Iteration and Fragmentation

Daniel Greco

That our mental lives are not transparent to us is one of the few Freudian doctrines that has been robustly confirmed by subsequent psychological research. Nevertheless, much work in the philosophy of language, as well as formal work in economics and computer science, assumes various “iteration principles” that can seem to conflict with this psychological truism. By “iteration principles” I have in mind principles like the following:

(KK) If S knows that P, then S knows that S knows that P.

(BB) If S believes that P, then S believes that S believes that P.¹

These principles tell us that certain epistemic and doxastic operators (knowledge and belief, in these cases) automatically iterate. While KK and BB are probably the most well known iteration principles of interest to philosophers, I hope that my discussion

¹Many writers find the unqualified versions of iteration principles listed above extremely implausible, on the grounds that they lead to infinite regress—if S knows that P, and an unqualified version of KK holds, then S knows that S knows that S knows that S knows...ad infinitum. I do not find this consequence nearly as implausible as do many philosophers, but for those who do, there are available readings of the principles that don’t imply actual regresses (but merely potential ones). For example, KK can be formulated so that it merely implies that one is always in a position to know that one knows that P whenever one knows that P, and BB can be formulated so that it merely implies that one is always disposed to come to believe that one believes that P, whenever one believes that P.
may shed light on other iteration principles as well.\footnote{For early discussions of KK and BB, see Hintikka (1962), who endorses both principles. For some discussion of some applications of iteration principles in philosophy, economics, and computer science, see Vanderschraaf and Sillari's (2001) \textit{Stanford Encyclopedia of Philosophy} article on “Common Knowledge.” While the article is about common knowledge rather than iteration principles directly, the formal frameworks it discusses that have been used for modeling common knowledge each essentially incorporate KK as an assumption. Without that assumption, it is very hard to allow for the attainability of common knowledge. In particular, I argue that certain popular objections to KK entail that common knowledge is never attainable in Greco (Forthcoming). I will not discuss iteration principles concerning rationality, evidence, justification, or epistemic obligation in this paper—though I do defend an iteration principle concerning epistemic obligation in Greco (2013)—but iteration principles can be formulated for each of these notions, and I suspect that in each case, the plausibility of the principles will be closely related to the plausibility of analogous principles for knowledge and belief.}

In this paper I’ll discuss a pair of paradigmatic examples of the opacity of the mental. Both can be seen as counterexamples to iteration principles like KK and BB. However, I’ll distinguish two strategies for making sense of the examples, and only on one of these strategies, which I’ll call the “anti-iteration” strategy, are the examples straightforward counterexamples to iteration principles. On the other strategy, which I’ll call the “fragmentation” strategy, matters are more complicated, and there is the possibility of defending versions of iteration principles like KK and BB while acknowledging that our mental lives are not transparent to us. While the anti-iteration and fragmentation strategies are not strictly speaking inconsistent with one another, they do compete, since they do similar explanatory work. That they compete is significant, as I suspect many philosophers have been sympathetic to both strategies, and have failed to notice that adopting the fragmentation strategy undercuts much of the motivation for adopting the anti-iteration strategy. After introducing the examples and the two different strategies for making sense of them, I’ll ultimately offer one argument in favor of the fragmentation strategy and against the anti-iteration strategy. I’ll argue that the fragmentation strategy provides a satisfying way of making sense of the opacity of the mental, while also saving what’s attractive about iteration principles.
1 The Examples

Versions of the first example have been discussed a great deal in both the psychological and philosophical literature.\(^3\)

**The Implicit Sexist:** John is an avowed anti-sexist. In particular, he is prepared to defend vigorously the equality of the sexes in intelligence. Yet, in a variety of contexts, John’s behavior and judgments are systematically sexist. Concerning the individual women he knows, John rarely thinks they’re as intelligent as the men he knows, even when John has ample evidence of their intelligence. In group discussions, John is systematically less likely to pay attention to and take seriously the contributions of women. On the rare occasions when he does judge a woman to have expressed a novel, interesting idea, he is much more surprised than he would have been if a man had expressed the same idea. Still, John is unaware of these dispositions, and he would deny that he had them if asked.

The second example is adapted from a discussion of Colin Radford’s (1966), though similar examples have been discussed by a number of epistemologists.\(^4\)

**The Unwitting Historian:** Jean insists that she knows nothing about English history. As a matter of fact, she studied English history in secondary school, but doesn’t recall taking the course. She hasn’t forgotten the content of what she learned, however. If you force her to guess as to matters such as when William the Conqueror landed in England, the dates of Queen Elizabeth’s reign, and so on, she’ll reliably respond correctly. But if told that her answers are correct, she’ll be quite surprised, as she takes herself to have

\(^3\)For a philosophical discussion of a more detailed example very much like the one below, see Schwitzgebel (2010, p.2). The psychological literature on implicit attitudes is vast, but the collected bibliography here is a good start: [https://implicit.harvard.edu/implicit/demo/background/bibliography.html](https://implicit.harvard.edu/implicit/demo/background/bibliography.html)

\(^4\)Such examples are often referred to as “quiz show knowledge,” e.g., by Feldman (2005).
Iteration and Fragmentation

no way of knowing these facts.

