UNSTABLE KNOWLEDGE, UNSTABLE BELIEF

Hans ROTT

ABSTRACT: An idea going back to Plato's *Meno* is that knowledge is stable. Recently, a seemingly stronger and more exciting thesis has been advanced, namely that rational belief is stable. I sketch two stability theories of knowledge and rational belief, and present an example intended to show that knowledge need not be stable and rational belief need not be stable either. The second claim does not follow from the first, even if we take knowledge to be a special kind of rational belief. 'Stability' is an ambiguous term that has an internally conditional structure.

KEYWORDS: knowledge, rational belief, subjective probability, stability

1. The Example

Consider the following story that ramifies into two alternative versions.

> Yesterday afternoon, at four o'clock, Sam looked out of his window and saw his neighbours Ann and Ben passing by (or so he thought). Sam could see the couple very clearly in the bright sunshine. It did not occur to him at all that he might mistake some other people for his neighbours. Still, he starts doubting now whether it was really Ann and Ben who he saw yesterday. Mia, a very serious and reliable person and a very close friend of Ann and Ben's, just told Sam that ...
>
> *(Version 1)* ... it wasn't Ann and Ben who were passing by. Mia did not want to give Sam more information, but there is no doubt that what she said is true. Sam knows Ann really well, so he is reluctant to call into question that he saw her. And exactly the same is true for Ben. Still Sam concludes, with some amazement, that it must have been another man or another woman who he saw passing by his window. As a matter of fact, the woman walking past his window was indeed Ann, but the man was Ben's twin brother Bob.
>
> *(Version 2)* ... Ann and Ben had to present their joint paper in a Graduate Workshop at the university at 4 p.m. yesterday. Since there is no question that what Mia said is true, it is doubtful whether Ann and Ben could have been in the neighbourhood at four o'clock. Sam reconsiders the situation, and even though he still thinks there is a fair chance that it was Ann and Ben who he saw, it does not appear unlikely to him that he mistook some other persons for them.

© LOGOS & EPISTEME, X, 4 (2019): 395-407

From the description of the scenarios, it is clear that Sam *fully believed* yesterday that Ann and Ben were passing by his window, and he was *fully justified* and *rational* in so believing. In addition, it seems that Sam in fact *knew* yesterday that Ann was passing by in version 1 of the story, notwithstanding his later retraction of the belief that he saw her. His view of Ann was completely unimpaired, he could recognise her clearly, and it was in fact Ann who he saw. His successful identification of Ann is not undermined by his bad luck with the (mis)identification of Ben.

We shall see that if these intuitions about Sam's propositional attitudes are right, the story goes against the idea of *stability* that some authors have suggested to be a necessary condition for knowledge or for rational belief. In the next section, I will briefly review a stability theory of knowledge fathomed by a number of recent authors. I then use Version 1 of our story to show that stability is not necessary for knowledge. In Section 3, I present a stability theory of rational belief recently proposed and developed by Hannes Leitgeb.[1] The second version of our story will then be employed to show that belief need not be stable either. Section 4 clarifies the relationship between the two kinds of theories by distinguishing various meanings of the predicate 'stable.' Assuming that knowledge entails rational belief, the existence of unstable knowledge seems to entail that stability cannot be a necessary condition for rational belief either. But Section 5 explains why such an inference would be fallacious. Version 2 of the story is indeed needed for my argument that rational belief need not be stable.

2. The Stability Theory of Knowledge

In Plato's *Meno*, stability is suggested as a feature of knowledge that makes it more valuable than merely true belief.[2] Contemporary epistemological writings have rarely considered stability as a part of the definition (or nature) of knowledge. In recent semantic modellings of epistemic states, by contrast, the stability condition has been the topic of considerable discussion. Stability is defined here with reference to a multitude of possible worlds, which makes it a modal concept. Referring to a then unpublished paper of Stalnaker's, Lamarre and Shoham provided an axiomatisation and a semantics reflecting the idea that "knowledge is

[1] Leitgeb's theory is a theory about *rational belief*, even if he frequently just calls it a theory about *belief* (see the titles of his works quoted below).

