Commitment issues in the naive theory of belief[1]

J Robert G Williams

Part I: Commitments and belief

Commitment and rationality.

Ada and Beth are arguing over whether the Butler murdered the Gardener. Beth steadfastly maintains the Butler's innocence. Ada pounces on Beth's latest admission:

> "Ada, you just said yourself that whoever was in the toolshed at midnight must have done it. We know it was one of the staff. And we knew that the person in the woodshed was over six feet tall. But remember, the Butler is the only person over six feet tall on the staff. You believe all this! So put the pieces together: it must have been the Butler".

At the point at which Ada makes this intervention, Beth does not believe that the Butler is the murderer. She does, however, believe all the following:

1. The murderer was in the toolshed at midnight.
2. The murderer was one of the staff.
3. The Butler is the only member of staff over six feet tall.
4. The person in the woodshed was over six feet tall.

From these four propositions it *follows* that the Butler is the murderer. Ada, in pointing this out, is revealing to Beth that Beth is already *committed* to believing that the Butler did it.

Beth has several options at this point. She can yield to Ada's pressure and form the belief that the Butler did it. She give up one of (1-4), making the ensuing commitment go away. She might, perhaps, stubbornly cling to (1-4) but refuse to live up to her commitments. Failing to live up to your doxastic commitments is a bad thing (it makes you less than perfectly rational), but is possible.

[1] The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 312938. It was presented at a Cambridge graduate conference in the philosophy of logic and mathematics, and at a CMM intramural workshop. I thank everyone who has given me comments on the paper, and especially Andy Egan.

There's a final option for Beth, but to see it I will shift to a second case. Beth initially doubts that there could be irrational numbers x and y such x^y is rational. Ada gets Beth to accept all the following:

i. $\sqrt{2}$ is an irrational number.

ii. If $\sqrt{2}^{\sqrt{2}}$ is a rational number, then there exist irrational numbers x and y such x^y is rational. (since $x = y = \sqrt{2}$ provides an example).

iii. $\sqrt{2}^{\sqrt{2}^{\sqrt{2}}} = \sqrt{2}^{\sqrt{2}\times\sqrt{2}} = \sqrt{2}^2 = 2$

iv. If $\sqrt{2}^{\sqrt{2}}$ is not a rational number, then there exist irrational numbers x and y such x^y is rational (since $x = \sqrt{2}$, $y = \sqrt{2}^{\sqrt{2}}$ provides an example).

Ada claims that, since Beth believes (1-4), Beth is now committed to believe:

> (R), that there exist irrational numbers x and y such x^y is rational.

Beth can now respond:

> "(R) is a *classical* consequence of things I believe, but I reject the underlying classical presupposition that a given number is either rational, or not rational. The proposition in question has not been shown to be an *intuitionistic* consequence of what I believe. I'm not convinced it follows from what I believe via rules I accept. So I'm not committed to it".

Ada replies:

> "The issue is not what rules of logic you accept. The issue is what the right rules are. The correct logic is classical, and the reasoning I gave you was valid. You are committed to the conclusion, whether you acknowledge that commitment or not".

Ada and Beth agree on the status of the argument according to the two logics in question: the conclusion is a classical but not an intuitionistic consequence of (i-iv). The conversation can now proceed in two ways. Beth can agree with Ada that what matters as far as her commitments go is not what logic she acknowledges, but what the logical facts are. In that case, Ada and Beth will have to agree to disagree about whether Beth is committed to believe R, as a reflex of their disagreement about the One True logic. The other option, however, is that Beth can insist that commitments (what the world must be like, given her beliefs) depend on the logic she acknowledges. If she can convince Ada of this, then Ada will agree with that Beth is not committed to believe R, even while Ada continues to believe R follows from what Beth believes.

The first option is an externalist conception of commitment. The second is an internalist conception. Both are well-defined relations, so the only question is what theoretical role each plays. Elsewhere, I've argued (in effect) that internalist commitment is what matters for rationality (Williams, 2017). But I will continue to keep both on the table, since they are subject to a common problem.

The puzzle of commitment

Ada accepts classical logic, but the classical rules that she accepts can be chained together in ways that she lacks the time, energy and inclination to run through. The same goes for the intuitionistic rules Beth accepts. So whether we think it is Accepted Logic or the One True Logic that defines commitment, a person's commitments-to-believe can far outstrip what they believe they are committed to. Complex mathematical truths—e.g. of classical or intuitionistic arithmetic—bear witness to the phenomenon.[2]

There is a puzzle about commitments-to-believe that afflicts any computationally bounded creature. Ada believes each of the axioms of classical arithmetic, and accepts classical inference rules. She also believes the negation of Fermat's Last Theorem, FLT (perhaps on the basis of some erroneous testimony). What now is she committed to believe?

The axioms of arithmetic that Ada believes entail, via classical rules that Ada accepts, FLT. So the totality of what Ada believes—the axioms and the negation of FLT—form a classically contradictory set. Classical logic also says that any proposition whatsoever follows from a contradictory set of propositions. So in virtue of (perfectly reasonably) believing the axioms and the negation of FLT, Ada is committed to believing any proposition whatsoever. This is so on both the internalist and the externalist conception of commitment-to-believe.

[2] In the picture of rationality I defended in my (2017), I was concerned to allow that an agent may not accept any named logic at all, but instead, in a clear-eyed way, be uncertain between then. Perhaps you are like this, being agnostic over whether Ada or Beth has locked onto the right logic. An analogue of the point still remains. One can be committed to believe-under-a-supposition things one doesn't in fact believe-under-that-supposition. *Under the supposition* that classical logic is right, you accept the same rules as Ada does, and so your commitments to believe-under-that-supposition far outstrip your current beliefs-under-that-supposition in ways that you haven't the time, energy or inclination to address.

