HAL
archives-ouvertes.fr

Reliability, Margin for Error and Self-Knowledge

Paul Egré

▶ To cite this version:

Paul Egré. Reliability, Margin for Error and Self-Knowledge. Duncan H. Pritchard

Vincent Hendricks. New Waves in Epistemology, Ashgate Publishing, 2006. <ijn_00000667>

HAL Id: ijn_00000667
https://jeannicod.ccsd.cnrs.fr/ijn_00000667

Submitted on 3 Feb 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Reliability, Margin for Error and Self-Knowledge*

Paul Égré†

CNRS - Institut Jean-Nicod, Paris

Knowledge is more than justified true belief. Plato was certainly the first to state the lesson when he had Socrates argue that knowledge (*epistemè*) is more than true belief (*alèthès doxa*), and even more than true belief with an account or justification (*logos*). The lesson was later enforced by Russell, who also argued that knowledge is more than true belief. Thus, you don't know that the name of the Prime Minister starts with a B if you believe that the Prime Minister is Balfour, and if the actual Prime Minister is in fact Bannerman (Russell 1912, 131). In his celebrated paper, Gettier produced even more convincing examples, by adding that there can even be a *good* justification to believe a true proposition, without that justification being *appropriate* for the truth of that proposition. Consequently, it seems that the knowledge of a proposition p is the true belief of that proposition, with a justification that is appropriate for the truth of that proposition. But what does it take for a justification to be appropriate for the truth of a proposition?

In recent work, Timothy Williamson has defended the controversial thesis that a conjunctive analysis of knowledge such as: "knowledge is true belief with an appropriate justification" is doomed to failure, for it might be that no good account of the notion of appropriate justification can be framed without resorting to the very concept of knowledge by which we started. The thesis of Williamson that there can be no reductive and non-circular analysis of knowledge in terms of belief with other conditions has been the main focus of attention of his most recent book, *Knowledge and its Limits*. Despite this, Williamson firmly agrees that knowledge is a form of reliable belief. Thus Williamson writes: "if one believes p truly in a case α, one must avoid false belief in other cases sufficiently similar to α in order to count as reliable enough to know p in α" (2000, 100). In Williamson's account of knowledge, this notion of reliability is cashed out in a series of principles, which Williamson calls margin for error principles, and which he relates to a more general notion of safety. The interest of these principles is twofold: first, they provide the basis for a modal analysis of the concept of reliability of knowledge in the framework of epistemic logic. Secondly, and more substantially, Williamson has argued that knowledge is not transparent, namely that one can know a proposition without knowing that one knows it, and margin for error principles play a decisive

*Forthcoming in V. Hendricks & D.H. Pritchard (eds.), *New Waves in Epistemology*, (Adelshot: Ashgate Publishing). Not quite the last version. Please quote from the published version.

†I thank the Editors for their invitation to contribute to this volume and for their patience in awaiting the manuscript. I wish to express my gratitude to Inge Arnesen, Johan van Benthem, Denis Bonnay, Jérôme Dokic, Julien Dutant, Pascal Engel, Philippe Schlenker, Josh Snyder and Timothy Williamson for the many discussions, inspiring suggestions, and stimulating exchanges without which the paper would simply not exist in its present form (except for remaining errors, which are only mine). Special thanks are due to Denis Bonnay and Philippe Schlenker for comments on the logical aspects. I also thank Hannes Leitgeb for communicating me his review of Williamson 2000, and several audiences in Paris, Stanford and Lisbon.

role for the validity of this thesis.

The present paper pursues two objectives: the first is an attempt to refine and systematize the modal analysis of the reliability of knowledge given by Williamson, and to delimit the scope of margin for error principles. The second is a criticism of Williamson's thesis that knowledge is not transparent, elaborating on previous work by Dokic & Égré (2004), and based on the intuition that knowledge is modular and that a representation of this modularity is needed at the logical level in order to avoid the paradoxical conclusions which result from Williamson's assumptions. The paper is structured in two main parts: in the first part, I distinguish several concepts of reliability and use these distinctions to evaluate the limits of application of margin for error principles. In the second part, I examine the consequences margin for error principles might have concerning the reliability of self-knowledge.

1 Reliability and Margin for Error

Knowledge is a form of reliable belief, or belief with a good justification. It is a problem, however, to specify in sufficiently general terms what counts as a reliable belief, or as a good justification. In this section, my aim is to clarify the notion of knowledge reliability, in relation to the analysis Williamson has given of the concept of reliability *qua* truth at relevantly similar cases. More specifically, the aim of this section will be to contrast several related notions of reliability, namely the notions of safety, robustness, sensitivity, and margin of error, and to discuss their generality. This clarification will prepare the ground for the closer examination of margin for error principles given in section 2.

1.1 Knowledge and reliable belief

In the *Theaetetus*, Plato imagined the following dialogue between Socrates and Theaetetus. Socrates asks: "When someone, at the time of learning writes the name of "Theaetetus", and thinks that he ought to write and does write "Th" and "e"; but, again meaning to write the name of "Theodorus", thinks that he ought to write and does write "T" and "e"– are we going to say that he knows the first syllable of your names?" (207e-208a, tr. B. Jowett, revised by myself).[1] To this question, Theaetetus agrees with Socrates that a negative answer should be given.

This little piece of dialogue between Socrates and Theaetetus concerns our knowledge of a particular concept. Socrates is concerned with what it takes to know "the first syllable of "Theaetetus" and "Theodorus"". On one interpretation, this could mean: "to know that "Theaetetus" starts with "Th-e" and to know that "Theodorus" starts with "Th-e"". Clearly then, someone who fails to believe that "Theodorus" starts with "Th-e" can't know any proposition where this condition enters as a conjunct. This interpretation, however, trivializes Socrates's point a bit. Another, more interesting way to interpret this passage is the following: according to Socrates, you can't really know that "Theaetetus" starts with "Th-e", if you believe that "Theodorus" start with "T" followed by "e". In other words: even if you give "Theaetetus" the correct spelling, your belief that "Theaetetus" starts with "Th-e" is not reliable enough to count as knowledge if it becomes false at a relevantly similar case like

[1] Note that "Th" is the English transliteration for the one Greek letter θ, and "e" the English transliteration for the Greek ε. Likewise for "t", which is the transliteration of the Greek τ. Plato refers to each of them by their names "theta", "ei" (another name for "epsilon", which might have also referred to the way the letter was pronounced in Greek) and "tau". See also the translation of Levett in Burnyeat (1990, 345, fn 61).

"Theodorus". What this little piece of dialogue between Socrates and Theaetetus brings to light, then, is nothing but the problem of what counts as a reliable belief.

Let us represent by t the name "Theaetetus", by t' the name "Theodorus", and by P the predicate "starts with "Th-e"". Using the vocabulary of epistemic logic, Socrates's requirement on knowledge may be expressed more formally in the following way:

(1) $KP(t) \rightarrow \neg B \neg P(t')$

that is you don't know that "Theaetetus" starts with "Th-e" if you believe that "Theodorus" does not start with "Th-e".

Another way of interpreting Socrates's requirement would be to impose, not only that one fails to believe that "Theodorus", for instance, starts with "T-e" instead of "Th-e", but also that one does hold the belief that "Theodorus" starts with "Th-e". In that case, Socrates's requirement may be stated by the following, more positive constraint, namely you don't know that "Theaetetus" starts with "Th-e" unless you believe that "Theodorus" starts with "Th-e":

(2) $KP(t) \rightarrow BP(t')$

The strength of this analysis rests crucially on the hypothesis that t and t' are relevantly similar names. To say that "Theaetetus" and "Theodorus" are relevantly similar, in this situation, may be to say that they share the same property of starting with "Th" followed by "e". To articulate both principles more explicitly, one may therefore write both principles in the following schematic form, where t and t' now are to be read as universally quantified variables (that is each of the formulas in the rest of this section is implicitly prefixed by $\forall t \forall t'$.):

(3) $KP(t) \rightarrow (P(t') \rightarrow \neg B \neg P(t'))$

(4) $KP(t) \rightarrow (P(t') \rightarrow BP(t'))$

The first condition says that, in order to know that t is P, one can't believe of some t' that is also P that it is not P. The second condition says that, in order to know that t is P, one has to believe of every t' that is P that it is P. These principles, however, seem to put very strong requirements on what it takes to know something. I know, for instance that 5 is a prime number. Do I believe, of any other number that is a prime, that it is a prime? This certainly is not something I believe explicitly, if we think of those very large numbers about which I have not the least idea whether they are prime or not. Hence condition (4) seems prima facie too strong, since it says: in order to know that a property holds of an object, one ought to have correct beliefs about its complete extension.

On the other hand, condition (3) might be easier to fulfill. Suppose there is some number of which I believe that it is not a prime. Then it ought not to be a prime. The standards by which I judge that 5 is a prime should also secure as true my belief that such and such distinct number is not a prime. Still, this rules out the possibility of being fallible while retaining knowledge. For instance, I know that Paris is a capital. But suppose I believe that Washington DC is not a capital (because I think New York City is the capital of the US): should that prevent my belief that Paris is a capital from constituting knowledge?

In practice, however, requirements as the ones we just stated certainly do hold of some particular instances. For instance, if someone believes that 5 is prime but does not believe

that 7 is prime, we would certainly deny her the knowledge that 5 is prime. In that situation, 7 is not only similar to 5 by the fact that it is a prime, it is also very close to 5 in the sequence of natural numbers. Thus, in the case of 5 and 7, most people would probably agree that a condition like (4) does hold, namely: you don't know that 5 is a prime if you don't hold the belief that 7 is a prime. Maybe the same condition applies to 11 with respect to 7, and then to 13 with respect to 11. However, it certainly does not hold of 577, which is also a prime, with respect to 5. The idea is that 577 is already quite far from 5 in the sequence of natural numbers in order to count as similar enough for a condition like (4) to apply. Likewise, I might temporarily hold the wrong belief that 1321 is not a prime, without this impugning on my knowledge that 5 is a prime, in which case even the weaker condition (3) would fail to hold.

