Inexact Knowledge with Introspection

Denis Bonnay, Paul Egré

Inexact Knowledge with Introspection

Denis Bonnay Paul Egré

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 euclidean. Correlatively, Williamson's margin for error semantics for inexact knowledge invalidates axioms 4 and 5. We present a new semantics for modal logic which is shown to be complete for $\mathbf{K45}$, without constraining the accessibility relation to be transitive or euclidean. 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. We formulate a generalized version of the semantics, called *token semantics*, 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 and 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 euclidean, which also makes them symmetric and transitive. Euclideanness 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).

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 euclidean, 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 euclidean.

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? Consider, for instance, a situation of approximate 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 Kripke model on Figure 1, where worlds are named by numbers, and where p represents the property of fitting in the box.

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) sees independent motivation to reject the princi-

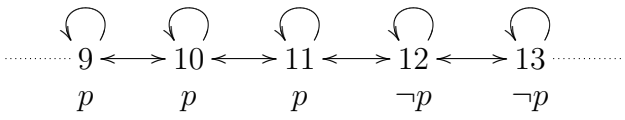

Figure 1: A structure of inexact knowledge

ple 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 (that is $\Box p \rightarrow \Box\Box p$) nor axiom 5 (that is $\neg\Box p \rightarrow \Box\neg\Box p$) is valid over all fixed margin models, by obvious failures of transitivity and euclideanness for the accessibility relation generated by 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, $\phi \rightarrow \Box\phi$ is valid (namely true everywhere in every fixed margin model) if and only if either ϕ is valid or $\neg\phi$ is valid (Williamson 1992b, 1994). A consequence of this general rule (which Williamson 1992b calls the *rule of margins*) is that only logical truths and logical falsehoods can be assumed to be known automatically if knowledge is subject to a margin of error. More specifically, positive introspection (and likewise negative introspection) will be valid if and only if either $\Box\phi$ is valid, or $\neg\Box\phi$ is valid, namely if ϕ is always known, or never. This puts a heavy restriction on the principle of positive introspection, which cannot be assumed to hold for contingently true propositions. 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:

[1] K is the axiom schema $\Box(p \rightarrow q) \rightarrow (\Box p \rightarrow \Box q)$. T is the schema $\Box p \rightarrow p$, and B is the schema $p \rightarrow \Box\Diamond p$. Williamson also presents a *variable* margin semantics, relative to which the logic $\mathbf{KT}$ is sound and complete.

knowledge can't be vague while obeying positive or negative introspection at the same time.

Furthermore, Williamson (2000) also offers a general argument showing that a contradiction can be derived from the principle of margin for error and positive introspection together with some fairly minimal assumptions about knowledge, or so it seems. This argument does not presuppose a Kripkean semantics for knowledge, so that it is independent from the logical problem of making introspection principles valid on non-transitive and non-euclidean frames. We shall first deal with the logical problem. 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 Kripkean semantics. We are therefore led to the following question:

Question. *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 answer positively to this Question in the first section of the paper, by presenting an alternative semantics for knowledge, which we call "centered semantics". We explain how it relates to Williamson general argument by making one of its premisses invalid, and we offer independent reasons to reject that premiss.

Section 3 discusses the problem of knowledge iterations and introduces a generalization of centered semantics which accounts for the fact that the trivialization of knowledge iterations might not occur as low as level 2 (one may know without knowing that one knows, or even know that one knows without knowing that one knows one knows, but it probably does not make sense to suppose that one knows that one knows that one knows without knowing that one knows that one knows that one knows). The last section extends our semantics to the multi-agent case and outlines some further applications to the analysis of common knowledge in situations of bounded rationality.

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 whether a formula with n nested operators is satisfied. In a situation of perceptual knowledge like the one pictured in Figure 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 $\mathcal{M} = \langle W, R, V\rangle$ (in which R is an arbitrary

relation over W), 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 $\mathcal{M}, (w, w') \nvDash_{\text{CS}} \phi$.
(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, it holds that $\mathcal{M}, 10 \vDash_{\text{CS}} \Box p$, $\mathcal{M}, 11 \vDash_{\text{CS}} \neg\Box p$, and $\mathcal{M}, 13 \vDash_{\text{CS}} \Box\neg p$, 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 rather straightforward to check that CS validates axioms K, 4 and 5, and that modus ponens and uniform substitution preserve validity. We give a proof for 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 same domain and valuation as $\mathcal{M}$, but in which $R'(w') = R(w)$ (where, by definition, $R(w)$ is the set of worlds w' such that wRw'). 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

[2] As pointed out to us by Ph. 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. A more detailed account of the connection of the present semantics with Kamp's semantics and other two-dimensional frameworks can be found in Bonnay & Egré (2008).

$\mathcal{M}, (w, w') \vDash_{\mathrm{CS}} \phi$ iff $\mathcal{M}', (w', w') \vDash_{\mathrm{CS}} \phi$. So if we suppose that ϕ is CS-valid but nevertheless such that $\mathcal{M}, (w, w') \nvDash_{\mathrm{CS}} \phi$, then we should have $\mathcal{M}', (w', w') \nvDash_{\mathrm{CS}} \phi$, that is $\mathcal{M}', w' \nvDash_{\mathrm{CS}} \phi$, 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 $w'Rw''$. $\mathcal{M} \vDash_{\mathrm{CS}} \Box p \rightarrow q$. But $\mathcal{M} \nvDash_{\mathrm{CS}} \Box(\Box p \rightarrow q)$, because $\mathcal{M}, (w, w') \nvDash_{\mathrm{CS}} \Box p \rightarrow q$.

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 with respect to 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 $\vDash_{\mathrm{CS}} \phi$ 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 $\vDash_{\mathrm{CS}} \phi$, by showing that $\mathcal{M}, w_0 \nvDash_{\mathrm{CS}} \phi$.

