The paradox of the proofreader

We've all been there. Rereading our papers, looking for mistakes, long past the point of diminished returns. I've gone through drafts line by line without catching a single error. Every claim looks good, and when I consider it, I accept it. But I still don't think my papers are error free. That's why I keep checking. I know my track record. At some point in the future I'm going to realize that something I wrote was wrong. So I'm not merely open to the possibility that my paper contains an error – I positively believe that it does. In fact, I *know* it does, by any ordinary standard for inductive knowledge.

This situation is reminiscent of Makinson's (1965) famous *preface paradox*: in the preface of their new book, an author thanks their colleagues for spotting errors in earlier drafts while assuming responsibility for the errors that no doubt remain. The paradox is meant to be that, on the one hand, the author's attitude seems entirely reasonable, but, on the other hand, it seems to involve knowingly having inconsistent beliefs. For the author believes every claim in their book and believes that the book contains an error, and this last claim seems tantamount to believing the negation of the conjunction of all the claims in the book – after all, the book contains an error if and only if that conjunction is false.

My own view, following Evnine (1999), is that the preface author's beliefs are only apparently inconsistent: believing that a book contains an error is not tantamount believing the negation of the conjunction of the claims in the book, and such book-length conjunctions are not propositions that ordinary people can even entertain, let alone believe. But nothing in what follows depends on this diagnosis of the preface paradox.

My goal here is to isolate a different, overlooked paradox, which I call the *proofreader paradox*. Unlike the preface paradox, which is a puzzle about what a person believes at a single time, the proofreader paradox is a puzzle about belief dynamics, and, in particular, about how to reconcile careful checking that uncovers no errors with the conviction that some errors have gone undetected. Section 1 present the puzzle. Section 2 offers a tentative diagnosis that brings together recent ideas of Yalcin (2018, 2021), Hoek (forthcoming), Friedman (2013, 2017), Holguín (2022), and Worsnip (2021) on the role of questions and structural rationality in constraining our opinions. Section 3 shows that a seemingly minor variant of the paradox, in which the authors memorizes the order of the claims in their book, raises a challenge for this diagnosis. Section 4 concludes with a suggestion for how the challenge might be met by using question-sensitive models of belief to distinguish explicit from implicit beliefs.

1 The paradox

The paradox was introduced, in effect, by Christensen (2004), in the course of arguing that preface cases generate puzzles for ordinary authors even if we grant that they cannot believe book-length conjunctions. 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 surely right that it would be irrational to come to believe that one's book contains no errors by following the procedure he describes. The proofreader paradox is to explain why.

Before tackling the normative question of where the procedure Christensen describes goes wrong, let us first consider the descriptive question: what do ordinary people actually do in cases like this?

I don't think Christensen is right that "only an agent hard-up for entertainment would embark on such a process". On the contrary, I embark on something very much like this process every time I correct the proofs of an article prior to publication: I check the claims in the paper one by one, numbered-line by numbered-line. Of course, when I do this, I don't end up believing that my papers are error free. So at some point I must depart from the reasoning Christensen outlines. But where? Every step of the reasoning looks impeccable.

Let us consider how far a typical author will follow the above reasoning. If they're like me, they'll happily accept that the first two claims in their book are true, and also that the first three claims in the book are true. But by the time they are a hundred claims in, they won't accept that the first hundred claims in the book are all true. So there is a last number n, between three and ninety nine, such that the author will accept that the first n claims in the book are true: when the author gets to the $n+1$st claim, they won't accept that the first $n+1$ claims in the book are true. Why the hesitation?

The answer depends on the details of the case. In some versions of the case, the $n+1$st claim will be the first one that the author finds doubtful. But that is

not the version of the case I am interested in. I want to consider a proofreader in the situation I have often found myself in, who accepts every claim they make in a book when they consider it but nevertheless believes that there is at least one error they haven't uncovered. So we may assume that the author accepts that the $n+1$st claim in the book is true.

Moreover – again, if the author is anything like me – then at the point when, having just considered the $n+1$st claim in the book, they both accept that it is true and suspend judgement on whether the first $n+1$ claims in the book are true, they will also *no longer accept that the first n claims in the book are true.* They will have lost their nerve.

I take it that this pattern is typical (although this is an empirical question, and I think the dynamics of acceptance and suspension in these cases would benefit from experimental investigation). For one thing, it aligns with our patterns of assertion. An author who didn't lose their nerve could sincerely assert: "The $n+1$st claim in the book is true. Also, the first n claims in the book are true. But I'm not sure whether the first $n+1$ claims in the book are true." This speech is utterly bizarre, suggesting that one's hesitance about the first $n+1$ claims goes along with a newfound hesitance about the first n claims.