What we ought to do, therefore, is relativize Socrates's requirements to relevant standards of similarity. This kind of relativization is needed if we want to explain that we are capable of knowledge without being omniscient (condition (4)), and also that we are capable of knowledge while being fallible (condition (3)). Thus, to use an example much similar to Socrates's original example, certainly most speakers of French who know how to write and read also know that the word "théâtre" starts with "th". But then, they should know that the adjective "théâtral", which is morphologically related, also starts with "th". Yet would we say of the same educated person who thinks that the word "thuya" is spelt "tuya" (and who knows the meaning of the word), that they don't really know that "théâtre" starts with "th" after all? Maybe some purists would say this, but most people certainly would not. What this suggests is that the standards by which we consider that a piece of knowledge should depend on some other crucial set of beliefs depend on what we count as relevantly similar cases. In this respect the word "thuya" is less similar to "théâtre" than "théâtral" is, not being morphologically related. As a consequence, a more adequate formulation of conditions (4) and (3) above might be:

(5) $KP(t) \rightarrow (t \sim t' \rightarrow (P(t') \rightarrow \neg B \neg P(t')))$

(6) $KP(t) \rightarrow (t \sim t' \rightarrow (P(t') \rightarrow BP(t')))$

where $t \sim t'$ means that t and t' are relevantly similar.

Conditions (3) and (4) correspond to normative principles about knowledge. Ideally, in order to reliably know that some property holds of an object, one ought to know the complete extension of that property. Realistically, however, to reliably know that some property holds of an object is to be able to discriminate whether the property holds or not of sufficiently similar cases. Thus, conditions (5) and (6) are relativized in a way which makes them descriptively more adequate. Condition (5) says that in order to know that t is an instance of P, then one should not have false beliefs about some t' which is also P and which is relevantly similar to t. Condition (6) requires that one should moreover have correct beliefs about any t' which is also a P and which is relevantly similar.

Once again, the standards which govern this notion of similarity may vary. For a demanding number theorist, for instance, the actual knowledge of the first 100 primes may be required to say that someone reliably knows that 5 is prime ("How can you know that 5 is prime if you don't know that 1321 is prime?"). More plausibly, however, a well-trained mathematician would probably consider that to know that 5 is prime is to master the definition of "prime" in such a way that conditions (3) and (4) are in principle satisfiable. Practically, however, what seems required is that someone who is able to apply Euclid's algorithm (for instance) to

the case of 5 is also able to apply it to "neighbouring" numbers, numbers for which the calculations don't take significantly more time. As we know, however, the distribution of prime numbers is such that, to continue to apply the algorithm, it takes more and more calculations in order to know that the next prime is indeed the next prime. If two primes count as similar enough when the calculations needed to establish that they are prime take about the same time, then this suffices to explain that we can know that n is prime without yet knowing which is the next prime.

By parity of analysis, we may wonder if we can say of someone who truly believes that 5 is a prime, but who wrongly believes that 6 is a prime, that they really know that 5 is a prime. Thus, to (3) and (4) there correspond two further conditions:

(7) $KP(t) \rightarrow (\neg P(t') \rightarrow \neg BP(t'))$

(8) $KP(t) \rightarrow (\neg P(t') \rightarrow B\neg P(t'))$

The first condition requires that in order to know that t is P, one should not have false beliefs about things that are not P. The second again is stronger, since it requires that one does believe then that they are not P. We know from logic that the latter condition is hard to come by, since the standards by which I judge that something is a P may sometimes not be sufficient to ascertain that something is not a P. If we think of P as the property of being a logical validity of first-order logic, for instance, a condition like (7) can be secured by the use of a complete proof method, but not a condition like (8), since there is no decision method for first-order logic.

In practice, however, we can get more realistic approximations of conditions (7) and (8) if we restrict them by reference to a range of similar cases. For instance, if I know that this formula is a logical validity, I ought to believe that this other formula which requires about the same number of proof steps is also a logical validity. And analogously for a sufficiently similar formula (in the same sense of similar) that is not a logical validity. Thus conditions (7) and (8) can be modified in the way we modified (3) and (4), namely:

(9) $KP(t) \rightarrow (t \sim t' \rightarrow (\neg P(t') \rightarrow \neg BP(t')))$

(10) $KP(t) \rightarrow (t \sim t' \rightarrow (\neg P(t') \rightarrow B\neg P(t')))$

Putting together conditions (5), (6), (9) and (10), we can summarize the point of our analysis in the following way: the more robust our knowledge that some object t is P, the more it will extend to similar cases that are P, and the more it will discriminate between similar cases that are not P. Those standards of similarity may vary, but here again, the more inclusive the relation of similarity will be, the stronger our knowledge should be too.

1.2 Reliability as safety vs. reliability as robustness

The remarks of the previous section were given in order to introduce the general analysis Williamson has presented of the notion of reliability of knowledge. Williamson writes that "Reliability resembles safety, stability, and robustness" and points out that "these terms can all be understood in several ways" (2000, 124). In this section, we shall actually refine the distinction between the notions of safety and the notions of robustness in the case of knowledge, on the basis of Williamson's analysis of the notion of reliability. The idea is quite

natural: safety of knowledge is defined by Williamson as avoidance of false belief for similar cases. We define robustness as propensity to true belief for similar cases. Both notions can be related in a natural way to Nozick's analysis of the reliability of knowledge in terms of sensitivity.

According to Williamson, "something happens reliably in a case α if and only if it happens (reliably or not) in every case similar enough to α" (2000, 124). The consequence Williamson draws in the case of knowledge is the following: "if one believes p truly in a case α, one must avoid false belief in other cases sufficiently similar to α in order to count as reliable enough to know p in α" (2000, 100). This condition can be expressed in propositional modal logic, in a way which establishes a close connection to our previous remarks on the mastery of concepts. Williamson's requirement can indeed be stated in the following way (letting t and t' denote arbitrary contexts – I shall talk indifferently of *contexts*, *cases* or *worlds* in what follows):

(11) $t \models Kp$ only if for every t' such that $t \sim t'$ and $t' \models p$, $t' \models \neg B \neg p$

(12) $t \models Kp$ only if for every t' such that $t \sim t'$ and $t' \models \neg p$, $t \models \neg Bp$

These two conditions correspond to the two conditions (5) and (9) of the previous section, here repeated as:

(13) $KP(t) \rightarrow (t \sim t' \rightarrow (P(t') \rightarrow \neg B \neg P(t')))$

(14) $KP(t) \rightarrow (t \sim t' \rightarrow (\neg P(t') \rightarrow \neg BP(t')))$

The difference between the two formulations is that we now identify objects with contexts, and properties with propositions (as is standard in modal logic). Following a suggestion originally made by J. van Benthem and investigated by I. Arnesen (see section 1.3 below), conditions (11) and (12) can be given yet a more compact form if we express the notion of similarity between contexts by means of an explicit closeness or neighborhood modality. We shall note $\Box$ this modality. Following Arnesen, we assume it is a normal operator, satisfying $t \models \Box p$ iff for all t' such that $t \sim t', t' \models p$. Using this operator, the previous conditions can be rephrased in the following way:

(15) $Kp \rightarrow \Box(p \rightarrow \neg B \neg p)$

(16) $Kp \rightarrow \Box(\neg p \rightarrow \neg Bp)$

Those two conditions correspond to Williamson's concept of safe belief (note that the strict conditionals in each of the consequents are equivalent if p is given a schematic reading, but (15) and (16) need not be). Condition (16), in particular, as Williamson observes (2000, 148-150), is closely related to one of the conditions set forth by Nozick in his analysis of knowledge.[2] According to Nozick, knowledge is true belief, which is counterfactually sensitive to truth. Thus, for X to know that p, X has to be in a state such that, if p were not the

[2]Williamson's concept of safety is even more closely related to the concept of safety introduced by E. Sosa (see Williamson 2000, 151), which Sosa contrasts with Nozick's notion of sensitivity and which he defends as a more plausible requirement on knowledge. See in particular Sosa (1999), where a belief is called safe if and only if it satisfies the contrapositive of Nozick's condition 3, that is: $Bp > p$ (X would believe p only if it were the case that p). Sosa therefore defines safety in terms of a counterfactual conditional, and not as a strict conditional.

case, X would not believe p. Representing the counterfactual conditional by the symbol $>$, Nozick's condition 3 can be stated $(\neg p > \neg Bp)$ and leads to the following requirement on knowledge (which I also refer to as "condition 3" eponymously):

(17) $Kp \rightarrow (\neg p > \neg Bp)$ (Nozick's condition 3)

Nozick's condition 3 is intended to exclude standard Gettier cases. In the case of Russell's example, for instance, supposing that X already believes the name of the Prime Minister is "Balfour", while it is in fact "Bannerman", it is plausible enough that if the name of the Prime Minister started with an "N", X would still believe that it is "Balfour", and therefore would still believe that it starts with a "B". Consequently, his belief displays a lack of sensitivity to truth, and X can't be said to know that the name of the Prime Minister starts with a "B" in the actual context.

Unlike Nozick's sensitivity condition, Williamson's condition is stated in terms of a strict conditional, with a neighborhood modality expressing similarity between contexts. Nozick's counterfactual conditional, if understood relative to a Lewis-Stalnaker semantics for conditionals, makes a distinct requirement on the possession of knowledge. Indeed, condition (16) can be expressed in contraposed form as: $Kp \rightarrow \Box(Bp \rightarrow p)$. This requires only that one's belief be true at all the worlds that are relevantly similar to the actual world where p is known. This is not sufficient to guarantee that one's belief will not be false at the closest worlds where $\neg p$ holds, however. Likewise, as Williamson notes, $Kp \rightarrow (\neg p > \neg Bp)$ could be true at the actual world, without $Kp \rightarrow \Box(\neg p \rightarrow \neg Bp)$ being true at that world: it suffices to imagine that $\neg Bp$ holds at the closest $\neg p$ worlds, without holding at all relevantly similar $\neg p$ worlds.

Williamson's safety conditions yield an analysis of the notion of reliability of knowledge which therefore differs from Nozick's. As we suggested with the examples of the previous section, a safe belief that p is a belief which can be persistently true, if not with respect to the whole set of worlds where p is true, at least with those p worlds that are relevantly similar to the actual world. As we tried to motivate by analogy with the case of concepts, one properly knows that a property holds of an object only if one is able to maintain that belief for a relevant set of objects that also instantiate the property, and also if one is able to exclude a relevant set of objects which do not fall under the property.