2 The Strategies

2.1 Anti-Iteration

According to the anti-iteration strategy, both examples above have the same abstract structure; in both cases, for some subject $S$, some propositional attitude $\phi$, and some proposition $P$, we have the following:

1. $S \phi$'s that $P$

2. $S$ does not $\phi$ that $S \phi$'s that $P$.

In the case of the unwitting historian, the relevant attitude is knowledge—Jean knows various facts about English history, without knowing that she knows them. In the case of the implicit sexist, the relevant attitude is belief—John believes various claims about the inferiority of women, without believing that he believes them. According to the anti-iterationist, examples like those of Jean and John are important in part because they show us that certain iteration principles for knowledge and belief are false; one can know something without knowing that one knows it, just as one can believe something without believing that one believes it.

2.2 Fragmentation

According to the fragmentation strategy, neither example can be adequately explained by sticking to just one, unitary attitude (either belief or knowledge), and holding that the subject of the example bears that attitude to some propositions but not others (e.g., to $P$, but not to the proposition that she believes or knows that $P$). Rather, in both cases, we need to appeal to at least two different attitudes—or perhaps different species of the same attitude—to give an adequate characterization of the subject’s mental state.
As it’s easier to see how the strategy might work in the case of the implicit sexist, I’ll start there.

The fragmentationist will hold that the concept of belief is too crude a tool to use to capture the implicit sexist’s mental state. Instead, the fragmentationist will distinguish various different belief-like attitudes, and hold that John bears some of them but not others to the claim that the sexes are equal in intelligence. One version of this strategy involves distinguishing between explicit and implicit belief. Roughly, one’s explicit beliefs are what one is willing to assert and defend in argument, while one’s implicit beliefs are what one treats as true when one is acting unreflectively and automatically. With this distinction in hand, we can describe John’s state by holding that he explicitly believes that the sexes are equal in intelligence, while he implicitly believes that men are smarter than women. John’s psychological state is “fragmented”; in an important sense, he’s not of one mind on the question of the equality of the sexes.

Other versions of the fragmentation strategy will not draw an implicit/explicit belief distinction, but will still hold that some distinction in the neighborhood is called for. For instance, we might follow Tamar Gendler in distinguishing between belief and “alief,” and hold that John believes that the sexes are equal, but alieves that they are not. Or we might hold that John’s case is one of “in-between believing,” in which John has some but not all of the characteristic dispositions of believing that the sexes are equal in intelligence, and some but not all of the characteristic dispositions of believing that men

---

5Another way of drawing an implicit/explicit belief distinction involves holding that explicit beliefs are explicitly represented in the head (maybe as a sentence in the belief box), while implicit beliefs are merely dispositional. This is an orthogonal distinction to the one I am drawing. As I am drawing the distinction, explicit beliefs can be merely dispositional (as long as the subject is disposed to assert and defend them in argument), and implicit beliefs can be explicitly represented in the head (as long as they are manifested in unreflective, automatic behavior).

6The fragmentation strategy is so-called for two reasons. First, according to the fragmentation strategy, belief and knowledge are not unitary attitudes, but instead they fragment into different but related sub-attitudes. Second, the overall psychological state of an individual who bears some but not all of these sub-attitudes towards some proposition can be thought of as a fragmented, compartmentalized state.

7See Gendler (2008a).
Iteration and Fragmentation

are smarter than women. Any strategy that involves distinguishing between various different belief-like attitudes (or aspects of belief), and holding that John bears some but not all of these attitudes towards claims about the equality of the sexes, counts as an instance of the fragmentation strategy.

For convenience, I’ll use the strategy of explaining cases like John’s by appeal to the implicit/explicit belief distinction as my paradigm example of the fragmentation strategy, though not much will hang on this; those who think that even the implicit/explicit distinction is still too crude, or omits important complexities (as I do), can substitute their preferred classification in my discussions of the fragmentation strategy.

