I'm mistaken

Jeremy Goodman

November 10, 2020

Abstract

If you believe that not everything you believe is true, it follows that not everything you believe is true. But it does not follow that your beliefs are inconsistent. I first explain how a logically omniscient and fully introspective thinker can consistently believe that they have false beliefs, and why such a thinker would be a counterexample to Robert Stalnaker's theory of belief and related theories of justification. I then consider whether the preface paradox shows that people like us have inconsistent beliefs, and argue that it does not. I then describe a case where people like us do have inconsistent beliefs and also believe that they have false beliefs. Despite appearances, it turns out that bounded rationality rather than modesty is responsible for the inconsistency. Both modesty and inconsistency raise important challenges for our best theories of belief, but the former does not beget the latter in idealized thinkers or in ordinary people.

1 The Puzzle of Modesty

I know that not everything I believe is true. For one thing, I am not perfectly rational. I don't pretend that my beliefs are completely consistent. But this failing is intuitively inessential to the fact that not everything I believe is true. If somehow I managed to have consistent beliefs, through a combination of good luck, better reasoning, and being less opinionated, I still wouldn't be infallible. I'd still sometimes misperceive, misremember, be misinformed, and draw inferences to false conclusions. And I'd still recognize this about myself.

I am also not perfectly introspective. I don't have perfect knowledge of what I do and don't believe. But this fact too is intuitively orthogonal to the fact that not everything I believe is true. Imagine I had the following idealized psychological profile. Whenever I consider the question whether a given proposition is true, I can tell how I am disposed to answer. If I am disposed to answer yes, I am also disposed to notice this about myself and thereby believe that I believe the proposition. And if I am not disposed to answer yes, I am likewise disposed to notice this about myself and thereby believe that I do not believe the proposition. So for any proposition I might consider, if I believe it, then I believe that I believe it, and if I do not believe it, I believe that I do not believe it. I am not

actually a creature like this. But if I were, I'd still be fallible. I'd still be wrong about some things, and know this about myself.

Say that a person is *modest* if they believe that they believe something false, *consistent* if the propositions they believe could all be true together, and *negatively introspective* if everything they don't believe is something that they believe they don't believe. Intuitively, it seems like it should be possible to be modest, consistent, and negatively introspective. Surprisingly, it isn't.

Here is why. Suppose I am consistent. So the set of propositions that I believe is such that, in some possible situation w, every member of that set is true. Suppose I am modest: I believe that I believe something false. It follows that, in w, I believe something false. Let p be some proposition that I falsely believe in w. p is a counterexample my being negatively introspective. If I believed it, it would have to be true at w, but it is isn't. And if I believed I didn't believe it, then I would have to not believe it at w, but I do. So I neither believe that p nor believe that I don't believe that p. A formal version of this argument is given in an appendix.

What should we conclude from this result? Since I don't think it requires too much idealization to imagine people like us who are consistent in addition to being modest, I think the moral is that people like us cannot be negatively introspective. It is not an accidental feature of our cognitive endowment that there are propositions that we neither believe nor believe that we don't believe.

This conclusion is not so surprising if we keep in mind that these propositions may be ones that we do not even understand. This point is relevant to our imagined idealized creature who, for every proposition they can consider, knows whether or not they believe it. We were wrong to conclude that such a creature would be negatively introspective, since not all propositions will be ones that it can even consider. The thought experiment establishes at most that a creature could be *weakly negatively introspective*, in the sense that every proposition they understand but don't believe is one that they believe they don't believe. And it is obviously consistent that a person be modest, consistent, and weakly negatively introspective, since it is consistent that a person understands nothing. (Note: here and throughout I use "consistent" both to describe logically consistent sets of claims and to describe people whose beliefs could all be true together.)

Below I will show something less obvious and more interesting: *a person who is modest, consistent, and weakly negatively introspective can also understand everything that is intuitively required of a logically and introspectively perfect being.* (Note: the remainder of this section is more technical and more abstract than the rest of the paper, and can be skipped up to the penultimate paragraph without loss of continuity.)

To work up to this result, let's start by reexamining our puzzle using some tools from modal logic. We will theorize in a formal language with the usual logical connectives and an operator B interpreted as "I believe that ...". In this formalism, the orthodox theory of belief is KD45, axiomatized as follows:

PL every propositional tautology

K $B(\varphi \to \psi) \to (B\varphi \to B\psi)$

D $B\varphi \rightarrow \neg B\neg\varphi$

4 $B\varphi \rightarrow BB\varphi$

5 $\neg B\varphi \rightarrow B\neg B\varphi$

MP If $\vdash \varphi \rightarrow \psi$ and $\vdash \varphi$, then $\vdash \psi$.

RN If $\vdash \varphi$, then $\vdash B\varphi$.

The combination of PL, K, MP, and RN encodes the idea that I am logically perfect (anything derivable using classical logic from things I believe and from principles of this very theory will be things that I believe); D encodes the idea that I don't have contradictory beliefs (if I believe something, then I don't believe its negation); 4 encodes the idea that I am positively introspective (if I believe something, I believe that I believe it); and 5 encodes the idea that I am negatively introspective (if I don't believe something, then I believe that I don't believe it). I say "encodes the idea" because the correspondences are not exact. This is because these principles are schemas. 5, for example, stands for infinitely many axioms, one for each sentence of the language φ. The claim that I am negatively introspective, by contrast, is a generalization about all propositions, including those not expressed by any sentence of our formal language. This distinction will be important in what follows.