Lemma 1. *Let $\mathcal{M}$ be a transitive and euclidean Kripke model and ϕ a modal formula. Then for every world w_0, $\mathcal{M}, w_0 \vDash \phi$ iff $\mathcal{M}, w_0 \vDash_{\mathrm{CS}} \phi$*

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 $\mathcal{M}, (w_0, w) \vDash_{\mathrm{CS}} \phi$. The lemma follows immediately, since $\mathcal{M}, w_0 \vDash_{\mathrm{CS}} \phi$ iff $\mathcal{M}, (w_0, w_0) \vDash_{\mathrm{CS}} \phi$. The only non-trivial case is $\phi = \Box\psi$. Since R is euclidean and transitive, it holds that wRw' iff w_0Rw' (Assume wRw'. If $w = w_0$, there is nothing to prove. If not, 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 euclideanness, 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', $\mathcal{M}, (w_0, w') \vDash_{\mathrm{CS}} \psi$, by induction hypothesis.
iff for all w' such that w_0Rw', $\mathcal{M}, (w_0, w') \vDash_{\mathrm{CS}} \psi$, by the property of R.
iff $\mathcal{M}, (w_0, w) \vDash_{\mathrm{CS}} \Box\psi$, by definition of $\vDash_{\mathrm{CS}}$.

□

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 $\vDash_{\mathrm{CS}}$ 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-euclidean models like the model of Figure 1.

The model of Figure 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 set $\mathcal{M}, w \vDash_{\text{CMS}} \phi$ 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. Moreover, $\vDash_{\text{CMS}} \phi$ iff every margin model CMS-validates ϕ. We know that Williamson's fixed margin semantics 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), and CMS-validity is closed under necessitation, modus ponens and uniform substitution, so CMS is sound for $\mathbf{S5}$. Completeness is just an adaptation of Lemma 1: given a reflexive euclidean model $\mathcal{M}$ of $\mathbf{S5}$, one can see it as a fixed margin model $\mathcal{M}^*$ with parameter $\alpha = 1$, setting $d(v, w) = 0$ iff $w = v$; $d(w, v) = 1$ iff wRv and $w \neq v$; $d(w, v) = 2$ iff not wRv. From Lemma 1 it can be checked that $\mathcal{M}, w \models \phi$ iff $\mathcal{M}^*, w \vDash_{\text{CMS}} \phi$.

□

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.

2.3 Centered Semantics and Williamson's paradox

Centered Semantics answers a question about modeling: it provides a framework alternative to standard Kripke semantics, in which inexact knowledge can safely coexist with introspection principles over Williamson's margin for error models. As we recalled in the first section, however, Williamson (2000) presents a direct argument against introspection, which is independent of modeling issues, and in particular which does not appeal to properties of the standard Kripke semantics. This suggests that centered semantics itself is in need of further philosophical motivation. How does centered semantics relate to Williamson's argument? In this section we review

the argument and explain its connection with Centered Semantics. As it turns out, Centered Semantics is not sound with respect to one of the premises of the argument, but we consider that there is independent motivation to reject the premise in question.[3]

Williamson's argument rests on a quantitative version of the situation we described in the first section. The scenario, presented in particular in chapter 5 of Williamson 2000, is the following. Mr Magoo sees a tree in the distance. He has some knowledge of its size, for example, he certainly knows that it is less than 1000 inches tall. But his visual abilities are limited: if the tree is exactly 600 inches tall, Mr Magoo cannot know that it is less than 601 inches tall, for the margin of error of his knowledge exceeds 1 inch. Now, the argument goes like this, where $s(t) < k$ stands for 'the size of the tree is less than k inches' and η is a parameter representing (an approximation of) Mr Magoo's margin for error:

(1) $K(s(t) \leq k)$

(2) $K((s(t) \geq k - \eta) \rightarrow \neg K(s(t) \leq k))$

(3) $KK(s(t) \leq k)$

(4) $K(s(t) < k - \eta)$

(1) is the assumption that Mr Magoo knows the tree to be less thank k inches tall. (2) says that Mr Magoo knows that his knowledge obeys a margin for error principle: thus Magoo can reflect on the limitations of his ability to judge heights, and can come to know that he is not able to tell the difference between a tree which is k inches tall and a tree which $k - \eta$ inches tall. (3) follows from (1) by positive introspection, and (4) follows from (2) and (3) by closure of knowledge under logical consequence. By repeating the argument, we can infer that Mr Magoo knows that the tree is less than k' inches tall for an arbitrarily small k', which is clearly absurd.

Let us put all options on the table. To resist the absurd conclusion, one could deny either assumption (1), or assumption (2), or positive introspection, or closure under logical consequence. Since there seems to be nothing wrong with granting that Mr Magoo should at least in principle be able to make the relevant inferences safely, denying closure of knowledge under logical consequence is not very attractive. Similarly, a concept of knowledge such that Mr Magoo could not be said to know that an elm tree which is reasonably close to him is less than 100000 inches tall would set unreasonable high requirements for knowledge to hold. So there are clearly some values of k such that (1) holds. There are only two options left: either reject (2) or drop positive introspection. Williamson argues for the second option. Not only should knowledge obey a principle of margin for error, but also one should attribute to the knower some knowledge of this limitation. The reason is that if η is sufficiently small – and, in the previous argument, it *can* be taken to be arbitrarily small – denying Mr Magoo the kind of self-knowledge expressed by (2) would be fairly unreasonable.

We want to go in a different direction. First, our view is that in the kind of scenario envisaged by Williamson, *a priori* rejection of positive introspection may be too drastic. Williamson grants

[3]For a more detailed examination of this argument, see Dokic & Egré (2004) and Egré (2006). A comparison between the approach of Dokic & Egré and the present one is given in Bonnay & Egré (2008).

Mr Magoo the knowledge expressed by (1). But then whenever an agent knows that ϕ, it seems that it should be at least possible for that agent to know that he knows ϕ (in the same way in which whenever an agent knows that ϕ and $\phi \rightarrow \psi$, it is at least possible for that agent to know that ψ, even if there are many cases in which real agents fail to draw the consequences of what they know). On the contrary, on Williamson's account, because of Mr Magoo's visual limitations, Mr Magoo *cannot* in the same way know that he knows some of the propositions that he knows. Sure enough, humans are probably not endowed with perfect introspective abilities, and positive introspection might be an idealization on a more mundane notion of knowledge. Williamson's argument, however, would show that *no* creature can exist who would share Mr Magoo's visual limitations and be endowed with the ability to systematically know that she knows when she knows. In other words, our point is that the rejection of (KK) is no less costly in principle than the rejection of (2).