The example exploits a structural vulnerability: if our commitments outstrip what we currently take our commitments to be, it will always be possible for us to make a mistake and believe contrary to our commitments. If Ada had unlimited cognitive resources, she could avoid getting into this situation simply by deriving the consequences of her beliefs and monitoring their consistency. But none of us are like this. Most of us, indeed, have hidden inconsistencies in our beliefs, which we try to iron out as they come to light.[3] So if commitment is as I have taken it to be, it seems that most of us are committed to believe absolutely every absurd proposition: that Trump is a great president, that the moon is made of green cheese, that we ourselves do not exist.[4]

So, on the one hand, it does very much seem that our commitments-to-believe things outstrip what we in fact believe. What use would be a theory of commitment on which this was not so? On the other hand, if our commitments explode in the way just sketched, this looks like an utterly undiscriminating, and so useless, notion.

What is to be done

The strategy of this paper is to locate a partner in crime—a theory which faces this kind of puzzle in its rawest form, and has grappled with it before. The extant criminal is the theory of belief that Stalnaker laid out in *Inquiry* (1984). Stalnaker's account is highly revisionary in various ways. In particular, if one believes p, and p entails q, then according to Stalnaker one believes q. By contrast, the puzzle above is a puzzle even for a naive theory of belief where no such closure properties are assumed. Since even on the naive theory what we are committed to believe is closed under an appropriate consequence relation, there are structural parallels.

Part II explains Stalnaker's account of belief, concentrating in particular on the foundational picture of the metaphysics of belief that motivates his conception, and the way that his approach allows for inconsistent beliefs.

In Part III, I examine the feasibility of adapting this strategy to solve the commitment problem. I also set out a rival foundational picture, closely related to Stalnaker's, that will avoid Stalnaker's arguments for his revisionary picture of belief. This rival metaphysics is suited to underpin a more naive picture of belief, including the structures needed for an adequate treatment of commitment.

[3] The preface paradox of Makinson (1965) gives one general route to this conclusion.

[4] I am not going to discuss an interesting alternative reaction: to model commitment not with the One True Logic, nor with Accepted Logic, but with some extremely weak special-purpose logic. For an example of this reaction, and references to the background literature, see e.g. Berto (forthcoming).

Part I: Stalnakerology.

Stalnaker on equivalence and omniscience

Stalnaker argues for the following:

> The most basic belief-relation that an agent stands in is to a “belief state”. The content of a belief state is a set of possible worlds (those possibilities that, for all the agent believes, could be actual). The agent believes a thing if and only if the content of a belief state of theirs entails it.

For all this says, agent may stand in the belief-relation to many belief-states. But irrespective of the number of belief states they are in, two striking consequences follow:

* **The Equivalence Property.** If p and q are equivalent (i.e. entail one another) then an agent believes p iff she believes q.
* **The Closure Property.** If q is entailed by p, then an agent believes q if she believes p.

Closure implies Equivalence, and has two more immediate consequences:

* **The Omniscience property**. If a subject believes anything, then for every necessary truth p, she believes p.
* **The Consistency property.** If p is any proposition that is not possibly true, then no subject who fails to believe at least one thing believes p.

There seem to be intuitively compelling counterexamples to all of these. Here are two. Against Equivalence: Lois believes that Clark Kent is sitting in front of her, but she doesn’t believe Superman is sitting in front of her. But since Clark Kent is Superman, the proposition that Clark is sitting in front of her entails and is entailed by the proposition that Superman is sitting in front of her. Against the Omniscience property: Lois doesn’t believe Fermat’s Last Theorem (FLT). Maybe she explicitly suspends judgement on the question, having not seen a convincing proof. But FLT is necessary, so by Omniscience, she believes it. Against the Consistency property: Lois may believe, on the basis of testimony or a mistake in reasoning, the negation of Fermat’s Last Theorem (or that water is not H2O, or that composition is not universal—substitute in the negation of your favourite necessary truth). Since Closure implies the other properties, it appears to be counterexampled three times over.

As Stalnaker understands belief, the modal concepts (entailment, necessity, possibility) relevant to the account of belief, and which feature in Closure and its siblings, are to be understand as expressing *metaphysical modalities.* For p to entail q is for q to obtain at every metaphysical possibility where p obtains. Some theorists have accepted the letter of the above, but advocated an alternative understanding of entailment—perhaps focusing on narrowly logical consequence. This helps around the edges, but it is easy to construct analogous equally compelling counterexamples.

Stalnaker is well aware of the surprising consequences of his theory of belief. He argues that this is accounted for by slack between two distinct things: true *ascriptions* of belief, and the facts about the belief-relation itself. What is said when we utter the words "Lois believes that Fermat's Last Theorem" might not be true—even if Lois does indeed stand in the belief relation to the proposition that the statement of Fermat's Last Theorem expresses. The account he gives of the way this slack arises will matter to us a great deal—we will look at if further, but for now, let it be a promissory note.

To understand Stalnaker's perspective, the question to ask is this: Why start from a model of belief that obviously has such odd consequences, such that one then has to scrabble around to make a case that it is not immediately refuted? One possible answer that I will not go into here is that this kind of theory of belief (and refinements of it, such as degrees of beliefs as probability functions defined over an algebra of possibilities) has proved theoretically fruitful—in economics, decision theory, linguistics, and so forth. This is a potentially excellent motivation, but exploring it would get us caught up in issues such as: to what extent such theoretical deployments concern ideal rather than real, people? To what extent should the features of these formal models (e.g. entailment) be glossed in the way Stalnaker advocates? Stalnaker's stated motivation in *Inquiry* is rather different. He describes a metaphysics of belief and argues his theory of belief is a consequence of it. The next two sections explore these arguments.