[2] See Casey Perin, "Knowledge, Stability, and Virtue in the Meno," *Ancient Philosophy* 32, 1 (2012): 15–34, and the references cited therein.

belief that is 'stable with respect to the truth.'"[3] Stalnaker seems to have been the first author presenting the idea of the stability analysis of knowledge:

> [... an agent] *a* knows that ϕ if and only if *a* believes that ϕ [...], *and that belief is robust with respect to the truth*. [...] More precisely, the proposition that *a* knows that ϕ is the set $\{w \in W$: for all ψ such that $w \in \psi$, $B_{a,w}(\psi) \subseteq \phi\}$.[4]

Here, $B_{a,w}(\psi)$ denotes the belief state of agent *a* in world *w* conditional on ψ, or more precisely, the posterior belief state of *a* that would be induced by learning ψ in *w*. In Stalnaker's paper, a belief state is simply the strongest proposition believed to be true, i.e., the set of worlds that the subject believes might be the actual world. It is important that the propositions ψ on which the belief state is to be conditioned in order to determine whether agent *a*'s belief that ϕ is stable are propositions that are true at *w*.

This stability analysis of knowledge is a simplified variant of the defeasibility analyses of knowledge prevalent in the 1960s and 1970s: where the former refers to a loss of belief, the latter refer to a loss of justification.[5] It is easier to give a semantic model of the loss of belief than to give one of the loss of justification. The stability analysis was later entertained and discussed by Rott,[6] Stalnaker[7] and Baltag and Smets,[8] but none of these authors has actually embraced it as a successful analysis of knowledge. Baltag and Smets occasionally use the term "Stalnaker

---

[3] Philippe Lamarre and Yoav Shoham, "Knowledge, Certainty, Belief, and Conditionalization," in *Principles of Knowledge Representation and Reasoning (KR'94)*, eds. Jon Doyle, Erik Sandewall, and Pietro Torasso (San Francisco, CA: Morgan Kaufmann, 1994), 415–424, here 418.

[4] Robert Stalnaker, "Knowledge, Belief and Counterfactual Reasoning in Games," *Economics and Philosophy* 12, 2 (1996): 133–163, here 146 and 155–156, notation adapted.

[5] The defeasibility analysis of knowledge is linked to philosophers like Annis, Harman, Klein, Lehrer, Paxson, Sosa and Swain. It has been criticised many times, but for some epistemologists, it still remains the most plausible approach to solving the Gettier problem; see Claudio de Almeida and João R. Fett, "Defeasibility and Gettierization: A Reminder," *Australasian Journal of Philosophy* 94, 1 (2016): 152–169.

[6] "A belief α is a piece of knowledge of the subject *S* iff α is not given up by *S* on the basis of any true information that *S* may receive" (Hans Rott, "Stability, Strength and Sensitivity: Converting Belief into Knowledge," *Erkenntnis* 61, 2–3 (2004): 469–493, here 471).

[7] "[...] define knowledge as belief (or justified belief) that is stable under any potential revision by a piece of information that is in fact true" (Robert Stalnaker, "On Logics of Knowledge and Belief," *Philosophical Studies* 128, 1 (2006): 169–199, here 187).

[8] What Alexandru Baltag and Sonja Smets, "A Qualitative Theory of Dynamic Interactive Belief Revision," in *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, eds. Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge (Amsterdam: Amsterdam University Press, 2008), 11–58, call "Stalnaker knowledge" is "belief that is *persistent under revision with any true information*" (13).

knowledge" (in scare quotes), but in general prefer calling what is defined by the stability analysis "safe belief."[9] Independently of each other, Rott and Stalnaker offered counterexamples against the stability analysis.[10]

Like defeasibility analyses, stability analyses have a problem with misleading evidence (or "misleading defeaters"). The first version of the story above is similar to the counterexamples advanced earlier against defeasibility and stability analyses, even though it would seem odd to call the information provided by Mia misleading. After his observation in the bright sunshine, Sam knew that Ann was passing by his window. Upon receiving the true, but belief-contravening information that it wasn't Ann and Ben who were passing by, however, Sam drops not only his false belief that he saw Ben, but also his true belief that he saw Ann. If this interpretation of the situation is correct, then knowledge need not be stable in the sense of the stability theory of knowledge.

3. The Stability Theory of Rational Belief

We now turn to the question whether stability is a necessary requirement for rational belief.[11] The claim that belief needs to be stable is surprising, because intuitively, and also according to the Platonic Socrates, stability or strength may just be features that *distinguish* knowledge from belief.[12]

---

[9] See Baltag and Smets, "Qualitative Theory," 13 and 27–29. They think that the stability condition is *too weak* for knowledge, and suggest that knowledge requires stability even upon receipt of arbitrary, possibly false information.

