$ -world, either *P* adds falsity to *Q* or $\neg P$ does, but not both.

Take the case where *P* = *the number of dragons is zero* and *Q* = *there is such a thing as the number of dragons*. The question is, does it hold in every number-of-dragons-less world (let’s say numberless world to save breath) that exactly one of *The number of dragons is zero*, *The number of dragons is not zero* adds falsity to *There is such a thing as the number of dragons*. The answer is pretty clearly yes: the first adds falsity in worlds with dragons—the presence of dragons is number-compatible falsity-maker for *The number of dragons is zero*—and the second adds falsity in worlds without dragons

Between these two extremes is a vast unexplored ocean of imperfect extricability. Take Wittgenstein’s arm-raising example, and look first at a world where I have not the slightest intention to raise my arm, let’s say because I am unconscious. Here it seems *I raise my arm* does add falsity to *My arm goes up*; unconsciousness would seem to falsify the first while still being compatible with the second. Does *I don’t raise my arm* add falsity? The question is whether *My arm goes up without my raising it* is false in this world for an arm-up-compatible reason. I don’t see that it is; it’s false because my arm didn’t go up, not because of anything about my raising it. Here then it looks like the remainder is evaluable. Move now to a world where I try to raise my arm but fail, because someone is holding the arm down. Is *I raise my arm* false for an arm-up compatible reason? Not on the face of it, it’s false simply because my arm doesn’t go up. Is *My arm goes up without my raising it* false for a reason detachable

from the fact that my arm doesn't go up? No, that too is false just because the arm doesn't go up. This is a world then where the remainder is unevaluable.

Not everyone finds these extricability issues as gripping as I do (not anyone, actually) so let's look briefly at three areas of possible philosophical relevance.

I mentioned that philosophy has tended to approach elusive contents from *below*, asking how they might be reached by conjoining weaker contents. What are the prospects for "analysis from above," in which we characterize a content by first overshooting it and then scaling back? The difference of course is that conjunctions are always well-defined, while remainders are not, if the subtrahend is imperfectly extricable from the minuend. And certainly there have been some spectacular failures here, for instance, the attempt to define "narrow" or "solipsistic" attitudes--what Dennett called the "organismic contribution" to the attitude--by subtracting away those aspects of externalist attitudes that pertain to the external world. But there have also been some successes, like *quasi-memory* and *having a visual experience as of a tomato*.¹ Or were they successes? The whole matter needs investigation.

A second place subtraction might help is with metaontology-- specifically the project of making the world safe for people who can't bring themselves to take (some) ontological questions seriously. One thing these people need is a convincing *objection* to ontological orthodoxy: to the view that there is a fact of the matter about whether my socks have a mereological sum, or there is such a thing as the number 9, or the amount of water in the basin. But what they also need is a model of how it could *be* objectively moot whether these things exist. Ideally it would be a model that helped to explain why some ontological questions seem mooter than others. I am tempted to think that the question "do so and so's exist?" is moot when and because the assumption of their existence is perfectly extricable from the hypotheses that make it. (.....)

One final appn, this one of a very different nature. I have been characterizing subtraction as a way of *cancelling* the subtrahend's content, as contrasted with *negating* its content. But on a certain view of negation, negation is itself just a cancellation device.² Here is Strawson in Introduction to Logical Theory (p2)

¹ Imperfect inextricability might be less of a problem than it seems. *A-B* doesn't have to be evaluable everywhere to be evaluable in our little corner of logical space.

² Priest, "Negation as Cancellation, and Connexive Logic" (Topoi 18, 1999: 141-8).

Suppose a man sets out to walk to a certain place; but when he gets half way there, he turns round and comes back again. This may not be pointless. But, from the point of view of change of position it is as if he had never set out. And so a man who contradicts himself may have succeeded in exercising his vocal chords. But from the point of view of imparting information, or communicating facts (or falsehoods), it is as if he had never opened his mouth...the standard function of speech...is frustrated by self-contradiction. Contradiction is like writing something down and erasing it, or putting a line through it. A contradiction cancels itself and leaves nothing.