To see the tension between KD45 and modesty, we need to review some standard modal model theory. We start with a set W and a binary relation R on this set, which I'll pronounce "sees". The intuitive interpretation is that W is the set of possible worlds and R is the relation of *doxastic accessibility*, where a world w *doxastically accesses* a world v just in case everything I believe in w is true in v. Note that, so interpreted, any world that sees itself is a world in which all of my beliefs are true. Now, to interpret our formal language, we first associate every atomic sentence with the set of worlds in which it is true. Then, for logically complex sentences, we say that $\neg\varphi$ is true at w if and only if φ isn't true at w, $\varphi \wedge \psi$ is true at w if and only if both φ and ψ are true at w, and $B\varphi$ is true at w if and only if φ is true at every world that w sees. Finally, we designate one world as actualized and say that a sentence is true in the model if and only if it is true at that world.

It is a standard result of propositional model logic that KD45 is the logic determined by the class of such models in which R is serial (every world sees some world), transitive, and euclidean (any two worlds seen from a given world see each other). Since in such models the worlds seen from a given world all see each other, they also see themselves. Interpreting R as doxastic accessibility, this means that, in every world compatible with what I believe, I have no false beliefs. Since every world sees some world, this means that there is some doxastically accessible world in which I have no false beliefs. In other words, I am not modest: I don't believe that something I believe is false. In fact, I am *immodest*: I believe that everything I believe is true.

So the possibility of being modest is in direct conflict with any theory according to which doxastic accessibility is serial and euclidean – and, indeed, with

any theory according to which every world doxastically accesses some world that doxastically accesses itself. (Similarly, the necessity of immodesty follows from any theory according to which every world doxastically access only worlds that doxastically access themselves.) This is noteworthy because some quite influential theories of belief imply that doxastic accessibility has these properties.

A prominent example is the picture of belief advocated by Stalnaker (1984), which Stalnaker (2006) uses to defend a KD45 logic of belief. The guiding idea is that belief is species of "indication", in the same sense where a tree's rings are an indication of its age. Grossly simplified, his view is that, just as a tree with n rings is in an internal state that, normally, trees are in only if they are n years old, likewise a person counts as believing those propositions that are true in all normal circumstances in which they are in the same internal state as they are actually in. The operative notion of "internal state" is rather involved and also highly context-sensitive; see Stalnaker (2019, p. 143). But for present purposes what matters is only that sameness of internal state is an equivalence relation. There is then a serial, transitive, euclidean relation between worlds – w sees v if and only if in v I am in the same internal state as I am in w and v is normal for me given that I am in that state – which, according to Stalnaker's theory, coincides with doxastic accessibility, at least for ideal thinkers.

A different route to the impossibility of modesty involves the interaction of knowledge and belief. For Stalnaker (2006, p. 192), when conditions are normal "belief and knowledge coincide". So when conditions are normal, I have no false beliefs, and hence am not modest, since being modest entails having a false belief. Stalnaker also holds that everything one believes is something one could know, and thereby believe, in normal conditions. This rules out modesty, since to be modest is to have a belief that, on his view, it is impossible to have in normal conditions.

A parallel challenge arises for certain theories of *justified* belief. For example, Smith (2016) and Goodman and Salow (2018) defend the view that one has justification to believe a proposition if and only if it is true in all the sufficiently normal circumstances in which one has the same evidence as one actually has. In other words, in w I have justification to believe all and only those propositions that are true in all worlds seen by w, where now w sees v if and only if in v I have the same evidence as I do in w and v is sufficiently normal given that evidence. Since seeing (so defined) is euclidean, every world sees only worlds that see themselves. So I have justification to believe that nothing I have justification to believe is false. Provided every world also sees at least one world (as these authors also hold), I do not have justification to believe that something I have justification to believe is false.

But the analogue of modesty for justified belief is no less plausible than modesty itself (at least for notions of justification, like the present one, such that having some justified false beliefs is commonplace). It is hard to deny that I have justification to believe – and indeed know – that I have at least one justified false belief, and hence have justification to believe at least one falsehood. This fact constitutes an important challenge to these theories of justification.

Let's return to the logical analysis of the puzzle. It is important not to con-

flate the claim (i) that KD45 is a correct theory about what a person believes, with the claim (ii) that what the person believes can be read off a serial, transitive, euclidean accessibility relation. We have seen that (ii) is incompatible with the person being modest, since it implies that they are consistent and negatively introspective. But as we will now see, (i) is compatible with the person being modest. I'll first describe an obvious strengthening of KD45 which implies that I am negatively introspective, and show that this theory is consistent with my being modest. I'll then describe what I think is a more interesting strengthening of KD45, which implies that I am weakly negatively introspective, and show that this theory is consistent with my being both modest and consistent.

To start, we need a way of formalizing the claims that I am negatively introspective and that I am modest. We do so by adding propositional quantifiers to our formal language. These two claims can then be respectively formalized as $\forall p(\neg Bp \rightarrow B\neg Bp)$ and as $B\exists p(Bp \wedge \neg p)$.