The metaphysics of belief as motivating equivalence

Stalnaker (1984) offers a theory of how a naturalistic world can contain content-bearing states. He advocates a "causal-pragmatic picture" of content. I understand this as follows. Let an *interpretation* of an agent be an abstract function which maps "states" of the agent into pairs of attitude-types and contents. Thus relative to an interpretation of Harry, we can say that Harry's state s1 is a belief with content p, and his state s2 is a desire with content q. Then I take it that Stalnaker proposes two conditions for an interpretation I of Harry to be correct:

(a) If I maps s to <belief, p> then s must be something that, optimally, Harry is in only if p, and optimally, he is in that state because of p or something that entails p;

(b) Harry is disposed to act in ways that would tend to satisfy the contents assigned by I to desire-states (i.e. those classified by I as desires) at worlds in which the contents assigned by I to belief states (i.e. those classified by I as beliefs) are true.[5]

The minimal version of the theory tells us that (a) and (b) are jointly sufficient, as well as necessary, for an interpretation to be correct. Stalnaker emphasizes that this is a first pass at articulating the causal-pragmatic picture and may need refinement. Nevertheless, he treats it as representative when drawing out consequences for belief.

Suppose we accept this view of the correctness of interpretations. Stalnaker then argues that it entails that the content of beliefs can be no more fine-grained than the conditional or causal-explanatory relations that feature in (a) and (b). As an example, consider the content that grass is green. In order to be in have a belief with this content, it must be that we are in some S such that:

(i) Under optimal conditions, we are in s only if grass is green, and

(ii) Under optimal conditions, we are in s because grass is green, or because of something that entails that grass is green

It follows from (i) and (ii) that:

(iii) Under optimal conditions, we are in s only if grass is green and Fermat's last theorem holds, and

[5] Relevant texts:
"To desire that P is to be disposed to act in ways that would tend to bring it about that P in a world in which one's beliefs, whatever they are, were true. To believe that P is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which P (together with one's other beliefs) were true." (op cit p.15).
"We believe that P just because we are in a state that, under optimal conditions, we are in only if P, and under optimal conditions, we are in that state because P, or because of something that entails P" (op cit p.18).
"Beliefs have determinate content because of their presumed causal connections with the world. Beliefs are *beliefs* rather than some other representational state, because of their connection, through desire, with action. Desires have determinate content because of their dual connection with belief and action. (op cit p.19)"

(iv) Under optimal conditions, we are in s because grass is green and Fermat's last theorem holds, or because of something that entails that grass is green and Fermat's last theorem holds.

But that is just to say that an interpretation satisfies clause (a) with respect to the content *that grass is green* just in case it does so for the content *that grass is green and Fermat's Last Theorem is true*. Nor will condition (b) knock out the FLT-involving content, since what matters is whether actions performed would promote the satisfaction of certain desires *in certain possible worlds* and the two candidate contents are true at precisely the same set of worlds. Stalnaker puts the situation thus: "We lack a satisfactory understanding, from any point of view, of what it is to believe that P but disbelieve that Q, where P and Q stand for necessarily equivalent propositions" (op cit p.24) His recommendation is that we accept the conclusion: belief-states are 'coarse grained', in the sense that they have the content p iff they have the content q, whenever p and q are equivalent.

Stalnaker's argument doesn't turn on the specific details of his account of correct interpretation. The same style of argument can be given for any account that has the form: (a) If I maps s to <belief, p> then $R_1(s,p)$; and (b) If I maps s to <desire,q> then $R_2(s,q)$, so long as the matrices R_1 and R_2 do not discriminate among necessarily equivalent propositions, i.e. are intentional rather than hyperintensional operators.[6] The counterfactuals, causal-explanatory relations and restricted modals of the causal-pragmatic account have this feature, according to Stalnaker, but so do many of its rivals. Stalnaker's underlying argument is that a good account of correct representation has to be built out of a basic set of "naturalistically respectable" resources, and those resources are intentional rather than hyperintensional. Given that, the coarse-graining conclusion follows as before.[7]

The conclusion of the argument is a form of equivalence, but it is not quite the form I originally laid out. The original principle talked of what the agent believes. The form above concerns the content of belief-states that the agent is in. In order to bridge the gap, we need a principle that

[6] More generally, the constraints take the form: if I maps each s1…si… to <A1,p1>…<Ai,pi>…, respectively, then R*(s1,p1,…si,pi…). Then the argument goes through so long as the relation R* is intentional rather than hyperintensional.

[7] Note that it is consistent with this argument that we can characterize an (extensional) belief relation between an agent's belief state and a structured proposition, conceived as Russellians or Fregeans favour. We may use the extensional belief relation bel(x,y), where y is some object (e.g. a structured proposition) to analyze truths about belief by claiming: x believes that p iff (Ey) (bel(x,y) and y is the proposition that p). But if propositions are Russellian structures or Fregean thoughts the expression "y is the proposition that p" is itself hyperintensional in the sense that substitution of cointensional sentences in the p-position can change its truth value. The Stalnakerian would request a naturalistic analysis of this notion.

connects the three-place relation between an agent, a state and its content, to a two-place belief-relation between an agent and a content-believed.

Belief, belief states and closure.

There are two salient options for connecting the content of belief states to facts about what an agent believes:[8]

(**Pointwise**): An agent believes that p iff they are in some state s, such that s is mapped by the correct interpretation to <belief, the proposition that p>
(**Holistic**): An agent believes that p iff they are in some state s, such that s is mapped by the correct interpretation to <belief, the proposition that q>, and q entails p.