More importantly, we consider that the step from the acceptance of the principle of margin for error to the acceptance of (2) is less obvious than Williamson seems to suggest. First of all, there is a potential equivocation in the assumption that η can be taken to be arbitrarily small. The point of Williamson's argument is that if there is an η such that for that η, the argument can be indefinitely iterated, then one can get $K(s(t) < 0)$. To that effect, what Williamson needs to assume is the following:

> (A) There is an η such that (2) is true after any number of iterations of the argument.

where of course η needs to be non zero and where by 'after any number of iterations of the argument', we allude to the fact that, once Mr Magoo has gone through the steps (1)-(4) of the reasoning, his knowledge has changed so that if the reasoning is repeated, it is repeated in a different context. As a consequence, a premiss like (2) might be true for a given value of η in a context (say the first time Mr Magoo goes through the reasoning) and false in another context (say the fifth time Mr Magoo goes through the reasoning). Keeping this in mind, compare (A) with the following assumption:

> (B) After any number of iterations of the argument, there is an η such that (2) can be taken to be true.

(B) is weaker than (A) and does *not* give rise to paradox. If what we have is a decreasing sequence η_1, η_2, η_3... of 'margins that can be known', Mr Magoo will be able to indefinitely improve on his knowledge by reflecting on his margins, but this does not imply that he will be able to attain knowledge that the tree is of size less than k' for an arbitrarily small k'.

The question then is the following: if we accept the principle of margin for error and the ability of Mr Magoo to reflect on his cognitive limitations, should we really accept the stronger assumption (A), as Williamson suggests? (B) does appear to be a reasonable assumption: no matter how many iterations of the previous argument have been performed, Mr Magoo's knowledge still obeys a principle of margin for error, and Mr Magoo could realize that it does, and find a very small η such that he knows (2) for that η. (A) is plausible in so far as the value of η can be assumed to be strongly context-independent. Presumably, if Magoo's visual capacities are

limited, then they come about with a constant margin of error in all contexts. On the other hand, a principle like (2) expresses more than the existence of a margin of error: it says something of the way Magoo estimates the dependence between his first-order knowledge and the value of a certain margin. In that case, the value of η could very well depend on the context in which Magoo makes his estimation. On our understanding of Mr Magoo's story, Mr Magoo comes to realize that his knowledge is constrained by a margin of error, and he uses this knowledge to gain new knowledge about the size of the tree. But by the very fact that he has gained new knowledge, the value of the margin is decreased. If indeed Magoo can gain knowledge from the margin principle used in (2), then there is no reason to think that his approximation will be constant at the next step. Thus, reasons to accept (B) will not count as reasons to accept (A) as well.

Let us make this idea more concrete. Imagine that Mr Magoo looks at a tree which is 80 inches tall, and that he feels entitled to assert that it is of size less than 120. This leaves a margin of 40 inches. Now Mr Magoo can perform the reflective step that Williamson describes, and thus realize that, say, if the tree had been only 105 inches tall, he would not have been able to assert that it was less than 120 inches tall (thereby estimating his margin to be of at least 15 inches). Along the lines of the previous argument, he comes to know that the tree is less than 105 inches tall. It is far from clear that Mr Magoo will again be in a position to realize that if the tree had been 90 inches tall, he would not have been able to tell that it was less than 105 inches tall – as he did, on the basis of his first estimation of the size of the tree *and* of his performing the reasoning step described by Williamson once. Our case is even stronger at the third step. Imagine that Mr Magoo has performed the inference twice for $\eta = 15$. As a consequence, he knows that the tree is less than 90 inches tall. Can he realize that if the tree is 75 inches tall (or less), he does not know that it is of size less than 90? Clearly not. This is not something he can come to know, simply because this is not true. By assumption, the size of the tree is actually 80 inches, and Mr Magoo now knows that the tree is less than 90 inches tall.

We do not wish to deny that reflection on one's margin can occur and yield new knowledge, as Williamson's argument suggests. But we see this as a dynamic process, which, as it makes it possible for Mr Magoo's to make better estimates of the size of the tree, results in lowering the greatest number x such that Mr Magoo can realize that his knowledge obeys a margin of at least x. Hence the fallacy in arguing from Mr Magoo's ability to reflect on his limitations to the truth of (A), even though (B) might well be true.

Consequently, we think that Wiliamson's argument should be taken as a *reductio* against (2) (more precisely, as a *reductio* against (A)) rather than as a *reductio* against (KK). This makes it all the more desirable to find a model of inexact knowledge in which introspection can hold even in cases in which the underlying relation of perceptual indiscriminability is not transitive. And among the principles involved in Williamson's argument, the one that this model should invalidate on our account is (2).[4] This is precisely what centered semantics does. By fixing what

[4]As an anonymous referee pointed out, it is possible to get Williamson's paradox in discrete cases: Mr Magoo might be taken to estimate the number of people in a stadium (see Williamson 1994 on perceiving a crowd). In that kind of situation, the previous discussion suggests that *at some point* (2) should no more considered to be true even for $\eta = 1$. Assume that Mr Magoo has been improving on his approximation of the crowd's size by reflecting on the margin for error affecting his visual counting. Then there is a point at which he will no longer be in a position to be confident that his margin is of at least 1 person. For all he knows, it might the case that he has been able to figure

the visual limitations of Mr Magoo are, we get a natural model of Mr Magoo's knowledge in the situation described by Williamson. Such a model is a continuous version of the model introduced in section 1, with worlds w_i representing a size of i inches, together with the natural valuation for statements of the form $s(t) = k$ and $s(t) < k$. Using centered semantics, this model validates the principle of margin for error, as well as (KK). Assume that Mr Magoo cannot discriminate between trees whose sizes differ by less than m inches. The model predicts that Mr Magoo knows the tree to be of size less than k whenever the actual size of the tree plus m is less than k. Though such a model validates the principle of margin for error for m, it does not validate knowledge of the principle for m (nor for any number less than m). This exemplifies what we have called the failure of model-necessitation in paragraph 2.2.[5]