Let KD45π be the result of adding standard axioms of quantificational logic to KD45, and closing the result under universal generalization (as described below), so that the claim that I am negatively introspective is a theorem of this theory. It turns out that KD45π is consistent with my being modest. This may come as a surprise, since we saw that there are no models of the standard sort – in which belief is interpreted using an accessibility relation – according to which I am both negatively introspective and modest. So showing that KD45π is consistent with my being modest requires a different kind of model.

Here is a simple such model. Identify worlds with real numbers between 0 and 1, identify propositions with sets of worlds, and consider the probability distribution over propositions corresponding to the Lebesgue measure on the unit interval (i.e., the function that in a natural way assigns every interval a probability equal to its length). In every world I believe the same propositions: those that have probability 1. It is straightforward to verify that this is a model of KD45π in which I am modest. (For an illuminating discussion of a broader class of models of KD45π of which this is an instance, see Ding (forthcoming).)

I don't myself think that this model offers a very compelling picture of how an idealized thinker might be modest, since it essentially involves a failure to be consistent: the propositions believed could not all be true together. But the model is still instructive. It shows that a prohibition on contradictory beliefs together with an assumption of logical omniscience is insufficient to ensure the consistency of our beliefs (in the sense of their possibly being true together) when infinitely many propositions are involved. It thereby highlights some counterintuitive consequences of the identification of belief with probability 1 assuming standard probability theory. The model is also a helpful point of comparison for assessing an alternative model of modesty I'll describe presently.

I will now show that being modest, consistent, and weakly negatively introspective is compatible with being logically and introspectively ideal in a quite demanding sense. To make this claim precise, we will add to our language an operator U interpreted as "I understand the proposition that ...". Now consider the following combined theory of belief and understanding, $\mathcal{U}$, axiomatized below. (Related principles are explored in the literature on combined theories

of knowledge and "awareness"; see Schipper 2015 for a review.)

PL every propositional tautology

K $B(\varphi \to \psi) \to (B\varphi \to B\psi)$

D $B\varphi \to \neg B\neg\varphi$

4 $B\varphi \to BB\varphi$

C4 $BB\varphi \to B\varphi$

5U $U\varphi \to (\neg B\varphi \to B\neg B\varphi)$

Dist $\forall p(\varphi \to \psi) \to (\forall p\varphi \to \forall p\psi)$

Vac $\varphi \to \forall p\varphi$ where p is not free in φ

UI $\forall p\varphi \to \varphi[\psi/p]$ where ψ is free for p in φ

U $B\varphi \to U\varphi$

UB $U\varphi \to BU\varphi$

CUB $BU\varphi \to U\varphi$

CL $(U\varphi_1 \wedge \cdots \wedge U\varphi_n) \to U\psi$
where ψ is built from $\varphi_1, \ldots, \varphi_n$ using $\neg, \wedge, B, U$ and quantifiers

MP If $\vdash \varphi \to \psi$ and $\vdash \varphi$, then $\vdash \psi$.

RN If $\vdash \varphi$, then $\vdash B\varphi$.

Gen If $\vdash \varphi$, then $\vdash \forall p\varphi$.

RE If $\vdash \varphi \leftrightarrow \psi$, then $\vdash \Phi \to \Phi[\psi/\varphi]$.

$\mathcal{U}$ differs from KD45 in a few ways. (i) Like KD45π, it includes standard quantificational principles Dist, Vac, UI, and Gen. (ii) 5 has been weakened to 5U, so that the theory implies that I am weakly negatively introspective but not that I am negatively introspective. (iii) U, CL, and RE imply that everything I believe, any logical combination of things I understand, and anything logically equivalent to something I understand are all things that I understand. (iv) UB states that, if I understand something, I believe that I understand it; it is the analogue of 4 for understanding. (v) C4 (which is a derived axiom of KD45), says that if I believe that I believe something, then I really do believe it; CUB says the same about what I understand.

I will now describe a model that shows that $\mathcal{U}$ is consistent with modesty. Let W be the set of rational numbers and let R be the less-than relation. In w, I believe all and only the propositions that are true in all worlds seen by w, and I understand all and only the propositions I believe and their negations. Given the symmetries of the model, we may choose any world to be the actualized one. It

is straightforward to verify that this is a model of $\mathcal{U}$ in which I am modest. And unlike the model of KD45π and modesty described above, I am also consistent, since the conjunction of everything I believe is possibly true.

Since in this model I am modest and consistent, I must not be negatively introspective. Counterexamples abound. For example, if w sees v, then the conjunction of everything I believe in v is a counterexample to my being negatively introspective in w. But since I am weakly negatively introspective in w, this and all other counterexamples to my being negatively introspective in w are propositions that I don't even understand in w. (The model also validates all closed instances of the 5 schema, for the same reason, provided every atomic sentence expresses a proposition that I understand in the actualized world.)

While the model establishes that being modest, consistent, and weakly negatively introspective is consistent with a strong theory $\mathcal{U}$ of belief and understanding, the picture it offers is unexpected. For it is surprising that being a modest idealized agent should involve contingency in what one understands. What might explain such contingency?