[10] Rott, "Stability, Strength and Sensitivity," 482–483, and Stalnaker, "On Logics of Knowledge and Belief," 190. The stability analysis had not been criticised either by Lamarre and Shoham, "Knowledge, Certainty, Belief, and Conditionalization," or by Stalnaker, "Knowledge, Belief and Counterfactual Reasoning in Games." Rott's and Stalnaker's examples are intended to show that the stability condition is *too strong*. Rott (Stability, Strength and Sensitivity," 476–477) points to a general reason for the failure of the stability analysis. He shows that a belief is stable (in the above sense) just in case it is more entrenched in the subject's belief state than *every* false belief. This is a requirement that seems very hard to meet: we probably have many false beliefs, some of them highly entrenched in our cognitive states. So meeting this requirement can hardly be a necessary condition for knowledge.

[11] The first authors to make the connection between the stability theories of knowledge and rational belief were Eric Raidl and Niels Skovgaard-Olsen, "Bridging Ranking Theory and the Stability Theory of Belief," *Journal of Philosophical Logic* 46, 6 (2017): 577–609.

[12] See Terry Penner, "Socrates on the Strength of Knowledge: Protagoras 351B–357E," *Archiv für Geschichte der Philosophie* 79, 2 (1997): 117–149, here 121: "Knowledge is strong while belief is *weak*." Also compare John Hawthorne, Daniel Rothschild and Levi Spectre, "Belief is Weak," *Philosophical Studies* 173, 5 (2016): 1393–1404, who argue that our everyday notion of belief is unambiguously a weak one.

According to Louis Loeb, however, David Hume held that the most essential elements of belief are steadiness and stability.[13] Every belief, qua belief, is *steady* or "infixed" by a belief-forming mechanism (the senses, memory, causal inference, custom or repetition), and it may as such be called *stable in a wider sense*. But not every belief is *stable in the narrower, proper sense* of the term: not every belief is steady in its influence on thought, feeling, the will and action. Steadiness makes for justification other things being equal, but only stability proper makes for justification all things considered.[14] According to Loeb's "more demanding" reading of the Hume's stability theory, rational (justified) beliefs have to be stable under *full* or *intense* reflection. Such reflection includes an assessment of the quality of one's belief-forming processes, as well as the elimination of incoherences among the beliefs that were infixed by the belief-forming mechanisms. Let us call this a *reflective* conception of stability.

I will not contest Loeb's account, neither as an interpretation of the historical Hume nor as a substantive analysis of belief. Instead I want to turn to a recent alternative approach championed by Hannes Leitgeb.[15] He assumes that the doxastic state of a subject includes both her categorical beliefs, represented by a single proposition, and her degrees of beliefs, represented by a probability function. He picks up on Loeb's interpretation of Hume, but his motivation can be traced even further back than to Hume. Leitgeb's initial project was to reconcile two things: (i) the so-called *Lockean thesis*, according to which rational belief *simpliciter* is tied to high probability above a certain threshold value *r*, and (ii) the logical closure and consistency of rational categorical beliefs.[16] The lesson from the lottery paradox seems to be that this is an infeasible project. But Leitgeb

[13] Louis E. Loeb, *Stability and Justification in Hume's Treatise* (Oxford: Oxford University Press, 2002) and Louis E. Loeb, *Reflection and the Stability of Belief: Essays on Descartes, Hume, and Reid* (Oxford: Oxford University Press, 2010).

[14] For this and the following, see Loeb, *Stability and Justification in Hume's Treatise*, chapter 3, and Loeb, *Reflection and the Stability of Belief*, 16–21 and Chapter 5. "A belief might fail to be steady in its influence owing to the presence of beliefs with which it conflicts, beliefs which [...] reduce its influence on the will and action. [...] I use the term 'stable' as a shorthand for 'steady in its influence on thought, passions, and action'" (Loeb, *Stability and Justification in Hume's Treatise*, 80, and Loeb, *Stability and Justification in Hume's Treatise*, 155–156).

[15] Hannes Leitgeb, "The Humean Thesis on Belief," *Proceedings of the Aristotelian Society, Supplementary Volume* 89, 1 (2015): 143–185, and Hannes Leitgeb, *The Stability of Belief: How Rational Belief Coheres with Probability* (Oxford: Oxford University Press, 2017).