When we lose our nerve, how much nerve do we lose? For all I have said, when the author gives up their belief that the first n claims in the book are true they will still believe that the first $n-1$ claims in the book are true. But this is not what happens. Not only will they no longer believe this, they will no longer believe that even the first half of the claims they have reviewed are true. After all, it would be bizarre to say "I'm not sure that there's no error in the first n claims in the book, but if there is then the error isn't in the first half of those claims." When we lose our nerve, we give up almost all of our beliefs about which numbers correspond to error free claims.

In this respect the situation is different from sorites cases. If we consider a sorites sequence of shorter and shorter people, the first of whom is seven feet tall, we will eventually come to someone who we will hesitate to describe as "tall". When this happens, we also hesitate to describe as "tall" the previous few people we just considered (who we previously did describe as "tall"); see Raffman (1994, 1996), Fara (2000). But we are still happy to describe a nontrivial initial segment of the sequence of the people we considered as "tall". Since proofreaders are no longer happy to say even that a nontrivial initial segment of the sequence of claims they have checked are error free, something different is going on. The proofreader paradox is not merely a manifestation of the sorites paradox.

Here is what seems to be going on. The author believes each claim in the book – that doesn't change. What changes are which numbers n are such that they believe that these are the number of a true claim in the book. I am assuming that, at the beginning, the author doesn't know of any claim in the book what its number is. This is slightly unrealistic – most of us know how our manuscripts begin – but inessential (we could imagine starting in the middle of the book rather than the beginning). As they proofread they learn new facts about which claims occur in what ordinal position: when they come to the ith claim in the book they learn that p_i is the ith claim in the book. Putting this together with

their belief in p_i they conclude that the ith claim in the book is true.

But an ordinary author will quickly forget which claims have which numbers. This doesn't stop them from continuing to believe that those numbers correspond to truths, but the basis for such beliefs is no longer the beliefs in the propositions in the book that these numbers correspond to. They are, rather, the author's memory of having recently checked the claim so numbered and having concluded it was true. But eventually, when they reflect on how many claims they have checked and it seems too many to reasonably believe that all are error free, they withhold that belief. In doing so, they also give up their belief that the ith claim in the book is true, for each $i \leq n$ such that they no longer remember which claim is the ith claim.

A similar diagnosis applies to people with unusually good memories who don't forget the numbers of their claims. Even if they know, for every $i \leq n$, which claim in the book is the ith one, this unusually robust but not superhuman factual memory won't give them a superhuman working memory. They will still not be able to entertain book-length conjunctions. We may suppose that n-length conjunctions are similarly beyond their ken. So such a person still cannot believe that the first n-claims in the book are true by holding in mind the conjunction of those claims, believing it, and then putting this belief together with knowledge that it is the conjunction of the first n-claims in the book. Such a person may still believe, of every $i \leq n$, that the ith claim in the book is true, but that is not the same as believing that the first n claims in the book are true.

With this description of proofreaders' beliefs on the table, let us finally return to the normative question: how *should* their beliefs evolve?

Here is why the case presents a puzzle. On the one hand, it seem as though the only reasonable belief dynamics must be something like the ones proofreaders actually have. Plausibly: we should accept the individual claims in our books; we shouldn't follow the procedure Christensen outlines to conclude that our books are error free; and at no point should we think that the first i claims are true and think that the $i+1$st claim is true without thinking that the first $i+1$ claims are true. On the other hand, it is plausible that the only situations in which we are ever required to give up beliefs which we previously rationally held are situations in which we get evidence against our previous beliefs. Indeed, it might seem not merely capricious but positively irrational to suspend judgment on a claim one previously believed in response to learning something with no evidential bearing on that claim. Yet this is exactly what one would have to be doing if one gave up one's belief that the first n claims in the book are true in response to the seemingly unrelated fact that p_{n+1} is the $n+1$th claim in the book, which is what it seems that real proofreaders actually do.

2 Diagnosing the puzzle

Let us begin by reviewing the most well-known account of rational belief that issues a clear verdict about the proofreader paradox. This is a normative version of the *Lockean thesis*, according to which a person should believe all and only the

propositions that have high enough probability given their evidence. This view predicts that the proofreader should adopt the bizarre stance discussed above: when they accept that the $n+1$st claim in the book is true yet suspend judgment on whether the first $n+1$ claims are true, they should continue to accept that the first n claims in the book are true. The proposition that the first $n+1$ claims in the book are true will be the first such proposition to have below-threshold probability, but the $n+1$st claim on its own has above-threshold probability, and the proposition that the first n-claims in the book are true continues to have above threshold probability, since the author's new evidence – that p_{n+1} is the $n+1$st claim in the book – does not lower its probability.