Here is one idea. Assume that the propositions we understand include those that are expressed by sentences of our language (either a public language or a language of thought). There are strong independent reasons to think that most of our sentences are *semantically plastic*: if the world were even slightly different from how it actually is, these sentences would have expressed different propositions from those they actually express (albeit often propositions with very similar truth conditions). In particular, there are reasons to think that sentences containing words like "believe", which are prone to vagueness, are generally plastic in this way – see Dorr and Hawthorne (2014). Since our agent is introspective, they presumably have such a word in their language. Since they are modest, any world compatible with what they believe is different from actuality. Putting these ideas together, it follows that the meaning of "believe" (or its analogue) in their mouth would have been at least slightly different in any world compatible with what they believe. So which proposition they understood via sentences containing "believe" (or its analogue) would be different too. While this suggestion is somewhat speculative, so too are intuitive judgments about the psychology of idealized agents. So I don't think the surprising predictions of the above model show that it isn't worth taking seriously.

Moreover, it isn't obvious that being logically and introspectively ideal actually demands anything as strong as satisfying the theory $\mathcal{U}$. Here are two reasons why. First, it isn't obvious that even ideally introspective beings will be weakly negatively introspective. Suppose I am unsure whether Alex and Bob are the same person. In fact, they are different people. I believe that Alex is tall but do not believe that Bob is tall. I also accept the controversial but popular philosophical position according to which, if Alex is Bob, then anyone who believes that Alex is tall thereby believes that Bob is tall. For this reason, despite both understanding and failing to believe the proposition that Bob is tall, I am unsure whether I fail to believe that Bob is tall. Intuitively, this does not reveal any logical or introspective shortcoming.

Second, it isn't obvious that even ideally introspective beings will be *posi-*

tively introspective. Let's grant that they will never be mistaken or uncertain about what they believe. Still, if I have never considered the question whether I believe that p, then even if I do believe that p, my failing to believe that I do would be neither a mistake nor a case of uncertainty. Being perfect at something does not imply having already done it: this is just as true for introspection as for anything else. Of course, there are views about self-knowledge where it is not tied to having done anything – Stalnaker's theory is a case in point. But we saw above that his theory is untenable in the present context, since it is inconsistent with modesty. By contrast, theories according to which self-knowledge is the result of a process of introspection do not imply that being perfect at introspection means already believing that you believe everything that you in fact believe; see Byrne (2018).

The crucial feature of the above model is that, although all of my beliefs are open to view, the fact that they are all of my beliefs is opaque to me, since I am open to the possibility that I am even more opinionated. The point is not that there is some particular opinion that I lack but am unsure whether I have: there is none, since being unsure requires understanding, and all such questions I understand are ones to which I know the answer. Rather, some propositions are simply beyond my ken, and even my ample logical and introspective powers do not tell me that I do not believe them.

Does the logical possibility of an idealized modest consistent agent, with the logical and introspective powers described above, resolve the puzzle with which we began? Only partially. It does vindicate the intuitive thought that consistency and introspectiveness are not a barrier to modesty in principle. But given how idealized such agents would be, their logical possibility does not clearly to speak to the question of whether ordinary modesty essentially involves any inconsistency in *our* beliefs or any inability to answer questions about what we do and don't believe. To address this question, we need to consider more closely real cases that lead people like us to believe that we have false beliefs. Let us now turn to such cases.

2 The preface

Most of the literature on how we should think about our own fallibility has focused on Makinson's (1965) "paradox of the preface". Makinson has us consider a typical academic who includes in the preface to their latest book an apology for the errors it undoubtedly contains. The author believes every claim made in the book, including this one; moreover, we may assume that every claim made in the book is one that the author correctly believes is made in the book. Following Makinson, such cases are usually taken to show that ordinary people often knowingly have inconsistent beliefs. This is a mistake.

If every claim not made in the book is one the author believes is not made in the book, then indeed the totality of what the author believes cannot be true. This is for the same reason that it is impossible to be modest, consistent, and negatively introspective, as explained in the last section. But not every

claim not made in the book will be one that the author believes is not made in the book. This is for the same reason that people like us cannot be negatively introspective: we don't believe propositions we don't even understand, and not every claim not made in the book is one such that the author even understands the proposition that it is not written in the book.

So for all we have said the author's beliefs are consistent. That is, for all we have said it is consistent with everything they believe that, in addition to containing each claim it in fact contains, and these claims each being true, the book also contains a claim it does not in fact contain, and that claim is false.

This point is usually overlooked in the literature on the preface paradox (with Evnine (1999) being a notable exception; see also Weatherson (2005)). Philosophers writing about the case usually assume that the author believes the negation of the conjunction of all of the claims in their book. Their beliefs would then clearly be inconsistent. But an ordinary author will not believe the negation of the conjunction of all the claims written in their book. The reason is not merely that this proposition is too complicated for an ordinary person to entertain. Even if the author were capable of entertaining it, it is not clear why they should believe it. After all, it is not a consequence of anything else we have supposed that they believe. It is a consequence of the fact that their book contains an error, which they do believe, together with the fact that every claim made in the book is either p_1, or p_2, or ..., where these are all the claims that are in fact made in the book. But why should we think the author believes this second claim, that every proposition in the book is one of these? Again, the issue is not merely that this proposition is too complicated for someone like us to entertain. Rather, the point is that ordinary book-writing does not involve keeping a mental inventory of the claims made in the book. Even if (as we're unrealistically assuming) the author hasn't forgotten making any of the claims in the book, the fact that these are the *only* claims made in the book is simply not a consequence of what they remember from writing it.