[16] Hannes Leitgeb, "The Stability Theory of Belief," *Philosophical Review* 123, 2 (2014): 131–171. The label "Lockean Thesis" is due to Richard Foley, "The Epistemology of Belief and the Epistemology of Degrees of Belief," *American Philosophical Quarterly* 29, 2 (1992): 111–124.

demonstrated that such a reconciliation is non-trivially[17] possible, provided that the subject's personal probability function is such that there is a proposition the probability of which does not sink below 0.5, *conditional on any information compatible with the subject's beliefs*. Leitgeb thus modifies the idea that Loeb finds in Hume, and requires stability not under reflection, but under (potential or actual) revision by new information. More precisely, he considers updates of the subject's actual beliefs by new information that is compatible with these beliefs. Here is what Leitgeb calls the *Humean thesis on rational belief*:

> It is rational to believe a proposition just in case it is rational to assign a *stably* high subjective probability to it (or to have a *stably* high degree of belief in it).[18]
>
> *The Humean Thesis Explicated*: If *Bel* is a perfectly rational agent's class of believed propositions at a time, and if P is the same agent's subjective probability measure at the same time, then for all ϕ:
>
> > ϕ is in *Bel* if and only if for all ψ, if ψ is possible both in the all-or-nothing sense (i.e., ψ is logically compatible with *Bel*) and the probabilistic sense (i.e., ψ has non-zero probability), then $P(\phi \mid \psi) > r$.[19]

Here $P(\phi \mid \psi)$ is the standard conditional probability of ϕ given ψ, defined as $P(\phi \cap \psi)/P(\psi)$, and r is a threshold parameter lying between 0.5 and 1. Conditionalising one's probability function P on a proposition ψ essentially means accepting ψ either actually or hypothetically. According to the Humean thesis, it is rational to believe a proposition ϕ just in case its probability remains high conditional on any proposition ψ that is doxastically possible for the agent: no such proposition ψ defeats the high degree of belief in ϕ.[20] The idea here is similar to

---

[17] 'Non-trivial' here means that there are beliefs with a probability below 1. This is equivalent to there being non-tautological beliefs, if the probability function is supposed to be regular. I assume that rational agents in general aim at having non-trivial belief sets.

[18] Leitgeb, "The Humean Thesis on Belief," 152.

[19] Leitgeb, "The Humean Thesis on Belief," 163, notation adapted and some more technical clauses replaced by ordinary-language formulations. On 159–162, Leitgeb reviews five alternative ways of making the generic idea of the Humean thesis precise. His option (b) which "would correspond to a kind of coherence theory of belief" (160) is closer to (Loeb's interpretation of) Hume than option (d) which Leitgeb ultimately embraces.

[20] Leitgeb's move of adopting the Humean rather than the Lockean thesis, i.e., of requiring r-stability rather than P-stability (which has the constant 0.5 in place of the parameter r), can be interpreted as reflecting the idea that the threshold value for the conditional probabilities should be the same as for the unconditional probability, i.e., it should be r rather than 0.5. I find this the most natural interpretation, but Leitgeb (personal communication) is ready to apply different thresholds to conditional and unconditional beliefs. For the ranges of Lockean and Humean thresholds that are suitable for a given proposition, see Hans Rott, "Stability and Skepticism in

that of the stability theory of knowledge, with the crucial difference that the latter refers to the (hypothetical or actual) acceptance of *true* propositions while the former refers to the (hypothetical or actual) acceptance of propositions *compatible with the subject's beliefs*.

The second version of our story shows, I submit, that the stability account based on the Humean thesis does not adequately capture the intuitive notion of rational belief. Sam was fully rational in believing that Ann and Ben were passing by when he looked out of his window (independently of whether it actually was Ann and Ben who he saw). The information that Ann and Ben have had an important obligation to present their joint paper at the workshop is consistent with Sam's belief that he saw the couple walking past his window, and indeed with his full body of belief. Sam knew, after all, that their scheduled presentation might have been put off. But the news about their commitment dramatically decreases the likelihood that it was Ann and Ben who he saw. So we have found a perfectly rational belief that has a rather low subjective probability when conditionalised on information compatible with Sam's full body of beliefs. This is a counterexample to the Humean thesis.