We can imagine, in this respect, stronger conditions than the ones considered by Williamson, namely:

(18) $Kp \rightarrow \Box(p \rightarrow Bp)$

(19) $Kp \rightarrow \Box(\neg p \rightarrow B\neg p)$

These two conditions (18) and (19) stand to (15) and (16) as (6) and (10) stood to (5) and (9) respectively. They now require that one does hold true beliefs at all relevantly similar worlds to the world where p is known, and not simply that one avoids false beliefs at such worlds. If we follow Williamson's analysis, safety is a negative concept, much like immunity to error. Positive conditions like (18) and (19) rather express some notion of robustness. To clarify the examples we introduced in the previous section, we could say that one has robust knowledge about prime numbers if one's knowledge about primes is sufficiently inclusive, both positively ($\Box(p \rightarrow Bp)$) and negatively ($\Box(\neg p \rightarrow B\neg p)$). Conversely, we could say that one has safe knowledge about prime numbers if one's knowledge is sufficiently immune to error,

both positively ($\Box(p \rightarrow \neg B\neg p)$) and negatively ($\Box(\neg p \rightarrow \neg Bp)$). Williamson says relatively little of this notion of robustness in his analysis, even though the definition we suggest is quite natural.

Safety and robustness are indeed dual notions. Robust knowledge is also safe knowledge, and unsafe knowledge can't be robust. Yet some knowledge may be safe without being robust. To get the correct entailment patterns, we need to make the assumption that beliefs are coherent in the following sense: someone who explicitly believes $\neg p$ ought not to be believe p (namely $B\neg p \rightarrow \neg Bp$). Figure 1 gives a representation of this duality ($\neg p$ and p can be exchanged uniformly to get the negative instance of Robustness and the positive instance of Safety).

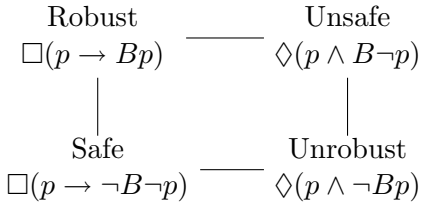

Figure 1: Safety and Robustness

An example of safe knowledge which is unrobust is given in a case in which Kp holds at world w, and yet such that there is a relevantly similar world w' in which p holds, and such that neither p nor $\neg p$ are believed at that world. This is a situation of agnosticism in which one's knowledge is true and immune to error, and yet is not as inclusive as it could be.

Interestingly, a fourth condition in Nozick's counterfactual analysis of the reliability of knowledge bears a connection to the notion of robustness we just introduced. This condition is expressed by Nozick as $(p > Bp)$ and makes the following requirement on knowledge:

(20) $Kp \rightarrow (p > Bp)$ (Nozick's condition 4)

Nozick's condition 4 states that, in order to know p, one has to believe p at all contexts which are alternative to the actual context and in which p also holds. Suppose for instance that X believes that the name of the Prime Minister is Bannerman, because he read it in the newspaper, and it is indeed Bannerman. Now, if the context were different, and the newspaper had mistakenly published "Balfour" instead of the correct name "Bannerman", X still ought to believe that the name of the Prime Minister is "Bannerman" and not "Balfour". This means that X's belief should be reliable in a way that makes it truth-sensitive even in cases in which the information channel goes wrong.

As emphasized by Williamson, Nozick's condition 4, unlike condition 3, is not adequately captured within the standard Stalnaker-Lewis semantics, since if knowing p entails believing p truly, the actual world is already the closest world where p holds. Williamson therefore suggested that Nozick's truth-conditions for the conditional $(p > Bp)$ should be equivalent to those of the strict conditional $\Box(p \rightarrow Bp)$, whenever p already holds at the world of evaluation (see Williamson 2000, 149). If this asymmetry in truth-conditions is granted (a variably strict conditional when the antecedent is false at the actual world, and a strict conditional when the antecedent is true at the actual world), then Nozick's condition 4 is formally equivalent to what we defined as robustness.

This common logical form should not hide some important conceptual differences, however. The meaning of Nozick's condition 4 would probably be more adequately expressed by reference to a particular "method of coming to believe", or to some verb expressing the way information is acquired by the believer. Let us write Hp to mean that X is given (truly or not) the information that p. We shall call "being told p" this property. The notion of robustness expressed through condition 4 entails that, in the case where I know p, if p were true and if I were to be told not p, I would *nevertheless* believe p. For instance, I know that 2+2=4. If someone were to tell me now that 2+2 is not equal to 4, I would still believe that 2+2=4. My knowledge that 2+2=4 is robust in that sense. Thus Nozick's fourth requirement, as stated in (20), may be expressed more explicitly by:

(21) $(Kp \wedge \neg H\neg p) \rightarrow ((p \wedge H\neg p) > Bp)$

That is: in order to know p in a case in which I'm not told not p, if I was told deceptively that it is not the case that p, I would still believe p. A belief of that kind is a belief that is robustly true, true against false appearances or deceiving evidence. This formulation is now compatible with Stalnaker-Lewis truth-conditions, and asks that one consider the most similar worlds to the actual world in which one's evidence about the same proposition is significantly different.

This notion of robustness concerns the *justification* of one's knowledge, and contrasts with the one we presented as a dual to Williamson's conception of safety, which concerns rather the *extension* of one's knowledge. Indeed, in the case of safety, Williamson describes the concept as saying that "if at time t on basis b one knows p, and at time t^* close enough to t on basis b^* close enough to b one believes a proposition p^* close enough to p, then p^* should be true" (2000, 102). This concerns the way in which one's actual knowledge is likely to be maintained for similar cases, assuming one's basis or justification is itself sufficiently close to the actual one. Nozick's condition 3, by contrast, asks to consider worlds in which one's evidence is likely to change with the world itself.

Analogously, the notion of robustness we introduced differs from condition 4 in much the same way in which safety differs from condition 3. Thinking back to the example of Socrates: my belief that "Theaetetus" starts with "Th" is robustly true if it is likely to elicit the belief that "Theodorus" also starts with "Th". This notion of robustness, we should note, is characteristic of any form of competence, or "knowing how": you are good at solving conics if, when faced with a new instance of a polynomial equation of the same degree, you are able to solve it in much the same way in which you solved the previous cases you were confronted with. This says something of the way one's actual justification for knowledge is likely to cover new cases, and not something about the way in which one would or would not revise one's beliefs if one were told that the evidence or method one actually uses is bad in the first place.

The results of this section can be summarized in the following way. First, the notion of knowledge reliability has two sides. We called the first safety, following Williamson, and the second robustness. As we argued, those two notions are dual to each other. This duality is already present in the analysis Nozick gave of the notion of the reliability of knowledge in terms of counterfactual sensitivity, but has a different meaning. Nozick's counterfactual analysis of the notion of knowledge reliability contrasts with Williamson's closeness analysis in so far as the counterfactual analysis deals with the strength of one's belief in the light of new or conflicting evidence, whereas the closeness analysis deals with the stability or potential extension of one's knowledge with respect to further data.

1.3 Margin for Error

The notions of safety and robustness which we discussed in the previous section are related to yet another concept of reliability, which plays an essential role in Williamson's epistemology, the notion of margin for error. The margin for error principle, as we will show in this section, can be seen as a strengthening of Williamson's safety condition. It also constitutes a generalization of the principle of factivity. More specifically, however, it was stated by Williamson in order to describe the notion of inexact or approximate knowledge.

Margin for error principles were first introduced by Williamson in his epistemic account of vagueness (Williamson 1992, Williamson 1994). Ordinary judgements involving scalar predicates, namely predicates like "red" or "bald", which can be mapped to a numerical scale, do not go by sharp distinctions. We know for sure that someone with fewer than 10 hair on their head is bald, and for sure that someone like Mick Jagger, whom we may suppose to have more than 100000 hair on his head, is not bald. The trouble lies with intermediate cases. According to the epistemic account of vagueness, vagueness comes from the limitation of our discriminative capacities. To say that our discriminative capacities are limited is to say that they go with a certain margin of error, which stands for the fine-grainedness with which we are able to make discriminations. Suppose for instance that only a difference of more than 1000 hair makes a difference in our perception of hairiness (or rather baldness, for that matter). Under the supposition that the cut-off point between bald and not bald is at 50000 hair, then we may not be able to tell that someone with 49500 hair on their head is bald. The reason is that someone with only 1000 more hair, namely 50500 hair, is not bald. In this way, we can explain our inability to say of individuals standing around the cut-off point between bald and not bald whether they are bald or not.

Conversely, whenever we know that someone with n hair on their hair is bald, then it has to be the case that someone with $n+1000$ hair on their head is still bald. Let us represent by $bald(n)$ the sentence that someone with n hair on their head is bald. Factivity of knowledge says that, in order to know that someone with n hair on their head is bald, it has to be the case that someone with n hair on their head is bald:

(22) $K(bald(n)) \rightarrow bald(n)$

By analogy, the margin for error principle can be seen as a generalization of factivity to neighboring cases. In the previous example, for instance, a margin for error of 1000 hair requires that:

(23) for all i such that $|n - i| \leq 1000$, $K(bald(n)) \rightarrow bald(i)$

Treating numbers as contexts and matching the predicate *bald* to the propositional atom p, this can be expressed in modal terms:

(24) $n \models Kp$ only if for all i such that $|n - i| \leq 1000$, $i \models p$

In Williamson (1994) and Williamson (1997), Gomez-Torrente (1997), and Graff (2000), margin for error principles like (24) are imposed at the semantic level on the knowledge operator by giving the epistemic accessibility relation a topological interpretation. Williamson defines a fixed-margin model as a quadruple $\langle W, d, r, V \rangle$, such that W is a set of worlds, V a propositional valuation over W, d a distance function between worlds of W, and r a positive real value. In such a model,