One might reply that an author with superhuman cognitive capacities would be able entertain the hypothesis that every claim in their book is either p_1, or p_2, or ..., and then learn that it is true by checking the hypothesis against every claim in the book, one by one, and finding no counterexamples. Such a thinker would then have inconsistent beliefs if they continued to believe every claim made in the book (including that the book contains an error). But this reply does nothing to suggest that ordinary writers of such prefaces thereby reveal themselves to have inconsistent beliefs. And the distinction between ordinary people and superhuman thinkers matters. We have good reason to think that some ordinary preface writers in fact believe what they write in their books, and need not be unreasonable in doing so. So if they thereby have inconsistent beliefs, this would be a reason to think that people can reasonably have inconsistent beliefs. The same cannot be said for unfamiliar superhuman thinkers. When they consider the negation of the conjunction of the claims in their book, they hold it in mind all at once, and not under any quantificational or ellipsis-involving guise. It is just not clear that such a creature would be reasonable in believing this proposition while also believing each of the claims whose conjunction it is the

negation of. (Here I agree with Smith (forthcoming), despite my worries about his theory's implications about modesty mentioned in the previous section.)

Christensen (2004, p. 38) sees things differently. He writes:

> It is undoubtedly true that ordinary humans cannot entertain book-length conjunctions. But surely, agents who do not share this fairly superficial limitation are easily conceived. And it seems just as wrong to say of such agents that they are rationally required to believe in the inerrancy of the books they write. Clearly, the reason that we think it would be wrong to require this sort of belief in ordinary humans has nothing to do with our limited capacity to entertain long conjunctions.

I agree with Christensen that if such an ideal agent were to shadow an ordinary human historian while they researched and composed a monograph, and hence shared the historian's evidence for all of the claims made in the book, then the agent, like the historian, would believe the book contains some errors, and might detect no internal tensions in the narrative, and so find every claim written in the book to be well-supported by the evidence. Perhaps this apprenticeship would even inspire them to become a historian themselves, and write a book of a similar kind, which they will acknowledge inevitably contain some errors. But what is not clear is that such an agent would believe every claim made in their book. In making flat assertions, they might simply be conforming to the norms of scholarly writing in the community of ordinary humans in which they are embedded. Perhaps their synoptic view of the content of the book would lead them to hedge their commitment to some claims where an ordinary historian would not. Maybe only the most secure claims in the book would be ones the agent fully believes, while the majority of the claims would be ones that the agent only believed to be highly probable. (Stalnaker (1984, p. 92-4) holds a version of this view about *ordinary* historians. He claims that they "accept" the claims in their books but don't fully believe everything they accept. Stalnaker (1991, p. 429) seem to think this response to the preface paradox extends to modesty in general. But that seems doubtful: even a historian who doesn't believe that they falsely believe anything written in their books, on the grounds that they don't believe everything they've written, surely still recognizes that they have some false beliefs about non-historical matters.)

I am not saying that it is obvious that such an idealized agent would *not* write the sort of books ordinary historians write and believe each of the claims in their book as perhaps some ordinary historians reasonably do. I am merely denying that it is obvious that such an idealized agent *would* sincerely and reasonably write such books. The preface paradox has been so compelling because it involves not an idealized agent, but an ordinary author. This is why our judgments about what the author believes and which of these beliefs are reasonable are relatively secure – we can put ourselves in the author's shoes. Parallel claims about hypothetical idealized authors are quite speculative by comparison. Christensen has a response to this criticism, which we will consider below. But first,

I want to present a new case in which no idealizations are needed to argue that ordinary modest people are inconsistent.

3 Test scores

Suppose you take a general knowledge test comprised of true/false questions. You're allowed to skip questions you are unsure about, and you do. Afterwards, you haven't forgotten what answers you gave. You then learn that you scored 99 correct out of 100 answers given. You continue to believe every answer you originally gave. This isn't unrealistic. After all, you may have done better than expected: suppose you've taken tests like this before and never done so well. Since you weren't unsure of your answers when you gave them, it is only natural that you still believe them after getting the good news.

Your beliefs are now inconsistent. Here is why. Let $p_1, \ldots, p_{100}$ be the answers you gave on the test. Here are some things you believe: the test asked whether p_1; you answered p_1; p_1 is true; ...; the test asked whether p_{100}; you answered p_{100}; p_{100} is true; you answered exactly 100 questions; you only answered 99 questions correctly. It is impossible that all of these propositions be true.

Furthermore, your beliefs are inconsistent without any of the propositions you believe being especially complicated. In particular, we haven't assumed that you believe that the only questions you answered were whether p_1 is true, ..., and whether p_{100} is true. The impossibility of all of your beliefs being true together is instead the result of your new beliefs about how many questions you answered and about how many you got right. (Here I'm tacitly assuming that, if p and q are different propositions, then, necessarily, whether p is true and whether q is true are different questions. But we could drop this assumption, since any two questions on the test are ones you also believe to be distinct.)