The fragmentationist will take a similar line concerning the case of the unwitting historian. Rather than treating the case as one in which Jean straightforwardly knows various facts, while straightforwardly failing to know that she knows them, the fragmentationist will hold that we need to draw a distinction between (something like) different species of knowledge, and to allow that they can come apart. After all, it’s not clearly correct to say that Jean straightforwardly knows the dates of Elizabeth’s reign; if given a test in which one is penalized for offering wrong answers, Jean will presumably leave questions about such matters blank, rather than providing the correct answers. Exactly

---

9 I intend the fragmentation strategy to encompass a wide range of views. In particular, views that don’t draw a distinction between different species of belief, but just hold that, e.g., beliefs are somehow relative to purposes or circumstances, would count. I have in mind views on which, e.g., the implicit sexist has straightforwardly inconsistent first-order beliefs—he both believes that men are smarter than women, and that men are not smarter than women—but that these beliefs are accessed for different purposes, or are activated in different circumstances. Why do I lump this sort of view together with views that distinguish implicit and explicit belief, or belief and alief? For two main reasons: (1) the view does not explain the case of the implicit sexist by appeal to a distinction between first-order and higher-order beliefs, and (2) the view does draw a distinction within the realm of belief that may play some of the same explanatory roles as an implicit/explicit or a belief/alief distinction—the distinction between the different purposes for which a belief can be accessed, or the different circumstances in which a belief can be disposed to be activated. Elga and Rayo (Ms.) propose a view along these lines, to which I am highly sympathetic, and which I take to be consistent with the strategy for salvaging iteration principles that I offer in the text.
10 The idea that we must distinguish between different species of knowledge is a familiar one in epistemology. Much of Ernest Sosa’s work depends on a distinction between animal knowledge and reflective knowledge. See Sosa (2007). The idea that some such distinction (or family of distinctions) may open up the possibility of defending iteration principles for knowledge, however, is not one that (to my knowledge) has been defended before.
how this approach should be implemented is not a question I'll take up here, though I'll try to say at least enough to suggest that some fragmentationist treatment of the case could be well motivated.

A central role of knowledge attributions is to explain successful action. For instance, we might explain how a traveller managed to reach her intended destination by pointing out that she knew which road led to it.\(^\text{11}\) A natural extension of this thought is the idea that knowing that \(P\) involves being in a state that would tend to explain certain successful actions (namely, those actions whose success depends on the truth of \(P\)).\(^\text{12}\) Once we accept this thought, however, a certain sort of fragmentationist picture is very naturally suggested. For someone might be in a state that would tend to explain some successful actions whose truth depends on \(P\), but not others. The unwitting historian is in such a state—her latent memories of English history explain why her actions in forced choice situations are regularly successful (i.e., why she regularly guesses right in such cases), and these are actions whose success depends on the truth of facts about English history. So arguably, she knows these facts in some sense. But whatever state she’s in, it does not tend to produce, and would not explain, other successful actions whose truth depends on facts about English history (e.g., actions in situations without the element of forced choice, and with penalties for guessing incorrectly). So arguably, she fails to know these facts in some other sense.\(^\text{13}\)

Someone taken with these thoughts might hold that rather than characterizing Jean’s epistemic state by saying what she knows full stop, we should admit that knowledge can be accessible for guiding various different actions, and that the very same piece of information (e.g., that Elizabeth succeeded to the throne in 1558) might be known and accessible to guide one sort of action, but not for another. Once we admit this,

----

\(^\text{11}\)See Gibbons (2001) for a nuanced discussion of the role of knowledge in explanations of intentional actions. See also Williamson (2000, pp.62-4).

\(^\text{12}\)Mellor (2012) defends a related idea; he argues, roughly, that the contents of belief are given by the success conditions for the actions that they (together with desires) cause.

\(^\text{13}\)For discussion of the idea that knowledge might be accessible for guiding some actions, but not others, see Stalnaker (1999, ch.13-14), and Elga and Rayo (Unpublished Manuscript).
we can characterize Jean’s state as one of knowing the dates of Elizabeth’s reign for certain practical purposes (e.g., answering questions about the dates of her reign in forced choice situations), but not others (e.g., just about all other situations). In other cases, we might distinguish between implicit and explicit knowledge (along the same lines as the distinction between implicit and explicit belief), and pursue the fragmentation strategy by holding that putative failures of knowledge iteration should be understood as cases involving a subject who has implicit but not explicit knowledge of some fact.

2.3 Competition

The anti-iterationist explains each case by identifying an attitude (e.g., belief) and explaining certain features of the case (e.g., John’s sexist behavior) in terms of the subject’s bearing that attitude towards a proposition (e.g., the proposition that women are less intelligent than men), while explaining other features of the case (e.g., John’s anti-sexist avowals) in terms of the subject’s not bearing that attitude towards a different, higher-order proposition (e.g., the proposition that John believes that women are less intelligent than men).

By contrast, the fragmentationist explains each case by identifying multiple attitudes, (e.g., explicit and implicit belief) and explaining certain features of the case in terms of the subject’s bearing one of the attitudes towards some proposition (e.g., the proposition that women are less intelligent than men), while explaining other features of the case in terms of the subject’s not bearing one of the other attitudes towards that very same proposition. A table may help make the contrast clearer. “B(P)” should be read as “John believes that P;” “I(P)” as “John implicitly believes that P;” and “E(P)” as “John explicitly believes that P.”