(25) $w \models K\phi$ iff for all v such that $d(w, v) \leq r,\ v \models \phi$

Every fixed-margin model can be seen as a standard Kripke model $\langle W, R, V\rangle$ in which the epistemic accessibility relation R satisfies: wRv iff $d(w, v) \leq r$. A limitation of this representation, however, is that it incorporates the margin for error principle directly at the semantic level, without giving it a corresponding syntactic form. To get an explicit version of the principle, a different move consists in using a closeness modality, as was done originally by Arnesen, following a suggestion of van Benthem, as we did already in the previous section. Instead of using a distance function, one can then use a similarity relation $\sim$ between worlds. To get the correspondence with fixed margin models, it then suffices to set: $w \sim v$ iff $d(w, v) < r$, for some fixed r. We then rephrase the principle by weakening the biconditional in (25) to a conditional:

(26) $w \models Kp$ only if for all w' such that $w \sim w',\ w' \models p$

This yields the van Benthem-Arnesen formulation of the principle, namely:

(27) $Kp \rightarrow \Box p$ (Margin for Error)

Margin for Error imposes that, in order to know that p holds, p hold in all contexts that are relevantly similar to the actual one. Williamson himself presents the margin for error principle as closely tied to the notion of reliability, by writing: "Where one has a limited capacity to discriminate between cases in which p is true and cases in which p is false, knowledge requires a margin for error: cases in which one is in a position to know p must not be too close to cases in which p is false, otherwise one's beliefs in the former cases would lack a sufficiently reliable basis to constitute knowledge" (2000, 17). We should note here that this notion of reliability is stronger than the notion of safety discussed in the previous section. We could in principle imagine that at world w, I know p, and that in some case w' very close to w, p is false, without my believing that p is true at w'. This would be an instance of safe belief in the sense previously discussed. However, that kind of safety seems compatible only with a notion of exact knowledge, capable of drawing sharp boundaries. When knowledge becomes inexact, Williamson's remark suggests that one's beliefs require more safety. This is the case by assuming Margin for Error, since Safety (as expressed by the schema $Kp \rightarrow \Box(Bp \rightarrow p)$) logically follows from Margin for Error in standard modal logic. Indeed, for $\Box$ a normal operator, $\Box(q \rightarrow r)$ follows from $\Box r$, hence:

(28) $(Kp \rightarrow \Box p) \rightarrow (Kp \rightarrow \Box(Bp \rightarrow p))$

The converse, however, does not hold. To derive Margin for Error from Safety, it is sufficient to assume a property like $Kp \rightarrow \Box Bp$, which expresses that whenever I know p, I believe p holds at all close cases. This condition, which we might call Persistency, may be characteristic of belief in situations of approximate knowledge. For instance, if I know that someone is less than 2 meters tall, maybe I also believe that she is less than 1,99 meters tall (although I don't know this for a fact). Together with Margin for Error, Persistency would entail that whenever Kp holds, p is both true and believed to be true at all sufficiently similar cases.

A condition like $Kp \rightarrow \Box Bp$ would be unreasonable to postulate in general, however, since it would entail $Kp \rightarrow \Box(\neg p \rightarrow Bp)$, namely that whenever p is known, p is believed

even at similar cases where p does not hold, a situation of knowledge that would not only be unsafe, but systematically deviant for close cases. For instance, in the case of knowledge about prime numbers, which is a variety of exact knowledge, this would predict of someone who knows that 5 is a prime that they also believe that 6 and 8 are prime, assuming the latter are relevantly close to 5. In other words, Persistency can be assumed safely for beliefs that already go about with a margin for error, but not for beliefs whose content is supposed to be exact.

This raises the question of the validity of margin for error principles in general. Williamson insists that a margin for error is required "where one has a limited capacity to discriminate between cases in which p is true and cases in which p is false". We could argue that any kind of knowledge is in principle subject to such limitations. We certainly have limited capacities to discriminate whether a number is a prime, or whether a formula of first-order logic is a validity or not, despite the fact the the latter is a recursively enumerable property, and the former even a recursive property. The reply, however, is that the kind of limitation in question is not comparable to the limitation we experience in perception. Predicates like "prime" or "logically valid" are not vague predicates, and we are able in principle to discriminate whether any number is a prime or not. Our knowledge can be exact and yet be subject to all sorts of practical limitations (of time, speed, memory and so on), without these limitations giving rise to categorial quandaries of the kind we experience for vague predicates.

This means that margin for error principles are specific of inexact knowledge, in the sense intended by Williamson. This contrasts with the notions of safety and robustness which we discussed earlier. These principles make sense also for the kind of exact knowledge involved in logic or arithmetic, as we tried to illustrate. There remains an area, on the other hand, for which the question is moot whether it can be a matter of exact knowledge, namely the area of self-knowledge and introspection. Williamson has argued that introspective knowledge is also subject to a margin for error, and for that reason that knowledge is not transparent or luminous in general. This will be the object of the next section to discuss the extent to which self-knowledge is subjected to the sort of limitation he claims.

2 The reliability of self-knowledge

Traditionally, knowledge is seen as a reflexive capacity satisfying the principle of positive introspection, which says that if one knows p, one knows that one knows p. This principle was defended in particular by Hintikka (1962) as one of the fundamental properties of knowledge: how could I know that 2+2=4 without knowing that I know it? Hintikka mentions a number of philosophers who have argued in favor of the same thesis. Spinoza probably remains the most famous, when he claims that whoever has a true idea knows that he has a true idea (*Ethics*, II, prop. 43).[3] Some counterexamples have been considered, but they remain inconclusive. For instance, a student who lacks self-confidence might know the answer, and yet fail to believe he knows. It is unclear, however, to what extent he really knows the answer, if his failure to believe he knows is in fact due to the belief that the answer might be wrong.[4] Hence the

[3]By "true idea", Spinoza means an idea that is "adequate", intending to rule out the case of ideas that would be true simply by accident.

[4]See Lewis (1996, 429) on this example. Lewis writes: "I even allow knowledge without belief, as in the case of the timid student who knows the answer but has no confidence that he has it right and so does not believe what he knows." Lewis does not say, however, whether or not knowing that one knows requires believing that one knows.

reliability of my believing that I know p seems inherited from the strength of my knowing p.

In Williamson (2000), Williamson presents an indirect argument against the principle of positive introspection, whose goal is to be radical. The argument rests on the idea that self-knowledge obeys a principle of margin for error, and on the observation that the conjunction of margin for error and positive introspection leads to paradoxical conclusions. The aim of this section will be to challenge the soundness of his argument, and to examine to what extent the knowledge one can gain by introspection is necessarily subject to the limitations claimed by Williamson. The first paragraph gives a presentation of Williamson's argument. In the next one, I present the criticism Dokic & Égré (2004) made of this argument by defending the idea that knowledge is modular, and that two forms of knowledge, which we might call perceptual and reflective, ought to be distinguished in order to state one of Williamson's premises. I will conclude by the discussion of some objections to the present account.

2.1 Williamson's paradox against luminosity

Williamson's argument against positive introspection is part of a more general criticism of the idea that knowledge should be *luminous*. Luminosity is the thesis that whenever I am in a given mental state, I know that I am in that mental state. For instance, luminosity says that if I am in the mental state of being cold, then I know that I am cold (or I am in a position to know that I am cold).[5] The principle of positive introspection is a particular case of the luminosity principle (granting that there are mental states of knowing, an assumption I shall not dispute here), since it says that, if I am in a state of knowing something, then I am also in a position to know that I know that thing. Since positive introspection is luminosity concerning knowledge itself, this means that in principle, one could accept the thesis that luminosity does not hold in general, while defending the idea that knowledge is nevertheless positively introspective. In *Knowledge and its Limits*, Williamson offers to challenge both claims. In chapter 4 of his book, he states a general argument against luminosity. In chapter 5, he states the same argument in the case of knowledge. Both arguments are given in the form of puzzles, and take the form of sorites arguments.

Williamson's general argument against luminosity trades on the assumption that our judgements on the occurrence or non-occurrence of specific mental states are not a matter of exact knowledge. Presumably, most of our mental states gradually appear and disappear, so that our judgements on their occurrence or non-occurrence go about with a certain margin of error. In the case of a mental state like "feeling cold", for instance, we have the following particular instance of margin for error:

(29) If at time t one knows that one feels cold, then at time $t+1$ sufficiently close to t, one feels cold.

Then, assuming one feels cold at t, and that this state is luminous, this entails that one knows one feels cold at t. But by the previous instance of margin for error, this entails that one feels cold at $t+1$, and by luminosity one knows that one feels cold at $t+1$. By induction, if one feels cold at some time, one should feel cold at all future times, which *de facto* is not the case. This leads Williamson to reject luminosity, given that margin for error seems a reasonable postulate in this context.

[5] Like Williamson, I shall talk indifferently of knowing that one knows or being in a position to know that one knows (see Williamson 2000, 95). Likewise, I use the notions of transparency and luminosity interchangeably in this paper (compare Williamson 2000, 24 and 95 where the terms are given essentially the same definition).

Williamson's rejection of positive introspection in chapter 5 rests on a similar argument. Williamson considers a myopic character, Mr Magoo, who observes a tree at some distance and makes judgements about its height. Magoo's knowledge is constrained by the following margin of error principle:

(30) If Magoo knows that the tree is not of size i, then the tree is not of size $i+1$

Williamson (2000, 115) makes the further assumption that 'Mr Magoo reflects on the limitations of his eyesight and ability to judge heights', so that the previous principle itself is known by Magoo. It is further assumed that Magoo's knowledge is closed under logical consequence, namely that if Magoo knows all the propositions of some set of propositions, he also knows any proposition that follows logically from that set.[6] If furthermore Magoo's knowledge is positively introspective, one reaches a paradoxical conclusion. Let us represent by p_i the sentence "the tree is i inches tall". The knowledge of Magoo is taken to obey the following principles:

(31)

(ME)	$K\neg p_i \rightarrow \neg p_{i+1}$
(KME)	$K(K\neg p_i \rightarrow \neg p_{i+1})$
(KK)	$K\phi \rightarrow KK\phi$
(C)	If ϕ follows logically from a set of propositions Γ, and for all members ψ of Γ, $K\psi$ holds, then $K\phi$.

