HAL
open science

A non-standard Semantics for Inexact Knowledge with Introspection

Denis Bonnay, Paul Egré

To cite this version:

Denis Bonnay, Paul Egré. A non-standard Semantics for Inexact Knowledge with Introspection. 2006. ijn_00000679v2

HAL Id: ijn_00000679

https://hal.science/ijn_00000679v2

Preprint submitted on 21 Jun 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.

A Non-Standard Semantics for Inexact Knowledge with Introspection

DENIS BONNAY
Université Paris 1, IHPST
denis.bonnay@univ-paris1.fr

PAUL ÉGRÉ
Institut Jean-Nicod, CNRS, Paris
paulegre@magic.fr

ABSTRACT.

Standard Kripke models are inadequate to model situations of inexact knowledge with introspection, since positive and negative introspection force the relation of epistemic indiscernibility to be transitive and euclidian. Correlatively, Williamson's margin for error semantics for inexact knowledge invalidates axioms 4 and 5. We state a non-standard semantics for modal logic which is shown to be complete for $\mathbf{K45}$, without constraining the accessibility relation to be transitive or euclidian. The semantics corresponds to a system of modular knowledge, in which iterated modalities and simple modalities are not on a par. We show how the semantics helps to solve Williamson's luminosity paradox, and argue that it corresponds to an integrated model of perceptual and introspective knowledge that is psychologically more plausible than the one defended by Williamson. A generalized version of the semantics is formulated, in which modalities are iteration-sensitive up to degree n and insensitive beyond n. The multi-agent version of the semantics yields a resource-sensitive logic with implications for the representation of common knowledge in situations of bounded rationality.

1 Inexact knowledge with introspection

Standard modal models for knowledge are commonly $\mathbf{S5}$ models in which the epistemic accessibility relation is an equivalence relation, namely a relation that is reflexive, symmetric and transitive. From an axiomatic point of view, reflexivity corresponds to the fact that knowledge is veridical, symmetry to the idea that if something is true, one knows one will not exclude it, and transitivity to the idea that knowledge is positively introspective, that is the property that whenever I know some proposition, I know that I know it. $\mathbf{S5}$ models can also be described as reflexive models that are euclidian, which also makes them symmetric and transitive. Euclidianity corresponds to the property of negative introspection, namely to the property that whenever I don't know, I know that I don't know. $\mathbf{S5}$ models are commonly used to represent situations of social knowledge, for instance in game theory, due to their well-known correspondence with partitional models of information (Osborne & Rubinstein 1994).

Proceedings of the Workshop on Rationality and Knowledge, ESSLLI 2006
Sergei Artemov & Rohit Parikh (editors)
Copyright © 2006, the author(s)

An important feature of these models is the fact that they represent a notion of precise or exact knowledge in the following sense: whenever an agent fails to discriminate between two worlds or situations w and w', any other situation which he fails to discriminate from w is also a situation which he fails to discriminate from w', and vice versa. In other words, even though one's knowledge is not necessarily as fine-grained as it should be, it is at least clear cut, since one's uncertainty is partitional. This contrasts with situations of imprecise knowledge, in which the relation of epistemic indiscriminability can fail to be transitive, as in cases of perceptual knowledge in which I can't discriminate between any two adjacent shades of color, and yet such that I can distinguish between shades of color that are non-adjacent. Such situations are equivalently described as situations in which one's knowledge fails to be euclidian, if it is assumed that one's failure to discriminate between worlds is at least symmetrical. In cases like these, one's uncertainty is no longer partitional, but rather fuzzy. Situations of this kind have been described as situations of *inexact* knowledge (Williamson 1992a), although the term "inexact" has also been used to refer to situations of false belief (failure of reflexivity), and the term "imprecise" preferred to talk of vague or fuzzy or approximate knowledge (Mongin 2002). In this paper, we shall use the terms "imprecise" and "inexact" interchangeably, and we shall focus on situations of vague knowledge for which one's accessibility relation, although reflexive and symmetric, fails to be transitive and euclidian.

The representation of situations of inexact knowledge is not as straightforward as one might expect. Indeed, how should we model situations in which one's knowledge is imprecise, and yet in which one wants to maintain properties like negative and positive introspection? A good indication that this is not obvious is provided by the existence of a general argument of T. Williamson against the idea that knowledge is positively introspective in general, based on the description of situations of approximate knowledge.

Consider, for instance, a situation of visual knowledge in which I am asked to distinguish objects by their sizes. From where I am, I can't discriminate between objects that differ from each other only by less than one centimeter. However, I can discriminate between objects that differ from each other by more than one centimeter. This is a situation where I can't discriminate between 10 and 11, nor between 11 and 12, but in which I can nevertheless discriminate between 10 and 12, so the relation of visual indiscriminability is reflexive and symmetric but non-transitive. Suppose further that I am asked to make judgements about whether the objects are small enough to fit in a certain box. Let us suppose that objects with size 10 and 11 can fit in, but that objects with size 12 and more cannot. This situation can be represented by the following Kripke model, where worlds are named by numbers, and where p represents the property of fitting in the box.

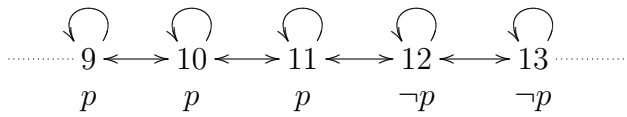

Figure 1.1: A structure of inexact knowledge

Let us represent $\Box\phi$ the proposition "I know that ϕ". By giving the modal operator its usual semantics, it is easily seen that $10 \models \Box p$, since all the worlds that I can't discriminate from 10

also satisfy p. This means that when I look at an object of size 10, I know that it will fit in the box. Likewise, $13 \models \Box\neg p$: looking at an object of size 13, I know it won't fit in the box. Intermediate cases, however, are cases for which I fail to know whether they fit in the box, namely $11 \models \neg\Box p$ and $11 \models \neg\Box\neg p$, and similarly for 12. Statements of higher-order knowledge, however, become problematic. Indeed, since $11 \models \neg\Box p$, it follows that $10 \models \neg\Box\Box p$. Therefore, although I know that an object of size 10 will fit in the box, I don't know that I know it.

