Lewis on iterated knowledge

Citation: Salow, Bernhard. "Lewis on Iterated Knowledge." Philosophical Studies 173, no. 6 (September 19, 2015): 1571–1590.

As Published: http://dx.doi.org/10.1007/s11098-015-0568-0

Publisher: Springer Netherlands

Persistent URL: http://hdl.handle.net/1721.1/106220

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Philos Stud (2016) 173:1571–1590
DOI 10.1007/s11098-015-0568-0

Lewis on iterated knowledge

Bernhard Salow[1,2]

Published online: 19 September 2015

Abstract The status of the knowledge iteration principles in the account provided by Lewis in "Elusive Knowledge" is disputed. By distinguishing carefully between what in the account describes the contribution of the attributor's context and what describes the contribution of the subject's situation, we can resolve this dispute in favour of Holliday's (2015) claim that the iteration principles are rendered invalid. However, that is not the end of the story. For Lewis's account still predicts that counterexamples to the negative iteration principle ($\neg Kp \rightarrow K\neg Kp$) come out as elusive: such counterexamples can occur only in possibilities which the attributors of knowledge are ignoring. This consequence is more defensible than it might look at first sight.

Keywords Epistemic logic · Epistemic contextualism · David Lewis

One of the most influential versions of epistemic contextualism is the one Lewis develops in "Elusive Knowledge".[1] Despite its influence, this account is not always well understood. One place where matters are particularly unclear is the status of knowledge iteration principles in Lewis's account. Several authors [including

[1] Lewis (1996). Blome-Tillman (2009, 2012, 2014) and Ichikawa (2011a, b, 2013) are recent defenders of (modified) versions of Lewis's account; commentators that pay close attention to Lewis's account in particular include Cohen (1998), Vogel (1999), Williams (2001), Williamson (2001), Schaffer (2004), Hawthorne (2004), and Douven (2005).

✉ Bernhard Salow
bs416@cam.ac.uk

[1] MIT, Cambridge, MA, USA

[2] Trinity College Cambridge, Trinity Street, Cambridge CB2 1TQ, UK

Williamson (2001, 2009), Holton (2003), and Greco (2014), who all trace the claim to Lloyd Humberstone] maintain that Lewis's account validates an S5 epistemic logic, which would mean that it is committed to implausibly strong iteration principles for knowledge; by contrast, Holliday (2015) maintains that the knowledge iteration principles are invalid in Lewis's system.

By distinguishing carefully between what is contributed by the conversational context of the agents attributing knowledge and what is contributed by the situation of the subject to whom knowledge is attributed, we can resolve this dispute in Holliday's favour: Lewis's system allows counterexamples to both the *KK* principle (that whenever someone knows something, they know that they know it) and what I will call the $K\neg K$ principle (that whenever someone doesn't know something, they know that they don't know it). However, we can also see that this is not the end of the story: counterexamples to the $K\neg K$ principle can only occur at worlds that the attributors of knowledge are ignoring. (No analogous result holds for the *KK* principle.) On the face of it, this surprising consequence of Lewis's account looks almost as implausible as the claim that the $K\neg K$ principle is valid. However, I will argue that there are ways of rendering the consequence acceptable.[2] Throughout the paper, I will try to draw more general lessons about the relationship between epistemic contextualism and the knowledge iteration principle, explaining why their interaction is both subtle and fruitful.

1 Lewis, formalized

Discussions of epistemic logic standardly proceed in a possible worlds framework, in which an agent X is said to know p at w if and only if every world accessible from w (under the accessibility relation associated with X) is a p-world. Lewis seems to proceed similarly. Consider, for example, his well-known summary of the account:

> X knows that P iff X's evidence eliminates every possibility in which not-P – Psst! – except for those possibilities that we [attributors] are properly ignoring. (1996, p.554)

This seems to translate quite straightforwardly into the traditional framework: we simply say that a world is accessible if it is neither properly ignored nor ruled out by X's evidence.[3] One would thus expect it to be relatively straightforward to distil a logic from Lewis's account. However, as we will see shortly, there are some pitfalls here to be navigated.

[2] I actually think that, in addition to it not being obviously false, there are positive reasons to want something like the Lewisian treatment of $K\neg K$ to be correct. For, as I hope to show in future work, it allows us to solve hard problems for the (thoroughly non-Lewisian) thesis, defended by Williamson (2000), that one's evidence consists of all and only the claims that one knows.

[3] This is not quite right as an interpretation of Lewis, since he uses 'possibilities' to mean something slightly different from possible worlds (1996, p. 552). To keep the formalization of his account manageable, I ignore that complication here.

To proceed with the approach just sketched, it is natural to look to 'frames' that consist of a set of worlds W, together with a specification of how Lewis's primitives behave at the various worlds; we can then see what happens when we define accessibility in terms of these primitives. Deciding on how to represent the primitives, however, requires some care. For Lewis's theory is, above all, a *contextualist* theory. This means that whether an attribution of knowledge correctly describes a situation depends on both features of the situation described and features of the context from which the attribution was made. However, only the features of the situation (the 'world of evaluation') will vary as we consider what an agent knows in different possible worlds; we are interested in the logic of 'knows' within a single context, and so whatever is supplied by context will remain fixed. Our frames thus need to represent the features of the situation as world-relative, but can represent the contributions of the context absolutely. Whether something is a feature of the situation described or of the context of ascription thus matters greatly to how our frames should represent it.

1.1 A natural mistake

How does this distinction between features of the context and features of the situation described apply to Lewis's account? The above summary of the account suggests that the correctness of knowledge attributions depends on two components: (i) what evidence the subject has, which we can represent by a relation E so that wEv iff v is compatible with the evidence X has in w, and (ii) a set S of possibilities that are not being properly ignored. The first of these is clearly a feature of the situation described; the second looks, at least at first sight, like a feature of the context—that's why it seemed natural to represent it absolutely, i.e. as a set rather than a function from possibilities to the set of worlds ignored at that possibility.

We will see shortly that this approach isn't textually plausible. Nonetheless, it is worth briefly exploring it, since it helps explain the appeal of the idea that Lewis's account vindicates an S5 logic. For the current proposal would see Lewis vindicate the iteration principles. Lewis views a subject's 'evidence' as her total phenomenal state, so that wEv if and only if the subject is in the same total phenomenal state in w and v; this makes E an equivalence relation. The obvious definition of R_K, the accessibility relation for our subject's knowledge, holds that wR_Kv if and only if wEv and $v \in S$, so that an agent knows p only if her evidence eliminates all the unignored p-worlds. And on this definition, R_K will be transitive and Euclidean.[4] We thus validate both the KK principle ($Kp \rightarrow KKp$) and the $K\neg K$ principle ($\neg Kp \rightarrow K\neg Kp$).

However, we don't quite vindicate a full S5 logic. The missing principle is the most basic one: that what is known must be true. For note that no world outside of S will be accessible to *any* world under R_K, not even to itself. R_K thus isn't reflexive, and so we do not validate the T principle ($Kp \rightarrow p$); in worlds outside S, people can

[4] To see that it's transitive, note that from xR_Ky and yR_Kz it follows that $z \in S$ and xEy and yEz. So $z \in S$ and xEz (since E is transitive), and so xR_Kz. To see that it's Euclidean, note that if xR_Ky and xR_Kz, then $z \in S$ and xEy and xEz. So $z \in S$ and yEz (since E is euclidean), and hence yR_Kz.

know things that aren't true there. This is a clear sign that something has gone wrong; the factivity of knowledge is not only epistemologically non-negotiable, but also a feature Lewis (1996, p. 554) specifically intended his account to vindicate.