Either would suffice to complete the argument of the previous section—so in order to finish that argument, we wouldn't need to take a stance. But distinctively, (Holistic) also entails Closure (and so would entail Equivalence, Omniscience, Consistency). I present in this section an argument for Holistic using materials that appear on p.82-83 of *Inquiry*—though I'm unsure how close it is to what Stalnaker originally intended.

I want to first show that we must understand clause (b) in a very particular way if the account is to work. Recall that for Stalnaker, an agent can be in an inconsistent belief states, that is, it can be the case that there is no possibility at which all the contents of all of Harry's belief states are true together. But now consider again the way Stalnaker characterizes the content of desires:

> (b) Harry is disposed to act in ways that would tend to satisfy the contents assigned by I to desire-states (i.e. those classified by I as desires) at worlds in which the contents assigned by I to belief states (i.e. those classified by I as beliefs) are true.

Suppose Harry's belief states are inconsistent. Then there are no worlds in which the contents (plural) assigned by I to his belief states are all true. Clause (b) is a counterfactual with an impossible antecedent. Such counterfactuals are standardly treated as vacuously true. Any assignment of desire content whatsoever would therefore satisfy this clause. But obviously some attributions of desire to agents with inconsistent beliefs are correct, and others wrong. Something has gone wrong. To avoid all this, I read clause (b) as follows:

[8] For Stalnaker, "the proposition that p" can be characterized modally: it is a set S of possible worlds such that w in S iff were w the case, p. This treatment of course presupposes the conclusion of the coarse-graining argument.

> (b*) Harry is disposed to act in ways that would tend to satisfy the content assigned by I to a desire-state at worlds in which the contents assigned by I to an associated belief state are true.

This starts with each individual belief state. For Stalnaker, that is always a consistent proposition, so we can non-vacuously look at how the agent acts in worlds where it is true. And now it makes sense, in principle, to fix on a proposition as that one which the agent's actions at those worlds tends to bring about. The proposal is then that that proposition is the content of a desire state paired with the belief state we started with.

With the proposal clarified, consider the following. An agent has an overriding desire for food and raises their hand because they believe each of the following:

(i) a person wearing a red hat will get food if and only if they raise their hand.
(ii) I am wearing a red hat

It's not true that they act as to satisfy their desires at all worlds in which (ii) is the case (at some of those worlds, people wearing red hats go hungry when they raise their hand. Nor is it the case that they act as to satisfy their desires at all worlds in which (i) is the case (at some of those worlds, they are wearing a blue hat, and so raising their hand is ineffective). So if belief states were conceived to pair one-to-one to beliefs, then Stalnaker's clause (b*) would not work at all. The content of the relevant belief state would have to entail both (i) and (ii). In general, the causal-pragmatic account will only fly if the content of a belief state is a "totalizing" content, entailing the content of all practical relevant beliefs.

The causal-pragmatic account forces us to a totalizing conception of the content of belief states. To connect this to belief, we then face an ugly choice. (Pointwise) would falsely predict that the agent does not believe they are wearing a red hat. So it is wrong, and absent any better suggestions, (Holistic) is the way to go.

Conclusion to Part II.

It would be pleasant task to pick apart the motivations for Equivalence and Closure set out above, and identify loopholes. What conception of belief would we motivated, for example, if we think that counterfactuals with impossible antecedents are not vacuous? And could we avoid closure if we, using the sophisticated resources of (Leitgeb 2017), deny the possibility of inconsistent beliefs altogether?

But my concern is to get back to the puzzles about commitment, so I simply finish this section with an observation about the relation between the two arguments discussed. As noted, Equivalence follows from Closure, so the conclusion of the second argument is stronger than the first (their premises are different, though, so it's not the case that the latter supercedes the former). But the first argument has an additional dimension of interest. As mentioned earlier, Stalnaker has a particularly strong version of (Equivalence) and (Closure), on which p entails q when it is metaphysically necessary that either not p, or q. Why not, ask some critics (e.g. Jago 2014), adopt the structure of this account, but make the relevant entailment relation much more demanding? If for example entailment was *analytic* entailment, then (Equivalence) would not have the disturbing Clark Kent-Superman implications, since it's not *analytic* that Clark is Superman. Or perhaps it could be even more demanding, with entailment requiring truth preservation at every *epistemic possibility* for the agent—where an agent's epistemic possibilities might include, for all we have said, worlds that are *logically* impossible. Now, the argument I extracted for Closure via (Holistic) is essentially schematic: an argument for closure of belief under *the relevant notion of entailment, whatever that might be.* So these considerations don't give Stalnaker much leverage against critics who want to preserve the formal structure of his model but implement it with a different modality. However, the arguments for Equivalence do provide such leverage.

Part III. Fragmented commitment, modified metaphysics

Having explored Stalnaker's theory of belief and its metaphysical motivations, I'm now go back to the original, naive picture from which I started. On this view, the beliefs of ordinary agents like us satisfy neither (Equivalence) nor (Closure). In virtue of what we believe, however, we incur commitments to believe other things—and that generated a puzzle, since when we have inconsistent believes (or inconsistent-by-our-own-lights beliefs) then commitment trivializes.

Commitment and fragmented belief

Suppose the set B contains all the propositions Harry believes. On the current, naive, conception of belief, B need not be closed under equivalence or entailment, but the set of all the propositions that Harry is committed to believing, C, is closed under entailment (and so equivalence). The relevant conception of entailment under which C is closed will depend on the details of the conception of commitment with which we are working. But C satisfies all four of the properties that on Stalnaker's conception, B itself enjoys.