This consequence is problematic, for one could insist that knowledge about one's visual knowledge is not constrained in the way one's visual knowledge is. In his account of inexact knowledge, by contrast, Williamson (1992a) turned this limitation of the standard semantics for knowledge into a negative argument against the principle of positive introspection. A model like the one we gave is a particular instance of what Williamson calls a *fixed margin model*. A fixed margin model, relative to a monomodal propositional language, is a quadruple $\langle W, d, \alpha, V\rangle$, where W is a set of worlds, α is a non-negative real number, d a metric over W (namely a function from $W \times W$ to $\mathbb{R}^+$, such that $d(w, w') = 0$ if and only if $w = w'$, $d(w, w') = d(w', w)$, and $d(w, z) \leq d(w, x) + d(x, z)$), and V is a valuation function over the atoms. The satisfaction clause for the $\Box$ is the expected one, namely $M, w \models \Box\phi$ iff for every w' such that $d(w, w') \leq \alpha, M, w' \models \phi$. The fixed parameter α corresponds to the notion of margin for error: at a world w, one knows ϕ if and only if ϕ holds throughout the worlds that are within the margin α, that is at all the worlds that are not discriminable from w. As Williamson shows, validity in fixed margin models is axiomatized by the normal logic $\mathbf{KTB}$, namely the logic of reflexive-symmetric frames, and neither axiom 4 nor axiom 5 is valid in fixed margin models, by obvious failures of transitivity and euclidianity for the distance function.[1]

Whether this consequence is welcome or unwelcome should depend on the notion of knowledge the semantics is intended to capture. For Williamson, the failure of positive and negative introspection in margin models is actually an important lesson that we should draw concerning the notion of self-knowledge in general. Indeed, Williamson insists that "where one has only 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" (2000: 18). If knowledge obeys a margin for error principle, then to suppose that positive introspection is valid is likely to give rise to paradox. In the previous scenario, for instance, in which one's margin for error is of 1 centimeter, $0 \models \Box p$, that is I do know that an object of size 0 will fit in the box. But if one assumes positive introspection to be valid, it also holds that $0 \models \Box\Box p$, and so $1 \models \Box p$. By repeated applications of the same rule, it follows that $i \models p$ for every $i \geq 0$, a plain contradiction if p does not hold universally in the model. Thus, it should follow from my knowledge that an object of size 0 will fit in the box that any object, whatever its size, will fit in the box. Putting together margin for error and positive introspection, we thus end up with a form of epistemic sorites, on the basis of which Williamson argues that positive introspection does not hold. Exactly the same reasoning can be performed if one assumes negative, instead of positive introspection.

This result does not depend on the model, and is even more general, since for every formula ϕ,

[1] T is the axiom schema $\Box p \to p$, and B is the schema $p \to \Box\Diamond p$. Williamson also presents a *variable* margin semantics, relative to which the logic $\mathbf{KT}$ is sound and complete. For lack of space, we do not consider it here, although the results of the following sections would carry over to it.

$\phi \rightarrow \Box\phi$ is valid in fixed margin models if and only if either ϕ is valid or $\neg\phi$ is valid (Williamson 1992b, 1994). In particular, positive introspection (and likewise negative introspection) is valid in fixed margin models if and only if either $\Box\phi$ is valid, or $\neg\Box\phi$ is valid. Thus, I know that I know ϕ (resp. don't know ϕ) at some world only if either I know ϕ at every world, or if I don't know ϕ at every world. As a result, if I know a proposition to hold at some world, and if that proposition is contingent, it is inconsistent to assume positive introspection in full generality (and similarly for negative introspection). This result if fairly dramatic, for it seems to show that whenever knowledge obeys a margin for error principle that applies non-trivially, knowledge can't be introspective unconditionally. The problem may be summarized in the following rough terms: *knowledge can't be vague while obeying positive or negative introspection at the same time*.

Contrary to Williamson, we don't consider this conclusion to be sound. From a psychological point of view, as argued by Dokic & Égré (2004) and Égré (forthcoming), Williamson's argument rests on the controversial assumption that all levels of knowledge obey the same kinds of margin for error. In particular, it presupposes that my visual knowledge and the knowledge I have about my visual knowledge both rest on the same discriminative capacities, an assumption we can't take for granted. From a logical point of view, moreover, the result is relative to the semantics given to $\Box$. As Williamson notes, the failure of positive and negative introspection in margin models reflects the non-transitive and non-euclidian character of the perceptual indiscriminability relation. Conversely, imposing positive and negative introspection "from the outside" forces knowledge to hold universally or nowhere throughout the model. But if one wishes both to model a notion of inexact knowledge, obeying margin for error principles, and yet to preserve the introspective properties, a possibility is to revise the standard semantics for knowledge. We are therefore led to following question:

Question 1. *Is there a non-standard semantics suitable to validate introspection (either positive or negative) and which would still be adequate to model the notion of inexact knowledge?*

We give a positive answer to this question in the next section. More precisely, what we are looking for is a semantics for knowledge based on the non-transitive and non-euclidian property of indiscriminability, but nevertheless adequate to support introspection (whether positive or negative). We argue that the semantics is also plausible from a cognitive point of view, namely that it corresponds to a system of modular knowledge, in which higher-order knowledge is not necessarily on a par with knowledge at the low level.

2 A Centered Semantics for knowledge

2.1 The new semantics

In the standard semantics, it takes 2 steps from a given world to check whether an iterative formula of the form $\Box\Box p$ holds at that world, and more generally it takes n transitions within a model to check for the satisfiability of a formula with n nested operators. In a situation of perceptual knowledge like the one pictured in Figure 1.1, this property is at odds with our intuition:

looking at an object of size 10, I know it will fit in the box, and yet the semantics predicts that I don't know that I know it, since an object of size 12 doesn't fit in the box. However, it seems that one's reflective knowledge should not depend on such remote epistemic alternatives. To restore that intuition, we define a "centered semantics" in which the epistemic alternatives relevant for iterated modalities remain the worlds accessible in one transition from the world of evaluation. In other words, every fact concerning the knowledge of the agent should be decided solely on the basis of worlds that are not distinguishable from that world, without having to move further along the accessibility relation. Given a model $\mathcal{M} = \langle W, R, V \rangle$, we first define the notion of satisfaction for couples of worlds, and extract the definition of satisfaction for single worlds:

Definition 1. *Satisfaction for couples of worlds:*

(i) $\mathcal{M}, (w, w') \vDash_{\text{CS}} p$ iff $w' \in V(p)$.
(ii) $\mathcal{M}, (w, w') \vDash_{\text{CS}} \neg\phi$ iff $\mathcal{M}, (w, w') \nvDash_{CS} \phi$.
(iii) $\mathcal{M}, (w, w') \vDash_{\text{CS}} (\phi \wedge \psi)$ iff $\mathcal{M}, (w, w') \vDash_{\text{CS}} \phi$ and $\mathcal{M}, (w, w') \vDash_{\text{CS}} \psi$.
(iv) $\mathcal{M}, (w, w') \vDash_{\text{CS}} \Box\phi$ iff for all w'' such that wRw", $\mathcal{M}, (w, w'') \vDash_{\text{CS}} \phi$.