We run into this problem with factivity because our logic is sensitive to how knowledge behaves in possibilities that are properly ignored. Since Lewis (1996, pp. 555–559) explains that such possibilities are neither actual nor salient, this sensitivity might seem excessive.[5] It can be avoided by redefining validity as truth at every not-properly-ignored-world in every model;[6] this would, in fact, allow us to vindicate a full S5 logic.[7] However, R_K won't be reflexive even on this revised approach, suggesting that the original problem has been hidden rather than solved. One way to bring this out is by considering what happens when we introduce other modal operators. For suppose we introduce an operator $\Box$ for metaphysical necessity. It seems plausible that some worlds outside S are metaphysically possible with respect to some worlds in S in at least some models. But then $\Box(Kp \rightarrow p)$ will not be a principle of the combined logic of knowledge and metaphysical necessity. This strikes me as no less serious than the original problem of allowing for *actual* factivity failures.

Our simple-minded approach, whilst hospitable to the iteration principles, thus has consequences which are both extremely unattractive and difficult to eliminate. The culprit seems to be the fact that the set of relevant possibilities that need to be eliminated is treated as something entirely supplied by context. For this means that the relevant possibilities cannot vary when we evaluate a knowledge attribution at different worlds; but this, in turn, implies that some possibilities aren't relevant to themselves, so that agents at those possibilities can eliminate all relevant ¬p-worlds (and thus know p) even though p is false. We thus fail to capture the factivity of knowledge.

1.2 Doing better

Fortunately, Lewis's discussion does not commit him to such an inadequate account. It is true that which possibilities are being *ignored* is settled by the context.

[5] In an unpublished manuscript, Julien Dutant identifies a "rigid interpretation" of Lewis's semantics, shows how it conflicts with the factivity of knowledge, and then considers a response analogous to this one. He observes that, even once we acknowledge such a response, the interpretation still predicts that the sentence 'someone could have known something false' could be true, which is the inspiration for the objection I offer below.

[6] A variant of this is more familiar in modal logic. We could move to 'model structures' $<W, E, S, w>$ which designate world $w \in W$ as the actual world. Since the actual world is never properly ignored, we would then want to impose the structural requirement that $w \in S$. When working with model structures instead of frames, it's also natural to redefine validity as truth at the designated world of every model. The resulting system is very similar to the one discussed in the main text; in particular, it validates S5 for essentially the same reason.

[7] Why? Let us say that v can be reached from w if there are worlds $u_1, \ldots u_n$ such that $wR_Ku_1, u_1R_Ku_2, \ldots u_nR_Kv$. Then truth in a model depends only on what happens in worlds that are either in S or can be reached from a world in S. Moreover, the definition of R_K ensures that all such worlds are themselves in S. Finally, R_K is an equivalence relation when restricted to S (though not outside it). Together, these facts ensure that we validate an S5 logic.

But Lewis defines knowledge in terms of *proper ignoring*, and it is far from obvious that it is the context which settles which ignorings are proper. In fact, when Lewis, in introducing the 'Rule of Actuality', explicitly discusses this issue, he asserts that propriety is (at least partially) determined by the world of evaluation:

> The possibility that actually obtains is never properly ignored. ...Whose actuality? Ours, when we ascribe knowledge or ignorance to others? Or the subject's? ...[T]he right answer is that it is the subject's actuality, not the ascriber's, that never can be properly ignored. (1996, p.554f)

"The subject's actuality" seems to be the world of evaluation;[8] so what can be properly ignored depends on what the world of evaluation is. We therefore need to reinterpret *S* to represent only what is contributed by the context. Plausibly, that is the set of worlds that are not *in fact* ignored by the attributors; this set will thus leave out worlds that are ignored but only improperly so. This is how '*S*' will be interpreted from here on in. In addition to this reinterpretation, we need to enrich our frames to represent directly all the features of the worlds that constrain what can be properly ignored relative to each of them.

What features are these? Lewis articulates the limits of proper ignoring by appeal to the Rules of Actuality, Belief, and Resemblance.[9] The information relevant to the Rule of Actuality is trivially represented in the frame, since every world is actual relative to itself. So the first addition is the notion of the subject's beliefs,[10] which we will need to implement the 'Rule of Belief' stating that "a possibility that the subject believes to obtain is not properly ignored" (1996, p. 555f). Following the standard formalization of belief, we can represent this by an accessibility relation R_B on worlds, where wR_Bv is understood as 'v is consistent with all of X's beliefs in w.'

The second addition required to constrain proper ignoring is that of relevant similarity, which we will need to implement the 'Rule of Resemblance':

> Suppose one possibility saliently resembles another. Then if one of them may not be properly ignored [in virtue of rules other than this rule], neither may the other. (1996, p.556)

[8] In the unpublished manuscript mentioned in footnote 5, Dutant argues that "the subject's actuality" might be construed instead as the (potentially counterfactual) world on which the conversation is focused; this would allow for context alone to determine propriety. I agree that such a reading is just about possible. But since it would leave us with the unsatisfactory account discussed in Sect. 1.1, and the context of the passage strongly suggests that Lewis is trying to rule out this variant account, I think it safe to assume that this is not how Lewis intended these remarks.

[9] What is the role of the 'permissive' rules, such as the Rules of Reliability, Method, and Conservatism (1996, pp. 558–559)? I have to confess to finding these rather puzzling. As I understand Lewis, any world that isn't being attended to is automatically ignored, and thus properly ignored if no 'restrictive' rule prevents this from happening. But then what role could there be for the permissive rules to play? One hypothesis is that they aren't *rules* about the propriety of ignoring at all, but are rather *empirical generalizations* about what kind of worlds are in fact ignored in ordinary contexts. Another thought, suggested to me by Bob Stalnaker, is that they function as constraints on what 'restrictive' rules Lewis would be willing to add to his account: they had better be consistent with it being proper, except in very specific circumstances, to ignore worlds in which our faculties and methods are unreliable.

[10] Or what the agent should believe, but I will set that complication aside.

Since it is context, rather than the world of evaluation, which determines which respects of similarity are salient, this can be represented by a binary relation 'C' (for 'closeness') with wCv read as 'w is close to/relevantly resembles v'. Crucially, we may not assume that C is transitive, since Lewis is at pains to distinguish between worlds resembling each other and worlds being connected by a chain of resembling worlds.

A full Lewisian frame is thus a 5-tuple $< W, E, S, R_B, C >$; such a frame does better at representing the information needed for an adequate formalization. For we can now define proper ignoring in a way which ensures that different possibilities are properly ignored relative to different worlds of evaluation. According to Lewis, the worlds not properly ignored relative to w are (i) w itself (to respect the Rule of Actuality), (ii) the worlds consistent with X's beliefs at w (to respect the Rule of Belief)[11] (iii) the salient worlds S (to respect the Rule of Attention), and (iv) any world close to those mentioned in (i)–(iii) (to respect the Rule of Resemblance).