My own view is that there are decisive reasons to reject Lockeanism that are independent of its counterintuitive predictions about cases like these. But here is not the place for a defense of this claim. In this section I will instead consider whether an alternative picture of rational belief can vindicate the dynamics of actual proofreaders' beliefs.

Here are the components of the diagnosis I will be exploring:

1. Limits in the complexity of the questions we can hold in mind motivates a model of belief as question sensitive (Yalcin, 2018);

2. There are nontrivial coherence constraints on how one's question-sensitive beliefs should fit together (Hoek, forthcoming);

3. These coherence constraints go beyond the rational requirements on the individual attitudes in question (Worsnip, 2021);

4. The rational requirements on beliefs are more permissive, and belief itself more voluntary, that is usually assumed (Holguín, 2022); and

5. Suspending judgment is an attitude in its own right subject to non-trivial rational requirements in connection to belief (Friedman, 2013, 2017).

This section motivates and informally presents the proposal. More precise definitions and models are given in an appendix.

The first idea is that belief is question sensitive. To *believe p relative to Q* is to have a total opinion about Q at least as strong as p; see Yalcin (2018). For now, let us assume that agents have beliefs relative to *all* (and only) the questions they can hold in mind; we will revisit this assumption in Section 4.

Second, following Hoek (forthcoming) (and *pace* Yalcin (2021)) there is a non-trivial *coherence* requirement on one's beliefs relative to different questions. Specifically, one's beliefs relative to coarse-grained questions should align with one's beliefs relative to more fine-grained questions. For example, if relative to the question *who is President and who is Vice President?* you believe that Biden is President and Harris is Vice President, then, if you believe anything at all relative to the more coarse-grained question *who is President?*, you should believe that Biden is President. And, conversely, anything you believe relative to a coarse-grained question is something you should also believe relative to any strictly more fine-grained question relative to which you believe anything at all.

The main idea I want to explore is that proofreaders' question sensitive belief states evolve as they do in order to remain (i) *coherent* (in the sense just described), (ii) *(locally) consistent* (in the sense that they don't have an inconsistent opinion on any individual question), (iii) *opinionated* (in the sense that they believe each claim in their book relative to the question whether it is true), and (iv) *modest* (in the sense that they don't believe relative to any question that the first $n + 1$ claims in their book are true).

First, some clarifications about the framework. We model question sensitive belief states as partial functions from questions to sets of possible worlds. An agent has beliefs relative to a question just in case their belief state is defined on that question. The value of this function for that question is the set of worlds in which everything they believe relative to that question is true. We assume that, relative to any given question, the agent's beliefs are closed under conjunction and logical equivalence, and that any (partial) answer to the question that is true in all worlds compatible with what the agent believes relative to that question is something that they also believe relative to that question.

It is important to distinguish the local consistency requirement (ii) from a stronger *global* consistency requirement that the totality of propositions an agent believes relative to *some question or other* is consistent. Moreover, in what follows I will assume that the proofreader is *not* globally consistent. This is because I will assume, first, that the proofreader believes that the book contains an error (relative to the question whether it does), and, second, that the book containing an error is equivalent to the negation of the conjunction of all the claims in the book. This second assumption is strictly speaking false, but it is a useful idealization since it allows us to restrict our attention to worlds that agree on which claims are written in the book. (This idealization would be illegitimate if our main concern was whether the author's beliefs are globally inconsistent. For as Evnine (1999) emphasizes, it is contingent what claims are written in the book, and as a result is it not clear that ordinary sincere preface writers thereby have inconsistent beliefs. But the idealization is harmless in the present context, for two reasons. The first is that, as Goodman (2022) shows, it is not hard to make a sincere preface author's beliefs globally inconsistent – for example, by having them count the claims in their book. The other, more important reason is that, unlike the preface paradox, the proofreader paradox is not primarily about consistency. For it does not depend on assuming that the author believes their book contains an error, but only on the claim they are modest in the sense of (iv). We attribute the error belief only to make the case more realistic, and we treat this belief as tantamount to a belief in the negation of the conjunction of the claims in the book only to make our models more tractable. Doing so is inessential to generating the puzzle and to the proposed solution.)