Definition 2. $\mathcal{M}, w \vDash_{\text{CS}} \phi$ *iff* $\mathcal{M}, (w, w) \vDash_{\text{CS}} \phi$

Clause (iv) of the definition accounts for the "centered" feature of the semantics, for it entails that for every w and w': $\mathcal{M}, (w, w') \vDash_{\text{CS}} \Box\phi$ iff $\mathcal{M}, (w, w) \vDash_{\text{CS}} \Box\phi$ iff $\mathcal{M}, w \vDash_{\text{CS}} \Box\phi$. This ensures that instead of looking at worlds that are two steps away to check whether $\Box\Box\phi$ is satisfied, one backtracks to the actual world to see whether $\Box\phi$ already holds there.[2] For instance, relative to the model $\mathcal{M}$ of Figure 1.1, it holds that $\mathcal{M}, 10 \vDash_{\text{CS}} \Box p$, $\mathcal{M}, 11 \vDash_{\text{CS}} \neg\Box p$, and $\mathcal{M}, 12 \vDash_{\text{CS}} \Box\neg p$, just as with the standard semantics. However, we now have: $\mathcal{M}, 10 \vDash_{\text{CS}} \Box\Box p$, and likewise for any further level of iteration. Interestingly, it also holds that $\mathcal{M}, 11 \vDash_{\text{CS}} \Box\neg\Box p$. More generally, the semantics validates both positive and negative introspection, and we can prove the following completeness theorem:

2.2 K45 is sound and complete with respect to CS

Given a Kripke model $\mathcal{M}$, we say that $\mathcal{M}$ CS-validates ϕ, and we write $\mathcal{M} \vDash_{\text{CS}} \phi$ if and only if for every world w of the model, $\mathcal{M}, w \vDash_{\text{CS}} \phi$. We call a formula ϕ CS-valid, and we write $\vDash_{\text{CS}} \phi$, if every model $\mathcal{M}$ CS-validates ϕ.

Theorem 1. K45 *is sound with respect to* CS.

Proof. It is straightforward to check that (CS) validates axioms K, 4 and 5, and that modus ponens and uniform substitution preserve validity. The only non-trivial case concerns the rule of necessitation. Suppose $\vDash_{\text{CS}} \phi$, but $\nvDash_{\text{CS}} \Box\phi$. So there is a model $\mathcal{M} = \langle W, R, V \rangle$ and a couple of worlds (w, w') such that $\mathcal{M}, (w, w') \nvDash_{\text{CS}} \phi$. Consider any model $\mathcal{M}' = \langle W, R', V \rangle$ with the

[2] As pointed out to us by P. Schlenker and J. van Benthem (p.c.), the double-indexed semantics we use is closely related to H. Kamp's 1971 semantics for the operator "Now", since the box allows to reset the index of evaluation to the initial world, in the same way in which "Now" resets the moment of evaluation to the moment of utterance.

same domain and valuation as $\mathcal{M}$, but in which $R'(w') = R(w)$. We first show by induction that for any formula ϕ, $\mathcal{M}, (w, x) \vDash_{\text{CS}} \phi$ iff $\mathcal{M}', (w', x) \vDash_{\text{CS}} \phi$. The atomic and boolean cases are straightforward. Consider $\phi := \Box\psi$. $\mathcal{M}, (w, x) \vDash_{\text{CS}} \Box\psi$ iff for every v such that wRv, $\mathcal{M}, (w, v) \vDash_{\text{CS}} \psi$ iff for every v such that $w'R'v$, $\mathcal{M}, (w, v) \vDash_{\text{CS}} \psi$ (by definition of R'), iff for every v such that $w'R'v$, $\mathcal{M}', (w', v) \vDash_{\text{CS}} \psi$ (by induction hypothesis), iff $\mathcal{M}', (w', x) \vDash_{\text{CS}} \Box\phi$. From this, it follows that $\mathcal{M}, (w, w') \vDash_{\text{CS}} \phi$ iff $\mathcal{M}', (w', w') \vDash_{\text{CS}} \phi$. So if we suppose that ϕ is CS-valid but nevertheless such that $\mathcal{M}, (w, w') \nvDash_{\text{CS}} \phi$, then we should have $\mathcal{M}', (w', w') \nvDash_{\text{CS}} \phi$, that is $\mathcal{M}', w' \nvDash_{\text{CS}} \phi$, and ϕ could not be valid.

□

The proof that the rule of necessitation preserves validity is slightly more complicated than the usual proof given for the standard semantics. The reason is that in the standard semantics, the rule of necessitation also holds within models: given a model $\mathcal{M}$, if $\mathcal{M} \models \phi$, then it follows that $\mathcal{M} \models \Box\phi$. Another way to put it is to say that necessitation is not only *frame-valid*, but also *model-valid* for the standard semantics. Relative to CS, however, necessitation is only frame-valid. Consider, for instance, a model $\mathcal{M}$ with three worlds w, w', w'' such that $V(q) = \{w, w''\}$ and $V(p) = \{w'\}$, and in which wRw' and $w'Rw''$. Clearly $\mathcal{M} \vDash_{\text{CS}} \Box p \to q$ but $\mathcal{M} \nvDash_{\text{CS}} \Box(\Box p \to q)$.

Theorem 2. $\mathbf{K45}$ *is complete with respect to* (CS).

Proof. We rely on the standard completeness proof for $\mathbf{K45}$: $\mathbf{K45}$ is sound and complete w.r.t. the class of transitive and euclidean frames. Let us assume for contradiction that $\mathbf{K45}$ is not complete for CS. This means that there is a sentence ϕ such that $\vDash_{\text{CS}} \phi$ but $\nvdash_{\mathbf{K45}} \phi$. By completeness, there is a transitive and euclidean model $\mathcal{M}$ and a world w_0 such that $\mathcal{M}, w_0 \nvDash \phi$. We show that this model contradicts $\vDash_{\text{CS}} \phi$, by showing that $\mathcal{M}, w_0 \nvDash_{\text{CS}} \phi$.

Lemma 1. *Let $\mathcal{M}$ be a transitive and euclidean model of modal logic* (ML) and ϕ an ML-formula. Then for every world w_0, $\mathcal{M}, w_0 \vDash \phi$ iff $\mathcal{M}, w_0 \vDash_{\text{CS}} \phi$

We show by induction on the length of the formula that for every world w in the submodel generated from w_0, $\mathcal{M}, w \vDash \phi$ iff $\mathcal{M}, (w_0, w) \vDash_{\text{CS}} \phi$. The lemma follows immediately, since $\mathcal{M}, w_0 \vDash_{\text{CS}} \phi$ iff $\mathcal{M}, (w_0, w_0) \vDash_{\text{CS}} \phi$. The only non-trivial case is $\phi = \Box\psi$. Since R is euclidean and transitive, it holds that wRw' iff w_0Rw' (Assume wRw'. By transitivity w is reachable from w_0, and wRw', so by transitivity again we have that w_0Rw'. Now assume w_0Rw'. Since w is reachable from w_0 and R is transitive, w_0Rw. By euclideanity, wRw' ensues). We then have:

$\mathcal{M}, w \vDash \Box\psi$ iff for all w' such that wRw', $\mathcal{M}, w' \vDash \psi$, by definition of $\vDash$
iff for all w' such that wRw', $\mathcal{M}, (w_0, w') \vDash_{\text{CS}} \psi$, by induction hypothesis.
iff for all w' such that w_0Rw', $\mathcal{M}, (w_0, w') \vDash_{\text{CS}} \psi$, by the property of R.
iff $\mathcal{M}, (w_0, w) \vDash_{\text{CS}} \Box\psi$, by definition of $\vDash_{\text{CS}}$.