We formalize this thought by defining an 'alternatives' function $A : W \to \mathcal{P}(W)$, which takes each world w to its alternatives, i.e. the possibilities not properly ignored relative to w. We first implement (i)–(iii) to define an impoverished function A^-, and then 'fill it in' to define an A which also respects (iv):

$$A^-(w) =_{def} \{w\} \cup \{v : wR_Bv\} \cup S$$
$$A(w) =_{def} \{u : \exists v \in A^-(w) \text{s.t.uCv}\}$$

We then use A together with E to define the accessibility relation for knowledge R_K in the natural way: for all worlds u and v,

$$uR_Kv \text{ if and only if } uEv \quad \text{and} \quad v \in \text{A(u)}.$$

The resulting system is essentially a special case of Holliday's (2015) formalization of Lewis.[12] Simplifying slightly, Holliday's frames are, in our notation, the triples $< W, E, A >$; the rule of actuality is built in by requiring that $w \in A(w)$. Our models are less general, because defining A in terms of S, R_B, and C imposes additional constraints.[13] Formally, this lesser generality will generate the surprising new result discussed in Sect. 2; and at an informal level, I hope that building up A in the way I have done (and making explicit the rival approach discussed in Sect. 1.1) helps clarify why this really is the right way to formalize Lewis.

[11] Given the above statement of the rule of belief, one might worry that this is much too strong: there, Lewis seems to say that a possibility believed to obtain isn't properly ignored, not that a possibility not believed not to obtain isn't properly ignored. But Lewis later clarifies that what he really means is that "a possibility may not be properly ignored if the subject gives it [...] a degree of belief that is sufficiently high," (1996, p. 556) and context makes clear that "sufficiently high" is usually far below .5 (as it has to be, since otherwise almost no reasonable agent will have a "sufficiently high" degree of belief in any single possibility). So 'the worlds consistent with X's beliefs' is a better approximation of Lewis's rule than 'the world (if there is one) uniquely consistent with X's beliefs.' It is nonetheless merely an approximation of what Lewis was after; one consequence of this choice will be that, contrary to Lewis's (1996, p. 556) explicit intentions, our formalization will not allow for knowledge without belief in cases like that of the reliable but underconfident examinee.

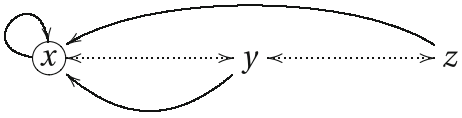
[12] Thanks to an anonymous referee for extremely helpful discussion on this point.

[13] Though Holliday (2013) considers imposing the constraint corresponding to the rule of belief.

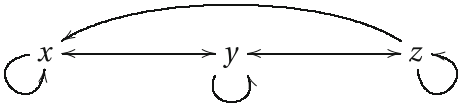
What, then, are the formal features of this system? Unlike the first attempt, it has no trouble accounting for the factivity of knowledge. And the way this account implements the rule of belief means that we *almost* validate the principle that everything known is believed.

(But only almost: as Ichikawa (2011a, p. 386) points out, Lewis's account implies that if a proposition p is entailed by an agent's evidence, she automatically knows p, regardless of whether she believes it. In fact, she can know p whilst believing its negation: while $A(w)$ will contain the $\neg p$-worlds compatible with the subject's beliefs, those will then be ruled out by her evidence, and thus no longer accessible under R_K. This is a bad result even if, like Lewis (1996, p. 556), we think that the connection between knowledge and belief is rather loose. But it seems to me an unavoidable feature of Lewis's thought that we know everything that is true in all the possibilities compatible with our evidence. Of course, we can reject this thought to preserve the link between belief and knowledge, e.g. by replacing E with $E \cup R_B$ in the definition of R_K. Alternatively, we can hold onto the Lewisian thought (and hence the original definition), and simply admit that, in so doing, we are restricting our attention to somewhat idealized agents who believe everything their evidence entails.[14] Since our models, as is standard, already build in a variety of similar idealizations, such as the assumption that agents always know and believe logical consequences of what they know and believe, I will opt for this simpler approach.)

As Holliday points out, however, this system does not provide a hospitable environment for the iteration principles. For consider the three world model on which (a) x is the only salient world, (b) x is the only world compatible with our agent's beliefs in any of the three worlds, (c) x resembles y and y resembles z but x does not resemble z, and (d) our agent's evidence at each of the worlds is compatible with her inhabiting any of them. These facts can be visually represented as follows, with continuous lines standing for R_B, dotted lines standing for C, and worlds in S occurring inside the circle (information about E, being trivial, is omitted):



Then, under R_K, x will access only itself and y, whilst y and z both access all three worlds. Letting continuous lines now stand for R_K, we can represent this as follows:



Now let p be a claim that is true in x and y, but false in z. Since p is true at both x and y, Kp will be true at x; but since p is false at z, Kp will be false at y. So KKp will fail at x even though Kp was true there, and so we have a counterexample to the KK

[14] That is, we require that, in all our models, wR_Bv entails wEv. Cf Holliday (2013).

principle. The same model also provides a counterexample to the $K\neg K$ principle, since $\neg Kp$ will be true at y and z but $K\neg Kp$ will be false at both.[15]

It is worth noting that these results are independently attractive. The $K\neg K$ principle in particular seems clearly invalid: someone who reasonably believes something false fails to know but doesn't know (and needn't be in a position to know) that he so fails. And Lewis seems to be trying to do justice to this thought. Thus he (1996, p. 554) describes his account as "'externalist'—the subject himself may not be able to tell what is properly ignored." But this is inconsistent with the iteration principles, since the subject could use his knowledge of what he knows to work back to what is being properly ignored.[16]

2 Elusive $K\neg K$

Getting clear on whether the iteration principles are valid in Lewis's system matters if we are interested in what Lewis thought. It also matters if we want to appeal to their status in Lewis's system either to bolster the plausibility of a principle [as Greco (2014) does in appealing to the claim that Lewis's system vindicates the *KK* principle] or to criticize Lewis's account [as Williamson (2001) does, in saying that Lewis's system vindicates the $K\neg K$ principle]. But there is also a more surprising reason for noting that Lewis's account does not, in fact, validate the iteration principles: the $K\neg K$ principle turns out to have a different but still unusual status in this system.

For suppose we may assume that, on any interpretation of 'knows', the agent in question always knows what her beliefs are.[17] Then we can show that the $K\neg K$ principle has no counterexamples in any of the worlds that are in fact salient to the attributors:[18]

Elusive K¬K. *For any $w \in S$ and proposition p, $\neg Kp \rightarrow K\neg Kp$ is true at w.*

Proof Suppose that $\neg Kp$ is true at $w \in S$. Then there must be some v at which p is false such that wR_Kv, which implies $v \in A(w)$. Now let u be any world such that

[15] It's worth noting that, while the counterexample to *KK* relies on the intransitivity of C, the counterexample to $K\neg K$ does not. For we can simply drop y from the example, rendering C irrelevant; the resulting model will validate *KK*, but $K\neg K$ will still fail at z.

[16] Moreover, Bob Stalnaker tells me that, while Lewis initially thought that his theory should satisfy an S5 logic, he became convinced of the implausibility of the $K\neg K$ principle whilst presenting early versions of "Elusive Knowledge". This change of heart coincided with the introduction of his extended discussion of the Rule of Actuality, and we saw earlier that this is the crucial passage warning us against the iteration-friendly formalization of Sect. 1.1.