Stalnaker allows that B contain inconsistent beliefs. Beliefs stand to Stalnakerian belief-states in a many-one relation. The set B* of contents of an individual's belief-states consists of several individually consistent but potentially mutually inconsistent propositions. B then contains all propositions entailed by any member of B*. B* therefore defines a certain structure over B—dividing it into various overlapping and individually consistent "fragments".

The analogous move for a fan of the naive picture of belief and commitment is to posit a parallel structure of internally consistent/mutually inconsistent fragments of C. We need not follow Stalnaker in thinking of a fragment as defined by something belief-like (with its own representational content)—that is part of Stalnaker's *explanation* of how the structure arises, but it is inessential. The minimal conception of a fragment is that of a space in which a subset of the agents beliefs are located and interact.

I propose to take nothing away from the naive picture of belief, but add an additional relation: a relation of co-belief. Strictly, this is a relation between beliefs, but it induces a relation among contents that are believed: contents can be co-believed or not. Maximal clusters of co-believed propositions are the fragments of our mind.

We now re-characterize commitment: we are committed to believe is anything that follows (in the logic we accept) from any collection of co-believed propositions. So long as we never co-believe mutually inconsistent propositions, commitment won't trivialize, even if we believe contradictory things. (There is an additional parameter to be factored in if we go for an internalist conception of commitment—I will set this out later, but for now, let us run with an externalist conception of commitment to keep things simple, so that notions like "consistency" are exogeneous).

With Stalnaker, I hold that the ideal believer would have a unified mind, where all beliefs are co-believed. This is why it's normatively bad to be inconsistent, even if you do not co-believe the inconsistent propositions. Let us say that we pool two fragments when we come to co-believe all the beliefs in each fragment. The problem with inconsistent beliefs, even if they are in different fragments, is that without changing your mind on one or the other, you cannot pool the relevant two fragments without things going badly wrong---becoming inconsistent within a single fragment.

With the revised account of commitment, it may appear we lose some of the original data. For example, it's pointed out to Harry that he believes p, and that he believes if p then q (and Harry accepts modus ponens). Nevertheless, Harry steadfastly suspends belief on q. This is terrible! And the simple theory of commitment would capture this by saying that Harry was failing to live up to his commitment to believe q. The revised account of commitment only says this in situations in which Harry's two original beliefs are co-believed. The accusation would be, though, that this is too weak. Harry's stance is terrible whether or not he co-believes the premises.

I think the ideal of unity just mentioned can explain what needs to be explained here. Let us divide it into two cases. The first case is Harry's belief that p and belief that if p then q are drawn from two fragments that are wildly inconsistent with one another. Then I am comfortable sticking to my guns: there's no sense in which Harry is committed to believing that q, and suspending may well be the best attitude for him to take. The situation is not a good one for Harry, but that can already be explained by pointing to the inconsistency of the two fragments, and we have already noted having inconsistent fragments is a way of being irrational, given the ideal of a unified mind. Appealing to commitment in this circumstance is redundant.

On the other hand, when two fragments are perfectly compatible with each other, the ideal of unity suggests that prima facie, they should be pooled. We should carefully distinguish between two situations: *that x is committed to p* and *that x would be committed to p if they were as they should be*. When Harry believes p and believes if p then q in two perfectly compatible but distinct fragments, then prima facie this is the case: if Harry were if he should be, he would be committed to q. So prima facie, given that Harry suspends on q, he is either not as he should be, or he is not living up to his commitments. So we again explained what is out of order in Harry's situation.

There are many cases intermediate between wild inconstancy and perfect compatibility. Sometimes we'd do better to integrate mutually consistent fragments by revising each for a more powerful overall account of the disparate data each reflects, rather than straightforwardly pooling. In situations where we have three jointly inconsistent but pairwise consistent fragments, it is not clear that pooling the pairs helps gets closer to the ideal of integration. So the normative truths about how and when to integrate a fragmented mind is messy. That's one reason for keeping them separate from the clean lines of the theory of commitment; but as we've seen, they can't be ignored.[9]

In endorsing this revised theory of commitment, not everything in the naive picture of belief can be preserved. For the naive picture, taken straight, would tell us that among the things we believe are straightforwardly contradictory propositions (propositions that are inconsistent even by our own lights, though we might not yet realize this). An example would be Lois's belief in the conjunction of the axioms of arithmetic and the negation of Fermat's Last Theorem. To borrow the Stalnakerian

[9] Thanks to Andy Egan here for questions that led to these paragraphs.

solution, then, we might also have to borrow some of his revisionism. And Stalnaker discusses strategies for reanalyzing these cases at length (1984, ch 4,5). In the case just mentioned, he denies that we believe the impossible conjunctive proposition—even when we believe the individual conjuncts. To be sure, we would be prepared to utter a conjunctive sentence that expresses this proposition, but even if we concede we believe that the sentence is true, this is not at all the same thing as believing the proposition it expresses.

If we have to follow Stalnaker in the reanalysis of apparent cases of inconsistent beliefs, this would be a significant concession, but it seems to me a price that would be worth paying. It tells us that the naive conception of belief *overgenerates* in some cases, attributing to us beliefs we don't really have. That is very far from the full-fat version of the Stalnakerian picture according to which the naive conception of belief dramatically *undergenerates*. I can live with a certain amount of revisionism around the edges here. I wouldn't want to play the card too often—for example, it seems clear to me that Graham Priest believes various explicit contradictions, without being committed to everything whatsoever. But that is why I favour the more internalized conception of commitment, on which beliefs with individually impossible content are perfectly kosher, as long as they're not impossible *by the lights of rules the agent accepts*.