Supposing Magoo knows that the tree is not i inches tall for some value i, we get the following derivation:

(32)

(i)	$K\neg p_i$, by hypothesis
(ii)	$K(K\neg p_i \rightarrow \neg p_{i+1})$, by (KME)
(iii)	$KK\neg p_i$, by (i) and (KK)
(iv)	$K\neg p_i, K\neg p_i \rightarrow \neg p_{i+1} \models \neg p_{i+1}$, by propositional reasoning
(v)	$K\neg p_{i+1}$, by (ii), (iii), (iv) and (C)

By induction, granted that Magoo knows that the tree is not of size 0, he knows that the tree is not of any positive size whatsoever, which can't be the case if knowledge is factive. Thus we reach a paradoxical conclusion with regard to a scale involving degrees of height, in the same way in which we reached a paradoxical conclusion with respect to times.

Williamson's general argument against luminosity and his specific argument against positive introspection exemplify a common structure which can be seen more perspicuously if each argument is rephrased within the framework of closeness modalities. Let us represent by p the proposition that one feels cold, and let us fix a model consisting of time points linearly ordered in which $t \sim t'$ iff $t' = t+1$. Condition (29) is then equivalent to:

(33) $t \models Kp$ only if for every t' such that $t \sim t'$, $t' \models p$

[6] Williamson assumes that this closure principle can be restricted to the propositions "pertinent to the argument". The closure assumption is certainly controversial in general, as was pointed out to me on several occasions, but it is a natural one to make if one assumes as underlying semantics a Kripke-Hintikka style semantics for knowledge, as we do below. In my opinion, a solution to Williamson's paradox does not call for a revision of that particular principle, even though one may wish to put it into question on other grounds.

This yields the van Benthem-Arnesen formulation of Margin for Error, namely $Kp \rightarrow \Box p$. To say that p is luminous is to say that p satisfies the principle $p \rightarrow Kp$. We thus get the immediate derivation:

(34) $\dfrac{\begin{array}{rl} Kp \rightarrow \Box p & \text{(Margin for Error)} \\ p \rightarrow Kp & \text{(Luminosity of } p\text{)} \end{array}}{p \rightarrow \Box p}$

Within the model, it therefore follows that if $t \models p$, then $t+1 \models p$, which yields the pernicious sorites progression. The argument against positive introspection can be restated in much the same way. Assuming margin for error holds uniformly of any complex proposition, it follows in particular that $KKp \rightarrow \Box Kp$, a higher-order form of margin for error which we may call reflective margin for error. The soritic conclusion follows from positive introspection along exactly the same lines:

(35) $\dfrac{\begin{array}{rl} KKp \rightarrow \Box Kp & \text{(Reflective Margin for Error)} \\ Kp \rightarrow KKp & \text{(Positive Introspection)} \end{array}}{Kp \rightarrow \Box Kp}$

The latter condition yields the pernicious sorites progression in the puzzle of Mr Magoo. Consider a model consisting of positive sizes that are linearly ordered. Each world i represents a world in which the size of the tree is of i inches. Suppose $i \sim j$ iff $|i-j| \leq 1$, giving a one-unit margin for error corresponding to a reflexive and symmetric similarity relation (a slight and natural generalization of condition (ME) in (31) above). Figure 2 represents such a structure of inexact knowledge, in which each circled area represents the margin of error around the middle-point, except of course for 0 where it contains only 0 and 1. The overlap between the areas explains that one gets a sorites progression. Let p_i denote the proposition that the tree is i inches tall. By definition, p_i is supposed to hold only at world i. Supposing the tree is 66 inches tall, it holds that $66 \models K\neg p_0$, from the assumption that, in this case, Magoo sees the tree is not of size 0. Then from $Kp \rightarrow \Box Kp$, $66 \models \Box K\neg p_0$, and so $65 \models K\neg p_0$. By repeated applications of the same condition, we reach the conclusion that $1 \models \Box K\neg p_0$, and so $0 \models K\neg p_0$, which can't be the case if knowledge is factive. The contradiction is even more direct, since if $1 \models K\neg p_0$, it will follow from Margin for Error that $1 \models \Box\neg p_0$ and therefore that one should have $0 \models \neg p_0$. The outcome will of course be the same, however large the size of the tree.

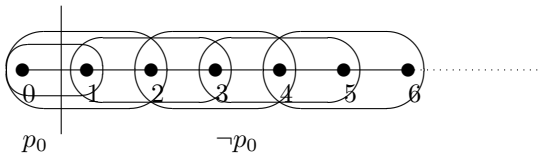

Figure 2: A structure of inexact knowledge (1)

This last example is closely related to Gomez-Torrente's observation that if a proposition like "a man with 0 hair on their head is bald", which can be represented by $B(0)$, is epistemically transparent (it is known, it is known that it is known, and so on), then one's

knowledge can't satisfy all the margin for error principles $\forall n(K^{m+1}B(n) \rightarrow K^m B(n+1))$, where K^m stands for m iterations of K (Gomez-Torrente 1997, Graff 2002). For otherwise it would follow that even a man with a million hair on his head is bald. The modal schema $Kp \rightarrow \Box p$ gives another representation of these margin for error principles. If we let the atom p represent the property of "being bald", to say that the proposition "a man with 0 hair is bald" is transparent is to assume that $K^n p$ would hold for all n at world 0 in a linear structure similar to that of Figure 2 (setting $i \sim j$ iff $j = i + 1$). From the normality of $\Box$, however, it follows that the schema $K^n p \rightarrow \Box^n p$ holds for all n, which would constrain p to hold at all worlds in the structure, including at world 1000000. More generally, $p \rightarrow \Box^n p$ holds for all luminous p, assuming Margin for Error. The idea that at least some propositions might be transparent has therefore led Gomez-Torrente and Graff to question the validity of margin for error principles, although they have not discussed the validity of positive introspection *in general* in situation of inexact knowledge: one may conceive indeed that positive introspection might hold for some restricted class of known propositions (which are then transparent), but not systematically.

Before discussing this point more thoroughly, we should note that a property like $p \rightarrow \Box p$ is a potential source of paradox, depending on the semantic status of p and the way the relation $\sim$ is defined within a knowledge structure. For instance, if p is to represent "I feel cold", and is supposed to hold for some finite set of thermometric degrees in a linear structure analogous to that of Figure 2, then assuming knowledge brings about the same kind of fixed margin for error, the property constrains the value of p to stretch out of its intended extension. On the other hand, if p is a permanent proposition, no paradox is threatening. To illustrate the point, let us suppose that p stands for "2+2=4", and that the scale of similarity is constituted by contiguous instants on the temporal line. Then p holds everywhere in the model. Since p holds everywhere, the margin for error principle $Kp \rightarrow \Box p$ is trivially satisfied, and the conclusion $p \rightarrow \Box p$ is harmless.

Thus whether there is a paradox or not depends on the nature of the property involved. As Leitgeb (2002, 200) notes, Williamson's anti-luminosity argument is rightly called a paradox, since it is a valid argument starting from plausible premises and yielding an implausible conclusion. The argument can be seen as a form of epistemic sorites. Importantly, however, Williamson insists that his argument does not rest essentially on the vagueness of predicates like "feeling cold" or "knowing", but on the fact that "we have limited powers of discrimination amongst our own sensations" (2000, 104). We can agree with him that vagueness is not essential, but it remains essential that the predicates or propositions involved do not hold universally (see Williamson 1994, 271-272). As the semantic version of his arguments given in (34) and (35) makes clear, a contradiction follows only if enough semantic assumptions are made on the interpretation of p and $\Box$ so that the schema $p \rightarrow \Box p$ gives rise to a pernicious induction.

The case is similar with the schema $Kp \rightarrow \Box Kp$. Take the same eternal proposition p and interpret closeness by temporal contiguity over instants ($t \sim t'$ iff $t' = t + 1$). This says that if I know that 2+2=4, then I know that 2+2=4 at all subsequent times. No paradox arises here. The situation is quite different if p stands for "I feel cold". For in the same model, if I know that I feel cold at time t, then at all times t' subsequent to t, I should know that at t' I feel cold. This is now very problematic, but on the other hand the problem seems partly related to the expressiveness of the modal language used to state the puzzle. If we shift to a first-order modal language (or to an extended modal language) in which instants can be named in the object language, so that $P(t)$ means that I am cold at time t, it would

be reasonable to rephrase the condition $Kp \rightarrow \Box Kp$ in the form $KP(t) \rightarrow \Box KP(t)$, in a way which permits to avoid hidden indexical reference to the time of evaluation. In that case, even if $\Box$ is given the same temporal interpretation, this leads to no paradox, since this entails:

(36) $t \models KP(t)$ only if for all $t' \sim t$, $t' \models KP(t)$

instead of the problematic:

(37) $t \models KP(t)$ only if for all $t' \sim t$, $t' \models KP(t')$

A condition like (36), which follows from the corresponding versions of Positive Introspection and Reflective Margin for Error, implies that for me to know at time t that P holds at t, it has to be the case that at every subsequent time t' I also know at t' that P holds at t: this now states a more reasonable property of self-knowledge, and not necessarily a pernicious condition, even when the property is a non-permanent property such as "feeling cold". The case is not satisfactory yet, however, since I can very well forget that I was feeling cold at some past time, even if at that time I knew that I was feeling cold, and knew that I knew it.

If we modify the basic margin for error principle $Kp \rightarrow \Box p$ in the same manner, the principle will be written: $KP(t) \rightarrow \Box P(t)$. Semantically, this will have the consequence that: if $t \models KP(t)$, then for all $t' \sim t$, $t' \models P(t)$. This condition becomes vacuous, however, since if at time t I know that P holds at t, then by factivity P holds at t, but then $P(t)$ remains true at any time subsequent to t. This modification of the margin for error principle seems now too weak. What this suggests, nevertheless, is that the basic margin for error principle and its reflective version may be given different interpretations. What I shall argue in the next section is that mental states such as "feeling cold" and "knowing that one feels cold" are not necessarily subject to the same margins for error in order to be known.

2.2 Reflective vs. perceptual knowledge

In this section I argue that margin for error principles do make sense for our knowledge of phenomenal properties, provided that these phenomenal properties do not pertain to the occurrence or non-occurrence of knowledge itself. The argument, originally put forward in Dokic & Égré (2004), rests on the idea that different kinds or methods of knowledge are involved in Williamson's argument against positive introspection, and that one should distinguish a notion of perceptual knowledge from a notion of reflective knowledge. Granted this distinction, one can argue that perceptual knowledge and reflective knowledge do not have the same reliability conditions, and that reflective knowledge, in particular, is not necessarily subject to the same kinds of margin of error as perceptual knowledge.