[17] Formally: $\forall x\forall y(xR_Ky \rightarrow \forall z(xR_Bz \leftrightarrow yR_Bz))$. Given Lewis's account, this claim can be true on every interpretation of 'knows' only if a difference in beliefs always makes for a difference in phenomenal state; Smithies (2014) develops a notion of 'phenomenal state' designed to have this feature, and argues that one's justification supervenes on what phenomenal state (in this sense) one is in, so this might be a way of incorporating the introspection assumption into a broadly Lewisian account. It's also worth noting that, even if we deny that agents in general always know what they believe, it is still interesting and surprising that the Lewisian account predicts our result to hold of those that do.

[18] Recall that the actual world may not be salient to the attributors; the result thus doesn't entail that the $K\neg K$ principle will be true.

wR_Ku. We will begin by showing that $v \in A(u)$: we argue that, since $v \in A(w)$, one of four conditions must hold, and that any of these are sufficient to ensure that $v \in A(u)$.

(i) $v \in A^-(w)$ because $v = w$. Since $w \in S$, this ensures that $v \in A^-(u)$.
(ii) $v \in A^-(w)$ because wR_Bv. Since wR_Ku, it follows from our introspection assumption that uR_Bv also. So $v \in A^-(u)$.
(iii) $v \in A^-(w)$ because $v \in S$. Then $v \in A^-(u)$ also.
(iv) $v \in A(w)$ but $v \notin A^-(w)$. Then there must be an $x \in A^-(w)$ such that vCx. But, then x must meet one of conditions (i)–(iii), and so $x \in A^-(u)$. So $v \in A(u)$ also.

So $v \in A(u)$. Since wR_Kv and wR_Ku, we have wEv and wEu, which implies uEv since E is an equivalence relation. So uR_Kv. So Kp is false at u also. Since u was an arbitrary world satisfying wR_Ku, it follows that $K\neg Kp$ is true at w. Since w was an arbitrary member of S and p an arbitrary proposition, this establishes the result. □

This result is extremely surprising; it seems to say that we can never attend to agents who are unaware of the fact that they fail to know something, that counterexamples to the $K\neg K$ principle are elusive. That sounds obviously false: the $K\neg K$ principle isn't just invalid, but subject to clear counterexamples which we have no trouble thinking about. I will argue shortly that things may not be quite so straightforward; but first, we should attempt to understand why Lewis's account has this kind of consequence.

Counterexamples to $K\neg K$ seem easy to come by: just pick an agent who has a belief which, while it looks good 'from the inside', falls short of knowledge because of an uncooperative environment. To have a concrete example, consider someone whose belief that the wall in front of her is red falls short of knowledge because the lighting is unreliable. Since the belief 'looks good from the inside' our agent must have evidence that rules out the kind of $\neg p$-possibilities that any would-be knower has to rule out, such as possibilities in which the wall is and looks yellow. Since, nonetheless, her belief doesn't amount to knowledge, there must be some other, more idiosyncratic, $\neg p$-possibilities, that are relevant to her because her actual environment is uncooperative, and which her evidence doesn't eliminate; in our example, these would be possibilities in which the wall is white but the lighting is misleading. (These possibilities might be either actual, or relevantly similar to the one that is actual; it doesn't matter which.) But now suppose, for *reductio*, that our agent's actual circumstances are salient. Then, according to Lewis, we will use 'knowledge' in such a way that *anyone* has to rule out these supposedly 'idiosyncratic' possibilities to count as knowing by the standards of the current conversation; for, by the rules of attention and resemblance, any would-be knower has to rule every possibility which is either salient or relevantly similar to one that is salient. And so the error possibilities cannot be idiosyncratic to our subject after all, contradicting our assumption. So, if a case like that of misleading lighting is salient, it cannot, after all, be a case in which our agent fails to know without knowing that she fails.

What is generating the result is thus the feature of contextualism that was also responsible for the weird consequences of the naive formalization in Sect. 1.1: that something contributed by the conversational context (now: the set S of salient possibilities) is independent of the world of evaluation. This means that, once a possibility (such as the possibility of misleading lighting) is in S, *any* would-be knower has to eliminate it, regardless of what his or her world is like. Since rational $K\neg K$ failures intuitively arise from error possibilities that are specific to the subject who fails to know, this has the surprising consequence that the error possibilities generating the counterexample to $K\neg K$ must not themselves be salient (or relevantly similar to possibilities that are salient.) And that's just another way of saying that counterexamples to $K\neg K$ occur only in possibilities that aren't in S.

Interestingly, we get no analogue of **Elusive K¬K** for the KK-principle; in fact, the model described in Sect. 1.2 already showed that KK can fail even at a salient possibility. This reveals quite how different the counterexamples to these two principles are on the Lewisian treatment. KK fails because C isn't transitive: someone's evidence can rule out all the worlds resembling the actual one, without thereby ruling out all the worlds resembling some world that resembles the actual one. By contrast, $K\neg K$ fails because agents sometimes reasonably think they can ignore possibilities which, because of facts specific to their actual situation, turn out to be relevant. Making the actual world salient, and thereby forcing it to be relevant no matter what, prevents the second of these but leaves the first untouched.

Now that we understand a little better why Lewis's account entails **Elusive K¬K**, we can turn to examine whether this is problematic. At first sight, it seems terrible. We *can* describe clear and concrete counterexamples to the $K\neg K$ principle; and **Elusive K¬K** seems to predict that we can't. But matters are not quite so straightforward. In Sects. 2.1 and 2.2 I will describe two ways in which Lewisians can respond. The first yields no ground at all, and argues that we can still do justice to the clear examples; the second is more conciliatory, taking **Elusive K¬K** to motivate a different conception of what it is to 'ignore' a possibility.[19] Each, I think, has promise; so the fact that Lewis's account entails **Elusive K¬K** doesn't refute that account.

2.1 The hard-nosed response

The Lewisian who wants to yield no ground has his work cut out for him. There are two natural ways of understanding Lewis's talk of 'ignoring'; and the prediction that $K\neg K$ failures happen only in ignored possibilities looks implausible on either one. The first way of understanding 'ignore' is more prominent in Lewis's discussion: a possibility isn't ignored if it is psychologically salient, if we are thinking or talking about it. But sometimes Lewis instead writes of which possibilities are compatible with our presuppositions; or, as I shall put it, which possibilities we take seriously.

[19] A more radically conciliatory response would give up on the thought that worlds that aren't ignored always need to be eliminated. To preserve any of the Lewisian spirit, we would then have to offer a different account of the role S plays in defining A^- or A.

And, as Blome-Tillman (2009, 2014) emphasizes, what is salient and what is taken seriously need not coincide. I tell you that the wall in the seminar room is red. You raise the worry that the lighting might have been misleading. When I discover that you have no special reason to think so, I tell you to stop being so tedious. Even though you have made the possibility of misleading lighting salient, I refuse to take it seriously and continue to presuppose that it does not obtain.

There are a number of independent reasons why understanding 'ignore' in terms of presuppositions is more attractive than understanding it in terms of salience.[20] To these we can add that this way of understanding 'ignore' helps reconcile **Elusive K¬K** with the possibility of clear counterexamples to $K\neg K$ when these counterexamples are thought of hypothetically. I claim to know that the wall is red. I agree that it's not impossible for the lighting to be unreliable and that, if it had been unreliable, my belief that the wall is red would have fallen short of knowledge without my knowing that it did. Perhaps I even agree that if, contra everything I believe, the lighting was unreliable this time, my actual belief falls short of knowledge even though I do not know that it does. But I continue to presuppose that the antecedents of these conditionals are false. So that speech is no counterexample to **Elusive K¬K** (when 'ignore' is understood as 'don't take seriously') since the possibility in which I locate the counterexamples to $K\neg K$, being inconsistent with my presuppositions, isn't in S.