<table>
<thead>
<tr>
<th>Attitudes Appealed To</th>
<th>Sexism Explained By</th>
<th>Anti-Sexism Explained By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Iteration</td>
<td>Belief</td>
<td>B(Women are inferior)</td>
</tr>
<tr>
<td>Fragmentation</td>
<td>Implicit and Explicit Belief</td>
<td>I(Women are inferior)</td>
</tr>
</tbody>
</table>

8
At this point I want to head off a potential confusion. One might think that there isn’t really a distinction between the anti-iteration and fragmentation strategies. For instance, we might simply identify first order belief with implicit belief, and second-order belief with explicit belief. We will be anti-iterationists (since we will accept that cases like the ones I’ve been discussing involve subjects who believe things without believing that they believe them), and fragmentationists (since, if pressed further, we will explain the failure of iteration in terms of a distinction between implicit and explicit belief). This would be a grave mistake, as the distinction between first-order attitudes (i.e., attitudes whose content does not itself concern attitudes) and higher-order attitudes (i.e., attitudes whose content concerns attitudes) cross-cuts the distinction between implicit and explicit attitudes (as well as other related distinctions fragmentationists might appeal to).

Consider Stalnaker’s (1999, p. 254) example of the “shrewd but inarticulate chess player,” who is “able to access information for the purpose of choosing a move even if she is unable to access that same information for the purpose of answering a question, or giving an explanation of why she moved as she did.” It may be that explaining her play requires attributing various higher-order attitudes to her. Perhaps she implicitly believes that her opponent doesn’t believe that his king is in danger, and that this explains why she attacks the king immediately, rather than holding back for fear of a hidden defense. Or perhaps she implicitly knows that her opponent knows that she knows that a rook trade would lead to a subtly hidden mate in six for her, and that this explains why she doesn’t offer such a trade. All of this, however, is compatible with her having no explicit attitudes concerning the benefits and drawbacks of any of these moves, at least on the assumption that explicit attitudes would be manifested in dispositions to answer questions about strategy. Perhaps if asked why she made any of these moves, she would simply say: “I don’t know, it seemed best at the time.”

While the case of the shrewd but inarticulate chess player is an extreme one, there are plausibly many real life cases that are approximations to it, in which implicit, inartic-
ulable attitudes whose content itself concerns attitudes play a crucial role in explaining subjects’ behavior. If we identify higher-order attitudes with explicit attitudes, and first-order attitudes with implicit attitudes, it becomes very difficult to describe cases like these; such cases are most naturally described as ones involving implicit, higher-order attitudes. The distinction between first-order attitudes and higher-order attitudes, then, is not the distinction between implicit attitudes and explicit attitudes.

Especially once the above point is appreciated, I suspect that the idea that some kind of fragmentationist approach is called for in explaining the examples I’ve been discussing will be relatively uncontroversial—few will think that the anti-iteration strategy is sufficient unto itself. My aim is not so much to convince you to take the fragmentationist approach seriously (I hope you already do), but to convince you that it undercuts some of the motivation for the anti-iteration approach, and that once we’re fragmentationists, we have some reasons to reject the anti-iteration strategy.

3 Fragmentation and Iteration

In this section, I’ll explain how the availability of the fragmentation strategy opens up theoretical space for defending iteration principles in the face of the opacity of the mental. In the next section, I’ll try to provide a reason to think that we should occupy that theoretical space.

The basic strategy for defending iteration principles by appeal to the fragmentation strategy is as follows. Cases of apparent iteration failure should really be understood as involving equivocations. The sense of “belief” in which it’s true that John believes that

\[\text{footnote}{For instance, when Tamar Gendler (2008b) defends her theory of alief as the most fruitful way of characterizing cases like that of the implicit sexist, the alternatives she argues against (as well as the account she ultimately defends) are all, on my taxonomy, versions of the fragmentation strategy. Moreover, many writers have found versions of the fragmentation approach to be necessary to characterize phenomena that nobody would try to explain via iteration failures. For instance, Lewis (1982), Stalnaker (1984), and others argue that we need something like the fragmentation approach above just to characterize the mental states of agents with inconsistent beliefs or preferences.}\]
Iteration and Fragmentation

men are smarter than women is different from the sense of “belief” in which John fails to believe that he believes that men are smarter than women. Once we distinguish different senses of belief (e.g., implicit and explicit), it may turn that iteration principles do hold for at least some (or perhaps all) of those senses, taken individually. For instance, it may be that if one explicitly believes something, then one explicitly believes that one explicitly believes it, even if one might fail to implicitly believe it (and also fail to implicitly believe that one implicitly believes it).\(^{15}\)