Suppose you know enough about yourself to know that your beliefs are inconsistent in this way. Then, like the preface author, you are modest. But whereas the cognitive limitations of the preface author arguably insulated them from having inconsistent beliefs, in the test case your cognitive limitations have the opposite effect of preventing you from reasonably extricating yourself from inconsistency. It would be unreasonable to suspend judgment on every proposition you answered on the test. That would be to relinquish far too many beliefs. But then which of these beliefs can you reasonably give up? Not always the ones your are least confident in. To see why, imagine a three-question test. You believe all your answers, but you are significantly more confident in the first. You then learn you answered only two questions correctly. Should you suspend judgment on your second or third answers? Not necessarily. For you might be certain that both are true if either is. You might then reasonably disbelieve your original answer to the first question, and increase your confidence in the two answers you were originally less confident in. (Aside: It is also notable that people like us are not able to reliably give up our beliefs in the answers we are least confident it in cases like this. For we have no way to identify those answers. Not forgetting your answers is one thing; being able to systematically recall them and compare

your relative confidence in them is quite another. Cf. Stalnaker (1999) and Elga and Rayo (forthcoming) on the role of limited recall in sustaining inconsistent beliefs.)

Belief-revision is holistic. This is why standard models of belief-revision involve plausibility-comparisons between possibilities that settle the truth values of all relevant propositions; see Grove (1988). But in the test case such possibilities are far too rich for us to get our minds around, let alone intelligently compare. Contrast Lewis (1982, p. 436), who "used to think that Nassau Street ran roughly east-west; that the railroad nearby ran roughly north-south; and that the two were roughly parallel". When he became aware of the inconsistency, he immediately and rationally adjusted his beliefs to eliminate it. The test-score-induced inconsistency isn't something you're capable of responding to in kind.

This may not be a bad thing, since the very cognitive limitations that sustain your inconsistent beliefs also prevent them from getting you into trouble. (Egan (2008) makes a related point in a slightly different context.) Deriving a contradiction from the relevant propositions in a standard proof system would take an astronomical amount of time and involve unmanageably long formulas. This is a consequence of the exponential growth of proofs of the pigeonhole principle as we increase the number of pigeons (see Razborov (2002)). By contrast, deriving a contradiction from n claims and the negation of their conjunction takes only n applications of conjunction introduction. (We'll explore one ramification of this contrast in the next section.) Hanging tough after receiving your score won't make you liable to draw any absurd conclusions through idle deduction.

There is much more to be said about how to think about the sort of inconsistent beliefs exemplified by the test case. But the important fact for present purposes is that the case does not implicate ordinary modesty in the explanation of how it is that people like us end up having inconsistent beliefs. For one thing, the list given above of things you believe that cannot all be true together does not include the claim that you have a false belief. It does include two propositions that together imply that you gave a false answer on the test. But even though in fact your current beliefs agree with your past answers, the proposition that one of your current beliefs is false is modally independent of the proposition that one of your past answers is false. The fact that you believe that one of your beliefs is false is a result of recognizing the inconsistency in your beliefs, but it is not a contributor to that inconsistency. The sort of inconsistency involved is rather the familiar – which is not to say unpuzzling – kind that persists because it is spread across too many of our beliefs for us to intelligently defuse.

To drive home the point that modesty is a red herring *vis-à-vis* inconsistency, notice that we can get a double dissociation of modestly and inconsistency by considering two minor variants of the case in which you are misinformed about your score on the test. First, suppose that instead of being told you scored 99 out of 100, you are told you scored 99 out of 99. Then your beliefs are still inconsistent, since there are 100 different propositions each of which you believe to be on the test while also believing that there were only 99 propositions on the test. But you needn't be modest – in fact, as far as the test goes, you are

immodest – and the inconsistency does not depend on your continuing to believe your original answers or on your believing that you have. Second, suppose that instead of being told that you scored 99 out of 100, you are told that you scored 100 out of 101. You will still be modest, for the same reason as in the original version of the case. But you need not be inconsistent: there may be a world in which the test had 101 questions, comprised of all of the questions you actually answered together with a further question that you didn't actually answer, in which you correctly answer all the former questions as you actually did but incorrectly answer the latter question.

4 Christensen's puzzle

With the test-score case in view, let us now return to Christensen's analysis of the preface. In response to the idea that the author's cognitive limitations insulate her from inconsistency, he writes:

> Surely an ordinary author who was paying attention could entertain the conjunction of the first two claims in her book, and recognize the material equivalence of this conjunction and the claim
>
> (1) The first two claims in my book are true.
>
> She would then be led by closure to believe (1). She could then easily entertain the conjunction of (1) and the third claim in her book. Our limited closure principle would then dictate believing that conjunction. Recognizing the equivalence of this believed conjunction with the claim
>
> (2) The first three claims in my book are true
>
> would lead, by similar reasoning, to belief in (2), and so on, until the belief in her book's inerrancy is reached. It must be granted that only an agent hard-up for entertainment would embark on such a process. But it is certainly not beyond normal cognitive capabilities, and the inerrancy belief seems no less irrational for having been arrived at by such a laborious route. (Christensen, 2004, p. 38-9)

Christensen is clearly correct that this kind of reasoning is possible for ordinary people. But his argument is unsound as it stands. Suppose the book contains 1,000 claims. Then at the end of this process, the author will believe that the first 1,000 claims in the book are true. But this is not inconsistent with each of these claims being in the book, being true, and yet the book containing a false claim. For it is consistent with these beliefs that the book contains more than 1,000 claims, the first 1,000 of which are true, but one of which is false.

Of course, this gap in the argument is easily filled. For the author can easily tell when they get to the end of the book and conclude that the book contains only 1,000 claims. But if they do this, then the processes of endorsing every claim in the book along the way was unnecessary. As the test-score case shows,

the author could instead have simply *counted* the number of claims in the book. That would have been enough for her to end up with inconsistent beliefs.