□

The lemma shows that the shift from the standard semantics to the centered semantics preserves satisfaction on the class of transitive and euclidean models. This does not mean that CS is

just a trivial rewording of the definition of satisfaction, because $\vDash$ and $\vDash_{\text{CS}}$ do not match in general: the previous proof rests in an essential way on the assumption that the accessibility relation is transitive and euclidean. The important fact is thus that our stock of models is now bigger: we have at our disposal not only the transitive euclidean models, but the full class of models, without having to relinquish the introspection principles. This includes, in particular, non-transitive and non-euclidian models like the model of Figure 1.1.

The model of Figure 1.1, it may be recalled, may also be seen as fixed-margin model $\langle W, d, \alpha, V\rangle$ with margin of error $\alpha = 1$. As a matter of fact, the completeness theorem we stated for $\mathbf{K45}$ with respect to the centered semantics can be turned into a completeness theorem for $\mathbf{S5}$ with respect to Williamson's fixed-margin semantics. Given a fixed margin model $\mathcal{M} = \langle W, d, \alpha, V\rangle$, we define a centered fixed-margin semantics (CMS), paralleling the definition of CS. The definition of satisfaction (for couple of worlds) is the same for the atomic and boolean cases, and becomes, for the $\Box$:

Definition 3. $\mathcal{M}, (w, w') \vDash_{\text{CMS}} \Box\phi$ *iff for every* v *such that* $d(w, v) \leq \alpha$, $\mathcal{M}, (w, v) \vDash_{\text{CMS}} \phi$

As before, we define $\mathcal{M}, w \vDash_{\text{CMS}} \phi$, iff $\mathcal{M}, (w, w) \vDash_{\text{CMS}} \phi$. A fixed-margin model $\mathcal{M}$ CMS-validates ϕ iff every world of the model CMS-satisfies ϕ; $\vDash_{\text{CMS}} \phi$ iff every margin model CMS-validates ϕ. We know that Williamson's fixed margin semantics (FM) is sound and complete for $\mathbf{KTB}$, and that (CS) is sound and complete for $\mathbf{K45}$. Putting together the results, we get:

Theorem 3. $\mathbf{S5}$ *is sound and complete with respect to* (CMS)

Proof. Every fixed margin model with parameter α can be seen as a reflexive symmetric standard model such that wRv iff $d(w, v) \leq \alpha$. From this, it follows that if $\vDash_{\text{CS}} \phi$ then $\vDash_{\text{CMS}} \phi$, and so $\mathbf{K45}$ is sound w.r.t CMS. Moreover, CMS validates T (and B), so CMS is sound for $\mathbf{S5}$. Completeness is just an adaptation of Lemma 1: given a reflexive euclidian model $\mathcal{M}$ of S5, one can see it as a fixed margin model $\mathcal{M}^*$ with parameter $\alpha = 0$, setting $d(w, v) = 0$ iff wRv; from Lemma 1 it can be checked that $\mathcal{M}, w \models \phi$ iff $\mathcal{M}^*, w \vDash_{\text{CMS}} \phi$.

$\Box$

From the standpoint of epistemic logic, $\mathbf{K45}$ can be seen as a system of introspective *belief*. With the inclusion of axiom T, $\mathbf{S5}$ is a system of introspective *knowledge* properly so called, and we take the completeness of $\mathbf{S5}$ with respect to the centered margin semantics to give a positive answer to the Question raised in the previous section.

3 Luminosity and knowledge iterations

3.1 Luminosity

Williamson's argument against positive (as well as negative) introspection is part of a more general argument against the so-called *luminosity* of mental states. Williamson calls a mental state luminous if and only if the occurrence of the state entails the knowledge that one is in that state (Williamson 2000, chap. 4). According to Williamson, no non-trivial mental state is luminous, a

non-trivial state being defined as a state that lasts for some time, but not all the time (2000: 107). This psychological claim rests on the idea that knowledge about one's mental states, in order to be reliable, obeys a margin of error, and is backed up by Williamson's result that $\phi \to \Box\phi$ is valid in (fixed or variable) margin semantics if and only if either ϕ is valid or $\neg\phi$ is valid. In centered semantics, however, it no longer holds that $\vDash_{\mathrm{CS}} \phi \to \Box\phi$ iff either $\vDash_{\mathrm{CS}} \phi$ or $\vDash_{\mathrm{CS}} \neg\phi$, as shown by the fact that $Kp \to KKp$ is CS-valid, but neither Kp nor $\neg Kp$ is CS-valid in the model of Figure 1.1. If, like Williamson, we admit that knowing can be a mental state, then this suggests that at least states of knowledge may be luminous without being trivial.

To be sure, consider close situations in which I'm asked whether I feel cold or not. Let the model of Figure 1.1 now represent a thermometric scale, where p stand for "I feel cold", with the assumption that I cannot perceptually discriminate between any two situations that differ only by 1°C. The model depicts a case in which I feel cold up to 11°C, and start not to feel cold from 12°C onward. Standard Kripke semantics, like Williamson's fixed margin semantics (with a margin of 1°C), predicts that at the world where the temperature is 12°C, I start not to feel cold, but don't know yet that I no longer feel cold. At the world where the temperature is 13°C, I know I don't feel cold, but don't know that I know this, due to the standard semantics: hence neither my feeling cold, nor my knowing that I feel cold is luminous. Using the centered semantics, it still holds that at 12°C I don't know yet that I start not to feel cold, but at 13°C I know it, and know that I know it. Relative to the centered semantics, the model of Figure 1.1 depicts a situation in which feeling cold is not a luminous condition, but in which my knowing that I feel cold is luminous. Thus we may agree with Williamson that *not all* mental states are luminous, but nevertheless disagree on the idea that *no* non-trivial mental state is luminous.