However, there are also clear counterexamples to $K\neg K$ that needn't be described hypothetically; these are most naturally described as cases in which the subject differs from the attributors. My friend Soraya says that the wall in the other room is red. But we know that the lighting in that room is unreliable. So it seems that we can rightly judge that Soraya fails to know but doesn't know that she so fails. After all, we know that (i) her belief, being formed in poor conditions, can't be knowledge, and (ii) she doesn't (and has no reason to) suspect, much less believe, that she doesn't know. In fact, she seems to think that she does know – otherwise she wouldn't have felt so confident in telling me the color of the wall. But her case is both salient to us and compatible with our presuppositions, since we believe it to be actual. Doesn't that refute **Elusive K¬K**?

Perhaps not. It does seem clear that we can judge that Soraya doesn't know but doesn't know she doesn't know. But it isn't clear that 'know' is interpreted relative to the possibilities salient to *us* throughout that judgement; and if it's not, the possibility of this judgement needn't conflict with the Lewisian result. For **Elusive K¬K** entails only that knowledge-relative-to-S behaves in line with $K\neg K$ throughout S; it makes no predictions about the behaviour of knowledge-relative-to-S', nor about principles which mix different interpretations of 'know'.

On Lewis's account, which relation is picked out by 'knows' depends on what possibilities are salient to, or taken seriously by, the speakers. In our example, Soraya is not, I assume, taking seriously the possibility that the lighting is odd—if she did take that possibility seriously, she wouldn't take herself to know that the wall is red. There are thus two senses of 'know' in play in the situation; since it

[20] See Hawthorne (2004) and Blome-Tillman (2009, 2014) for discussion.

takes more to know in our sense than in Soraya's, I will use 'know$_{hi}$' to name the relation 'know' refers to when the contextual parameter is filled with the possibilities *we* take seriously, and 'know$_{lo}$' for the relation it refers to when the contextual parameter is filled with the possibilities *Soraya* takes seriously.[21] **Elusive K¬K** then entails only that if Soraya doesn't know$_{hi}$, she knows$_{hi}$ that she doesn't know$_{hi}$; and I will show that the Lewisian has principled reason to deny that this conflicts with our intuitive judgement that Soraya fails to know without knowing that she does.

Let us begin by looking at what Soraya knows or believes about what she knows$_{lo}$ and knows$_{hi}$ about the wall. It seems pretty clear that she believes that she knows$_{lo}$ that the wall is red. That belief is why Soraya is inclined to say that the wall is red, and that she knows this, when talking with people that she takes to share her epistemic standards.[22] The belief is false, since the fact that the lighting is actually unreliable means that Soraya has to rule out possibilities with misleading lighting even to know$_{lo}$. In spite of being false, however, the belief is perfectly reasonable: had the environment been more cooperative, Soraya wouldn't have had to rule out possibilities with misleading lighting to know$_{lo}$; and Soraya has no reason to suspect the lack of cooperation.

A belief that she knows$_{hi}$ that the wall is red is quite a different matter. After all, it's clear from the meaning of 'know$_{hi}$' that one doesn't know$_{hi}$ that the wall is red unless one can rule out the possibility of misleading lighting, no matter how dissimilar such worlds are from the actual situation. And Soraya can tell that she is in no position to rule out possibilities with misleading lighting. A belief that she knows$_{hi}$ that the wall is red would thus be a highly unreasonable belief for her to have; and since Soraya (like all subjects satisfying the idealizations implicit in our reconstruction of Lewis) is rational, she doesn't have such unreasonable beliefs.

This last point can be strengthened. Since it is clear to Soraya that she can't rule out possibilities in which the lighting is misleading, she is well aware that she *doesn't* know$_{hi}$ that the wall is red. Or, at least, she is aware of this if she has ever thought about what she knows$_{hi}$ at all; and, in keeping with our Lewisian idealizations, we shall assume that she has.[23] So we have that Soraya believes that she doesn't know$_{hi}$ that the wall is red, and that this belief (being based purely on

[21] This may be a little misleading, since, as I argue later, it's not very intuitive to think that our *standards* for knowledge are higher than Soraya's, which is what the notation suggests.

[22] In saying this, we can be neutral on whether this is the belief expressed by her utterance, as it might not be if her conversational partners do not, in fact, take the same things seriously as she does. See DeRose (2004) for discussion.

[23] One might worry that this is in tension with our stipulation that Soraya is ignoring the possibility of misleading lighting; for if she is, how could she even articulate what it takes to know$_{hi}$? If 'ignoring' is understood in terms of presuppositions, the worry is easily dissolved, since Soraya can think about the possibilities of misleading lighting when determining what she knows$_{hi}$ without taking them seriously; that is, presumably, what most contextualists do when they agree that they know very little by sceptical standards. If 'ignoring' is understood in terms of salience, the worry has more bite; but we can still imagine that Soraya reflected earlier about what she would know$_{hi}$ in various situation, and that those earlier beliefs, which do not feature amongst her conscious thoughts when she is looking at the wall, are sufficient to constitute a belief that she does not know$_{hi}$ that the wall is red.

introspection into her evidence and a priori reasoning) amounts to knowledge in every relevant sense.

(At this point, it might start to seem as though our idealizing assumption—that Soraya's beliefs are consistent and include everything entailed by her evidence—is pulling a lot of weight. But it would, I think, be a mistake to blame the surprising **Elusive K¬K** on the strength of these idealizations. For we also want to say that Soraya's case is one in which she fails to know but is in no position to know that she so fails. Yet, even if Soraya were less ideal than we have been assuming, the above considerations would still suggest that she is at least in a position to know that she doesn't know_{hi} that the wall is red.)

Here, then, are the natural predictions of the Lewisian account:

(a) Soraya doesn't know_{lo} that the wall is red.
(b) Soraya believes that she knows_{lo} that the wall is red.
(c) Soraya doesn't believe/know_{lo}/know_{hi} that she doesn't know_{lo} that the wall is red.
(d) Soraya doesn't know_{hi} that the wall is red.
(e) Soraya does not believe that she knows_{hi} that the wall is red.
(f) Soraya believes/knows_{lo}/knows_{hi} that she doesn't know_{hi} that the wall is red.

Do these allow us to recover the obvious natural language judgements, such as 'Soraya thinks she knows that the wall is red'? They do, if we combine them with a surprising claim about how the context-sensitivity of 'know' is resolved when the word occurs embedded in an attitude ascription. For in order to get the obvious judgement to come out true, we have to say that 'know', when embedded under 'Soraya thinks that', means know_{lo} – even when said by us, with our high standards. More generally, we have to say that when 'know' is embedded in an attitude ascription, the contextual parameter relative to which it is interpreted is supplied not by the context of utterance, but by something like the private context of the subject of the attitude ascription.