Yet while the tedious chain of self-congratulations is not needed for an ordinary author to end up with inconsistent beliefs, it is essential to their arriving at a conclusion that *blatantly contradicts* their beliefs. As in the test-score case, the inconsistency involved when the author merely counts the claims in the book is extremely diffuse, spread over thousands of logically independent propositions. Deriving a contradiction from such propositions is computationally infeasible. By contrast, it is trivial to derive a contradiction from the claims that the first 1,000 claims in the book are true, that the book contains 1,000 claims, and that the book contains an error. Whatever one thinks about the rationality of diffuse inconsistency, it is clearly irrational to have such blatantly inconsistent beliefs.

Christensen's argument therefore presents a new and different puzzle. Let us agree with him that the conclusion of the imagined chain of reasoning – namely, that the first 1,000 claims in the book are all true – is not something that it would be reasonable for the author to believe on the basis of that reasoning. The puzzle is that every step in their reasoning seems impeccable. This is a new puzzle because it purports to show not that ordinary authors currently have inconsistent beliefs (they needn't), but rather that there is a sequence of trivial observations and inferences available to them which seem individually reasonable but ultimately and foreseeably lead to an unreasonable conclusion that also blatantly contradicts some of their current belief. Note that this sequence of new beliefs is not formed merely on the basis of deduction. It involves empirical investigation too, both in seeing what claims are made at what places in the book, and also in monitoring the progress of this census. This non-deductive component is essential, since even if the author began with inconsistent beliefs (for example, by knowing how many claims are in the book), deducing a contradiction from those beliefs would be medically impossible.

I am not sure what the solution to this puzzle is. But here is a tentative suggestion, modeled on Makinson's own tentative solution the preface paradox. He proposed that "Even though each individual belief expressed by our author [...] is rational, the collection of all his beliefs is not. If the author is to have a rational *set* of beliefs he must change them" (p. 207, emphasis original). While I am not sure I understand Makinson's proposed distinction between the rationality of each member of a set of beliefs and the rationality of that set of beliefs taken as a whole (a distinction he himself describes as "perhaps unwelcome"), I do think I understand a parallel distinction between the reasonableness of each member of a sequence of actions and the reasonableness of performing the entire sequence of actions. Perhaps this is what is going on with Christensen's puzzle. Each step in the author's reasoning is beyond reproach, but the reasoning is unreasonable taken as a whole. Note that this diagnosis is compatible with every chain of purely deductive inference involved in the reasoning being beyond reproach, since, as we saw, the paradoxical reasoning essentially involves more than mere deduction; in particular, it requires observation and introspective monitoring to be regularly interspersed along the way.

5 Conclusion

Let's take stock. We first saw that there is no logical incompatibility between being a logically and introspectively ideal agent and believing that you have false beliefs. The puzzle is not logical but metaphysical. What could rational belief be so that it has the structure it would need to have for there to be a modest ideal agent? Certain otherwise attractive theories of belief and justification will have to be rejected.

We then considered whether ordinary modest people thereby have inconsistent beliefs, and saw that modesty does not beget inconsistency. Roughly, this is because what makes ordinary people's beliefs inconsistent in the relevant cases also makes their beliefs inconsistent in cases that clearly have nothing to do with modesty. Suppose you stand at the door and greet everyone who attends your high-school reunion. Hundreds of people show up. But you have a great memory – everyone who comes is someone who, later that night, you remember greeting. Unbeknownst to you, a devious epistemologist has doctored your meticulous attendance records. Your records indicate, and you believe as a result, that fewer people attended than actually came. We saw that, in cases like this, your beliefs will be inconsistent. But modesty is neither here nor there.

Of course, modesty is not completely unrelated to inconsistency. We are modest when we believe that our beliefs are inconsistent. And when our beliefs are inconsistent, we are often in a position to recognize this fact. But this connection between modesty and inconsistency is superficial. It is like the connection between believing that there is a crustacean in front of you and there being a lobster in front of you: often, when there is a lobster in front of you, you are in a position to recognize this fact, and so will also believe its obvious consequence, namely that there is a crustacean in front of you.

The non-surveyability of belief has played a crucial role in our investigation, both in the case of ideal reasoners and in the case of ordinary people. Ideal reasoners can be modest and consistent only by failing to believe, of the totality of their beliefs, that it is the totality of their beliefs. And ordinary people's inability to survey their beliefs prevents inconsistencies that are spread across many independent beliefs from wreaking havoc in deductive reasoning.

This raises the question of the extent to which we are able to survey our beliefs. Books and tests are special in this regard. The relevant claims are neatly ordered and unchanging, so we can leisurely go through them one by one. Our body of beliefs as a whole is not like this, and being ideally rational does not change this fact. So the hypothesis that ideally rational creatures would have different patterns of beliefs about the claims in their books or their answers on tests, by being either less modest or more cautious, does not imply that they would be generally immodest or unopinionated.

Modesty and inconsistency are puzzling enough without being run together. The literature on the preface paradox has encouraged the impression that the former involves the latter. It does not. I'm sure my beliefs aren't all consistent, but not because I know this paper inevitably contains some errors.