3 Luminosity and knowledge iterations

3.1 Luminosity

Williamson's argument against positive (as well as negative) introspection is part of a more encompassing 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.[6] 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 $\phi \rightarrow \Box\phi$ 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 $\vDash_{\mathrm{CS}} \phi \rightarrow \Box\phi$ iff either $\vDash_{\mathrm{CS}} \phi$ or $\vDash_{\mathrm{CS}} \neg\phi$, as shown by the fact that $\Box p \rightarrow \Box\Box p$ is CS-valid, but of course neither $\Box p$ nor $\neg\Box p$ is CS-valid, as witnessed by the model of Figure 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 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 have started 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, given

out the exact number of people in the stadium.

[5]Note however that such a model is static: it does not account for the dynamic process of improving on one's margins we described, nor for the validity of assumption (B). Modeling the dynamics of improvement on one's margin lies beyond the scope of the present paper. We are currently working on such a model, based on the acceptance of (B) in a setting bringing together dynamic epistemic logic and centered semantics.

[6]More accurately, Williamson calls trivial a condition that "obtains in some cases and not in others" (2000: 107). For simplicity, we consider time points instead of cases. For Williamson, "a case is like a possible world but with a distinguished subject and time" (2000: 52).

the structure of the model,[7] it follows from the standard semantics that I know I don't feel cold, but don't know that I know this: 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 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 in principle one should be able to decide upon empirical grounds. The reason for saying so is that Williamson's semantics and centered semantics make sufficiently precise and opposite predictions regarding the occurrence of higher-order knowledge for the kind of discrimination task we have been discussing here. In order to test the two theories empirically, however, one would need to map the kind of idealized model of discrimination used in Figure 1 to an experimental layout with a clear criterion for the occurrence and non-occurrence of knowledge and higher-order knowledge. Pending such results, however, and sticking to the normative perspective, 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 one 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, and the general form of our satisfaction clauses is $\mathcal{M}, qw \vDash_{\text{TS}} \phi\ [n]$, where 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. Tokens are marked at the right of formulae in our formalism, but it is important

[7] It may be argued that the problem we are dealing with does not call for a revision of the semantics, but rather, that it stems from the particular choice of the model. The objection is perfectly to the point, but as a matter of fact, centered semantics can be seen precisely as operating an implicit transformation on the model. Technically, this is reflected by the Unpacking Theorem proved in the Appendix: in particular, a one-dimensional non-transitive structure like that of Figure 1 can be unpacked into a multi-dimensional transitive structure. From a conceptual point of view, however, we think there are good reasons to change the semantics and work with a non-transitive structure, rather than shift to unpacked models that are transitive. We discuss this point extensively in Bonnay & Egré (2007), where we offer a detailed comparison of centered semantics with Halpern's 2004 treatment of inexact knowledge and non-transitivity in two-dimensional logic.

to bear in mind that they are part of the metalanguage, and not of the (modal) object language. The idea behind the use of tokens is simple. From an epistemological point of view, tokens can be seen as metarepresentational resources which the agent uses to compute iterations of knowledge or belief. More precisely, the number of tokens available specifies the threshold beyond which iterations of knowledge or belief cease to make a difference (more on this below). From a model-theoretic point of view, tokens will appear quite natural if the reader thinks of modal formulae as little automata scanning the models over which they are evaluated (following Blackburn et al. 1999: xii). Initially, a formula is evaluated with respect to a positive number of tokens.[8] For each non-trivial move of the automaton in a model, corresponding to the evaluation of a box or diamond, a token is spent. By a non-trivial move, we mean a move that leads from a world to a distinct world along the accessibility relation. Thus, for reflexive moves, no token is spent, as reflected by the first half of clause (iv) below (the number of tokens spent is 1 if w' is distinct from w, and 0 otherwise).[9] When all tokens have been spent, and the subformula to be evaluated starts with a modality, the automaton gets one token back, but backtracks to the previous position reached along the sequence of moves, and the evaluation continues, as expressed by the second half of clause (iv).

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', $\mathcal{M}, qww' \vDash_{\text{TS}} \psi\ [n-k]$, where $k = 1$ if $w \neq w'$ and $k = 0$ if $w = w'$,
- or $n = 0$ and $\mathcal{M}, q \vDash_{\text{TS}} \Box\psi\ [1]$.

The main consequence of clause (iv) in the above definition is that the number of tokens restricts the space of worlds relevant for the evaluation of a formula: with only one token, one will not reach worlds that are further than one step away from the initial world of evaluation, as was the case with centered semantics; similarly, with n tokens, one can only reach worlds that are at most n steps away from the initial world of evaluation. Going back to the model of Figure 1 above, it is easy to check for instance that $10 \vDash_{\text{TS}} \Box\Box p\ [1]$, since 12 is an alternative that is never reached with only one token, and 9, 10 and 11 are all p-worlds. However, $10 \nvDash_{\text{TS}} \Box\Box p\ [2]$, since we have $10, 11, 12 \nvDash_{\text{TS}} p\ [0]$. By contrast, $9 \vDash_{\text{TS}} \Box\Box\Box p\ [2]$, since 9, unlike 10, is more than two steps away from 12.

[8] Thus if q is empty, $\mathcal{M}, qw \vDash_{\text{TS}} \Box\psi\ [0]$ is left undefined in Definition 4 below. This does not matter in practice, because, in all the semantics we are interested in, the evaluation of a formula starts with a non-zero number of tokens, so that if the condition $\mathcal{M}, qw \vDash_{\text{TS}} \Box\psi\ [0]$ is encountered in the evaluation process, q will always be non-empty.