Similarly, in the case of knowledge, we might hold that the sense of “knows” in which Jean knows facts about English history is different from the sense of “knows” in which Jean fails to know that she knows such facts. Depending on what sort of distinction in the notion of knowledge the fragmentationist draws, she might ultimately defend some or another iteration principle for knowledge. One way of doing this involves taking the approach mentioned earlier and holding that knowledge should be thought of as involving a three place (at least) relation between a person, a proposition, and a set of practical purposes to which the knowledge can be put. We might then defend an iteration principle for knowledge along the following lines. We might hold that if one knows (for some set of purposes \(\phi\)) that \(P\), then one knows (for purposes \(\phi\)) that one knows (for purposes \(\phi\)) that \(P\), but one might still fail to know (for some other set of purposes \(\psi\)) that \(P\). The notion of knowing for a set of purposes isn’t totally clear, but the example of Jean should help. Jean has stored information about English history, but it can only be accessed to complete certain tasks (answering questions about English history) and not others (everything else). With this distinction in hand, we can defend a version of the KK principle by holding that once we specify a set of purposes to which information might be put, knowledge will iterate—apparent failures of KK involve information that is available for certain purposes and not others; they do not involve

\(^{15}\) That explicit belief should iterate is extremely implausible if by “explicit belief” we mean something like “belief that is explicitly represented in a sentence in the language of thought.” But that’s not what I mean by “explicit belief.” See footnote 4.
first-order information (e.g., P) but not higher-order information (e.g., that one knows that P) being available for one and the same set of purposes.

Granted, it’s hard to get a pre-theoretical grip on purpose-relative higher-order knowledge, so it’s hard to have much by way of strong pre-theoretical intuitions as to whether iteration principles for it are plausible.\textsuperscript{16} Don’t worry about that for now; I don’t intend too much to hang on the particular examples of purpose-relative knowledge, or implicit versus explicit belief. The basic idea is that once we think that some sort of fragmentation strategy is called for in making sense of cases like those of Jean and John, there’s at least the possibility of holding that some (or many) of the notions into which belief and knowledge split will in fact obey iteration principles, and that apparent failures of iteration will involve cases in which different senses of belief or knowledge are failing to be distinguished. Whether that possibility should be taken seriously is a question I’ll take up in the next section.

Incidentally, the strategy I’m suggesting is in a very similar spirit to one pursued by Rose and Schaffer (Forthcoming), and Buckwalter, Rose, and Turri (Ms.). Both of those papers involve using a fragmentationist approach to belief (in particular, a distinction between occurrent and dispositional belief in the former paper, and a distinction between “thick” and “thin” belief in the latter) to defend a principle concerning belief from counterexamples, by arguing that apparent counterexamples to the principle (at least when the principle is properly framed) involve a sort of equivocation between different senses of belief. Moreover, among the counterexamples they discuss are cases like that of the unwitting historian (in their terms, the “unconfident examinee”). The principle those papers discuss, however, is not an iteration principle, but the principle that knowledge entails belief—if one knows that P, then one believes that P. It’s certainly possible to endorse a fragmentationist defense of the view that knowledge entails belief

\textsuperscript{16}What would it be to know (relative to the task of answering questions about English history) that you know (relative to the task of answering questions about English history) when Elizabeth reigned? It’s not at all obvious, though of course if the above mentioned version of the KK principle holds, it doesn’t take anything more than being in a state like Jean’s.
Iteration and Fragmentation

without endorsing a fragmentationist defense of iteration principles—after all, the view
that knowledge entails belief is a good deal more widely held than the KK or BB prin-
ciples. Still, the strategies complement each other nicely; I suspect that for those initially
sympathetic to the principle that knowledge entails belief but unsympathetic to iteration
principles, seeing how both can be threatened by the same sorts of counterexample and
defended against those counterexamples in the same sort of way may ultimately make
iteration principles seem more plausible.

On the view that combines fragmentation and iteration, it’s still true that our mental
lives are not transparent to us—the view I am exploring is not a neo-Cartesian view.
But the fact that our mental lives are not transparent to us is explained as the fact that
our overall psychological states are fragmented, or compartmentalized. For instance,
we can have implicit knowledge or beliefs concerning matters, without having explicit
knowledge or belief that we have implicit knowledge or beliefs concerning those mat-
ters. Characterizing the phenomenon of mental opacity needn’t involve holding that any
particular propositional attitude fails to obey some iteration principle, let alone holding
that no philosophically significant iteration principles are true.