I will not press the point further here, but I think we might escape even the cost just mentioned. At the end of this essay, I suggest that acceptance-of-rules, just like beliefs, will belong to some fragment of an agent's mentality. So while the agent might accept a rule of explosion, simpliciter, that acceptance need not figure in every fragment of her mentality. This would allow an agent accept classical rules, accept the conjunction of arithmetical axioms and negation of FLT, but deny that the acceptance of explosion and the conjunctive belief are in the same fragment. Indeed, in principle, beliefs that are bundled with no non-structural rules at all will contribute nothing at all (beyond themselves) to the agent's commitments. This might be a reasonable description of a belief which is a mere idee fixe, utterly insensitive and isolated from the rest of the agent's cognitive economy.

Co-belief and rationality

Stalnaker's model of belief is a special case of the co-belief strategy. If we accepted the Stalnakerian picture, a belief that p and a belief that q stand in the co-belief relation when there is some belief-state whose content entails both p and q. And we could also accept the definition of commitment to believe—it would just turn out that the picture predicts that we already believe anything we are committed to believe.

If we, as I have suggested, strip away the explanatory framework that supports Stalnaker's implementation of co-belief, the danger will be that the result is just ad hoc, a way to bar the monster of trivializing commitment, without independent motivation or foundation.

The best answer to this would be to find independent theoretic roles that co-belief plays, which would give us confidence that our new terminology picks up on a real doxastic phenomenon. It is very plausible that the theory of *suboptimal* rationality of agents, on the naive conception of belief, will already need to deploy a notion of co-belief. The following example, inspired by one given in Rayo (2013), illustrates this.

Suppose Harry and Sally both believe these four things: the patio is square, that it has one side of 8m, another of 7m, and its area is 49 metres square. These beliefs are mutually inconsistent (perhaps together with side-premises I have suppressed). Harry, going to the fencing store, buys 28m of fencing. Sally buys 30m of fencing. They each buy 49 metres worth of paving.

In either case, we can explain and rationalize their behaviour by appealing to subsets of their beliefs. In each case their paving-purchasing is rationalized by the area-belief. Sally's fence-purchasing is rationalized by her beliefs about the length of the long and short side together. Harry's fence-purchasing is rationalized by his beliefs about the length of the short side and his belief about its shape. The problem of course is that in each case we could pick some *other* subset of their beliefs that would rationalize a very different action. The beliefs about the lengths of sides of patio would support more paving. Sally has all the beliefs that led Harry to purchase a different length of fencing, and vice versa.

When an agents beliefs are perfectly consistent, we don't have to worry about what to ignore when rationalizing behaviour (or indeed, belief formation). But in cases of inconsistency, rival explanations are all in the offing. We *better* rationalize behaviour if we have more structure in description of the agents, so that we can explain Sally's choice to buy 30 *rather than* 28 metres of fencing. Co-belief could do this job for us, by identifying psychological difference between Sally and Harry: Sally co-believes that sides are 8m and 7m respectively, Harry does not. Harry co-believes that one side is 7m, and that the patio is square, and Sally does not. The contrast in their actions then can be rationalized by the difference in the ways the beliefs they share are differently integrated.

I have not provided a theory of suboptimal rationality that appeals to co-belief. I assert, rather, that there needs to be such a thing. Not surprisingly, the motivations for this echo what some

Stalnakerians say about fragmented belief and how it ties to action. But my point is that even on a naive non-Stalnakerian conception of belief, structure beyond belief will be required to explain what acts are or are not rationalized. I reject the accusation, therefore, that it is in any way ad hoc to appeal to co-belief in dissolving our problem of commitment.

Rival foundations.

As we saw in part II, Stalnaker's theory of belief was motivated by a causal-pragmatic metaphysics of belief. This metaphysics both predicts properties such as Equivalence and Closure, and also found a place for what we are now calling co-belief, the structure that in Stalnaker's framework is provided by the content of the underlying belief states.

I will finish by sketching a different metaphysics of belief, one that is a cousin of Stalnaker's own, but which is compatible both with the naive conception and with a Stalnakerian conception of belief. I will argue that it also provides foundations for co-belief—in a way that other rivals to the causal-pragmatic picture may struggle to do. This rival metaphysics is the form of Radical Interpretation I have developed in detail elsewhere (Williams 2018, MS).

Like the version of the causal-pragmatic picture I set out earlier, Radical Interpretation starts taking as given a space of abstract interpretations, which for present purposes we take to be assigning attitude-content pairs to states. One among these interpretations is correct (or, if indeterminacy arises, then perhaps a cluster will be co-correct). I will develop it in a way consistent with the naive conception of belief, and so I endorse an analogue of (Pointwise), viz:

(P) Harry believes that p iff Harry is in some state s such that the correct interpretation assigns to s the pair <belief, p>.

The causal-pragmatic metaphysics of belief consisted in two filters that an interpretation I must pass to be counted as correct. To repeat, these were:

(a) If I maps s to <belief, p> then s must be a state such that: optimally, Harry is in only if p, and optimally, he is in that state because of p or something that entails p;
(b) Harry is disposed to act in ways that would tend to satisfy the contents assigned by I to desjre-states (i.e. those classified by I as desires) at worlds in which the contents assigned by I to belief states (i.e. those classified by I as beliefs) are true.

I substitute the following:

> (RI) I is a correct belief/desire interpretation of x iff I is the belief/desire interpretation which best rationalizes x's dispositions to act in the light of x's experiences.