I will revisit the plausibility of this linguistic claim shortly. For now, we should simply note that, if it is correct, it also reconciles our example with **Elusive K¬K**. It is natural for us to judge that, even though Soraya doesn't know that the wall is red, she doesn't know that she doesn't know this; this seems to be in tension with **Elusive K¬K** because we are attending to and taking seriously Soraya's situation. However, if the above linguistic claim is correct, the tension is illusory. For our judgement then amounts to the observation that Soraya doesn't know_{hi} that she doesn't know_{lo} that the wall is red. And the Lewisian description of the situation vindicates that judgement: Soraya has no reason to suspect that she doesn't know_{lo} that the wall is red. **Elusive K¬K** entails only that Soraya knows_{hi} that she doesn't know_{hi} that the wall is red. And, as we saw above, that is actually a plausible thing to say about the situation.

This reconciliation relies on a linguistic hypothesis: that when 'know' is embedded in an attitude ascription, the contextual parameter relative to which it is interpreted is supplied not by the context of utterance, but by something like the private context of the subject of the attitude ascription. If this were a feature not

shared by other context-sensitive vocabulary, this would be an implausible consequence of the Lewisian account. But, fortunately for the Lewisian, there is independent reason to think that this kind of behaviour is actually quite common. For consider two other expressions which are naturally treated as context-sensitive: 'fun' and 'might'. It looks as though, usually, the contextual parameter (a standard of taste or evaluation, a body of information) is provided by the context of utterance: when we say that something is fun, we mean that it is fun *for us*, and when we say that something might be true, we mean that its truth is compatible with the information available *to us*. However, when these expressions are embedded in belief attributions, this natural treatment seems to go wrong. Consider:

(1) Soraya thinks that roller-coasters are fun.
(2) Soraya thinks that it might be raining in Abidjan.

Intuitively, (1) is true whenever Soraya thinks that roller-coasters are fun *for her*; she might be well aware that we abhor them, so that 'Soraya thinks that roller-coasters are fun for us' is definitely false. Similarly, (2) is true even when Soraya knows that we are better informed about the weather in Abidjan than she is, and thus suspends judgement on whether, for all you and I know, it might be raining in Abidjan. This suggests that, when they occur embedded in attitude ascriptions, the parameter for these expressions is usually supplied not by the context of utterance but by a derived context which is particularly sensitive to the subject of the embedding verb. And that is exactly the same as what our Lewisian wants to say about 'know'.[24]

It's worth emphasizing that this line of reasoning cannot be used to defend the stronger claim that the $K\neg K$ principle is valid. Our reasoning shows that the example described needn't be a counterexample to the claim that, if someone doesn't know$_{hi}$ that p, they know$_{hi}$ that they don't know$_{hi}$ that p. But the case is a genuine counterexample to the claim that, if someone doesn't know$_{lo}$ that p, they

[24] The thought that context-sensitive expressions embedded in attitude ascriptions are not simply interpreted relative to the context of utterance is quite familiar; see e.g. Stalnaker (1988) for a classic articulation and defence. It is frequently applied by contextualists to handle embeddings under 'says that' or 'believes that'; see e.g. Cappelen and Hawthorne (2009).

This strategy does face an important challenge with embeddings under factive attitude verbs such as 'knows' [cf Weatherson (2008) and Lasersohn (2009, pp. 369–372)]. For it seems to predict that we could say 'Soraya knows that roller-coasters are fun', even though we hate them (provided only that we think that Soraya loves them and knows that she does), which is clearly incorrect. We thus need to supplement the simple shifting story with a, perhaps pragmatic, account of why knowledge ascriptions seem to entail the proposition which their complement would have expressed had it not been embedded. But note that simply denying that embedding under 'knows' (unlike embedding under 'believes') shifts the parameter is also implausible. For we can say 'Soraya knows roller-coasters are fun' even if we know that she (falsely) believes that we hate them.

A less optimistic reaction to these problems is to conclude that they sink contextualism about such terms as 'fun' or 'might', and should push us towards relativism or expressivism instead. But then it seems like we could equally well rehabilitate a broadly Lewisian account of 'knows' in a relativist or expressivist framework. Abandoning the contextualist aspect of Lewis's account for relativism or expressivism seems to preserve all the applications Lewis makes of his contextualism; and it may have independent advantages, as claimed by MacFarlane (2005) for relativism and Chrisman (2007) for expressivism.

know$_{lo}$ that they don't know$_{lo}$ that p. For, in the case described, Soraya doesn't know$_{lo}$ that the wall is red—there are worlds that relevantly resemble the actual one in which it isn't (for all we've said, the actual world is such a world), and even knowing$_{lo}$ requires that one rule those out. But she (reasonably enough) thinks that she does know$_{lo}$ that the wall is red, and thus doesn't know$_{lo}$ that she doesn't know$_{lo}$ this. So the $K\neg K$ principle for know$_{lo}$ (and thus the general $K\neg K$ principle) is refuted by the example; it's just that, since the attributor's use of 'know' does not refer to know$_{lo}$, this does not refute the more modest claim that the $K\neg K$ principle for the relation attributors pick out with 'know' can fail only in cases which are ignored by those attributors.

How convincing is this hard-nosed response? I think that it is most attractive when the difference in what is presupposed by subject and attributors intuitively amounts to a difference in epistemic standards. By Soraya's standards, one does not, in general, have to verify that the lighting is good in order to use one's vision to know what colour an object is. By our standards, one does have to rule out such possibilities. Soraya knows that she doesn't know by our standards. But she reasonably (though falsely) believes that she knows by hers.

However, not all cases in which some attributors attend to a $K\neg K$ failure are intuitively described as cases in which their standards differ from the subject's. In fact, even the case of Soraya needn't be described as such. Perhaps we do not use 'know' in such a way that people need to, quite generally, rule out possibilities with misleading lighting before they can know the colour of an object. We think that many people know the colours of lots of things despite never performing such checks. We just also know about Soraya's specific situation, we know that the lighting in that specific room is unreliable, and thus want to deny knowledge to her in particular. If that is the situation, it doesn't seem as natural to describe us and Soraya as differing in standards; hence it also doesn't seem as natural to reconcile the case with **Elusive K¬K** by appeal to the fact that 'know' means something different for us than it does for Soraya.

(One might hope that such cases cannot arise: by the rule of resemblance, if the attributors attend to *any* possibilities in which the lighting is misleading, every subject has to rule out all of them before she can be said to know. But such a liberal application of the rule of resemblance would be disastrous, at least if 'ignoring' is understood in terms of presupposition.[25] When I was 10, someone stole my bicycle, so that it wasn't where I left it when I went to look for it. Since I know this, there are bike theft possibilities which are consistent with what I presuppose in almost any conversation. It had better not follow that 'know', in my mouth, is so stringent that I say something false whenever I claim of someone that she knows where her bike is.)

It should be noted that, even if it doesn't seem particularly natural, the hard-nosed strategy still applies in the cases where attributors and subject intuitively share standards. Since Soraya is ignoring the possibility that the lighting in this particular room is misleading, and we are not, the Lewisian theory predicts that we

[25] If we understand 'ignoring' in terms of salience, we cannot handle the cases of hypothetical $K\neg K$ failures described above, since (i) a scenario is salient even if it is discussed only hypothetically, and (ii) subject and attributor attend to all the same possibilities in that case.

use the word 'know' differently—even if, in some intuitive sense, our epistemic standards are the same. We can thus still appeal to the different interpretations of 'know' to reconcile the case with **Elusive K¬K** along the lines indicated above. Doing so is not *ad hoc*, because the Lewisian theory predicts quite independently that these two different interpretations will both be in play. If there is something uncomfortable about the response, then, this is not because it is unnatural by the Lewisian's own lights. Rather, the response draws our attention to a feature of the Lewisian account, that the range of possible interpretations might not correspond to the range of epistemic standards, which some may find unattractive. In the next section, I explore what happens to **Elusive K¬K** when we try to revise the Lewisian account to avoid this feature. It turns out that this yields a different, but also quite attractive, way of learning to live with **Elusive K¬K**.