If we are to reconcile global inconsistency with both local consistency and coherence, the questions relative to which the author has beliefs cannot be closed under fusion (where the *fusion* of some questions is the question *which answers to any of these questions are true?*). In particular, the author must have no beliefs whatsoever relative to the fine-grained question *is* p_1 *true, and is* p_2 *true, and . . . ?*, where these are all the claims in the book. For if they did have

beliefs relative to this question, coherence would require them to believe each p_i relative to it (since they believe each p_i relative to the more coarse grained question *is p_i true?*) and also to believe the negation of the conjunction of all p_i relative to it (since they believe this relative to the question of whether that negated conjunction is true), in violation of local consistency. Fortunately, it is not implausible that the author does fail to have any beliefs relative to this question, since ordinary agents cannot hold in mind such complicated questions.

For tractability we will now introduce a special class of question sensitive belief states, which are easy to specify and are guaranteed to be coherent. These belief states are *generated* from a set of propositions the agent *groks*, a subset of these propositions that the agent *accepts*, and a number κ which we may think of as a measure of the complexity of the questions that the agent can hold in mind. The agent has beliefs relative to Q just in case, for some set X of at most κ propositions they grok, Q is the question *which members of X are true?* (The neologism 'grok' is chosen to indicate that how best to understand the attitude involved here is debatable; we will return to this issue later.) Relative to that question, the agent believes all members of X which they accept, and we let their total belief state be the least opinionated one that satisfies these constraints and is coherent; see the appendix for a precise definition. We will assume that the proofreader's beliefs are at all times generated in this way; for concreteness, we let $\kappa = 3$.

We can now give a toy model of how the proofreader's beliefs evolve. It suffices to specify which propositions they grok and which of these they accept, as they sequentially consider the claims in their book. We assume that they always accept (and so grok) each claim in their book and also that the book contains an error. But they are quick to forget which claims occur where in the book: although they grok all propositions of the form *p_i is the jth claim in the book*, the only such propositions they accept are that *p_i is the ith claim in the book* and (for $i \geq 2$) that p_{i-1} is the $i-1$st claim in the book, where the ith claim is the one they are currently considering. The proofreader also groks the proposition that the ith claim in the book is true, that the first i claims are true, and (for $i \geq 2$) that the first $i-1$ claims are true. For now, let us assume that the abovementioned propositions are the only ones the proofreader groks.

When the proofreader considers claim i, they will accept that it is true. And if $2 \leq i \leq n$, then they will also accept that the first $i-1$ claims in the book are true. But when they consider the $n+1$st claim in the book, they will no longer accept that the first n claims in the book are true. The belief state generated from this pattern of groked and accepted propositions for $\kappa = 3$ satisfies the opinionatedness, locally consistency, coherence, and modesty requirements discussed above; see Proposition 13(a) in the appendix.

A key feature of this proposal is that, when the proofreader comes to the $n+1$st claim in the book and accepts that it is true, coherence and modesty require that they stop accepting that the first n claims in the book are true. (*Proof*: Suppose the proofreader continued to accept that the first n claims in the book are true upon accepting that the $n+1$st claim is true. By coherence, they would then believe both of these propositions relative to the question *which of those*

two propositions is true?, since they have beliefs relative to this question since it is the fusion of two propositions they accept and hence grok. So they would also believe that the first $n+1$ claims are true relative to this question, since belief relative to individual questions is closed under conjunction and logical equivalence. This contradicts our assumption of modesty.) This disposition to remain coherent and modest explains why the proofreader stops accepting that the first n claims are true.

This proposal raises two questions: why does the proofreader remain modest rather than accept that the first $n+1$ claims are true, and why isn't it irrational for them to stop accepting that the first n claims are true? Let us consider these in turn.

There are a number of reasons why the proofread may fail to accept that the first $n+1$ claims are true. For example, perhaps once $n+1$ claims are involved, the probability that they are all true becomes too low for it to be possible for the agent to *know* that they are all true, and the proofreader is disposed to conform to a knowledge norm on belief. This diagnosis is suggested by probabilistic question sensitive models of knowledge and rational belief; see Goodman and Salow (2021, 2023). A different and complementary explanation is that the agent *suspends judgment* on the question whether the first $n+1$ claims in the book are all true, where this is a positive stance on the question and not a mere lack of belief relative to it (Friedman, 2013), and the proofreader is simply conforming to a non-belief norm on suspension (Friedman, 2017).