[9] This feature of the semantics was absent from the first version of token semantics in Bonnay & Egré (2006). See the discussion in the next subsection about the reasons for this modification.

By fixing the number of tokens relevant in the evaluation of arbitrary formulae, one obtains a spectrum of semantics:

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 $\mathcal{M}, w \vDash_{\text{TS}} \phi\ [n]$.

As expected, centered semantics and Kripke semantics correspond to opposite ends of the spectrum. Centered semantics corresponds to token semantics with only one token, and Kripke semantics can be seen as token semantics with an infinite number of tokens, namely such that a formula of modal depth n systematically reaches n steps deep in a model:

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

Besides, one can check that for a formula ϕ of modal depth n (namely whose maximum number of nested modalities is n) and a model $\mathcal{M}$, $\mathcal{M}$ satisfies ϕ with respect to TS(n) if and only if $\mathcal{M}$ satisfies ϕ with respect to the standard Kripke semantics.

3.3 Axiomatization and Completeness

Token semantics aims at capturing the idea that knowledge iterations are trivialized *at some level*. According to positive introspection, knowing implies knowing that one knows. Here we are interested in weaker versions of this principle, according to which, for instance, knowing that one knows implies knowing that one knows that one knows, without positive introspection itself being valid. Thus the desired axioms for token semantics consist in weakened versions of 4 and 5. Moreover, as things go with standard doxastic and epistemic modal logics, we would like to get from these axioms either a logic of belief, in which T is not valid or a logic of knowledge in which it is.

As we have shown in the previous section, CS validates 4 and 5, but this is no longer true for every TS(n) semantics. As for CS, we say that a formula ϕ is TS(n)-valid if and only if for every Kripke model $\mathcal{M}$ and world w, $\mathcal{M}, w \vDash_{\text{TS}(n)} \phi$. It is easy to see that 4 and 5, though TS(1)-valid, are not TS(n)-valid for $n \geq 2$ (again, by looking at the model of Figure 1, we see that $10 \vDash_{\text{TS}(2)} \Box p \wedge \neg\Box\Box p$). Let $\Box^n$ be short for $\Box ... \Box$, n times. The generalized form of axiom 4 which is correlated to the family of semantics TS(n) is the following:

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

Intuitively $(4.n)$ is like just like 4, but for the fact that positive introspection is guaranteed only for knowledge claims with n iterations of the knowledge operator. $(4.n)$ 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). $(4.n)$ is valid in TS(n). We state and prove this for the particular case of $n = 2$ and for the Diamond version of our axiom, which is of course equivalent to the Box version. Fully general soundness and completeness theorems will be stated

later on and are proved in the Appendix. We focus on a particular case here in order to give a more concrete view of the interaction between tokens and modalities. In what follows, a *pointed* model $\mathcal{M}, w$ designates a model $\mathcal{M} = \langle W, R, V \rangle$ with a distinguished world $w \in W$.

Fact 2. *For all formulas ϕ and every $n \geq 1$: $\mathcal{M}, qww \vDash_{\text{TS}} \phi\ [n]$ iff $\mathcal{M}, qw \vDash_{\text{TS}} \phi\ [n]$.*

Proof. By induction on the complexity of ϕ. □

Fact 3. *For all pointed models $\mathcal{M}, w$:* $\mathcal{M}, w \vDash_{\text{TS}} \Diamond\Diamond\Diamond p \rightarrow \Diamond\Diamond p\ [2]$

Proof. Assume $\mathcal{M}, w \vDash_{\text{TS}} \Diamond\Diamond\Diamond p\ [2]$. From the satisfaction clause for $\Diamond$, there are two possibilities to consider, depending on whether the world accessed from w is identical to w or distinct:

- Case 1: $\mathcal{M}, ww \vDash_{\text{TS}} \Diamond\Diamond p\ [2]$. By Fact 2 this implies $\mathcal{M}, w \vDash_{\text{TS}} \Diamond\Diamond p\ [2]$.
- Case 2: $\mathcal{M}, ww' \vDash_{\text{TS}} \Diamond\Diamond p\ [1]$ where $w' \neq w$. Again, there are two possibilities:
  - $\mathcal{M}, ww'w' \vDash_{\text{TS}} \Diamond p\ [1]$. By Fact 2 this implies $\mathcal{M}, ww' \vDash_{\text{TS}} \Diamond p\ [1]$, hence $\mathcal{M}, w \vDash_{\text{TS}} \Diamond\Diamond p\ [2]$.
  - $\mathcal{M}, ww'w'' \vDash_{\text{TS}} \Diamond p\ [0]$ with $w'' \neq w'$. By the satisfaction definition, when the number of tokens is zero, this means that $\mathcal{M}, ww' \vDash_{\text{TS}} \Diamond p\ [1]$. Again, this implies as required that $\mathcal{M}, w \vDash_{\text{TS}} \Diamond\Diamond p\ [2]$ holds.

□

For the purpose of completeness however, $(4.n)$ turns out to be too weak and we shall use the following stronger version:

$(4.n')$ $(\neg p_1 \wedge \Diamond(p_1 \wedge \neg p_2 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Diamond\Diamond r)\ldots))) \rightarrow \Diamond(p_1 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Diamond r)\ldots))$

$(4.n')$ is constructed in the following manner. Its antecedent consists in a sequence of length $n-1$ of diamonds ensuring that each world 'reached' through a diamond satisfies an atom p_i ($i \in \{1, ..., n-1\}$) which was false at the world at which the diamond is evaluated, and it says that at the end of this sequence, $\Diamond\Diamond r$ is true. Its consequent consists in a sequence of length $n-1$ of diamonds marked with the same positive atoms, and it says that at the end of this sequence, $\Diamond r$ is true. When $n = 1$, we deal with an empty sequence of diamonds both in the antecedent and in the consequent, so to speak. Therefore we are left with $\Diamond\Diamond r \rightarrow \Diamond r$, which is nothing but 4.