2.2 A conciliatory response

We attend to the possibility that the lighting next door is misleading; in fact, we positively affirm that possibility. Soraya ignores it. Yet, none of us are inclined to generally take seriously such misleading lighting; and all of us are inclined to do so when we have particular reason to be suspicious. There is thus a clear similarity between our standards and Soraya's, making it somewhat odd that the Lewisian theory predicts that 'know' means something different relative to our different contexts.

It will help to dig a little deeper into where, intuitively, the Lewisian theory goes wrong. I suspect that the problem is that there are really two very different reasons we have for taking possibilities seriously. Some we take seriously because our standards require us to: you just don't qualify for the kind of state we're interested in unless you have ruled these out. Others we take seriously just because we have particular reasons to think that they obtain. Only the former reflect our standards, and so only those who differ in what possibilities they take seriously for the former reason should be classified as using 'know' differently.

Interestingly, this is something like a converse to the *Problem of Known Presuppositions* discussed by Blome-Tillman (2012). Suppose that I'm in a 'high stakes' situation: it really matters to me whether the bank will be open this Saturday, because my paycheck needs to be paid in before Monday if I want to avoid disastrous results.[26] In fact, it matters so much that I'm initially inclined to take seriously that the bank has changed its weekend opening hours during the last month, which was the last time I checked. However, I am now looking at the bank's website, and can see that the opening hours haven't changed, so I stop taking that possibility seriously. Nonetheless, I am inclined to say 'Omar doesn't know that the bank will be open tomorrow' when all he has to go on is that it opened on Saturdays a month ago; and this is true even if Omar, being in a low-stakes situation, believes the bank to be open tomorrow. In this case, my standards seem to make relevant a possibility which, because of the particular evidence I have, I don't take seriously

[26] Cf DeRose (1992).

(in the sense that it is not compatible with my presuppositions); in the wall case, my particular evidence makes me take seriously a possibility (that the lighting next door is misleading) which my standards usually allow me to ignore.

We can solve both problems at once if we interpret 'ignoring' not in terms of which possibilities we take seriously (i.e. are compatible with our presuppositions), but rather in terms of which possibilities we consider *ordinary* or *normal*. When the stakes are high, I take possibilities in which the bank changes its opening hours to be sufficiently ordinary to be worth worrying about, regardless of whether I have evidence that allows *me* to rule it out. Conversely, I might think of all cases of misleading lighting as abnormal despite having evidence that a particular such case has actually occurred. So, in the wall case, we attributors can agree with Soraya that only possibilities with ordinary lighting are normal, so that 'knowledge' means the same relative to our context and hers.

We thus avoid the somewhat counterintuitive feature of the Lewisian account that the hard-nosed defence relied on. In doing so, we make room for a different way of responding to **Elusive K¬K**. For that principle says that counterexamples to $K\neg K$ can only occur in worlds that are 'ignored' by the attributors of knowledge, however that is spelled out. If 'ignoring' is understood in terms of presupposition or salience, that seems implausible, so that an extended reconciliation along the lines outlined in Sect. 2.1 is called for. But if 'ignored' is interpreted as meaning simply 'is considered abnormal', the result is not so surprising. When things are normal, rational beliefs amount to knowledge; it is only when the environment is abnormally uncooperative that they do not, leading to $K\neg K$ failures. **Elusive K¬K** thus no longer seems threatening.[27]

The cost of responding in this way is that, unlike the notion of a presuppositions or of a possibility being salient, the notion of what attributors consider to be ordinary or normal remains somewhat unclear and does not feature elsewhere in our theories. But I do not here want to adjucate between the costs and benefits of the two responses I have suggested. The important point is that, between them, they show that **Elusive K¬K** is, initial appearances to the contrary, no *reductio* of a broadly Lewisian approach to 'knowledge'. The result is *prima facie* problematic if we interpret S so that attending to a world or treating it as a candidate for actuality automatically places it in S. Given such an understanding of S, however, the theory straightforwardly predicts that subject and attributors will often use 'know' differently, thus enabling the Lewisian to endorse the hard-nosed response without being *ad hoc*. If, on the other hand, we interpret S so that something more than salience or being a serious candidate for actuality is required to place a world in S, it

[27] Perhaps there will still be potential counterexamples in cases where attributors and subject do, intuitively, differ in their standards. Suppose that we are sceptics, refusing to dismiss any possibilities as abnormal. Should we describe ordinary people as failing to know without knowing that they fail? If so, such an ascription will have to be handled via the 'shifting' strategy developed in Sect. 2.1. But I actually have rather mixed feelings about this case; it strikes me as fairly natural to say that ordinary people, at least those that have encountered sceptical worries, do know that they don't *really* know, while a similar claim sounds absurd to me in the case of Soraya (provided we hold fixed that, in Soraya's case, the attributors don't *generally* take misleading lighting seriously). If that's right, it suggests that shifting, while perhaps possible, isn't obligatory, which would make trouble for the hard-nosed response.

is no longer clear that there is anything even *prima facie* implausible about **Elusive K¬K**. Either way then, the Lewisian needn't be worried.

3 Conclusion

The aim of this paper has been to investigate the status of the knowledge iteration principles according to the account of knowledge given by Lewis in "Elusive Knowledge". In Sect. 1 I showed how we could both (a) explain the wide-spread impression that Lewis's account vindicates the iteration principles and (b) confirm that, in fact, Holliday (2015) is right to maintain that the account invalidates them both; the key is to be careful to distinguish which parts of the account describe the dependence of knowledge attributions on the attributor's context and which parts describe the dependence of knowledge attributions on the subject's situation. In Sect. 2 I argued that, once this ground has been cleared, there is more to be said: while the $K\neg K$ principle is invalid, counterexamples to it are, in a certain sense, elusive, since they never occur in salient possibilities. I then argued that this consequence is, initial impressions to the contrary, quite defensible.

There are two novel lessons from this discussion that deserve to be highlighted, one general and one specific. The general lesson is that epistemic contextualism interacts in subtle and surprising ways with the knowledge iteration principles. The reason is that the contribution of context doesn't vary with the world of evaluation; it is therefore held fixed when we evaluate what is known at different worlds, and hence held fixed when we evaluate what is known at worlds compatible with the subject's actual knowledge. If we aren't careful, this can make iteration implausibly easy, as on the account discussed in Sect. 1.1. And even if we are careful, it leads to highly surprising theorems like **Elusive K¬K**. The connection is complicated somewhat by the fact that, as noted in Sect. 2.1, contextualists can cite precedents for holding that the contextual parameter with respect to which an embedded knowledge attribution is interpreted need not always be the one provided by the context of utterance. But this further complication doesn't show that there aren't interesting interactions between contextualism and iteration principles; rather, it shows that the interaction may be quite complex.