Let us now consider whether it is irrational for the proofreader to suspend judgment on whether the first n claims are true, which a moment before they accepted. Although this seems capricious considered on its own, it is explicable as a way of maintaining coherence. Moreover, there are good reasons arising from reflections on the question-sensitivity of belief for thinking that how much we believe, though constrained in certain ways by one's subjective probabilities, is not mandated by those probabilities in the way the Lockean thesis dictates. In particular, Holguín (2022) powerfully argues that the weak notion of belief expressed by 'thinks that' in ordinary English is highly permissive – we may believe that a horse will win a race simply because they are the horse we think is most likely to win, but we may also choose to withhold belief even if we think the horse is an overwhelming favorite. And if Goodman and Holguín (forthcoming) are right, then this voluntarism extends to stronger notions of belief such as the one expressed by 'be sure' in ordinary English, since thinking is a necessary condition on these stronger attitudes. The rational permissibility of suspension is then largely independent of the precise notion of belief at issue.

The picture that emerges is one where the proofreader's beliefs shift in order to maintain coherence within the bounds of what is rationally permitted. This fits the general picture of rational requirements defended by Worsnip (2021), who argues that the structural requirements of rationality, to do with how our attitudes fit together, are not reducible to rational requirements on individual attitudes, saying which individual attitudes we must and mustn't have.

3 The revenge of the memorizer?

How robust is this diagnosis of the proofreader paradox? This section explores this question by considering what happens when we try to extend the diagnosis to proofreaders with unusually good memories. It turns out that this raises a new challenge: unlike ordinary proofreaders, memorizers cannot be opinionated, coherent, and modest provided they have beliefs relative to all questions that are simple enough for them to hold in mind.

Suppose that, unlike a typical author, our proofreader has memorized the number of each claim in their book: for all i they accept that p_i is the ith claim in the book. Accommodating these additional beliefs requires only a slight tweak to the model from the last section. It requires a tweak because, when the proofreader gets to claim n in the book, they would be now be immodest if they still believed that the first n claims in the book are true, since coherence would then require them to believe that the first $n+1$ claims in the book are true relative to the question *are the first n claims true, and is p_{n+1} the $n+1$ claim in the book, and is p_{n+1} true?* (Unlike the forgetful proofreader, who will have no view before looking whether the $n+1$st claim is true, the memorize already believes this relative to *is p_{n+1} the $n+1$ claim in the book, and is p_{n+1} true?*) But the general diagnosis remains tenable; the proofreader simply has to loses their nerve a little sooner; see Proposition 14(b).

The problem arises when we consider all propositions of the form *the first i claims in the book are true*. In the model outlined in the last section, the proofreader only groks a few of these propositions at a time, and which ones they grok depends on which claim in the book they are currently considering. For example, when they are considering the eighth claim in the book, they won't have any beliefs relative to the question *are the first five claims in the book true?* *Prima facie*, this is an unnatural feature of the model, since all such propositions are ones the proofreader perfectly well understands and is capable of holding in mind and having opinions about. Is it an essential feature of the model?

In the case of an ordinary forgetful proofreader, who knows the ordinal position of at most a few claims in their book, it is inessential. In modeling their beliefs, we can modify the model from the last section by adding all propositions of the form *the first i claims in the book are true* to the set of propositions they grok without their beliefs collapsing into inconsistency or immodesty; see Proposition 13(b). The problem is that this isn't true for the memorizer: their being opinionated, coherent, and modest depends on their not grokking all such propositions; see Proposition 14(a), which proved essentially as follows.

Fact. *If the proofreader (a) is coherent, (b) for all i, accepts that p_i is true, accepts that p_i is the ith claim in the book, and grok the proposition that the first i claims in the book are true, and (c) has beliefs relative to the question* which members of X are true? *for every set X of at most three of the propositions mentioned in (b), then the proofreader must, for all i, believe that the first i claims in their book are true (relative to the question* are the first i claims in the book true?*).*

Proof: We argue by induction. Trivially, the author believes that the first 0 *claims in their book are true relative to the question* are the first 0 claims in the book true? *Now consider the question* are the first i claims in the book true, and is p_{i+i} the $i+1$st claim in the book, and is p_{i+1} true? *The proofreader groks all three relevant propositions (since acceptance entails grokking), and there are only three of them, so they have beliefs relative to this question. And they believe each of these propositions relative to the polar question whether it is true (the first two because they accept them; the third by the induction hypothesis). So coherence requires them to believe the conjunction of these propositions relative to the question just mentioned, and hence also believe the logically equivalent claim that the first* $i + 1$ *claims in the book are true, and hence (again by coherence) believe this relative to the more coarse grained question* are the first $i + 1$ claim in the book true?