The facts mentioned on the right hand side of (RI) are the metaphysical foundations of belief. They include facts about experience, action and the relations of rational update in the light of experience and rational choice in the light of belief and desire. Facts about how a subject is disposed to act were also part of the basis on which Stalnaker grounded belief, in his clause (b). Facts about experience are not appealed to by Stalnaker explicitly. But there is a parallel: the appeal to experience is what does duty for Stalnaker's clause (a).

(Just to briefly motivate this. Consider the belief that all emeralds are green, formed, defeasibly, on the basis of local evidence about the colour of observed emeralds. In any naturalistic, non-representational notion of optimality, such a belief (or any belief state whose content entails it) can exist while some far-flung emeralds are not green. So I think that the causal pragmatic picture overreaches when it tries to use optimal covariation characterize the content of arbitrary beliefs. My strategy, in developing (RI), is to use the kind of naturalistic tradition on which Stalnaker is drawing in (a) to naturalize perceptual evidence, which is an *input* to (RI), and so one of the things that constrains the assignment of beliefs in general. I think similar resources can be deployed to naturalize the content of action-guiding states—basic intentions and the like—which are also appealed to in (RI). See Williams (2019) for the full story).

At a strategic level, to this point (RI) has appealed to the same kind of resources as the causal-pragmatic picture, but arranging them differently within the account. The appeal to rationalization however marks a significant departure—but even here there is precedent within the causal-pragmatic picture. Clause (b) is the principle that (on the correct interpretation) one acts so as to realize one's desires given what one believes (=in worlds where the contents of one's belief is true). This is an approximation of a principle of instrumental rationality, and so to satisfy (b) is, at first-pass, to be instrumentally rational. But the demand for rationality in (RI) requires much more than this—in order for the account to be tenable, it must pick out a relation of *substantive* rationality. To be substantively rational, the agent needs to possess good reasons for their beliefs, and good reasons for acting as they do. Such demands may include, but go well beyond the kind of 'structural' patterns of instrumental rationality that the causal-pragmatic picture (b) approximates. Committed naturalizers like Stalnaker will have legitimate worries that the appeal to a body of normative truths about good reasons for belief and action will frustrate their project—and I can offer them little comfort. This is not the causal-pragmatic picture, but a rival.[10]

Relitigating Equivalence and Closure.

Stalnaker's argument for Equivalence was based on the naturalistic relations that appear in his (a) and (b) being *intensional* resources. Very roughly: if you say that the content of state b is some p such that in optimal conditions, you would be in b if and only if p, then the condition is met by one proposition if and only if it is met by every proposition necessarily equivalent to it. Throwing in more conditions on p will do no good unless they generate a hyperintentional context which means necessary equivalents can no longer be substituted *salva veritatis*. Stalnaker's bet is that naturalistically-acceptable resources will never do the job.

I favour comparable treatments of the content of experience (and of action-guiding states). I think this style of argument does show that their content is coarse-grained. But given (RI), the case of belief and desire is far more delicate. It turns on what theory of rationalization is given.

There are theories of rational belief update and rational choice that work with coarse-grained contents. David Lewis, one source for my (RI), favoured this kind of theory, with the assumption that a Bayesian theory of belief update on which (degrees of) belief have sets of possible worlds as their contents, are updated by coarse-grained facts about what a subject has experienced, and rationalize actions under coarse-grained descriptions (Lewis 1974,1986). Feed that model of rationality into (RI), and the argument for Equivalence goes through. However, we only get coarse-grained content out because we fed in a coarse-grained theory of rationality. I wish to deny Equivalence, and all I need to do in order to make (RI) compatible with this is feed in an account of rationalization that, just as one would naively think, generates hyperintentional contexts.

Here is one example: it might be that one has good reason to form a belief that a is F when one is in a certain kind of information-link with an object, one which inter alia generates the

[10] Stalnaker can hope to get away with the appeal to (b) because he hopes that the appeal to independent constraints on (a) will have already filtered out deviant interpretations of what the agent believes. The role for (b) is primarily to fix the content of desires. So one of the reasons I feel impelled to put more weight, and make more substantive, the notion of rationality in (RI) is that, as reported above, I am pessimistic about whether any plausible successor to (a) can bear the weight placed upon it.
There is a version of radical interpretation that does try to run with a thin notion of rationality (something like the "structural" rationality that I talked about earlier)—but this is in my view demonstrably hopeless, even if we can help ourselves to pretty much as much naturalistically-fixed representational content for experience and action as one likes (cf. Williams 2016).

perceptual content *that a is F.* But one may not have good reason to believe *that b is F* in these same circumstances, even when b and a are one and the same item, simply because the concept *b* involves a mode of presentation of the item which is tied to a numerically distinct information link. This is all compatible with the claim that the perceptual content itself is coarse-grained—one might indifferently describe the perceptual content as that a is F, or that b is F. The structural point is that justification (believing for good reasons) can involve more than a relation between the (assumed to be coarse-grained) perceptual content and the belief formed in response. Justification can require that the belief be formed *through a certain route* that inter alia involves a state with that perceptual content. There's nothing here that seems particularly outlandish so far as epistemology is concerned. But if things work this way, then the relata of rationalization include, at the belief-end, fine-grained content.[11]

I earlier went through an argument for (Holism) that was based on Stalnaker's clause (b). As noted, (RI) is a kind of generalization of (b), but the argument for Holism does not generalize. Again, what mattered to Stalnaker's argument were the particular details of the form of instrumental-rationality connection he assumed. In particular, it was set up in such a way that specific relation became vacuous if the beliefs involved were inconsistent. So long as our notion of substantive rationality allows for rankings of actions as better or worse relative to belief states that can be consistent or inconsistent, there's no presupposition of consistency baked into (RI) in the way there was into (b). I do not think that the argument generalizes.