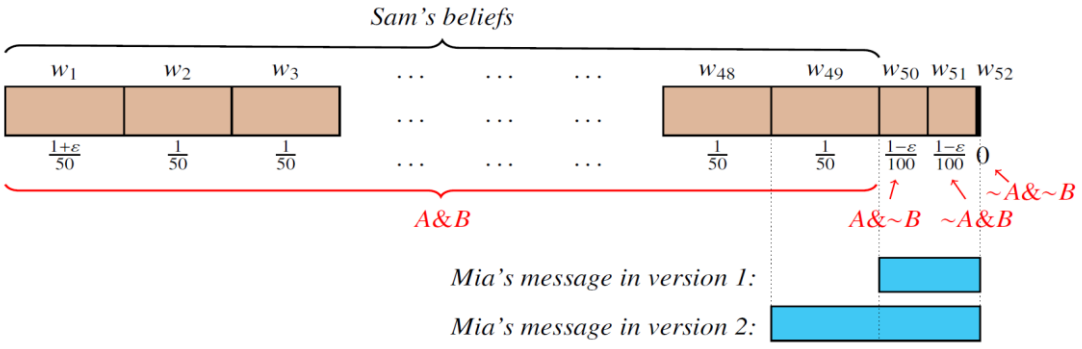

Fig. 1: Sam's doxastic state yesterday and Mia's messages

This example is meant to be compelling because it is intuitively plausible. Still, it will be reinforced and provide a better service as a counterexample if we can reproduce it in terms of the formal model used by Leitgeb. Suppose that Sam's doxastic situation yesterday looked as follows (see Fig. 1). He considered 52 worlds

the Modelling of Doxastic States: Probabilities and Plain Beliefs," *Minds and Machines* 27, 1 (2017): 167–197.

as possible that we call w_1, w_2, …, w_{52}. In w_1, …, w_{50}, Ann passes Sam's window, in w_{51} and w_{52}, she doesn't. In w_1, …, w_{49} and w_{51}, Ben passes Sam's window, in w_{50} and w_{52}, he doesn't. In w_{49}, …, w_{52}, Ann and Ben have an important obligation elsewhere (such that at least one of them ought to be present), in w_1, …, w_{48} they don't. Suppose further that Sam assigned the following subjective probabilities to the worlds he considered possible: for some very small positive real number ε, $P(w_1) = \frac{1+\varepsilon}{50}$, $P(w_2) = \ldots = P(w_{49}) = \frac{1}{50}$, $P(w_{50}) = P(w_{51}) = \frac{1-\varepsilon}{100}$ and $P(w_{52}) = 0$. Assuming that Sam can think of 49 different ways his neighbours' passing by might have come about, this is a natural representation of Sam's belief state yesterday.[21]

The proposition $\{w_1, \ldots, w_{49}\}$ is the only non-trivial *P*-stable set and may thus be taken to qualify as the proposition characterising Sam's initial beliefs. Sam's conditional probability that Ann and Ben passed by, *given* the information that they had an important obligation elsewhere, is above 0.5, but only slightly so. If we take ε = 0.1, for instance, it is 0.526. As long as Leitgeb assumes that the threshold for belief is set to 0.5, he can still recommend as rational the belief that Ann and Ben passed by, since $\{w_{49}\}$ is the only non-trivial *P*-stable set of Sam's subjective probabilities conditionalised on the information that Ann and Ben had an important obligation (i.e., conditionalised on the proposition $\{w_{49}, w_{50}, w_{51}, w_{52}\}$). But a posterior probability of 0.526 is low, arguably too low to support belief *simpliciter*. Belief appears to require at least a moderately *high* probability, one that lies significantly above 0.5. As a consequence, Sam loses his belief that Ann and Ben were passing by yesterday.

The following *Preservation condition* may be viewed as a qualitative analogue of Leitgeb's stability condition: If a proposition is consistent with a subject's current beliefs, she should give up none of her current beliefs on accepting or on hypothetically assuming that this proposition is true. Preservation is one of the basic conditions of Alchourrón, Gärdenfors and Makinson, and it has almost universally been accepted in the belief revision literature.[22] But the second

[21] There are of course many alternative representations of Sam's belief state that make sense. But it is important to stress here that a single natural way of formally fleshing out the informal example is sufficient for establishing that it can serve as a serious counterexample to Leitgeb's theory. And I claim that my formal precisification is a natural one. Two potential objections do not strike me as compelling. First, there is no reason to suppose that Mia's message introduces a context change that forces a refinement of the partition of all possibilities. Second, nothing depends on there being a world with zero probability; the example could easily be modified in such a way that w_{52} has positive probability.