Which semantics is more plausible from a psychological point of view is a question that should be decided upon empirical grounds and that we shall leave open for psychological investigation. That being said, it may be argued that our centered semantics is too quick to make knowledge insensitive to iterations. There may be situations, for instance, in which my knowing p is not sufficient to warrant my knowing that I know p, even though my knowing that I know p is sufficient to warrant any further level of iteration. We could imagine, for instance, that at 13°C I just become aware that I am not cold, but that in order for me to become aware of this awareness, the temperature should reach at least 14°C. More generally, we may conceive like Williamson that the higher-order awareness we have of our perceptual states comes in degrees which co-vary with the intensity of the perceptual stimulus, but only up to a point, from which iterations become insensitive. In the following subsection, we state a semantics which makes room for that possibility, and which actually allows to set the collapse between modalities at any arbitrary level.

3.2 Token semantics

The semantics, which we call "token semantics" for short, is a parameterized version of centered semantics. Satisfaction is defined with respect to a sequence of worlds and a number of tokens: q is short for an arbitrary sequence of worlds, qw for an arbitrary sequence with last item w, and n is an arbitrary number of tokens.

Definition 4. *Token satisfaction:*

(i) $\mathcal{M}, qw \vDash_{\text{TS}} p\ [n]$ iff $w \in V(p)$.
(ii) $\mathcal{M}, qw \vDash_{\text{TS}} \neg\phi\ [n]$ iff $\mathcal{M}, qw \nvDash_{\text{TS}} \phi\ [n]$.
(iii) $\mathcal{M}, qw \vDash_{\text{TS}} (\phi \wedge \psi)\ [n]$ iff $\mathcal{M}, qw \vDash_{\text{TS}} \phi\ [n]$ and $\mathcal{M}, qw \vDash_{\text{TS}} \psi\ [n]$.
(iv) $\mathcal{M}, qw \vDash_{\text{TS}} \Box\psi\ [n]$ iff
- $n \neq 0$ and for all w' such that wRw', $qww' \vDash_{\text{TS}} \psi\ [n-1]$
- Or $n = 0$ and $q \vDash_{\text{TS}} \Box\psi\ [1]$.

Definition 5. *Let n be such that $1 \leq n \leq \omega$ (we assume $\omega - 1 = \omega$).* TS(n) is the modal semantics defined by the following satisfaction relation: $\mathcal{M}, w \vDash_{\text{TS}(n)} \phi$ iff $\mathcal{M}, w \vDash_{\text{TS}} \phi\ [n]$.

Thus, a token is spent for each move along the accessibility relation. When all tokens have been spent, accessible worlds are those which were accessible when there was no token left but one. As intended, centered semantics and standard semantics come out as special cases:

Fact 1.
- $\mathcal{M}, w \vDash_{\text{CS}} \phi$ *iff* $\mathcal{M}, w \vDash_{\text{TS}} \phi\ [1]$
- $\mathcal{M}, w \vDash \phi$ *iff* $\mathcal{M}, w \vDash_{\text{TS}} \phi\ [\omega]$

Now, what happens with the introspection principles? CS validates 4 and 5, but this is not true for every TS(n) semantics: it is easy to see that 4 and 5, though TS(1)-valid, are not TS(n)-valid for $n \geq 2$. But there is a generalized form of these principles which is correlated to our family of semantics. Let $\Box^n$ be short for $\Box...\Box$, n times. We consider:

$$(4n) \quad \Box^n p \rightarrow \Box^n \Box p$$

Intuitively $4n$ is like just like 4, but it works only at a minimum distance of n steps from the starting point. $4n$ can be used to capture the cognitive limitation corresponding to the inability to distinguish meta-representations involved in self-knowledge at levels beyond n: it says that knowing that one knows etc...(n times) implies knowing that one knows etc...($n+1$ times). Similarly, 5 can be generalized to:

$$(5n) \quad \Diamond^n p \rightarrow \Diamond^{n-1} \Box \Diamond p$$

For the purpose of completeness, $(4n)$ and $(5n)$ are too weak, however: some TS(n) validities are not theorems of the system $\mathbf{K4n5n}$. As a matter of fact, $(4n)$ and $(5n)$ turn out to be particular instances of two more general schemata, namely:

$$(4n^*) \quad \Diamond(p_1 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_n \wedge \Diamond r)...) \rightarrow \Diamond(p_1 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Diamond r)...)$$

$$(5n^*) \quad \Diamond(p_1 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Diamond r)...) \rightarrow \Diamond(p_1 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Box\Diamond r)...)$$

Theorem 4. *For every integer n,* $\mathbf{K4n^*5n^*}$ *is sound and complete with respect to* TS(n).

Proof. We just sketch the completeness proof, which generalizes the method of Theorem 2. First, note that $4n^*$ defines the class of n-transitive frames, *i.e.* the class of frames satisfying:

$\forall x_1, ..., x_{n+2}\ ((x_1Rx_2 \wedge ... \wedge x_{n+1}Rx_{n+2}) \rightarrow x_nRx_{n+2})$

Similarly, $5n^*$ defines the class of n-euclidean frames, *i.e.* the class of frames satisfying:

$\forall x_1, ..., x_{n+2}\ ((x_1Rx_2 \wedge ... \wedge x_nRx_{n+1} \wedge x_nRx_{n+2}) \rightarrow x_{n+2}Rx_{n+1})$

$4n^*$ and $5n^*$ are Sahlqvist formulas. Therefore, by the Sahlqvist completeness theorem (see Blackburn & al. 1999, c. 4), $\mathbf{K4n^*5n^*}$ is complete for the class of n-euclidean and n-transitive frames. The following generalization of Lemma 1 holds:

Lemma 2. *Let $\mathcal{M}$ be an n-transitive and n-euclidean model and ϕ an arbitrary sentence. Then for every world w_0, $\mathcal{M}, w_0 \vDash \phi$ iff $\mathcal{M}, w_0 \vDash_{\mathrm{TS}(n)} \phi$*

The proof follows that of Lemma 1. Let w be a world reachable from w_0 in $n-1$ steps: since R is n-euclidian and n-transitive, it holds that the R-successors of any R-successor of w are exactly the R-successors of w.

□

The present generalization of centered semantics thus offers a compromise between the standard semantics (up to degree n) and the non-standard semantics (beyond n). The following yields the corresponding triviality results for the modalities:

Theorem 5. *Let ϕ be any basic modal sentence: for all $n \geq 1$, there is a sentence ϕ_n of degree at most n, such that ϕ is* TS(n)-equivalent to ϕ_n.

Proof. ϕ_n is obtained from ϕ by syntactic transformation. Let ψ be a maximal subformula of ϕ embedded under $n-1$ modalities (*i.e.* all subformulas containing ψ are embedded under fewer than $n-1$ modalities). By well-known results, ψ can be turned into a formula ψ' of degree at most 1 which is $\mathbf{K45}$-equivalent to ψ. Replace ψ by ψ'. We get ϕ_n by applying this operation to all maximal subformulas embedded under $n-1$ modalities. It follows from the definition of TS(n) that ϕ_n is TS(n)-equivalent to ϕ.