In order to get an intuitive idea of the meaning of $(4.n')$, as opposed to $(4.n)$, let us have a closer look at the case where $n = 2$:

$(4.2')$ $(\neg p_1 \wedge \Diamond(p_1 \wedge \Diamond\Diamond r)) \rightarrow \Diamond(p_1 \wedge \Diamond r)$

First note that (4.2) can be derived from (4.2′) in a normal modal logic. We replace p_1 by $\Diamond\Diamond p$ in (4.2′), and r by p. We get $(\neg\Diamond\Diamond p \wedge \Diamond(\Diamond\Diamond p \wedge \Diamond\Diamond p)) \rightarrow \Diamond(\Diamond\Diamond p \wedge \Diamond p)$. This implies $(\neg\Diamond\Diamond p \wedge \Diamond\Diamond\Diamond p) \rightarrow \Diamond\Diamond p$, which, by propositional reasoning, yields $\Diamond\Diamond\Diamond p \rightarrow \Diamond\Diamond p$. On the contrary (4.2) does not imply (4.2′) in normal modal logic.[10] Now the specific content of (4.2′) over (4.2) is that if it is conceivable that it is conceivable that r from the point of view of a certain epistemic alternative, marked by the atom p_1, then it is conceivable that r from that same point of view. By contrast, (4.2) only says that if it is conceivable that it is conceivable that r from the point of view of a certain epistemic alternative, then there is a *possibly different* epistemic alternative from the point of view of which it is conceivable that r.

Similarly, axiom 5 generalizes to:[11]

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

5 is the principle of negative introspection: if I do not know p, I know that I do not know it. $(5.n')$ is a form of negative introspection which holds only for knowledge claims with n iterations of the epistemic modality: as a first approximation, it says that considering as possible that one considers as possible etc... (n times) implies considering as possible that one considers as possible etc... ($n-1$ times) that one knows that one considers p as possible. This is only a first approximation because, as before with $(4.n')$, $(5.n')$ holds with respect to a chain of marked epistemic alternatives. In particular, for the case $n = 2$, it says that if there is an epistemic alternative to the actual world in which I consider r as possible, then I know that I consider r as possible from the point of view of the same epistemic alternative.

Note that the 'simple' version of $(5.n')$ would be:

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

Contrary to what happens for $4.n'$ with respect to $4.n$, the more complex axiom $(5.n')$ does not imply the simpler $(5.n)$, and $(5.n)$ is not TS(n)-valid. To see this, consider again the case $n = 2$ and look at the following model, represented in Figure 2 :

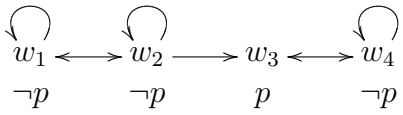

Figure 2: A model invalidating (5.2)

Clearly, $w_2 \models_{\text{TS}} \Diamond\Diamond p$ [2], but $w_2 \not\models_{\text{TS}} \Diamond\Box\Diamond p$ [2]. To see the latter, one has to check that $w_2, w_1 \not\models_{\text{TS}} \Box\Diamond p$ [1] (since $w_2, w_1, w_1 \not\models_{\text{TS}} \Diamond p$ [1]), that $w_2, w_2 \not\models_{\text{TS}} \Box\Diamond p$ [2] (since

[10] Consider the following frame $\mathcal{F} = \langle\{w_0, w_1, w_2, w_3, w_4\}, \{\langle w_0, w_1\rangle, \langle w_1, w_2\rangle, \langle w_2, w_3\rangle, \langle w_0, w_4\rangle, \langle w_4, w_3\rangle\}\rangle$. (4.2) is clearly valid over $\mathcal{F}$, but (4.2′) is not. To see this, take $V(p_1) = \{w_1\}$, $V(r) = \{w_3\}$, check that $(\mathcal{F}, V), w_0 \models \neg p_1 \wedge \Diamond(p_1 \wedge \Diamond\Diamond r)$ but $(\mathcal{F}, V), w_0 \not\models \Diamond(p_1 \wedge \Diamond r)$.

[11] As before, when $n = 1$, we mean (5.1') to be just 5. In the antecedent of (4.1'), we have an empty sequence of Diamonds and p_i atoms or negated atoms followed by $\Diamond r$. In its consequent, we want an empty sequence of Diamonds and p_i atoms followed by $\Box\Diamond r$. Thus what we actually have is $\Diamond r \rightarrow \Box\Diamond r$, which is nothing but 5.

$w_2, w_2, w_1 \nvDash_{\text{TS}} \Diamond p$ [1]), and finally that $w_2, w_3 \nvDash_{\text{TS}} \Box\Diamond p$ [1] (since $w_2, w_3, w_4 \nvDash_{\text{TS}} \Diamond p$ [0]). This establishes that the schema (5.2) is not TS(2)-valid over arbitrary structures.

The normal logic resulting from the inclusion of the schemata $(4.n')$ and $(5.n')$ to the system K axiomatizes token semantics:

Theorem 4. $\mathrm{K}(4.n')(5.n')$ *is sound and complete with respect to* $\mathrm{TS}(n)$.

K$(4.n')(5.n')$ is not a logic for knowledge, but rather a logic of introspective belief, because it does not have $\Box p \rightarrow p$ as a theorem; it is easy to check that the axiom T is not TS(n) valid by considering a non-reflexive model. As with standard Kripke semantics, we get a logic for knowledge by restricting ourselves to the class of reflexive frames:

Theorem 5. $\mathrm{KT}(4.n')(5.n')$ *is sound and complete with respect to* $\mathrm{TS}(n)$ *over the class of reflexive frames.*