This argument is blocked so long as we have available a notion of "best rationalization" that allows that the best rationalization of an agent is still not perfectly rational, allowing for inconsistent beliefs in particular. In a previous section, I made the case that a structure of co-belief, dividing the agent's mind into a set of individually consistent but potentially mutually inconsistent fragments is part of this. Supposing our theory of rationality appeals to this structure, we should really be enriching our conception of what the space of background interpretations is like. It should not only take states and classify them by type and content; it needs also to specify the co-belief structure. Facts about co-belief can then be grounded in exactly the way that facts about belief will be: by being part of the best rationalization of the agent's actions, in light of their course of experience.

I conclude that radical interpretation, though perfectly consistent with a Stalnakerian account of belief (this being exemplifed in Lewis's version of it) can also provide principled foundations for the

[11] For a discussion of demonstrative thought and modes of presentation that drives this, see Dickie 2015, and the discussion in Williams 2019.

naive conception of belief, as well as the needed co-belief structure involved in my pseudo-Stalnakerian dissolution of the problem of commitment.

Internal commitments.

There is one loose end to tie up. Ever since I posited co-belief as the solution to problems of commitment, I've been talking as though facts about consistency and inconsistency are given from above, exogeneously to the agent's mentality. This would be so an an external conception of commitment-to-believe, but the notion of commitment that I find interesting is the internal one—I also think the costs of maintaining that clusters of mutually co-believed propositions are consistent, which is part of the price of my pseudo-Stalnakerian proposal, is far higher on the external conception. On the internal conception of commitment, the relevant notion of consistency and inconsistency are fixed by attitudinal states of the agent—their acceptance of this or that rule. If facts about what the agents accepts are themselves internally consistent, then this may not matter. But if the agent accepts contradictory rules—or if some fragment of their beliefs and behaviour is best rationalized by a different logic to another—then we should be looking for an account of the mentality of the agent which bundles together suitable acceptance states along with beliefs to generate the maximally rational fragments. This, I think, poses no new problems of principle. It may well mean that we can no longer unambiguously judge whether clusters of co-believed propositions are mutually consistent or inconsistent, since we have no cluster-transcedent and acceptance-based notion of consistency to appeal to. But such facts play no load-bearing role in this theory. The attitudes of accepting-a-rule are intentional states for which a metaphysical foundation should be provided. But the same gambit applies: insofar as they are already caught up in our theory of rationality, (RI) already provides their metaphysical basis.

Conclusion.

I started with a puzzle that afflicts our ordinary naive thinking about belief and commitments to believe. Drawing inspiration from a very different conception of belief, we can dissolve the puzzle. So long as agents do not believe things that are *by their own lights* inconsistent, the extra structure of co-belief allows us to characterize their commitments. We have reason to believe that co-belief should show up in a theory of (suboptimal) rationality. And because of this, the kind of metaphysics of belief that I favour will provide the necessary foundations for it.

The theory of belief-and-commitment that I offer, and the foundations thereof, are Stalnakerian and yet not Stalnakerian. They echo and learn from Stalnaker's account of belief. They are subject to some of the same costs, and follow analogous escape routes to some analogous difficulties. Those of us who defend the naive picture of belief in the way described will be motivated to join forces with the Stalnakerians to make progress on shared and tricky issues, such as: how exactly to spell out a theory of suboptimal rationality via fragmentary mental states. Ultimately, one's resistance to the full-fat Stalnakerian theory might weaken, once one appreciates how much of it one is going to have to parallel.

I think the defender of the naive theory of belief should resist the temptation to throw in the towel, however. Particularly when it comes to understanding clear-eyed logical uncertainty, I find the Stalnakerian model to give too few resources—it does not allow me to capture adequately the attitude I adopt when I am agnostic over whether Ada or Beth has the right logic (Williams 2017). A version of Stalnaker's theory with a different modal base—where content is modelled by sets of doxastically possible worlds rather than sets of metaphysical possibilities—would do the job, but as Stalnaker persuasively argues, what is a minor switch from a formal point of view would require entirely different metaphysical foundations, and would require giving up Stalnaker's motivating foundational picture.

If what I've said here is on the right tracks, there is plenty more to be done. There is the theory of suboptimal rationality. There is generalizing what has been said here to the kind of framework I defended in (2017), which both involves degrees of belief and generalizes the notion of "acceptance of a rule" to a broader category of modal attitude (accepting-as-doxastically-necessary, accepting-as-doxastically-possible). And much more could and should be said about the way that co-belief and acceptance-of-rules arise out of Radical Interpretation—to say that (RI) as I have characterized it provides foundations for these attitudes is not yet to say that the predictions that ensue are extensionally adequate.

Bibliography

Berto "Simple hyperintensional belief revision" Erkenntnis, forthcoming.
Dickie Fixing Reference, OUP 2015.
Jago The Impossible: an essay on hyperintensionality OUP, 2014.
Leitgeb, H. The Stability of Belief, OUP 2017.
Lewis "Radical interpretation" Synthese 27, (1984).
Lewis On the Plurality of Worlds OUP 1986
Makinson, D. C., 1965, "The Paradox of the Preface", Analysis, 25: 205–207.
Rayo The construction of logical space OUP 2013.
Stalnaker, Robert Inquiry, MIT Press, Cambridge MA, 1984.
Williams, J Robert G. "Representational scepticism: the bubble puzzle". Philosophical Perspectives, 2016.
Williams, J Robert G "Rational Illogicality" Australasian Journal of Philosophy 2017.
Williams, J Robert G "Normative Reference Magnets", Philosophical Review, 2018.
Williams, J. Robert G. The Metaphysics of Representation. Oxford University Press 2019.