□

In the same way in which we formulated a centered version of Williamson's margin semantics with parameter 1, for any n we could formulate a centered margin semantics with parameter n, yielding an analogous completeness result for $\mathbf{KTB4n^*5n^*}$, corresponding to a modular logic of inexact knowledge with introspection, standard up to n, and non-standard beyond.

3.3 Multi-Dimensional Modal Logic

Before closing this section, we would like to make the premises of our response to Williamson's argument against the introspection principles even more explicit, in order to give a better sense of the use of our token semantics and to compare it to one alternative formulation in the framework of multi-dimensional modal logic. What we saw initially is that the standard Kripke semantics for modal logic is not adequate to model the notion of inexact knowledge with introspection. The main premises of the argument leading to this conclusion may be recalled here:

1. The principles of positive and negative introspection are valid in Kripke structures if and only if the relation of accessibility is transitive and euclidian.

2. Situations of inexact knowledge are characterized by the fact that the underlying relation of perceptual similarity is not transitive or euclidian.

3. To adequately model the notion of inexact knowledge using Kripke structures, one may identify the relation of epistemic accessibility with the relation of perceptual similarity.

Premise 1 states a mathematical fact about the standard Kripke semantics. Premises 2 and 3 are more of methodological nature. Premise 2 serves a definitional purpose and and should be common to most, if not all, accounts of the notion of inexact knowledge. Premise 3, on the other hand, is not as obvious. It is a natural assumption to make in so far as the relation of perceptual similarity induces a relation of epistemic uncertainty. But it is not obvious precisely because the epistemic uncertainty linked to higher levels of knowledge need not to be based on the same relation of perceptual similarity. We shall call assumption 3 the *raw import* hypothesis, since it is tantamount to "importing" the relation of perceptual similarity directly into the models, and therefore to identifying the epistemic states with the perceptual alternatives themselves.

Williamson accepts premises 2 and 3 in his logic of inexact knowledge. Since his logic rests on a standard Kripke semantics, he is led to reject the introspection principles, in virtue of the fact stated in 1. Like Williamson, we assumed premises 2 and 3 in our treatment of the notion of inexact knowledge. Unlike him, however, we think that at least some of the scenarios that he considers should be compatible with the introspection principles. Our goal, therefore, has been to challenge premise 1, by showing the relativity of the correspondence result to the standard semantics. One may wonder, however, about the logical status of premise 3. Wouldn't it be possible to make a different move and to change the models, instead of shifting to our non-standard semantics?

The answer is positive. Token semantics, and centered semantics as a particular case thereof, are indeed quite reminiscent of traditional multi-dimensional modal logics (MDML). In TS, satisfaction is defined in a non-standard way with respect to sequences of worlds. The semantics for MDML is the standard Kripke semantics, but models have one special feature: the states have some inner structure – usually they are tuples or sequences over some base structure – and the accessibility relation between the states is partly determined by that inner structure (Blackburn & al., 1999, p. 459). The framework of MDML can in fact be used to embed (TS) in a standard setting, thereby providing an alternative way of looking at TS-models as shorthand for bigger Kripke models, describable by means of standard satisfaction clauses.

Definition 6 (n-model unpacking)**.** *Let $\mathcal{M} = \langle M, R, V\rangle$ be a Kripke model and n be a number of tokens. The n-unpacking of $\mathcal{M}$, $UP_n(\mathcal{M})$, is a Kripke structure $\langle M \times M \times \{0, ..., n\}, R', V'\rangle$ with*

- $(w, w', i)R'(v, v', j)$ *iff either* $i > 0$, $j = i - 1$, $v = w'$ *and* $w'Rv'$, *or* $i = j = 0$, $v = w$ *and* wRv'.

- $(w, w', i) \in V'(p)$ *iff* $w' \in V(p)$

Theorem 6. *For any formula ϕ, $\mathcal{M}, w \vDash_{\mathrm{TS}(n)} \phi$ iff $UP_n(\mathcal{M}), (w, w, n) \vDash \phi$.*

Proof. By induction on ϕ.

□

Theorem 6 therefore shows that assumption 3 above can be put into question as much as assumption 1. What can be achieved with token semantics can indeed be achieved in a multi-dimensional setting, keeping the semantics standard. In order to do this, however, one needs to construct a different accessibility relation on the basis of the relation of perceptual indiscriminability – these constructions are precisely what is defined as the n-unpacking of a model: in an unpacked model, the epistemic states encode not only the relation of similarity between perceptual alternatives, but also the cost of metarepresentations.

Despite this, we do not think that TS should be considered as a mere trick in order to get at the "true" Kripke models involved in situations of inexact knowledge. The reason is that raw import speaks for itself. It remains indeed very natural to think that the accessibility relation, construed as a relation of epistemic indistinguishability, should be just the same as perceptual similarity in contexts in which the primary information the agent receives comes from his sensory perception. Moreover, raw import is certainly the simplest way of building Kripke models on the basis of a relation of perceptual similarity. In this respect, TS offers a simple way of making inexact knowledge compatible both with introspection principles and raw import.

To complete this discussion, it would be worth discussing the relationship between TS and other variations over the standard modal semantics that also rely on the idea of sequential evaluation. For lack of space, we only mention two results here. First, token semantics can be shown to be a particular case of hypermodal logic in the sense of Gabbay (2002), in which modalities receive different interpretations depending on where they occur in a formula. Likewise, it is possible to use the extended modal logic IFML of Tulenheimo (2004) to get a satisfaction-preserving translation of the modal formulas, using backward-looking modalities.

4 The multi-agent case and common knowledge

In the previous sections we discussed the problems of iterations of knowledge in the case of a single agent only. In this section, we present a generalization of token semantics to the case of several agents. Interestingly, the generalization casts light on a paradox of common knowledge analogous to Williamson's luminosity paradox for the multi-agent case.

We should note, to begin with, that there is a major conceptual difference between the case of a single agent and the case of several agents with respect to iterations of knowledge. In the mono-agent case, we argued that in a situation such as the one entertained by Williamson, it is natural to suppose that knowing that one knows depends only on whether one knows, and not on further external features of the world. To use a vocabulary common in the philosophy of mind: higher-order knowledge *supervenes* on first-order knowledge only (at least with respect to CS), and first-order knowledge in turn supervenes on a limited number of states of the model. In the multi-agent case, things are likely to be different. Indeed, b may know p without a knowing that b knows; and likewise, a may know that b knows without b knowing that a knows that b

knows, and so on. A well-known example of a situation of that kind is the coordinated attack problem, where two generals commanding distinct divisions will launch an attack only if each one is "absolutely sure that the other will attack with him" (Fagin & al. 1995: 176). General a sends a message to b to say he plans to attack at dawn. General b receives it and answers to a to acknowledge the message. But then a has to answer to b to give him assurance that he knows that b received the message, and so on and so forth. A situation like this one is a situation in which common knowledge is never attained. Equivalently, it may be described as a situation in which no iteration of knowledge is ever made at no cost.