I look at a chair in my office and I wonder how tall it might be. I am confident that it is less than 2 meters tall. Maybe I know this partly on the basis of inference: I know from my last medical record that I am less than 2 meters tall, and I know visually that I am taller than the chair. Still, my judgement that I am taller than the chair makes essential use of my vision and is subject to a margin for error: I consider that the chair is significantly smaller than me to conclude that it is less than 2 meters tall. Indeed, if I now look at the bookshelves and wonder whether that particular shelf is fixed above or below 2 meters from the ground, I am no longer confident about whether the top of my head might reach the shelf or not. However, I am confident that I know that the chair is less than 2 meters tall: I believe I know that the chair is less than 2 meters tall. Moreover, I also know that I believe that I know this. And

upon reflection, I know that I know that I believe I know this. It seems to me I could go on indefinitely. This kind of reflective knowledge, the knowledge I have upon my own beliefs, does not seem to me to fade off in principle, and does not seem subject to the same kind of margin of error as my knowledge that the chair is less than 2 meters tall.

What this example suggests is that the limitations of my visual knowledge, or of knowledge acquired on the basis of my visual perception, are not on a par with the limitations of my reflective knowledge, namely knowledge about my own epistemic states and capacities. In Dokic & Égré (2004), this distinction between perceptual and reflective knowledge was introduced upon close inspection of Williamson's specific argument against positive introspection, as given above in its syntactic version (32). The puzzle rests on the condition (KME), namely $K(K\neg p_i \rightarrow \neg p_{i+1})$, which expresses the knowledge Magoo has of the limitations of his visual knowledge. Thus (KME) expresses Magoo's knowledge that his visual knowledge obeys (ME), namely $K\neg p_i \rightarrow \neg p_{i+1}$. The main difference between (ME) and (KME) is that (ME) is a non-iterative principle, in which $\neg p_i$ expresses a phenomenal property concerning the outside world ("the tree is not i inches tall"). Accordingly, (ME) expresses a property of Magoo's visual knowledge. By contrast, (KME) can't express a property of Magoo's visual knowledge only: for then the principle should mean that Magoo *knows visually* that if he knows visually that the tree is not i inches tall, the tree is not $i+1$ inches tall. But is this higher-order form of knowledge really a kind of visual knowledge? This does not seem plausible. The reason seems fairly general. Suppose I see that it's raining outside the window. Then I know visually that it's raining. And presumably, I also know that I know that it's raining. But my knowing that I know is not visual knowledge, at least its content is sufficiently distinct that we postulate this form of knowledge to be of a different kind.

If this analysis is correct, this suggests that Williamson's argument, although formally valid, rests on a potential equivocation, by neglecting a property of modularity of our knowledge system. The principle (KME) ought to be rephrased in terms of two knowledge modalities. Let K_π express perceptual knowledge and K reflective knowledge, then (KME) is more adequately expressed as (KME'), namely:

(KME') $\quad K(K_\pi\neg p_i \rightarrow \neg p_{i+1})$

To see whether this gives us a way out of the paradox, however, it is necessary to give a plausible reformulation of the other principles involved in the puzzle. A minimal adjustment consists in maintaining the supposition that K is closed under logical consequence, and adding the following bridge principle (KK'), namely:

(KK') $\quad K_\pi\phi \rightarrow KK_\pi\phi$

(C') If ϕ follows logically from a set of propositions Γ, and for all members ψ of Γ, $K\psi$ holds, then $K\phi$.

(KK') is a mixed variant of positive introspection: in the scenario of Mr Magoo, this means that if Magoo knows visually p, then he is in a position to know reflectively that he knows visually p. The interest of this reformulation is that it still gives rise to a derived rule, but not as strong as the original one:

(38) (i) $K_\pi \neg p_i$, by hypothesis
(ii) $K(K_\pi \neg p_i \rightarrow \neg p_{i+1})$, by (KME)
(iii) $KK_\pi \neg p_i$, by (i) and (KK)
(iv) $K_\pi \neg p_i, K_\pi \neg p_i \rightarrow \neg p_{i+1} \models \neg p_{i+1}$, by propositional reasoning
(v) $K \neg p_{i+1}$, by (ii), (iii), (iv) and (C')

In this case, $K_\pi \neg p_i$ entails $K \neg p_{i+1}$. Thus, supposing Magoo knows visually that tree is not of size 0, he knows reflectively that the tree is not of size 1 either, but no further propagation need arise. This means that Magoo can know reflectively that the tree is not of size $i+1$ when he sees that it is not of size i, but his knowledge that the tree is not of size $i+1$ is not necessarily an item of direct perceptual knowledge.[7] This gives a fair illustration of the idea that perceptual knowledge and reflective knowledge correspond to distinct modules. The case of visual illusions (like the Müller-Lyer illusion) shows that one can fail to see that two lines are equal, for instance, while knowing, through a different information channel, that those two lines are equal. This does not count as direct evidence for the present argument, of course, but this should at least support the plausibility of the distinction between different forms of knowledge. In the same way as a visual illusion can persist despite one's knowing that it is an illusion, it is conceivable that one fails to see that some object is not of size $i+1$ while knowing that it is not of that size from reflection on one's visual abilities. We should note, moreover, that the new derived rule, albeit strong, implies that one has a correct grasp of the limitation of one's perceptual knowledge through margin for error.

2.3 Logics for modular knowledge

In order to make the articulation between reflective and perceptual knowledge more explicit, it is appropriate to give more details about the definition of a system of modular knowledge. I sketch three distinct approaches here. All of them rest on the idea that some mechanism of syntactic restriction is needed to model the distinction between perceptual knowledge and reflective knowledge.

To get a system of combined knowledge richer than the one we just presented on the basis of Williamson's principles, the first option is to define an axiomatic system of two epistemic modalities such that:

(i) K is an **S4** modality
(ii) K_π satisfies axiom **T**, namely $K_\pi \phi \rightarrow \phi$
(iii) (KK') holds, namely $K_\pi \phi \rightarrow K K_\pi \phi$
(iv) The system is closed under modus ponens and K-generalization
(v) Only non-epistemic sentences can occur in the scope of K_π.

Let us call **KK′** this system. The syntactic restriction in (v) is intended to reflect the fact that perceptual knowledge is not iterative, and that it applies only to phenomenal properties. Likewise, since the system is not closed under K_π-generalization, it won't follow from Kp, for instance, that $K_\pi Kp$. Conversely, since K is an **S4** modality, it satisfies the principle of positive introspection. This assumption is natural to make about reflective knowledge, if the notion of reflective knowledge is seen as a form of higher-order knowledge which can be self-applied.

[7] See Dokic & Égré (2004) for a concrete illustration of this situation with the puzzle of the Glimpse, a perceptual variant of the Surprise Examination Paradox stated by Williamson (2000, chap. 6).

Since the system is closed under K-generalization, $\mathbf{KK'} \vdash K(K_\pi\phi \rightarrow \phi)$ for every appropriate ϕ. It follows from that and from (KK') that $\mathbf{KK'} \vdash K_\pi\phi \rightarrow K\phi$ for every appropriate ϕ. This, we should note, does not go counter to the distinction between perceptual and reflective knowledge, provided the converse does not hold. If we consider the extension of $\mathbf{KK'}$ in which K_π is a normal operator, and such that the system is closed under K_π-generalization and uniform substitution, we get a normal system which is sound and complete for the class of frames $\langle W, R, R_\pi\rangle$ such that R is reflexive and transitive, R_π is reflexive, and $xR_\pi z$ whenever xRy and $yR_\pi z$. It is easy to show that $K\phi \rightarrow K_\pi\phi$ is not derivable in that system by means of an appropriate counter-model, and consequently that it is not derivable in the weaker system $\mathbf{KK'}$.

The main merit of a system of knowledge like $\mathbf{KK'}$ is that the two forms of knowledge, perceptual and reflective, are related in such a way that any perceptual content can be taken as input to knowledge of a different kind, without any collapse between the two modalities. A main shortcoming of this approach, on the other hand, is that margin for error principles have to be added explicitly, as above with the axiom (KME'), and that we have no explicit semantics for the operator K_π, which is supposed to be non-normal.

To remedy this, a second option, suggested by J. van Benthem (p.c.), would be to modify the principles used in (35) to give the semantic version of Williamson's paradox. Instead of making an explicit distinction between perceptual and reflective knowledge, one can keep a single knowledge operator and add a closeness modality, and restrict the Margin for Error principle "to just the case of atomic propositions, thought of as observational (whose values then have to obey some geometrical constraints), while leaving matters open for arbitrary assertions ϕ with iterated knowledge modalities" (van Benthem, p.c.). This is tantamount to making an implicit distinction between perceptual and reflective knowledge, but this enables us to keep a standard semantics for knowledge.

Arnesen 2004 presents a normal system of modal logic with two operators K and $\Box$ whose axiomatic correspondent can be restricted in this way. A margin for error model is a quadruple $\langle W, R, \sim, V\rangle$ such that:

$M, w \models K\phi$ iff for every w' such that $wRw', M, w' \models \phi$

$M, w \models \Box\phi$ iff for every w' such that $w \sim w, M, w' \models \phi$

In Arnesen's approach, the relation $\sim$ is taken to be a similarity relation, reflexive and symmetric, but not necessarily transitive, just as the epistemic accessibility in Williamson's fixed-margin models (see Williamson 1994 and Graff 2002). Thus $\Box$ is supposed to satisfy the **T**-axiom $\Box\phi \rightarrow \phi$ and the Brouwersche or **B**-axiom $\neg\phi \rightarrow \Box\neg\Box\phi$. In Arnesen's original system, the Margin for Error principle $K\phi \rightarrow \Box\phi$ was unrestricted, giving rise to the problematic principle $K\phi \rightarrow \Box K\phi$. Restricting it to atomic propositions, or even to non-epistemic propositions, permits to block that derivation. Thus we may call $\mathbf{K}\Box^*$ the bimodal system consisting of:

(i) An **S4** basis for K
(ii) A **BT** basis for $\Box$
(iii) The principle $K\phi \rightarrow \Box\phi$, restricted to a distinguished set of non-modal sentences ϕ.