[22] Among the very few authors arguing against Preservation are Charles B. Cross, "Belief Revision, Nonmonotonic Reasoning, and the Ramsey Test," in *Knowledge Representation and Defeasible Reasoning*, eds. Henry E. Kyburg, Ronald P. Loui, and Greg N. Carlson (Boston:

version of our example may also serve as a counterexample to Preservation. Proceeding on the supposition that categorical beliefs derive from probabilities, we have just fleshed out the example in such a way that Sam appears to be fully rational in dropping his belief that it was Ann and Ben who he saw yesterday.[23]

4. Analysis: What 'Stability' May Mean

The theories reviewed make use of different ideas of stability that are specialisations of a more general concept. The general stability scheme is this. A state property X is stable under the state transformation Y just in case the following holds: for all states S, if S has the property X and undergoes a transformation (of the kind) Y and no other transformation is performed on S, then the state $S' = Y(S)$ has the property X, too.

The states S^a we want to consider in the following are mental states of a rational agent a. For any proposition ϕ, let the property X^ϕ of a state S^a be that in S^a, a has a certain propositional attitude of the epistemic or doxastic kind with respect to ϕ. That the state S^a has the property X^ϕ means that agent a *X*es that ϕ in S^a, where '*X*es' stands for verbs such as 'knows,' 'believes,' 'rationally believes,' 'expects,' 'surmises,' 'doubts,' 'wonders,' 'is certain,' 'is convinced,' 'assigns a high subjective probability,' 'entertains (the idea),' etc.

The property X^ϕ of S^a is called *stable under reflection* just in case the following holds: if a *X*es that ϕ in state S^a and then reflects about the system of propositions *X*ed by herself (and nothing else happens), then α still *X*es that ϕ after having finished her reflections. The property X^ϕ is called *stable under updating (by eligible information)* just in case the following holds: if a *X*es that ϕ in S^a and then accepts an eligible piece of information ψ (and nothing else happens), then a still *X*es that ϕ in the updated state $S^a * \psi$.

Kluwer, 1990), 223–244, here 232–234; Wlodzimierz Rabinowicz, "Stable Revision, or is Preservation Worth Preserving?" in *Logic, Action, and Information: Essays on Logic in Philosophy and Artificial Intelligence*, eds. André Fuhrmann and Hans Rott (Berlin: de Gruyter, 1996), 101–128, here 105–106; and Richard Bradley, "Restricting Preservation: A Response to Hill," *Mind* 121, 481 (2012): 147–159, here 155–156.

[23] Hanti Lin and Kevin T. Kelly, "Propositional Reasoning that Tracks Probabilistic Reasoning," *Journal of Philosophical Logic* 41, 6 (2012): 957–981, here 964, call Preservation 'Accretion' and give a Gettier-style example that on the face of it resembles the probabilified second version of our story. However, I find their example unconvincing since they give no argument for their claim that "the strongest proposition we accept is the disjunction of 'Nogot' with 'Havit,' namely 'somebody'." Their example is also criticised by Leitgeb, *The Stability of Belief*, 187.

According to Loeb,[24] to whom Leitgeb makes essential reference, stability under reflection is what Hume was after. Stability under updating covers the stability theories of knowledge and rational belief introduced above if we specialise '*X*' to 'believes' and to 'assigns a probability above *r*,' respectively. For the definition of stability under updating, we still need to specify when to regard a piece of information as eligible. 'Eligible' is used as a generic term here that is supposed to cover different interpretations of stability. We focus on the two interpretations that shape the stability theories of knowledge and rational belief sketched above and call a proposition ψ (i) *eligible for knowledge* iff ψ is true; and (ii) *eligible for belief* iff ψ is compatible with the subject's current beliefs (i.e., iff ψ is not belief-contravening).

5. No Direct Route from the Instability of Knowledge to the Instability of (Rational) Belief

Do we really need two versions of our example? It is part of almost all contemporary epistemology that knowledge is a kind of rational belief. Though this is an assumption that clearly does *not* follow from the two stability theories, let us suppose it is true for the purposes of the following considerations. On this hypothesis, the fact that knowledge need not be stable seems to entail straightaway that rational belief need not be stable either. It looks as if this can be established simply by reasoning by way of a Bocardo inference:

Some pieces of knowledge are unstable.	(major premise)
All pieces of knowledge are rational beliefs.	(minor premise)
Some rational beliefs are unstable.	(conclusion)

The Bocardo scheme has been recognised as valid ever since Aristotle's syllogistics. But this particular inference is fallacious for two reasons. First, 'stability' is a syncategorematic predicate that may mean different things when applied to knowledge and when applied to belief. This is indeed the case with the stability theories of knowledge and belief: they involve different propositional attitudes and different notions of eligibility. Although both theories employ the notion of stability under updating, what makes a piece of information eligible is truth in the case of knowledge and compatibility with the subject's beliefs in the case of belief.