4 Knowledge and Assertion

Even if I’m right that the fragmentation strategy opens up the possibility of defending
iteration principles, is there any reason to seriously consider that possibility? In this
section, I’ll present one reason to think that the answer is yes.17

Many writers have thought that some sort of principle linking knowledge and as-
sertion is true. For instance, Timothy Williamson defends the idea that the rule that

17 Of course, even if the reason I present is a good one, it has to be weighed against other reasons to
reject iteration principles. For example, Timothy Williamson’s (2000, ch. 4.5) arguments against KK
do not depend on an appeal to examples like that of the implicit sexist, or the unwitting historian, and
so are not obviously threatened by the existence of the fragmentation strategy I’ve been discussing. I do
address Williamson’s arguments against KK in other work, however; see Greco (Forthcoming).
one should assert $P$ only if one knows $P$ is constitutive of the practice of assertion; a linguistic practice must be governed by such a rule if the practice is to be one of assertion (Williamson, 2000, ch. 11). Others have defended similar but related views, sometimes concerning what it takes for assertions to be permissible (rather than in terms of constitutive rules of the practice of assertion), and sometimes focusing not on knowledge, but on related attitudes (e.g., justified beliefs that one knows).

Many writers defend views linking knowledge and assertion by appeal to the infelicity of Moore-paradoxical utterances like the following:

$$P,$$ but I don’t know that $P$.

On views like Williamson’s, the oddity of such assertions is explained by the fact that in making them one must be violating the norms of assertion: even if both conjuncts are true of some subject, they cannot both be known by that subject, since one only knows the first conjunct if the second conjunct is false (and so not known). On views linking knowledge and assertion, while Moore-paradoxical assertions may be true, they cannot be felicitous, since they must violate the central norm of assertion. As David Sosa (2009) has observed, however, in the absence of some iteration principles, this explanation does not generalize to explain the infelicity of other utterances that seem to be in the Moore-paradoxical family.

Among the cases he considers—which he calls “dubious assertions”—are the following:

1. $P$, but I don’t know whether I know that $P$.
2. $P$, but I doubt that I know that $P$.
3. $P$, but I believe that I don’t know that $P$.

If the KK principle fails—if one can know without knowing that one knows—then it’s hard to see why such utterances should be infelicitous. After all, if one knows
that \( P \), but doesn’t know that one knows that \( P \), what could be wrong with dubious assertions like the ones above? How else should one express one’s first-order knowledge, while acknowledging one’s ignorance of whether one knows? But such utterances are infelicitous in much the same way that the original Moore-paradoxical sentences are. In case this isn’t clear, consider the following dialogue:

Alice: When did Queen Elizabeth die?
Bob: She died in 1603.
Alice: How do you know that?
Bob: I didn’t say I know it.
Alice: So you’re saying you don’t know when Queen Elizabeth died?
Bob: I’m not saying that either. I’m saying she died in 1603. Maybe I know that she died in 1603, maybe I don’t. Honestly, I’ve got no idea. But you didn’t ask about what I know, did you? You just asked when she died.

While it’s true that I haven’t given any background to the case, I find it very hard to imagine any way of fleshing it out that would make Bob’s utterances seem felicitous.\(^\text{18}\) Sosa takes the infelicity of utterances like Bob’s to cast doubt on knowledge/assertion links, or at least on the strategy of defending them by appeal to the oddity of Moore-paradoxical utterances. But this is because he doesn’t take seriously the possibility that some version of the KK principle might be true. If we do take that possibility seriously, a natural explanation of the infelicity of utterances like Bob’s—one that is compatible with independently attractive knowledge/assertion links—suggests itself.

\(^\text{18}\)Matthew Benton (2013, p.356) offers an example in which a conjunction of the form “\( P \), but I don’t know whether I know that \( P \)” is supposed to sound felicitous. Speaking for myself, the conjunction in his example doesn’t strike me as any more acceptable than what Bob says in the mini-dialogue above. Benton also suggests that even if we grant that such conjunctions are infelicitous, there are strategies for explaining their infelicity that don’t involve appealing to KK. While fully engaging with Benton’s arguments would take me far afield, I’ll just report that the strategies that Benton suggests seem to me to have various drawbacks that the strategy of appealing to a version of KK does not. In particular, given the views Benton suggests, speakers will find themselves in a sort of awkward dilemma whenever they know that \( P \) without knowing that they know. In such cases, while they will be able to permissibly assert that \( P \), if their permission to assert that \( P \) is challenged, they will not be able to permissibly defend themselves. It strikes me as implausible that our conversational norms allow for such situations.
If some version of KK holds, then it’s plausible that whenever one is entitled by the norms of assertion to assert that \( P \), one is also entitled by the norms of assertion to assert that one knows that \( P \). The infelicity of Bob’s assertions could then be explained along much much the same lines as the standard explanation of the infelicity of Moore paradoxical utterances—if Bob was entitled to assert that Queen Elizabeth died in 1603 (because he knew it), then he must have been unnecessarily reticent in refusing to go on to assert that he knew this, since the sort of knowledge that entitles one to assert that \( P \) iterates, and so also entitles one to assert that one knows that \( P \).