Relative to $\mathbf{K}\Box^*$, K also denotes general knowledge, just as K in the system $\mathbf{KK'}$. The operator implicitly refers to a form of perceptual knowledge in all the cases where it applies

to some proposition for which Margin for Error will hold, and it refers to a form of reflective knowledge in all cases where it takes in its scope sentences already containing occurrences of K.

I sketch, finally, a third option, which would be to go along with only one knowledge modality, without adding a closeness modality to the language, but to treat differently the semantics of sentences of the form $K\phi$, depending on the syntactic status of ϕ. Thus, we could try to define a non-standard semantics such that: (i) if ϕ is a purely propositional formula (meaning a formula not containing any occurrence of K), then given a Kripke model $\langle W, R, V\rangle$, $w \models K\phi$ iff for every w' such that wRw', $w' \models \phi$ according to the standard rules; (ii) if ϕ is a formula of the form $K\psi$, then $w \models K\phi$ iff $w \models K\psi$. We thus give iterative formulas and non-iterative formulas a distinct treatment. For instance, let us imagine a structure of inexact knowledge like Figure 3, where the numbers represent degrees on a thermometric scale. Like the cells of Figure 2, the arrows represent the relation of tactile indiscernibility, which is reflexive and symmetric, but not transitive. Let p represent "feeling cold", and let us suppose that the cut-off point between feeling cold and not feeling cold is between 1 and 2, so that p holds only at 0 and 1.[8]

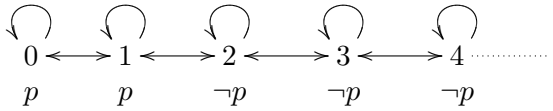

Figure 3: A structure of inexact knowledge (2)

According to the semantics, $0 \models Kp$, $3 \models K\neg p$, but $1 \not\models Kp$ and likewise $2 \not\models K\neg p$. Thus, at 0 I know that I feel cold, at 3 I know that I don't feel cold, but at intermediate cases I don't know whether I feel cold or not. The usual semantics would predict that $0 \not\models KKp$, since $0R1$ and $1 \not\models Kp$. However, according to the revised semantics, since $0 \models Kp$, it follows that $0 \models KKp$. Likewise, $3 \models KK\neg p$. The structure represents a situation in which knowledge about coldness does obey a margin for error, but such that knowledge can satisfy positive introspection.

The semantics is still incompletely specified, however, since it does not say how to evaluate $K\phi$ when ϕ does contain an occurrence of K but does not start with K. We could stipulate that (iii) if ϕ is of the form $\neg K\psi$, then $w \models K\phi$ iff $w \models \phi$. That would make knowledge negatively introspective. With respect to the previous model, condition (iii) predicts that $2 \models K\neg K\neg p$ and $2 \models K\neg Kp$, that is I know that I don't know whether I feel cold or not. Finally, it remains unclear how we should evaluate mixed formulae like $K(p \vee Kp)$ and $K(p \wedge Kq)$. In the case of a conjunctive formula containing an occurrence of K, it seems reasonable to stipulate that $w \models K(\phi \wedge \psi)$ if and only if $w \models K\phi$ and $w \models K\psi$, just as in the case of purely propositional formulas. The case of a disjunction containing an occurrence of K is more problematic, however. We might stipulate further that a disjunction containing an occurrence of K is known if at least one of the disjuncts is known. The rationale for such

[8]We should note that "feeling cold" can be ambiguous, due to the ambiguity of the expression "feeling". It can mean being cold simpliciter, or *feeling that* one is cold in the sense of being aware that one is cold. Being cold simpliciter is just the occurrence of a phenomenal property, which we may represent by p, as opposed to being aware that one is cold, which we should represent by Kp. Like Williamson, we take "feeling cold" in the first, non-reflective sense, just like "having hope", or "being in pain". Like Williamson we conceive that one might feel cold, have hope, or be in pain without being aware yet that one is in such mental states.

a rule is given by the case of epistemic formulas like $K(Kp \vee K\neg p)$: this means that I know that I know which of p or $\neg p$ holds. But intuitively, this should entail that I know that I know p, or that I know that I know $\neg p$. Furthermore, the semantics validates formulas like $K(\neg Kp \vee p)$ (that is $K(Kp \rightarrow p)$) and $K(Kp \vee \neg Kp)$ (that is $K(Kp \rightarrow Kp)$).

Although the rule for disjunction looks like an intuitionistic rule, it does not hold for purely propositional formulas like $p \vee \neg p$. The problem, however, is that the semantics makes wrong predictions in the case of more complicated mixed formulas like $[(p \wedge (Kp \vee \neg Kp) \vee \neg p \wedge (Kp \vee \neg Kp)]$. Although this formula is a tautology, the semantics predicts that it is known provided p is known or $\neg p$ is known. This suggests not only that the semantics is too stipulative, but also that a more principled representation should be given of the notion of perceptual content. In particular, the semantics should not cut across tautological contents depending on whether they are expressed by means of modal or non-modal formulas.

A more adequate attempt to formulate a semantics capable of validating the principle of positive introspection without giving rise to Williamson's paradox was suggested to me by Denis Bonnay. The idea is to have a semantics which will validate the principle of positive introspection without thereby making the corresponding accessibility relation transitive. Given a Kripke structure $\langle W, R, V\rangle$, we simultaneously define two notions of satisfaction, one for single worlds and one for couples of worlds, namely $w \models \phi$ and $(w, w') \models \phi$. The clauses for atoms and Boolean connectives are the usual ones in the simple case. In the case of a couple, the satisfaction of atoms is evaluated relative to the second component, and the Boolean clauses are also straightforward. Thus we have:

(i) For a propositional atom p, $(w, w') \models p$ iff $w' \in V(p)$.
(ii) $(w, w') \models \neg\phi$ iff $(w, w') \not\models \phi$.
(iii) $(w, w') \models (\phi \wedge \psi)$ iff $(w, w') \models \phi$ and $(w, w') \models \psi$.
(iv) $(w, w') \models (\phi \vee \psi)$ iff $(w, w') \models \phi$ or $(w, w') \models \psi$.

The interesting clauses are the clauses for the epistemic operators, that is:

(a) $w \models K\phi$ iff for every w' such that $wRw', (w, w') \models \phi$.
(b) $(w, w') \models K\phi$ iff $w \models K\phi$.

Clause (a) corresponds to the standard semantics for knowledge. Together with clause (b), however, it ensures that instead of looking at worlds that are two steps away to check whether $KK\phi$ is satisfied, one backtracks to the current world to check whether $K\phi$ is satisfied there.

Relative to the structure of Figure 3, it is easily seen that $0 \models Kp$, since $(0, 0) \models p$ and $(0, 1) \models p$, and hence $0 \models KKp$, for $(0, 0) \models Kp$ and $(0, 1) \models Kp$. But it is still the case that $1 \not\models Kp$, since $(1, 2) \not\models p$. A formula will be called valid if it holds at every single world of every structure according to the semantics. It can be checked that the semantics validates the principle of positive introspection: given a structure in which $w \models K\phi$, if $w \not\models KK\phi$, then there is an accessible w' such that $(w, w)' \not\models K\phi$, and so $w \not\models K\phi$, against the hypothesis. The semantics also validates negative introspection. Take a structure in which $w \models \neg K\phi$. Then if it were not the case that $w \models K\neg K\phi$, there would be an accessible world w' such that $(w, w') \models K\phi$, and so we would have $w \models Kp$, a contradiction.

Like the previous semantics, this semantics also validates a formula like $K(Kp \rightarrow p)$ (although not necessarily $K\phi \rightarrow \phi$, which means that the system should primarily be seen as a system of introspective belief). Unlike the previous semantics, however, it makes $K[(p \wedge (Kp \vee \neg Kp) \vee \neg p \wedge (Kp \vee \neg Kp)]$ valid, without entailing $Kp \vee K\neg p$. The case of mixed formulae may still give rise to too strong predictions, however, since $K(\phi \vee K\psi)$ entails $K\phi \vee K\psi$. We may wonder whether this is plausible in full generality. We should note here that this entailment

does not hold in a normal modal logic like **K**, nor in the stronger system **T**. However, it holds in a stronger system like **S5**, as a consequence of the principle of negative introspection. Thus, someone who wants to deny the entailment from $K(\phi \vee K\psi)$ to $K\phi \vee K\psi$ may wish to reject negative introspection. One can note here that the operator K also satisfies Kripke's distribution axiom $K(\phi \rightarrow \psi) \rightarrow (K\phi \rightarrow K\psi)$, and that the semantics is sound for the rule of necessitation, in the sense that if ϕ holds in every structure at every world according to the present semantics, then so does $K\phi$. It can be checked that the semantics is also sound for the rule of uniform substitution, which means that the semantics is sound more generally for the normal logic **K45**, and we may conjecture that the system **K45** is also complete for this revised semantics, a point I leave for further investigation.[9]

A question we have not solved here is whether it is possible to give a natural semantics which would tolerate margin for error and licence positive introspection without also validating negative introspection. What may be pointed out, however, is that the assumption of negative introspection is just as plausible as the assumption of positive introspection for the scenarios considered by Williamson in his attack on luminosity. Indeed, Williamson considers a process in which "one thoroughly considers how hot or cold one feels" (2000, 97), and it seems at least plausible, in such a case, to conceive that I don't know whether I feel cold or not while nevertheless knowing that I don't know. This does not mean, of course, that we should argue for the plausibility of the principle of negative introspection in all situations of knowledge alike, due to well known examples of false beliefs or unawareness for which the principle breaks down.[10]

2.4 Reply to some objections

If our analysis of Williamson's paradox is correct, it means that the principle of margin for error and the principle of positive introspection can coexist, provided each one is referred to the appropriate kind of knowledge. In the first logic we just discussed, this separation is effected by supposing that perceptual knowledge is not iterative, so that a fortiori it can't be positively introspective, and by supposing that reflective knowledge, which is positively introspective, does not bring about a margin of error. Likewise, in the system $\mathbf{K}\Box^*$, there is a class of non-epistemic propositions ϕ whose knowledge requires a margin for error, but for which iterations of knowledge are not constrained in the same way. In our revised semantics for knowledge, finally, iterated modalities are not given the same treatment as non-iterative ones. To rebut any suspicion of adhocity, we need to examine whether our claim of modularity is sufficiently motivated. To do this, we shall address three objections.