There clearly are, however, situations where common knowledge is much easier to achieve. Suppose a two player card game where each agent receives a card that the other can't see. Each player knows their own card, and each player knows that each player knows their own card, and so on. By contrast to the coordinated attack problem, this is a situation where common knowledge is attained statically. From a model-theoretic point of view, this corresponds to the fact that the information "x knows his own card" is actualized at every world of the Kripke model representing the different possible situations. There are, however, static situations where common knowledge fails and which nevertheless tend to create puzzles. An example is provided by the puzzle of Consecutive Numbers, where two agents each are given a positive number without knowing the number of the other (see van Ditmarsch & al. 2003). The rule is that the numbers are consecutive. Hence, it is common knowledge that the numbers are consecutive, and that every player knows their own number. A situation where a receives number 2, and b number 3 may be described by the following Kripke structure, where the states are coded by the distribution of numbers to the players:

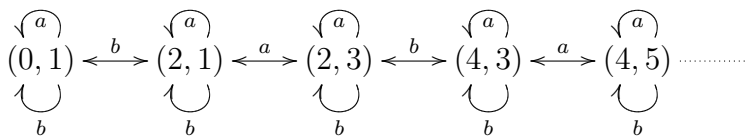

Figure 1.2: Consecutive Numbers

Thus (2,3) represents the state where a has a 2 and b has a 3, and analogously for the other states. Common knowledge can be represented by the operator $C_{a,b}\phi$, standing for the infinitary conjunction $(\phi \wedge \Box_a\phi \wedge \Box_b\phi \wedge \Box_a\Box_a\phi \wedge \Box_b\Box_b\phi \wedge \Box_a\Box_b\phi \wedge \Box_b\Box_a\phi \ldots)$. More compactly, common knowledge can be written in terms of the operator of shared knowledge, $E_{a,b}\phi$, standing for the conjunction $(\Box_a\phi \wedge \Box_b\phi)$, by letting $C_{a,b}\phi$ stand for the infinitary conjunction $(\phi \wedge E_{a,b}\phi \wedge E_{a,b}E_{a,b}\phi \wedge \ldots)$. In the standard semantics, $M, w \models E_{a,b}\phi$ if and only if ϕ holds at every world w' that belongs to $R_{a\cup b}(w)$, where $R_{a\cup b}$ denotes the union of the epistemic accessibility relations R_a and R_b of a and b. Likewise, $M, w \models C_{a,b}\phi$ if and only if ϕ holds at every world that belongs to the reflexive transitive closure of $R_{a\cup b}$. Thus, in the above structure and relative to state (2,3), it can be checked that for any number $n \geq 3$, it is not common knowledge between a and b that their number is less than n. For instance, in the situation in which a holds a 2 and b holds a 3, the semantics predicts that it is not common knowledge between a and b that their numbers are less than 100000. Indeed, there is a path from $(2, 3)$ to $(100000, 100001)$ such that the sentence $(\Diamond_a\Diamond_b)^{49999}100000_a$ holds at $(2, 3)$.

This situation is generally a source of bewilderment (for instance the first time it is presented to students). To motivate it, one generally reasons as follows: in the case in which a has a 2 and conceives that b might have a 3, a considers possible that b considers possible that a has a 4, namely $\Diamond_a \Diamond_b 4_a$. Likewise, it is conceivable for a that b thinks that a thinks b might have a 5, namely $\Diamond_a \Diamond_b \Diamond_a 5_b$. The same reasoning, it seems, can be maintained by adding one level of iteration, and one does not see any good reason to stop. Intuitively, on the other hand, it seems safe to say that a knows that a's number is less than 100000, b knows it too, and a knows that he knows, and so on and so forth. The intuitive reason is that a and b both know that their own number is far below 100000, and each knows that the number of the other is far below 100000, and each one knows the other knows this, indefinitely.

The problem, at this stage, is to determine to what extent this intuition may be relied upon: does the standard semantics really make too strong predictions about the worldly notion of common knowledge? Or is the intuition that common knowledge can be attained in this scenario a kind of cognitive illusion? It is hard to adjudicate between these two options. Formally, however, one may note that the puzzle of common knowledge for consecutive numbers is exactly analogous to the puzzle which concerns the failure of positive introspection for Williamson's scenario. Indeed, if we consider the union $R_{a \cup b}$ of the epistemic accessibility relations of a and b, we can see that the structure of Figure 1.2 is isomorphic to the structure of inexact knowledge of Figure 1.1. Like R_a and R_b, $R_{a \cup b}$ is reflexive and symmetric, but it is not transitive. Since common knowledge amounts to taking the reflexive transitive closure of $R_{a \cup b}$, the operators $E_{a,b}$ and $C_{a,b}$ will be equivalent if one forces $E_{a,b}$ to satisfy $E_{a,b}\phi \rightarrow \phi$ and $E_{a,b}\phi \rightarrow E_{a,b}E_{a,b}\phi$. In the case of consecutive numbers, it is shared knowledge between a and b that their numbers are less than 100000, and this knowledge is reflexive, but not transitive, so that "positive introspection" fails in the model for $E_{a,b}$. By analogy to the single-agent case, what we are wondering is whether it would make sense to ask for something like a notion of positive introspection for shared knowledge, without forcing the underlying relation to be transitive.

To answer this problem, we can generalize our token semantics to the case of several agents. Just as there was only one token parameter for one agent, it is natural to use n token parameters when there are n agents. Given a multi-agent model $M = \langle W, R_1, ..., R_n, V \rangle$, we note m_i the number of tokens initially available to agent i.

Definition 7. *Token satisfaction for n-agents:*

(i) $\mathcal{M}, qw \vDash_{\text{TS}} p\ [m_1, ..., m_n]$ iff $w \in V(p)$.
(ii) $\mathcal{M}, qw \vDash_{\text{TS}} \neg\phi\ [m_1, ..., m_n]$ iff $\mathcal{M}, qw \nvDash_{\text{TS}} \phi\ [m_1, ..., m_n]$.
(iii) $\mathcal{M}, qw \vDash_{\text{TS}} (\phi \wedge \psi)\ [m_1, ..., m_n]$ iff $\mathcal{M}, qw \vDash_{\text{TS}} \phi\ [m_1, ..., m_n]$ and $\mathcal{M}, qw \vDash_{\text{TS}} \psi\ [m_1, ..., m_n]$.
(iv) $\mathcal{M}, qw \vDash_{\text{TS}} \Box_i\psi\ [m_1, ..., m_n]$ iff
- $m_i \neq 0$ and for all w' such that wR_iw', $qww' \vDash_{\text{TS}} \psi\ [m_1, \ldots, m_i - 1, \ldots, m_n]$
- Or $m_i = 0$ and $q \vDash_{\text{TS}} \Box_i\psi\ [m_1, \ldots, 1, \ldots, m_n]$.