The ambiguity of the stability predicate is not deeply hidden, but it is worth emphasising, and it indeed prevents version 1 of our story from being suitable as a

[24] Loeb, *Stability and Justification in Hume's Treatise*, chapter 3.

counterexample to the stability theory of rational belief. But can't we perhaps find a more sophisticated concept of eligibility that is suitable for both knowledge and belief? I do not want to exclude this possibility. But even if the search for such a universally applicable notion of eligibility were successful, the inference above would still fail to go through. As we have seen, 'stable' is not a primitive predicate, but has an intrinsically conditional structure where the propositional attitude involved occurs both in the antecedent and the consequent of the relevant conditional. Consequently, 'unstable' has a conjunctive structure in which the propositional attitude involved occurs twice, once positively and once negatively. If we make the logical structures explicit, we realise that it is inadequate to represent the proposed inference as a Bocardo like this:

$$\frac{\exists\phi\,(\mathit{knows}(\phi)\ \&\ \mathit{unstable}(\phi))\qquad \forall\phi\,(\mathit{knows}(\phi)\supset \mathit{rbelieves}(\phi))}{\exists\phi\,(\mathit{rbelieves}(\phi)\ \&\ \mathit{unstable}(\phi))}$$

In its deeper structure, the inference above instantiates a scheme that is indeed logically invalid—even if we could avail ourselves of a notion of eligibility that is suitable for both knowledge and belief:

$$\frac{\exists\phi,\psi,S\,(\mathit{knows}(\phi,S)\ \&\ \mathit{eligible}(\psi,S)\ \&\ \sim\mathit{knows}(\phi,S{*}\psi))\qquad \forall\phi,S\,(\mathit{knows}(\phi,S)\supset \mathit{rbelieves}(\phi,S))}{\exists\phi,\psi,S\,(\mathit{rbelieves}(\phi,S)\ \&\ \mathit{eligible}(\psi,S)\ \&\ \sim\mathit{rbelieves}(\phi,S{*}\psi))}$$

Back to our example. The first version does not show that rational belief is unstable. Sam, I claim, initially *knew* and thus *rationally believed* that Ann was passing by. He does not believe that she was passing by any more after having received the true information that it wasn't Ann and Ben who were passing by. But the information he received from Mia was incompatible with his beliefs. So while it was eligible for knowledge, it wasn't eligible for rational belief.

The second version of the example, in contrast, does illustrate the instability of rational belief. Here the information provided by Mia is eligible for both knowledge and belief. We could actually have used this version as a counterexample to the stability theory of knowledge. However, since it is a lot more complicated than the first version (witness Fig. 1), the latter is of independent value in making a simple non-probabilistic case against the stability of knowledge.

6. The Stability of Knowledge and Belief Themselves

The stability theories outlined above define *knowledge* as stable true belief and *rational belief* as stably high probability. By so doing they do not immediately answer the question whether knowledge and rational belief *themselves* are stable, that is, stable under updating by propositions that are eligible in the suitable sense. In this final section, I identify some sufficient conditions for this being true, on the basis of the theories in question.

We begin with knowledge, as conceived by the qualitative stability theory. That agent *a* knows that ϕ in state S, in symbols $knows_a(\phi, S)$, has been defined by

$believes_a(\phi, S)$ and for all ψ, if $true(\psi)$, then $believes_a(\phi, S*\psi)$).

We want to show that knowledge is stable under eligible updating, that is:

If $knows_a(\phi, S)$ and $true(\psi)$, then $knows_a(\phi, S*\psi)$.

So suppose that $knows_a(\phi, S)$ and $true(\psi)$. We need to show that, first, that $believes_a(\phi, S*\psi)$ and, second, that for all χ, if $true(\chi)$, then $believes_a(\phi, S*\psi*\chi)$. Now it follows from the definition of $knows_a(\phi, S)$ that $believes_a(\phi, S*\psi)$, which gives us the first claim.