Note that the admittedly obscure notion of 'grokking' plays no essential role in this proof: it plays only an instrumental role in characterizing some simple questions relative to which the proofreader has beliefs.

Of course, like an ordinary author, the memorizer clearly shouldn't and presumably wouldn't believe that the first i claims in the book are true for every i. In this way the memorizer presents an important challenge to the above diagnosis of the proofreader paradox.

Intuitively, what is going on is that cognitively bounded agents like us are able to exploit the ordinal structure of numbers to entertain generalizations about intervals of numbers even when the parallel generalizations about the set of propositions corresponding to those numbers are beyond our cognitive powers. This generates a tension between opinionatedness, coherence, and modesty when the our beliefs take a stand on the number-to-proposition correspondence, as the memorizer's beliefs do.

Note that a parallel issue arrises for a more familiar kind of memorization, where the author doesn't know the number of each claim in their book but can reproduce the book from memory: they know what the first claim is, what the last claim is, and, for all claims in between, what claim precedes it and what claim follows it. The puzzle recurs if instead of propositions of the form *the first i claims in the book are true* we consider propositions of the form *all propositions preceding and including* p_i *are true*. There should again be a greatest i for which the author will accept the corresponding proposition. A parallel argument to the one given above then shows that such modesty is inconsistent with the author grokking all such propositions, accepting each claim in the book and each of the aforementioned facts about the relative order of the claims in the book, and having a coherent belief state generated from these basic attitudes in the aforementioned way.

4 Explicit beliefs as polar beliefs

The memorizer challenges our earlier diagnosis of the proofreader paradox in two ways. First, as just discussed, it poses a challenge for understanding the author's belief dynamics in terms of the interacting pressures of opinionatedness, consistency, coherence, and modesty. But it also poses a challenge to our

diagnosis of the proofreader paradox as unlike the preface paradox in concerning the dynamics of rational belief as opposed to the consistency of one's synchronic beliefs: opinionatedness and coherence threaten to force the author to believe that their book is error free before they have even proofread it.

I want to suggest that the model of the memorizer mentioned in the last section (in which they remain coherent and modest by failing to grok most propositions of the form *the first i claims in the book are true?*) isn't as unnatural as it first appears. It is only unnatural if we think of the propositions one groks as the claims one understands. But that was never an available interpretation, since in generated belief states agents typically believe many propositions that they do not grok, relative to non-polar questions (i.e., questions not of the form *is p true?*). How then should the notion be understood?

Suppose we think of the propositions one groks as those one either accepts, rejects (i.e., accepts the negation of), or suspends judgment on. (This would require tweaking our models so that the agent to grasp the negation of every proposition they grasp, but this modification doesn't change the models' predictions about the agent's beliefs.) Think of these as the propositions the agent has an *explicit* doxastic attitude towards: to *explicitly believe* a proposition is to believe it relative to the polar question of whether it is true, and to *suspend judgment* on a proposition is to have trivial beliefs relative to the polar question of whether it is true.

Beliefs relative to non-polar questions can be thought of as *implicit* beliefs, in at least one (of many possible) ways of drawing that distinction. This is reminiscent of the standard way of reconciling the representational theory of thought, according to which propositional attitudes are mediated by relations to stored mental representations, and dispositional theories of belief, according to which we count as believing propositions that we don't explicitly represent when these are obvious consequences of propositions we do believe by explicitly representing. Following Field (1978), the standard reconciliation is that we explicitly believe propositions that are expressed by mental representations playing the right role in our cognitive architecture ('mentalese sentences in our belief-box', to use the familiar jargon), and we implicitly believe propositions that are expressed by mental representations that are in some sense obvious consequences of these explicitly encoded representations. We can see the present models as implementing this idea: if a proposition is a Boolean combination of at most κ propositions one groks and is entailed by the subset of those propositions one accepts, then one counts as implicitly believing the proposition because, for small κ, it counts as an *obvious* consequence of the propositions one accepts.

In using this conception of grokking to model the memorizer in a way that essentially relies on them failing to grok all propositions of the form *the first i claims in the book are true*, it becomes crucial to distinguish suspend*ing* judgment, in the sense of the transition from acceptance to suspended judgment, and what we might call *discombobulation*, in the sense of the transition from accepting a proposition to one's belief state being *undefined* on the question whether that proposition is true, in which case one neither accepts nor denies *nor suspends judgment* on the proposition in question. The idea is that when

the memorizer suspends judgment on whether the $n+1$st claim in the book is true, having previously accepted that the first n-claims in the book are true, they are discombobulated – while for all we have said they suspend judgment on the question whether the first n claims in the book are true, there will be some $i \leq n$ such that their belief state is no longer defined on the question *are the first i claims in the book true*.