Strawson seems to be suggesting that a speech starting with A and following up with $\sim A$ has no effect on the conversational score: it's as if the speaker never opened his mouth. Is it just me, or is Strawson totally wrong about this? Even if we grant that the later $\sim A$ erases the earlier assertion of A , why think that A returns the favor, erasing the later assertion of $\sim A$?

Strawson's suggestion that negation is, or can be, a cancellation device raises an interesting question. What *does* one say to wipe the slate clean after making an assertion one then thinks better of? What goes in for X in the update rule

$$A + X = \text{nothing asserted?}$$

$\sim A$ is too strong, we saw; it leaves us with something still asserted.³ $A \vee \sim A$ is too weak. It might in some contexts carry an *implicature* that A is withdrawn, but it doesn't itself constitute a withdrawal of A . What we need, it seems, is something *weaker* than $\sim A$ but not so weak as to be a logical truth. I know of only one form of words that cancels A cleanly: *it might be that $\sim A$* . The equation we're looking for is

$$A + \diamond \sim A = \text{nothing asserted.}$$

This is interesting because we know of another operation that returns us to the nothing-asserted state, viz the operation of *subtracting* A from our earlier statement that A .

$$A \text{ minus } A = \text{nothing asserted.}$$

Putting those two equations together we get

$$\text{adding } \diamond \sim A = \text{subtracting } A,$$

³ *I shouldn't have said that* A is too strong as well, for it leaves the speaker still committed to a claim about what she should have done.

or, moving the negation,

adding $\diamond A =$ subtracting $\neg A$.

That's just the shell of a theory of "might," but one worth exploring, I think, because of the help it gives with two puzzles.

(1) I mentioned epistemic modality in the first lecture: the kind of modality expressed by words like "might" and "maybe." I complained that the usual sort of view -- "Bob might be in his office" is true in my mouth iff my information (or information available to me) is consistent with his being there -- seems on the face of it to get the subject matter wrong; the subject matter is Bob and his office, not the extent of my information.⁴ It was unclear at the time how any theory of "might" could avoid shifting the subject matter, but now we see how the thing might be possible. Negating A doesn't change its subject matter, and disavowing something as opposed to asserting it doesn't change the subject matter either; and attaching "might" on the present theory is ringing those two changes in sequence; it's disavowing the negation of A .

(2)⁵ The following argument is invalid: *It might be the case that $\neg A$, therefore $\neg A$.* An one-premise argument X therefore Y is invalid, one would think, iff the conclusion can be false while the premise is true, in other words if there is a possible scenario where $\neg Y \& X$. In this case that means a possible scenario where $A \& \diamond \neg A$. And there is no such scenario. The problem here isn't just unassertability, for unassertible hypotheses can still be hypothesized, say, in the antecedent of a conditional. And it makes no sense to say, *If it rained last night, but it might not have rained, then the clothes will be wet.* It all looks quite different on the cancellation account. *It might be the case that $\neg A$, therefore $\neg A$* is invalid, not because $\diamond \neg A$'s truth doesn't force $\neg A$ to be true, but because not asserting A doesn't force me to assert that $\neg A$. $A \& \diamond \neg A$ is incoherent, even as a supposition, because the instructions it gives are to suppose that A and also not suppose that A .

Summing up: Content-*parts* are not properly so-called unless there is such a thing as the remainder when A 's part B is subtracted from it. There would seem to be plenty of cases where there is *not* such a thing, including Wittgenstein's example: the remainder when we subtract my arm going up from my raising it. The dilemma is resolved by saying $A-B$ exists, but is undefined in

⁴ A: Bob might be in his office, Samantha might be on the plane, ... B: Get over yourself!

⁵ Yalcin's paradox.

some or all of the region where B is false. There is still the question of what the remainder proposition $A-B$ is. $A-B$ is usefully construed as the R that best extrapolates B beyond A . This suggests a way of constructing $A-B$ that yields a proposition with the desired properties. (In particular, though I didn't prove it, $A-B$ is a disjoint part of A , a part that does not overlap B . Possible applications include analysis from above, metaontology, and epistemic modality.