The first objection concerns the evidence we have that perceptual knowledge is not iterative. Do we have sufficient empirical evidence that visual knowledge, for instance, is not iterative? Or more precisely, that meta-knowledge about one's visual knowledge is not itself visual? Likewise, if we think back to the example of "feeling cold", my knowing that I feel cold goes through a specific perceptual channel: what grounds do we have to claim that meta-knowledge about this kind of tactile knowledge is not itself tactile? This objection was formulated in very clear terms by J. Snyder (p.c.), who notes that: "at the eye doctor, examining the eye charts, I discover that I cannot read letters below a certain size. This certainty seems to be perceptual knowledge of the limitations of my perceptual capabilities"

[9] Denis Bonnay and I are still working on the logic at the present moment, and I shall leave a more detailed presentation of the system for further work.

[10] See Williamson (2000, 23) for such a case of false belief, entailing the unawareness of one's ignorance.

(Snyder, p.c.). To be sure, let us imagine that I perceive a letter of which I can't tell whether it is an O or a D. I then conclude: "even if it is a D, I can't exclude visually that it is an O". Let us imagine further that I can see nevertheless that the letter is either a D or an O. Formally, this scenario seems to support the following statements: $K_\pi(D \vee O)$, and moreover $K_\pi(D \rightarrow \neg K_\pi \neg O)$, that is, "I see that, if it's a D, then for all I see it might be an O".

The reply, however, is that we should distinguish more firmly between visual experience and judgement. My judgement that even if it is a D, I might see an O, is sustained by my visual experience. However, this is not sufficient to affirm that the content of this judgement is itself visual. It seems more plausible to say that this judgement about my visual limitations is a piece of reflective knowledge, acquired on the basis of my visual experience. I don't "see" that I might be seeing an O when there is a D, in the same way in which I see that it's an O or a D.

This brings us to the second objection, which concerns the evidence we have that so-called reflective knowledge is not subject to margin for error. In Dokic & Égré (2004), this point is motivated by reference to the notion of ascent routine, as defined by Gordon (1995). The transition from the visual experience that p to the judgement "I see p" is an example of such an ascent routine. Ascent routines seem to be reliable methods of self-knowledge without a margin for error. At any rate it is highly dubious that they should rest on the same kinds of margins as are involved at the primary level. The transition from my sensory experience of feeling cold to the judgement "I feel cold" is another example of ascent routine. Undoubtedly, one needs to feel cold with a sufficient intensity to consciously feel cold, and this supports Williamson's claim that knowing that one feels cold requires a margin for error. But in a situation in which I have reached that intensity, and am aware that I feel cold, my judgement "I feel cold" will in turn support the fact that I know that I know that I feel cold. Thus, even if my primary knowledge is not simultaneous with my sensory experience, it is consistent to assume that the transition to higher levels of knowledge is simultaneous with the occurrence of my primary knowledge.

A similar point is made by Leitgeb (2002, 203) in his review of Williamson (2000), when he writes that we could imagine "the phenomenal system and the cognitive system...tuned to each other", in such a way that luminosity might hold. Leitgeb calls feeling cold a phenomenal property, as opposed to knowing that one feels cold, which he calls a cognitive property. As Leitgeb rightly emphasizes, each property goes with a distinct scale of similarity, since the similarity between shades of feeling cold is not the same as the similarity between belief states. To say that luminosity might hold is to imagine that the scales could match each other in such a way that the transition from not feeling cold to feeling cold would be matched by a simultaneous transition from not knowing that one is feeling cold to knowing that one is feeling cold. Taking things one level higher up, what we are suggesting here is that the metacognitive system (knowing that one knows) and the cognitive system (primary knowing) are likely to be tuned to each other in that sense, even if the cognitive system and the phenomenal system are not.[11]

[11]It does not matter for the argument whether there is an actual "metacognitive" system distinct from the "cognitive" system, although this is consistent with the idea that perceptual knowledge and reflective knowledge may correspond to distinct modules. One can nevertheless talk of a scale of similarity for the property of "knowing that one knows", distinct from the scale of similarity for the property of knowing simply. What we are saying is that these two scales are more likely to match each other or be calibrated, to use Leitgeb's vocabulary, than the cognitive scale (knowing that one is cold) and the scale attached to the phenomenal property (being cold).

The vision of self-knowledge defended by Williamson is different in this respect. For Williamson, my knowing that I feel cold is a phenomenal property in much the same way in which my feeling cold is a phenomenal property. More accurately, the former is to my knowing that I know what the latter is to my primary knowing. This is why Williamson insists on the idea that each new iteration of knowledge should bring about an additional margin for error. In this respect, however, we can only support the cautious claims made by Leitgeb when he writes that whether or not our cognitive system might be tuned to the phenomenal system in a way which supports luminosity is an empirical question, which should be investigated partly by empirical methods. A priori, however, the cognitive system and the metacognitive system (granting that distinction) are more likely to be tuned to each other than the cognitive system and the phenomenal system, since the formers involve the same kind of similarity.

This leads directly to the third objection, finally, which concerns the generality of this reply to Williamson's anti-luminosity argument. Here we focussed on Williamson's specific argument against the principle of positive introspection. If this analysis is correct, what we are saying is that the mental state of knowing p might be luminous, even if p itself is not a luminous condition. Feeling cold, among other phenomenal properties, is not necessarily luminous, since I can be in a state where I am cold without knowing that I am cold. So far, we agree with Williamson that not every mental state is luminous. Where we disagree with Williamson is on the idea that "we have no cognitive home", namely on the idea that *no* mental condition is luminous. Indeed, a case in which I know that I am feeling cold is also a state in which I am in a position to know that I know that I feel cold, even if for me to perceive that I am cold at the first level, I need a sufficient margin for error. At any rate, Williamson has not proved that knowledge about one's knowledge involves a margin for error, in the same way in which he has made plausible the idea that most situations of perceptual knowledge are situations of inexact knowledge which rely on a margin of error.

3 Conclusion

Williamson's anti-luminosity argument is not only challenging and intriguing, but the emphasis Williamson puts on margin for error principles is probably one of the most inspiring suggestions that have been made in recent years in the field of formal epistemology. One important reason for this, as we have seen, is that the principle of margin for error generalizes the notion of factivity, and that it brings the notion of reliability into an area in which the notions of truth and logical consequence have received most of the attention for a long time. Another reason is that margin for error principles take epistemic logic closer to the domain of cognitive psychology. By affirming against Williamson that his argument neglects an important dimension of modularity of our knowledge, the intention of this paper is also to make a further step in this direction. Although the claim that knowledge is modular should be fairly uncontroversial, more work needs to be done to give an adequate logical representation of the interaction between the notions of perceptual knowledge and reflective knowledge which we have distinguished, and to provide empirical evidence for the well-foundedness of the distinction.

References

[1] Arnesen Inge (2004), *Core Logic - Additional project, Period C, part 2*, manuscript, ILLC, Amsterdam.

[2] van Benthem, J. (p.c.), email message to P. Égré, March 15, 2004.

[3] Burnyeat M. (1990), *The Theaetetus of Plato*, with a translation of Plato's *Theaetetus* by M. J. Levett, revised by M. Burnyeat, Hackett Publishing Company.

[4] Dokic J. & Égré P. (2004), "Margin for Error and the Transparency of Knowledge", Technical Report, Archives électroniques de l'Institut Jean-Nicod, submitted for publication.

[5] Égré P. (2004), *Attitudes propositionnelles et paradoxes épistémiques*, PhD. Dissertation, University Paris I Panthéon-Sorbonne, IHPST.

[6] Gettier E. (1963), "Is Justified True Belief Knowledge?", *Analysis*, 23, pp. 121-3.

[7] Gomez-Torrente, M. (1997), "Two Problems for an Epistemicist View of Vagueness", in E. Villanueva (ed.), *Philosophical Issues*, 8: *Truth*, Ridgeview, Atascadero, pp. 237-245.

[8] Gordon R. (1995), "Simulation without Introspection or Inference from Me to You", in Davies M. & Stone T. (eds.), *Mental Simulation*, Oxford, Blackwell.

[9] Graff D. (2002), "An Anti-Epistemicist Consequence of Margin for Error Semantics for Knowledge", *Philosophy and Phenomenological Research*, 64, pp. 127-42.

[10] Hintikka J. (1962), *Knowledge and Belief*, Ithaca, Cornell University Press.

[11] Leitgeb H. (2002), Review of Timothy Williamson, *Knowledge and its Limits*, *Grazer Philosophische Studien*, 65, pp. 195-205.

[12] Lewis, D. (1996), "Elusive Knowledge", repr. in D. Lewis, *Papers in Metaphysics and Epistemology*, Cambridge Studies in Philosophy, c. 25, pp. 418-445.

[13] Nozick R. (1981), *Philosophical Explanations*, Belnap, Harvard.

[14] Russell B. (1912), *The Problems of Philosophy*, Oxford.

[15] Snyder Josh (p.c.), email message to P. Égré, July 7, 2004.

[16] Sosa E. (1999), "How to Defeat Opposition to Moore", *Philosophical Perspectives*, 13, on Epistemology, ed. by J. Tomberlin.

[17] Spinoza, *Ethics*, edited and translated by G.H.R. Parkinson, Oxford Philosophical Texts, 2000.

[18] Williamson T. (1992), "Inexact Knowledge", *Mind*, 101, pp. 217-42.

[19] Williamson T. (1994), *Vagueness*, Routledge.

[20] Williamson T. (1997), "Replies to Commentators", in E. Villanueva (ed.), *Philosophical Issues*, 8: *Truth*, Ridgeview, Atascadero, pp. 255-265.

[21] Williamson T. (2000), *Knowledge and its Limits*, Oxford University Press.