Obviously there is much more to be said about this way of thinking about explicit beliefs as polar beliefs. For example, for coherent agents, it runs counter to many of the ideas in Holguín (2022) and Goodman and Salow (2021) on the probabilistic structure of beliefs relative to non-polar questions. I hope to explore these questions in further work.

Appendix

Definition 1 (Partitions)**.** A *partition of* W is any set of disjoint non-empty subsets of W whose union is W. Let $\Pi(W)$ be the set of all partitions of W. We call the members of a partition its *cells*.

Definition 2 (Refinement)**.** Where $Q_1, Q_2 \in \Pi(W)$, Q_1 *refines* Q_2 iff every cell of Q_2 is a union of cells of Q_1; they are *comparable* iff one refines the other.

Definition 3 (Questioning)**.** $\pi_W : \mathcal{P}(\mathcal{P}(W)) \to \Pi(W)$ s.t. $\pi_W(X) = \{p \subseteq W : \exists w \in p, \forall v \in W(v \in p \leftrightarrow \forall q \in X(w \in q \leftrightarrow v \in q))\}$. Intuitively, π_W takes a set of propositions (modeled as subsets of W) and returns the question *which of these propositions are true?* (modeled as a partition of W).

Definition 4 (Belief States)**.** A *belief state (over* W*)* is a pair $\langle \mathcal{Q}, D \rangle$ such that

$\mathcal{Q} \subseteq \Pi(W)$, and

$D : \mathcal{Q} \to \mathcal{P}(W)$ such that, for all $Q \in \mathcal{Q}$, $D(Q) = \bigcup X$ for some $X \subseteq Q$.

Definition 5 (Local Consistency)**.** A belief state $\langle \mathcal{Q}, D \rangle$ is *locally consistent* iff $D(Q) \neq \emptyset$ for all $Q \in \mathcal{Q}$.

Definition 6 (Coherence)**.** A belief state $\langle \mathcal{Q}, D \rangle$ is *coherent* iff, for all $Q_1, Q_2 \in \mathcal{Q}$, if Q_1 refines Q_2, then $D(Q_2) = \bigcup\{q \in Q_2 : q \cap D(Q_1) \neq \emptyset\}$.

Definition 7 (Generation)**.** For $\kappa \geq 1$ and $B \subseteq A \subseteq \mathcal{P}(W)$ such that $\emptyset \notin B$ and, if $p \in B$, then $W \backslash p \notin B$, we say that $\langle A, B, \kappa \rangle$ *generates* the belief state $\langle \mathcal{Q}, D \rangle$ (over W) where:

1. $\mathcal{Q} = \{\pi_W(X) : X \subseteq A, |X| \leq \kappa\}$,
2. $D_0(Q) = p$ if $Q = \{p, W \backslash p\}$ and $p \in B$; otherwise $D_0(Q) = W$,
3. $D_{i+1}(Q) = \bigcup\{q \in Q : q \cap D_i(Q') \neq \emptyset$ for all $Q' \in \mathcal{Q}$ comparable with $Q\}$,
4. $D(Q) = \bigcap_i D_i(Q)$.

2 ensures that the agent believes everything they accept relative to the corresponding polar question. 3 and 4 ensure that these beliefs percolate between the overlapping questions in $\mathcal{Q}$ to achieve coherence in the least opinionated way.

Proposition 8. If a belief state is generated, then it is coherent.

Proof sketch: $D_{i+1}(Q) \subseteq D_i(Q)$, since every question is comparable with itself. So $D_i(Q)$ will reach a fixed point, since Q is finite and $D_i(Q) \in \{\bigcup X : X \subseteq Q\}$. Let j be a stage by which D_i reaches its fixed point for both Q_1 and Q_2. If Q_1 and Q_2 were a counterexample to coherence, then clause 3 would ensure that either $D_j(Q_1) \neq D_{j+1}(Q_1)$ or $D_j(Q_2) \neq D_{j+1}(Q_2)$, contradicting our fixed point assumption. So Q_1 and Q_2 are not a counterexample to coherence.

We now define our model. Assume that the book contains 1000 claims.

Definition 9 (Worlds)**.** $W = \mathcal{P}(\{1, \ldots, 1000\}) \times \{f : f \text{ a permutation of } \{1, \ldots, 1000\}\}$. $\langle X, f\rangle$ is a world in which, for all $1 \leq i \leq 1000$,

1. p_i is true iff $i \in X$, and
2. $p_{f(i)}$ is the ith claim in the book.