Appendix

A person is consistent if the things they believe could all be true together: that is, if it is possible that everything they actually believe be true. To formalize this claim, we enrich our language with an operator $\Box$ ("it is necessary that ...") and @ ("it is actually the case that ..."). I will show that, given some weak assumptions about the interaction of necessity, actuality, and propositional quantifiers, being modest, consistent and negatively introspective is impossible. The assumptions are that @ is a rigidifying operator that commutes with quantifiers and Boolean connectives within the scope of modal operators and quantifiers:

RIG $\varphi \leftrightarrow \Box @\varphi$

$@_\forall$ $\Box(@\forall p\varphi \leftrightarrow \forall p@\varphi)$

$@_\rightarrow$ $\Box\forall p(@(\varphi \rightarrow \psi) \leftrightarrow (@\varphi \rightarrow @\psi))$

$@_\neg$ $\Box\forall p(@\neg\varphi \leftrightarrow \neg@\varphi)$

Modesty, consistency and negative introspection are then formalized as follows:

MOD $B\exists p(Bp \wedge \neg p)$

CON $\Diamond\forall p(@Bp \rightarrow p)$

NI $\forall p(\neg Bp \rightarrow B\neg Bp)$

Combining NI with RIG, $@_\forall$, $@_\rightarrow$ and $@_\neg$, in that order, we derive:

$$\Box\forall p(\neg@Bp \rightarrow @B\neg Bp)$$

Combined with CON, this implies that everything I actually believe being true is not merely possible, as CON states, but also compossible with everything I believe being something I actually believe:

$$\Diamond(\forall p(@Bp \rightarrow p) \wedge \forall p(Bp \rightarrow @Bp))$$

And given MOD and RIG, this is in turn compossible with my actual modesty:

$$\Diamond(\forall p(@Bp \rightarrow p) \wedge \forall p(Bp \rightarrow @Bp) \wedge @B\exists p(Bp \wedge \neg p))$$

But this is an impossibility: the first two conjuncts imply that everything I believe is true, while the first and third conjuncts imply that something I believe is false. One cannot be modest, consistent, and negatively introspective.

Note that $@_\forall$ is essential to this argument. Intuitively, it encodes the assumption (tacit in our informal exposition of the puzzle) that we are considering only possibilities in which every proposition is an actually existing proposition. To see why we need this assumption, notice that if we take the model of $\mathcal{U}$ described in section 1 and reinterpret the propositional quantifiers so that, at w, they range only over subsets of W that are "understood" at w, the result is a model of $\mathcal{U}$ in which I *am* modest, consistent, and negatively introspective. $@_\forall$ fails in this model: although there could be something I actually neither believe nor believe I don't believe, there couldn't actually be something I neither believe nor believe I don't believe.

References

Alex Byrne. *Transparency and Self-Knowledge*. Oxford University Press, 2018.

David Christensen. *Putting Logic in its Place*. Oxford University Press, 2004.

Yifeng Ding. On the logic of belief and propositional quantification. *Journal of Philosophical Logic*, forthcoming.

Cian Dorr and John Hawthorne. Semantic plasticity and speech reports. *Philosophical Review*, 123(3):281–338, 2014.

Andy Egan. Seeing and believing: perception, belief formation and the divided mind. *Philosophical Studies*, 140(1):47–63, 2008.

Adam Elga and Agustín Rayo. Fragmentation and information access. In Cristina Borgoni, Dirk Kindermann, and Andrea Onofri, editors, *The Fragmented Mind*. Oxford University Press, forthcoming.

Simon J. Evnine. Believing conjunctions. *Synthese*, 118:201–27, 1999.

Jeremy Goodman and Bernhard Salow. Taking a chance on KK. *Philosophical Studies*, 175(1):183–96, 2018.

Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–70, 1988.

David Lewis. Logic for equivocators. *Noûs*, 16(3):431–41, 1982.

D. C. Makinson. The paradox of the preface. *Analysis*, 25(6):205–7, 1965.

Alexander Razborov. Proof complexity of pigeonhole principles. In Werner Kuich, Grzegorz Rozenberg, and Arto Salomaa, editors, *Developments in Language Theory 2001*, volume 2295 of *Lecture Notes in Computer Science*, pages 100–116. Springer, 2002.

Burkhard K. Schipper. Awareness. In Hans van Ditmarsch, Joseph Y. Halpern, Wiebe van der Hoek, and Barteld Kooi, editors, *Handbook of Epistemic Logic*, pages 77–146. College Publications, 2015.

Martin Smith. *Between Probability and Certainty: What Justifies Belief*. Oxford UP, 2016.

Martin Smith. The hardest paradox for closure. *Erkenntnis*, forthcoming.

Robert C. Stalnaker. *Inquiry*. Cambridge, MA: MIT Press, 1984.

Robert C. Stalnaker. The problem of logical omniscience, I. *Synthese*, 89(3): 425–40, 1991.

Robert C. Stalnaker. The problem of logical omniscience, II. In *Context and Content*, pages 255–73. Oxford University Press, 1999.

Robert C. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128:169–99, 2006.

Robert C. Stalnaker. Contextualism and the logic of knowledge. In *Knowledge and Conditionals: Essays on the Structure of Inquiry*, pages 129–48. Oxford University Press, 2019.

Brian Weatherson. Can we do without pragmatic encroachment? *Philosophical Perspectives*, 19(1):417–43, 2005.