The proof of Theorems 4 and 5 is deferred to the Appendix, in which basic model-theoretic tools for the study of token semantics are developed. One could ask whether standard axiomatization results for Kripke semantics generalize to token semantics in the way suggested by Theorem 5. In particular, the B axiom $p \rightarrow \Box\Diamond p$ standardly characterizes the class of symmetric frames. Is it the case that KB$(4.n')(5.n')$ axiomatizes TS(n) over the class of symmetric frames? The answer is negative. Look again at the model of Figure 1. It is symmetric; however, consider the substitution instance of B in which p is replaced by $\Diamond\Diamond\neg p \wedge \Box p$. Clearly, $10 \vDash_{\text{TS}} \Diamond\Diamond\neg p \wedge \Box p$ [2], but $10 \nvDash_{\text{TS}} \Box\Diamond(\Diamond\Diamond\neg p \wedge \Box p)$ [2]. For this to hold, it would have to be the case that $10, 9 \vDash_{\text{TS}} \Diamond(\Diamond\Diamond\neg p \wedge \Box p)$ [1]. However, for $x = 8$ and $x = 10$, we have $10, 9, x \nvDash_{\text{TS}} \Diamond\Diamond\neg p \wedge \Box p$ [0], and likewise $10, 9, 9 \nvDash_{\text{TS}} \Diamond\Diamond\neg p \wedge \Box p$ [1], since in each case the number of tokens left is not sufficient to reach a $\neg p$-world, as would be required for $\Diamond\Diamond\neg p$ to hold.

Thus KB$(4.n')(5.n')$ is not correct for TS(n) over the class of symmetric frames.[12] Where does this failure come from? Standardly, the symmetry of the accessibility relation guarantees that B *and* its substitution instances are valid, because a world w which satisfies a formula ϕ will also count as a witness for ϕ after a $\Box\Diamond$-step. In token semantics, this is no longer true, because a $\Box\Diamond$-move has a cost – two tokens for two non-trivial moves. Hence, the fact that $\mathcal{M}, w \vDash_{\text{TS}} \phi$ $[n]$ in a given model $\mathcal{M}$ does not guarantee that $\mathcal{M}, w \vDash_{\text{TS}} \phi$ $[n-2]$. The problem did not arise for T with respect to the class of reflexive frames because clause (iv) of token semantics ensures that trivial (reflexive) moves are performed at no cost. Hence $\mathcal{M}, w \vDash_{\text{TS}} \Box\phi$ $[n]$ does imply $\mathcal{M}, w \vDash_{\text{TS}} \Box\phi$ $[n]$ when w is accessible from itself, no token being spent.[13]

[12] As a matter of fact, $\phi \rightarrow \Box\Diamond\phi$ is TS(n) valid for $n > 2$ over symmetric frames when ϕ is an atom. This implies that the logic of TS(n) over the class of symmetric frames does not satisfy closure under substitution and hence is not normal. Because the model of Figure 1 is reflexive, the same holds for KTB$(4.n')(5.n')$ with respect to the class of reflexive and symmetric frames.

[13] In the original version of token semantics, presented in Bonnay and Égré (2006), we did not make the qualification that reflexive moves come at no cost, and obtained a distinct axiomatization for the corresponding semantics. As a consequence, adding to our axioms the axiom T did not provide a correct axiomatization of our semantics over the class of reflexive frames. We had overlooked this fact, which jeopardized the possibility of using token semantics for a logic of knowledge (as opposed to belief). We owe to the anonymous reviewer of the JPL the remark

Now, another question arises: would it be possible to modify token semantics so as to make B and its substitution instances valid over the class of symmetric frames? The idea would be to resort to the same trick that makes for the validity of T and its substitution instances over the class of reflexive frames. Clause (iv) in the definition of satisfaction would have to be changed to something like:[14]

(iv*) $\mathcal{M}, qw \vDash_{\text{TS}} \Box\psi\ [n]$ iff:

- $n \neq 0$ and for all w' such that wRw', $\mathcal{M}, qww' \vDash_{\text{TS}} \psi\ [n + k - 1]$, where:
  - $k = 2$ if q is a non-empty sequence ending with w',
  - $k = 1$ if $w = w'$,
  - $k = 0$ otherwise.
- or $n = 0$ and $\mathcal{M}, q \vDash_{\text{TS}} \Box\psi\ [1]$.

Of course, $(4.n')$ and $(5.n')$ would have to be modified in order to reflect the fact that cycles of length at most 2 come at no cost.

This way of looking at the problem suggests the possibility of an even more general formulation of token semantics, in which cycles in general would come for free. The satisfaction clause would then be something like:[15]

(iv**) $\mathcal{M}, qw \vDash_{\text{TS}} \Box\psi\ [n]$ iff:

- $n \neq 0$ and for all w' such that wRw', $\mathcal{M}, qww' \vDash_{\text{TS}} \psi\ [n + k - 1]$, where k is the length of the cycle from w' to its previous occurrence in the sequence, that is the number of steps between the occurrence of w' at the end of the sequence qww' and its closest antecedent occurrence in that sequence (k being zero if there is no such previous occurrence).
- or $n = 0$ and $\mathcal{M}, q \vDash_{\text{TS}} \Box\psi\ [1]$.

In that case too, $(4.n')$ and $(5.n')$ need to be modified in order to obtain completeness, so as to reflect the fact that cycles of arbitrary length come at no cost.

Developing token semantics along the lines of clauses (iv*) or (iv**) would be all the more urgent if we took B to be a desirable axiom for a logic of knowledge. However, the B axiom is at best unobvious for knowledge, as Williamson admits of the operator "it is clear that" in his logic of clarity (1994: 272). For that reason, Williamson (1994) proposes a variable margin semantics for knowledge which does validate T but no longer retains B. When the $\Box$ operator

that our claim concerning the possibility to get a logic of knowledge was incorrect as it stood. This remark resulted in the present version of token semantics, which is now adequate as a basis for a logic of knowledge.

[14] The actual definition should take care of intertwined cycles.