This is compatible with a fragmentationist explanation of cases like that of the unwitting historian. In that case, while there’s some temptation to say that Jean knows that \( P \) without knowing that she knows, it doesn’t seem to be a case where the sort of state that underwrites permissible assertion fails to iterate. Given Jean’s situation, she ought not assert that she knows facts about English history. But she also ought not assert facts about English history (even if she does know them in some sense, due to her latent memories enabling her to reliably albeit unwittingly answer correctly when forced to guess). Whatever the norms of assertion are, Jean would be violating them if she made outright assertions concerning English history, just as much as she would be if she claimed to know facts about English history.\(^{19}\)

Of course, Jean does not violate any norms when she offers the correct answers to questions about English history in forced choice situations. So if my claims in the previous paragraph are correct, there must be some way to distinguish what Jean is doing when she gives answers in forced choice situations—which she is perfectly within her rights to do—from the “outright assertions” which she should not make. Luckily, I don’t think this distinction is so hard to draw. One way to get at the distinction involves thinking of the contrast between the normative consequences of “assertions” in forced choice situations, and the normative consequences of typical or outright assertions. If,

\(^{19}\)I refer to “outright assertions” to contrast them with hedged ones such as the following: “If you put a gun to my head, I suppose I’ll guess that Elizabeth died in 1603.”
outside of a forced choice situation, you ask me whether $P$ and I respond by uttering a sentence of the form “yes, $P$.” I thereby incur certain debts. In particular, if you rely on my utterance, but it turns out that $P$ is false, I owe you some sort of apology or compensation—I have invited you to trust me, and you have a legitimate complaint against me if it turns out that I don’t know what I’m talking about.\footnote{See Hinchman (2005).}

Things are quite different in forced choice situations. Imagine you are helping me prepare for a state bar exam, and are asking me questions from the end of a chapter in a review book. Let’s say you ask me about the conditions under which an agreement constitutes a legally valid contract, and I, not having studied enough, respond incorrectly. Next, however, rather than checking my answer against the key in the back of the book (as you should), you rush out of the room and over to the office of a business associate of yours, with whom you enter into an agreement that you falsely believe, on the basis of my answer to the review question, to be a legally valid contract. Months later, your business associate reneges on the agreement, and you sue him for breach of contract. You lose your suit, since the agreement was not a legally valid contract. After all this has transpired, you get angry with me for misleading you as to the conditions under which an agreement constitutes a legally valid contract. In this case, it’s quite clear that you have no legitimate complaint against me. While I may have uttered a claim of the form “an agreement constitutes a valid contract under such-and-such conditions” in your presence, this clearly was not an assertion of the usual sort; I did not invite you to trust me, and my “assertion” was clearly not governed by anything like a norm to the effect that I ought only “assert” what I know.

I claim that it is only in situations like the one just discussed—situations in which a knowledge norm is not in effect, and in which the speaker does not invite her audience to rely on her claims—that Jean, our unwitting historian, can permissibly “assert” claims about English history. I put “assert” in scare quotes because I think that such speech acts
are fruitfully distinguished from genuine assertions, which are governed by a knowledge norm, and in which a speaker invites her audience to rely on her word.21

If we take seriously a fragmentationist view concerning knowledge, we can allow that there’s some kind of knowledge/assertion link, explain the infelicity of dubious assertions like those considered by Sosa and illustrated in the Alice/Bob dialogue, and still account for apparent failures of knowledge iteration principles. We can hold that the species of knowledge that underwrites permissible assertions does iterate, while granting that it is possible to know that \( P \) in some sense without having the sort of knowledge that underwrites permissible assertion—this is what’s going on in cases like that of the unwitting historian.

This view on which there are constitutive (or perhaps merely normative) knowledge/assertion links, and on which the species of knowledge for which such links hold is one that iterates, seems to me to do a better job of explaining a wide range of phenomena concerning knowledge and assertion than do rival views. Views that deny the existence of any interesting knowledge/assertion links have a difficult time explaining the infelicity of the original Moore-paradoxical utterances, as well as other phenomena appealed to by defenders of such links.22 Views that accept knowledge/assertion links but deny iteration principles for knowledge (such as Williamson’s) have a difficult time explaining the infelicity of dubious assertions. If accepting iteration principles for knowledge left us unable to account for cases like that of the unwitting historian, then we’d face a dilemma—we could either give a satisfactory account of such cases, or instead give a satisfactory explanation of the infelicity of dubious assertions. But if we adopt the fragmentation strategy for explaining such cases, we can have our cake and eat it too. On this view, there are different species of knowledge, and iteration principles may hold for some (or even all) such species, (this is how we explain the infelicity of dubious assertions).