It seems that the only way to prove the second claim is to take an arbitrary true sentence χ and show that the state $S*\psi*\chi$ supports all beliefs supported by $S*(\psi\&\chi)$ and that $\psi\&\chi$ is eligible, i.e., true. Since both ψ and χ are true, so is $\psi\&\chi$. That the beliefs supported by $S*(\psi\&\chi)$ are included in the beliefs supported by $S*\psi*\chi$ is a condition well-known in the theory of iterated belief revision. It is satisfied, among others, by irrevocable revision (also known as radical revision) and by lexicographic revision (also known as moderate revision); but it is not satisfied, for instance, by natural revision (also known as conservative revision) and restrained revision.[25] So knowledge in the sense defined by the stability theory of knowledge is stable if either irrevocable or lexicographic belief revision is employed, but knowledge need not be stable if any other method if iterated revision is employed.

Let us now look at rational belief, as conceived by the probabilistic stability theory. That agent *a* in state S rationally believes that ϕ, in symbols $rbelieves_a(\phi, S)$, has been defined by

$hiprob_a(\phi, S)$ and for all ψ, if $compatible_S(\psi)$, then $hiprob_a(\phi, S*\psi)$.

[25] For the four methods, compare Hans Rott, "Preservation and Postulation: Lessons from the New Debate on the Ramsey Test," *Mind* 126, 502 (2017): 609–626. Notice that since both ψ and χ are true, they are compatible with each other. Notice also that we need to take belief-contravening revisions into account here, too. It is not guaranteed that χ is consistent with $S*\psi$.

We want to show that rational belief is stable under eligible updating, that is:

If $rbelieves_a(\phi, S)$ and $compatible_S(\psi)$, then $rbelieves_a(\phi, S*\psi)$.

So suppose that $rbelieves_a(\phi, S)$ and $compatible_S(\psi)$. We need to show that, first, $hiprob_a(\phi, S*\psi)$ and, second, that for all χ, if $compatible_{S*\psi}(\chi)$ then $hiprob_a(\phi, S*\psi*\chi)$. It follows from the definition of $rbelieves_a(\phi, S)$ that $hiprob_a(\phi, S*\psi)$, which gives us the first claim.

The only way to prove the second claim seems to take an arbitrary sentence χ that is compatible with $S*\psi$ and show that the state $S*\psi*\chi$ assigns a high probability to all propositions that are highly probable in state $S*(\psi\&\chi)$ and that $\psi\&\chi$ is eligible, i.e., compatible with S. Since both ψ is compatible with S and χ is compatible with $S*\psi$, it is plausible to assume that $\psi\&\chi$ is indeed compatible with S. Thus, by the definition of $rbelieves_a(\phi, S)$, we get $hiprob_a(\phi, S*(\phi\&\psi))$. If the probabilities assigned in doxastic states are changed by ordinary Bayesian conditionalization when the input or assumptions are consistent with those states, then changing a state first by compatible ψ and then by compatible χ yields identical probabilities to changing the state only once by compatible $\psi\&\chi$. This gives us $hiprob_a(\phi, S*\psi*\chi)$, as desired. Thus on the assumptions made, rational belief is indeed stable. Other ways of changing probabilities by compatible input or assumptions may give different results.

7. Conclusion

I have presented a stability theory of knowledge (discussed by Stalnaker, Lamarre and Shoham, Rott, and Baltag and Smets) and a stability theory of rational belief (embraced by Leitgeb), which have not been compared in the literature before. It was shown that these theories make use of a general concept of stability which can be differentiated into two distinct species. Using two versions of a concrete example, I argued that (i) knowledge need not be stable, and that (ii) rational belief need not be stable either, in the senses intended by the two theories. The two claims are independent of each other. Even on the supposition that knowledge is a particular kind of rational belief, the existence of unstable knowledge does not entail the existence of unstable rational belief, due to the logical structure of the general stability scheme and an ambiguity in the meaning of the predicate "stable."[26]

[26] Acknowledgements. I'd like to thank Tim Kraft, Hannes Leitgeb, Eric Raidl, Niels Skovgaard-Olsen and audiences in Regensburg, Stockholm, Munich, Paris, Maastricht and Dortmund for instructive comments on earlier versions of this paper. I am also grateful to the Swedish Collegium for Advanced Study in Uppsala for providing me with excellent research conditions while part of this paper was written.