These interactions are worth studying for their own sake, as I've done here. But they also highlight an under-explored difference between contextualist views and their *subject sensitive invariantist* cousins.[28] These two approaches diverge most obviously when we consider third-personal knowledge ascriptions, where ascriber and subject come apart, and those divergences have been discussed in some detail. They may also diverge when it comes to counterfactual or temporal embeddings, again because the contribution of context won't vary as we shift the world (or time) of evaluation, while the contribution of the subject's situation will. To these known divergence we should now expect to add a third: the two approaches should make

[28] See Hawthorne (2004) and Stanley (2005) for subject sensitive invariantist views, and detailed discussion of their relation to contextualism.

different predictions for iterated knowledge attributions. And this is exactly what we find here, since no analogue of **Elusive K¬K** would hold if, in a subject-sensitive invariantist spirit, we replaced the contextually supplied S with a relation R_S representing which possibilities are salient (to the subject, or the attributors, or anyone else) from each world. I have not attempted a systematic evaluation of which position does better with respect to this divergence; but I have argued that, initial impressions to the contrary, contextualists needn't be overly worried.

This brings me to the more specific lesson of our discussion. I have shown that Lewis's account entails **Elusive K¬K**; very roughly, the claim that counterexamples to the $K\neg K$ principle can occur only in possibilities that are being ignored. Somewhat less roughly, rational subjects can fail to know, in the sense of 'knowledge' used by some attributors, without knowing that they fail to know *in this sense*, only if they inhabit possibilities which those attributors are ignoring. Whilst no doubt unexpected, I have argued that this consequence is not so surprising as to be a *reductio* of the Lewisian account. But it is still surprising enough, I think, to be epistemologically significant. Consider, for example, the Williamsonian E=K thesis that one's evidence consists of all and only the propositions one knows. Since the $K\neg K$ principle is non-negotiably false, this will mean that the iteration principles for 'evidence' will fail; and this, in turn, leads to counterexamples to otherwise plausible 'reflection principles'.[29] By maintaining that counterexamples to the $K\neg K$ principle occur only in ignored possibilities, we may be able to ease this tension. Under-described as it is, such an application remains a promissory note. But it is one that we can only even think about writing as a result of the present discussion.

Acknowledgments I'm grateful to Kevin Dorst, Julien Dutant, Jeremy Goodman, Sophie Horowitz, Brendan de Kenessey, Justin Khoo, Harvey Lederman, Ginger Schultheis, Alex Silk, Declan Smithies, Jack Spencer, Jonathan Vogel, Roger White, Steve Yablo, and one anonymous referee for helpful comments and discussion. I'm especially grateful to Bob Stalnaker and a second anonymous referee, whose critical yet sympathetic comments have improved the following discussion immeasurably, with respect to both numerous specific details (too many to acknowledge individually) and overall structure.

References

Blome-Tillman, M. (2009). Knowledge and presuppositions. *Mind*, *118*, 241–295.

Blome-Tillman, M. (2012). Contextualism and the problem of known presuppositions. In J. Brown & M. Gerken (Eds.), *Knowledge ascriptions*. Oxford: Oxford University Press.

Blome-Tillman, M. (2014). *Knowledge and presuppositions*. Oxford: Oxford University Press.

Cappelen, H., & Hawthorne, J. (2009). *Relativism and monadic truth*. Oxford: Oxford University Press.

Chrisman, M. (2007). From epistemic contextualism to epistemic expressivism. *Philosophical Studies*, *135*, 225–254.

Christensen, D. (2010). Rational reflection. *Philosophical Perspectives*, *24*, 121–140.

Cohen, S. (1998). Contextualist solutions to epistemological problems: Scepticism, Gettier and the lottery. *Australasian Journal of Philosophy*, *76*, 289–306.

DeRose, K. (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research*, *52*, 913–929.

[29] This includes both the standard diachronic reflection principles, as discussed by Williamson (2000, Chap. 10) and Weisberg (2007), and synchronic 'rational reflection' principles, as discussed by Christensen (2010), Williamson (2011), Elga (2013), Horowitz (2014), and Lasonen-Aarnio (2015).

DeRose, K. (2004). Single scoreboard semantics. *Philosophical Studies*, *119*, 1–21.
Douven, I. (2005). Lewis on fallible knowledge. *Australasian Journal of Philosophy*, *83*, 573–580.
Elga, A. (2013). The puzzle of the unmarked clock and the new rational reflection principle. *Philosophical Studies*, *164*, 127–139.
Greco, D. (2014). A puzzle about epistemic akrasia. *Philosophical Studies*, *167*, 201–219.
Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford: Oxford University Press.
Holliday, W. (2013). Response to Egré and Xu. In J. van Benthem & F. Liu (Eds.), *Logic across the university: Foundations and applications*. London: College Publications.
Holliday, W. (2015). Epistemic closure and epistemic logic I: Relevant alternatives and subjunctivism. *Journal of Philosophical Logic*, *44*, 1–62.
Holton, R. (2003). David Lewis's philosophy of language. *Mind and Language*, *18*, 286–295.
Horowitz, S. (2014). Epistemic akrasia. *Nous*, *48*, 718–744.
Ichikawa, J. (2011a). Quantifiers and epistemic contextualism. *Philosophical Studies*, *155*, 383–398.
Ichikawa, J. (2011b). Quantifiers, knowledge and counterfactuals. *Philosophy and Phenomenological Research*, *82*, 287–312.
Ichikawa, J. (2013). Basic knowledge and contextualist "E = K". *Thought*, *2*, 282–292.
Lasersohn, P. (2009). Relative truth, speaker commitment, and control of implicit arguments. *Synthese*, *166*, 359–374.
Lasonen-Aarnio, M. (2015). New rational reflection and internalism about rationality. In T. Szabo Gendler & J. Hawthorne (Eds.), *Oxford Studies in Epistemology* (Vol. 5). Oxford University Press.
Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, *74*, 549–567.
MacFarlane, J. (2005). The assessment sensitivity of knowledge attributions. In T. Szabo Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 1). Oxford: Oxford University Press.
Schaffer, J. (2004). Scepticism, contextualism, and discrimination. *Philosophy and Phenomenological Research*, *69*, 138–155.
Smithies, D. (2014). The phenomenal basis of epistemic justification. In J. Kallestrup & M. Sprevak (Eds.), *New waves in philosophy of mind*. London: Palgrave Macmillan.
Stalnaker, R. (1988). Belief attribution and context. In R. Grimm & D. Merill (Eds.), *Contents of thought*. Tuscon, AZ: University of Arizona Press.
Stanley, J. (2005). *Knowledge and practical interests*. Oxford: Oxford University Press.
Vogel, J. (1999). The new relevant alternatives theory. *Nous*, *33*, 155–180.
Weatherson, B. (2008). Attitudes and relativism. *Philosophical Perspectives*, *22*, 527–544.
Weisberg, J. (2007). Conditionalization, reflection, and self-knowledge. *Philosophical Studies*, *135*, 179–197.
Williams, M. (2001). Contextualism, externalism, and epistemic standards. *Philosophical Studies*, *103*, 1–23.
Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.
Williamson, T. (2001). Comments on Michael Williams 'scepticism, contextualism, and discrimination'. *Philosophical Studies*, *103*, 25–33.
Williamson, T. (2009). Reply to Stephen Schiffer. In P. Greenough & D. Pritchard (Eds.), *Williamson on knowledge*. Oxford: Oxford University Press.
Williamson, T. (2011). Improbable knowing. In T. Dougherty (Ed.), *Evidentialism and its discontents*. Oxford: Oxford University Press.