To illustrate how the semantics works, let us go back to the puzzle of Consecutive Numbers, supposing that each agent has only one token. One can check that, just as in the standard semantics:

$M, (2,3) \vDash_{\text{TS}} \Box_a \Box_b (0_a \vee 2_a \vee 4_a)\ [1,1]$
$M, (2,3) \vDash_{\text{TS}} \Box_b \Box_a (1_b \vee 3_b \vee 5_b)\ [1,1]$

Thus, a knows that b knows that a has an even number no greater than 4. Likewise b knows that a knows that b has an odd number no greater than 5. Unlike in the standard semantics, however, we have:

$M, (2,3) \vDash_{\text{TS}} \Box_b \Box_a \Box_b (0_a \vee 2_a \vee 4_a)\ [1,1]$
$M, (2,3) \vDash_{\text{TS}} \Box_a \Box_b \Box_a (1_b \vee 3_b \vee 5_b)\ [1,1]$

and likewise for any further level of embedding, since for any two-step history q from $(2,3)$, we have for instance $M, (2,3)q \vDash_{\text{TS}} (0_a \vee 2_a \vee 4_a)\ [0,0]$. Thus, if we let p stand for the proposition that "a and b each have a number no greater than 5", it holds that $M, (2,3) \vDash_{\text{TS}} E_{a,b}p\ [1,1]$, and for every m, $M, (2,3) \vDash_{\text{TS}} (E_{a,b})^m p \rightarrow (E_{a,b})^{m+1} p\ [1,1]$, and thus $M, (2,3) \vDash_{\text{TS}} C_{a,b}p\ [1,1]$. More generally, we can prove the following trivialization result for common knowledge in the case of two agents (the generalization to n agents is straightforward), assuming $\Box_i \phi \rightarrow \phi$ for each agent i:

Theorem 7 (Trivialization for two agents)**.** $\vDash_{\text{TS}} (E_{a,b})^{2n} \phi \rightarrow C_{a,b} \phi\ [n,n]$

In the case where the agents don't have the same number of tokens, the result can be stated by likewise considering the sum of the number of tokens available to each agent.

As in the single agent case, we can interpret this result to mean that when two agents have the same available resources, and will stop to perform computations from some point onward, common knowledge supervenes only on shared-knowledge of level n. From a conceptual point of view, the question remains open whether what this accounts for is a cognitive illusion regarding common knowledge, or whether this gives a characterization of how common knowledge is actually achieved. We lack space for a careful discussion of this issue, but we should note that the semantics is neutral between these two interpretations. Importantly, however, our trivialization result does not imply a modification of the syntactic definition of common knowledge. The operator of common knowledge is just what it used to be, namely an infinitary operator. Whichever of the two interpretations of our trivialization result one may end up favouring (illusion of common knowledge, or actual common knowledge), we thus believe the token semantics makes better sense of the notion of common knowledge with respect to the concept of bounded rationality. As in the one-agent case, the tokens can be seen as the resources that the agents will spend to compute metarepresentations, up to the point where they start to make iterations for free. The standard semantics is only a particular case of this situation, where the agents have to spend a new token for each new level of iteration.

5 Conclusions

The results of this paper should convince us that logics of introspective belief and knowledge like **K45** and **S5** are not tied intrinsically to the representation of a notion of *exact* knowledge. From a conceptual point of view, they show that Williamson's luminosity paradox can

be solved without abandoning the introspective principles, nor the original motivation for margin of error principles. More generally, the semantics here presented casts a new light on the problem of knowledge iterations, both at the individual level and at the social level. As it turns out, Williamson's luminosity paradox and the puzzle of Consecutive Numbers both belong to a broader family of epistemic sorites (including, in particular, the Surprise Examination paradox), and it would be interesting to look for further applications in this area. From a model-theoretic point of view, token semantics can be seen as a resource-sensitive logic, allowing to bound the number of moves necessary to check for satisfiability in a model. The sort of parameterization introduced here can in principle be applied to other varieties of modal operators, and may be extended to richer logics with combined modalities. Further variations on the idea of tokens as epistemic resources are also conceivable, in particular in a game-theoretical perspective, or to control other kinds of moves in a model.[3] We leave these different issues for future research.

Acknowledgements

We thank Johan van Benthem, Jérôme Dokic, Olivier Roy, Gabriel Sandu, Philippe Schlenker, Benjamin Spector, Yanjing Wang, and two anonymous ESSLLI referees for a number of helpful comments and suggestions. We also thank several audiences in Paris, Rennes, Geneva, Portland and Stanford.

[3]We are indebted to J. van Benthem, O. Roy and M. Sevenster for these various suggestions.

Bibliography

Blackburn P., Rijke M. de, Venema Y. (1999), *Modal Logic*, Cambridge Tracts in Theoretical Computer Science, 53.

van Ditmarsch H., van der Hoek W. & Kooi B. (2003), Lecture Notes on *Dynamic Epistemic Logic*, ESSLLI 2003, Vienna.

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.

Égré P. (forthcoming), "Reliability, Margin for Error and Self-Knowledge", forthcoming in D. Pritchard & V. Hendricks (eds.), *New Waves in Epistemology*, Ashgate.

Fagin R., Halpern J., Moses Y., Vardi M. (1995), *Reasoning about Knowledge*, MIT Press.

Gabbay D. (2002), "A Theory of Hypermodal Logics: Mode Shifting in Modal Logic", *Journal of Philosophical Logic* 31: 211-243.

Kamp H. (1971), Formal Properties of "Now", *Theoria* 37: 227-273.

Mongin P. (2002), "Modèles d'Information et Théorie de la Connaissance", Course Notes, Ecole Polytechnique, feb. 2002, Laboratoire d'Econométrie.

Osborne M.J. & Rubinstein A. (1994), *A Course in Game Theory*, MIT Press.

Tulenheimo T. (2004), *Independence-Friendly Modal Logic*, PhD. Dissertation, Philosophical Studies from the University of Helsinki.

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

Williamson T. (1992b), "An Alternative Rule of Disjunction in Modal Logic", *Notre Dame Journal of Formal Logic*, vol. 33 (1), 89-100.

Williamson T. (1994), *Vagueness*, Routledge.

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