Definition 10 (Propositions)**.** We can define various propositions, modeled as subsets of W, as follows (glosses in square brackets):

$p_i := \{\langle X, f\rangle : i \in X\}$ [claim p_i is true]

$\#_{i,j} := \{\langle X, f\rangle : f(i) = j\}$ [p_j is the ith claim in the book]

$[i, j] := \{\langle X, f\rangle : \{f(k) : i \leq k \leq j\} \subseteq X\}$ [the ith to jth claims are all true]

$e := W \backslash [1, 1000]$ [the book contains an error]

We now define various sets of propositions that will figure as 'grokking' sets and 'acceptance' sets in generating belief states.

Definition 11 (Grokking sets)**.**

$A_0 := \{p_i : 1 \leq i \leq 1000\} \cup \{\#_{i,j} : 1 \leq i, j \leq 1000\} \cup \{e\} \cup \{[i, i] : 1 \leq i \leq 1000\}$

$A_i := A_0 \cup \{[1, i]\} \cup \{[1, i-1] : 2 \leq i\}$

$A^+ := A_0 \cup \{[i, j] : 1 \leq i \leq j \leq 1000\}$

Definition 12 (Acceptance sets)**.** For some constant $n \geq 2$,

$B_0 := \{p_i : i \leq 1000\} \cup \{e\}$

$B_i := B_0 \cup \{[i, i]\} \cup \{[1, i-1] : 2 \leq i \leq n\} \cup \{\#_{i,i}\} \cup \{\#_{i-1,i-1} : 2 \leq i\}$

$B'_0 := B_0 \cup \{\#_{j,j} : 1 \leq j \leq 1000\}$

$B'_i := B'_0 \cup \{[i, i]\} \cup \{[1, i-1] : 2 \leq i \leq n-1\}$

Proposition 13 (The Forgetful Proofreader)**.**

(a) The belief state generated from $\langle A_i, B_i, 3\rangle$ is locally consistent, and the agent believes $[1, i]$ relative to $\pi_W(\{[1, i]\})$ iff $i \leq n$.

(b) The same is true if we replace A_i with A^+.

Proposition 14 (The Memorizer)**.**

(a) The belief state generated from $\langle A^+, B'_0, 3\rangle$ is not locally consistent.

(b) The belief state generated from $\langle A_i, B'_i, 3\rangle$ is locally consistent, and the agent believes $[1, i]$ relative to $\pi_W(\{[1, i]\})$ iff $i \leq n$.

References

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

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

Delia Graff Fara. Shifting sands: An interest relative theory of vagueness. *Philosophical Topics*, 28(1):45–81, 2000.

Hartry Field. Mental representation. *Erkenntnis*, 13(July):9–61, 1978.

Jane Friedman. Suspended judgment. *Philosophical Studies*, 162(2):165–181, 2013.

Jane Friedman. Why suspend judging? *Noûs*, 51(2):302–26, 2017.

Jeremy Goodman. I'm mistaken. Unpublished manuscript, 2022.

Jeremy Goodman and Ben Holguín. Thinking and being sure. *Philosophy and Phenomenological Research*, forthcoming.

Jeremy Goodman and Bernhard Salow. Knowledge from probability. In Joseph Y. Halpern and Andres Perea, editors, *Theoretical Aspects of Rationality and Knowledge 2021 (TARK 2021)*, pages 171–186, 2021.

Jeremy Goodman and Bernhard Salow. Epistemology normalized. *Philosophical Review*, 132(1):89–144, 2023.

Daniel Hoek. Minimal rationality and the web of questions. In Dirk Kindermann, Peter van Elswyk, and Andy Egan, editors, *Unstructured Content*. Oxford University Press, forthcoming.

Ben Holguín. Thinking, guessing and believing. *Philosophers' Imprint*, 1(1-34), 2022.

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

Diana Raffman. Vagueness without paradox. *Philosophical Review*, 103(1): 41–74, 1994.

Diana Raffman. Vagueness and context-relativity. *Philosophical Studies*, 81 (2-3):175–192, 1996.

Alex Worsnip. *Fitting Things Together: Coherence and the Demands of Structural Rationality*. New York: Oxford University Press, 2021.

Seth Yalcin. Belief as question sensitive. *Philosophy and Phenomenological Research*, 97(1):23–47, 2018.

Seth Yalcin. Fragmented but rational. In Cristina Borgoni, Dirk Kindermann, and Andrea Onofri, editors, *The Fragmented Mind*, pages 159–80. Oxford University Press, 2021.