21One way of putting the above is that genuine assertions, but not “assertions” in forced choice or exam review situations, are “second-personal” speech acts. See Darwall (2006).
22For instance, the infelicity of “lottery assertions.” See Williamson (2000, ch. 11).
assertions) but they do not hold across species of knowledge; knowing in one sense does not entail knowing that you know in some other sense (this is how we explain the case of the unwitting historian).23

While I’ve focused on iteration principles for knowledge in this section, to the extent that we accept such principles, I suspect it will be hard to resist accepting parallel principles for belief. Many writers accept that knowledge requires belief—one knows that \( P \) only if one believes that \( P \).24 While not forced upon us, it’s tempting to think that if that’s right, then if some species of knowledge automatically iterates, then whatever species of belief is required for that species of knowledge will automatically iterate as well. Here’s one way of spelling out the idea. Reversing the traditional direction of analysis, Williamson (2000, p.46) has suggested that we might think of believing that \( P \) as, roughly, being in a state subjectively indistinguishable from knowing that \( P \) (this would explain why knowledge requires belief, as all states are subjectively indistinguishable from themselves). Suppose that’s right. I’ll try to show, given this assumption and an iteration principle for knowledge, that an iteration principle for belief will hold as well, at least for logically competent agents who accept that knowledge requires belief.

Suppose an agent believes that \( P \)—by our assumption, this amounts to her being in a state subjectively indistinguishable from knowing that \( P \). If we accept an iteration

---

23I have argued that taking seriously the fragmentationist view can let us reconcile prima facie attractive claims about (1) the infelicity of dubious assertions, (2) knowledge-assertion links, and (3) iteration principles for knowledge. But I suspect that if this strategy works, it will also help us resolve analogous puzzles that arise when we focus on belief, rather than assertion. Many writers have suggested that discussions of Moore’s paradox that focus narrowly on assertion and conversational infelicity risk failing to explain what’s wrong with believing (but not asserting) Moore-paradoxical claims. (E.g., Heal (1994).) But this section would have proceeded much the same had I focused on knowledge-belief links (e.g., Williamson’s claim that knowledge is the norm of belief), and the infelicity of dubious beliefs (e.g., believing something, while regarding it as an open question whether one’s belief is knowledge). In that case too, we face a puzzle if we take dubious beliefs to be somehow defective, and also take iteration principles for belief and knowledge to fail. Allowing that such principles (properly interpreted) may be true gives us more resources for explaining what’s defective about believing something while doubting whether one’s belief is knowledge.

24Granted, this principle has certainly been denied. See Myers-Schulz and Schwitzgebel (Forthcoming) for a recent denial. My view, however, is that at least some of the motivation to deny this principle comes from a failure to appreciate the resources of the fragmentationist. See Rose and Schaffer (Forthcoming), and Buckwalter, Rose, and Turri (Ms.).
principle for knowledge, then she must also believe that she knows that \( P \), since she will also be in a state subjectively indistinguishable from knowing that she knows that \( P \) (as our iteration principle for knowledge will entail that knowing that \( P \) and knowing that one knows that \( P \) are the very same state). But if she is logically competent, she will believe that she knows that \( P \) only if she also believes that she believes that \( P \), since as we’ve already seen, knowing that \( P \) requires believing that \( P \). We assumed that an agent believes that \( P \), and derived the conclusion (via an iteration principle for knowledge, and an assumption about the links between belief and knowledge) that she believes that she believes that \( P \). I don’t intend this as a knockdown argument that if an iteration principle for knowledge holds, then an iteration principle for belief \textit{must} hold as well. I merely intend to show that they sit nicely together, and accepting such a principle for knowledge but no parallel principle for belief is awkward, even if ultimately consistent.

In this section I’ve argued that considerations concerning assertion and Moore paradoxical sentences motivate accepting some iteration principle for knowledge, and that once we’ve done so, we face strong pressure to accept some parallel such principle for belief as well.

5 Conclusions

I’ve argued for two main claims in this essay. The first is that it’s at least not \textit{obvious} that iteration principles for knowledge and belief fail. Apparently obvious counterexamples to such principles can be interpreted along independently well-motivated fragmentationist lines, and once done so, there is at least the possibility that some interesting iteration principles for belief and/or knowledge might hold. The second is that there is some reason to think that this possibility is realized—that some interesting iteration principles for belief and knowledge \textit{do} hold. The reasons I gave to think this are that (1) iteration
principles for knowledge let us make sense of some otherwise puzzling data concerning warranted assertion, and (2) if iteration principles for knowledge hold, iteration principles for belief probably do too. My hope is that, taken together, these considerations motivate taking iteration principles for knowledge and belief more seriously than they are often taken in the contemporary philosophical literature.

**Bibliography**


Myers-Schulz, Blake, and Eric Schwitzgebel. Forthcoming. Knowing that P Without Believing that P. *Noûs*.


———. 2010. Acting Contrary to Our Professed Beliefs or the Gulf Between Occurrent Judgment and Dispositional Belief. *Pacific Philosophical Quarterly* 91:531–553.