[15] As before, a special clause should be added to take care of intertwined cycles. The general idea of adding conditions in the satisfaction clause that depend on the sequence of worlds that has been explored has been investigated from an abstract perspective in Gabbay (2002). Our semantics are special cases of the kind of semantics that Gabbay considers.

is interpreted as "I know that", the B axiom is much like a factivity constraint, but lifted to the complex operator "I consider it possible that I know", since it is equivalent to $\Diamond\Box p \rightarrow p$. Intuitively, however, it may happen that I think that I know some proposition for sure, which will turn out to be false. Despite this, the B axiom is not obviously incorrect either (in the scenarios, in particular, in which one would like to preserve both 5 and T). So in the same way in which Williamson defines a special semantics that does not validate B over symmetric structures,[16] our point here is that there is a way to recover its validity over symmetric structures in the present framework if one wishes to do so. Like Williamson, therefore, we leave the B axiom to a status of neutrality with respect to the issues we are focussing on here. From a logical point of view, a noteworthy difference between B and axioms like 4 and 5 is that not both the antecedent and the consequent of B are modalized, and so despite its iterative form, B is not an axiom concerning self-knowledge in exactly the same way in which 4 and 5 are. From an epistemological point of view, therefore, the reasons we have to preserve weaker forms of the schemata 4 and 5 do not commit us to accepting the validity of B.

3.4 Knowledge iterations

The perspective offered by TS(n) on knowledge iterations is valuable in two respects, both with respect to the epistemological conception of introspection, and with regard to the phenomenon of higher-order vagueness.

Regarding introspection, we may say that CS corresponds to a "Cartesian" view of introspection, in so far as higher-order knowledge follows from first-order knowledge at no cost. Moreover, CS is a full logic of epistemic transparency, since it validates both positive and negative introspection without restriction. By contrast, the standard semantics, and with it Williamson's semantics for knowledge, is closer to a "Lockean" conception, where each iteration of knowledge is an act of the mind that comes at a cost.[17] Token semantics offers a compromise between these two conceptions, since iterations come for free at level n only. Thus, even on a Lockean view, one may wish to preserve a weakened form of the introspection principles 4 and 5, rather than simply dispense with them altogether. Indeed, even if one upholds the view that knowing that one knows is essentially more difficult than simply knowing, one may still consider possible that iterations of knowledge stop making a difference at some point beyond two or more iterations. A hint that this may be so is provided by the difficulty of ascribing knowledge beyond two levels of iterations in ordinary language (thus, a sentence like "he knows that he knows, but he does not know that he knows he knows" sounds nearly contradictory). The alternative view is to consider that each new level of knowledge is intrinsically more difficult to attain than the previous ones, as Williamson seems to consider, and that our reluctance to make knowledge ascriptions of the form "he knows^n that p but he does not know^{n+1} that p" beyond $n = 1$ is then a side-effect of some limited capacity to compute metarepresentations. Consistently with these two views of higher-order knowledge, tokens can serve as indications either of some intrinsic collapse in the

[16] See Williamson 1994: 272: "Since metrics are by definition symmetric (d(x,y)=d(y,x)), the failure of the (B) axiom in variable margin model may seem surprising".

[17] We are indebted to P. Engel for this suggestive comparison. See also Williamson 2000, ch. 4 on the discussion of the Cartesian view.

hierarchy of knowledge, or of some extrinsic limitation in the way we compute knowledge iterations. Whichever of these two interpretations one favours, however, an important point is that the determination of the critical value n at which iterations of knowledge are trivialized (whether *de jure* or *de facto*) is in part a matter of empirical investigation. To that extent, both the purely Cartesian view of knowledge and the purely Lockean view of knowledge may appear to be idealizations, and the strength of token semantics is make a bridge and leave a gradient between them.

Correlatively, the parameterization afforded by TS casts light on an aspect of the notion of inexact knowledge commonly referred to as higher-order vagueness. First-order vagueness corresponds to the situation in which something is neither clearly p, nor clearly not-p. Higher-order vagueness concerns the iteration of the adverb "clearly": when something is not clearly p, is it clearly not clearly p? If the answer to this question is negative, then it means there is room for second-order vagueness, since one cannot draw the line between the things that are clearly p and the things that are not clearly p. Williamson's margin for error semantics was designed in particular to deal with higher-order vagueness, since in Williamson (1994) the $\Box$ is interpreted as "it is clear that". This is one of the reasons why, in Williamson's original approach, iterations of knowledge are seen as a process of "gradual erosion" (Williamson 1994: 228). If we adopt the same interpretation here, then it may be objected that CS allows only for first-order vagueness, and not for higher-order vagueness. Likewise, for every n, TS(n) makes room for n-order vagueness, although not for $n + 1$-order vagueness. A detailed discussion of higher-order vagueness is out of the scope of this paper, but we can point out that in practice n-order vagueness is probably sufficient to account for higher-order vagueness.[18] Also, we should not expect the operator "it is clear that" to have exactly the same logical properties of the operator "I know that", even in situations of inexact knowledge. Indeed, one may accept the introspection principles for "I know that" and still acknowledge the phenomenon of higher-order vagueness pertaining to the operator "it is clear that" (see for example Halpern 2004).

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.

4.1 A puzzle about common knowledge

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). In the

[18] We discuss the phenomenon of higher-order vagueness in greater detail in Bonnay & Egré (2007).

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. 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:

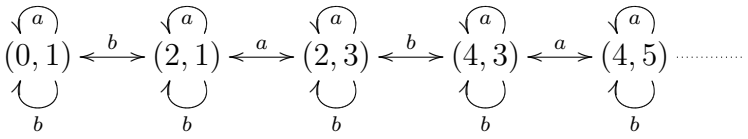

Figure 3: Consecutive Numbers

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 3 is isomorphic to the structure of inexact knowledge of Figure 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 $E_{a,b}\phi \rightarrow E_{a,b}E_{a,b}\phi$. 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.

4.2 Multi-Agent Token Semantics

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 types of tokens, one type for each of the n agents. Given a multi-agent model $M = \langle W, R_1, ..., R_n, V \rangle$, we note m_i the number of tokens of type i, that is the number of tokens which are initially available to agent i ($m_i > 0$). In the single agent case, satisfaction has been defined with respect to a sequence of worlds. In the multi-agent case, we shall need to enrich this structure in order to keep track of the tokens that are spent as the evaluation proceeds.

Satisfaction is therefore defined with respect to a sequence of ordered pairs of the form (w, k), where $w \in W$ and $k = 0$ if no tokens have been spent in order to reach w (the move was along a reflexive arrow), and $k = i$ if the move made to reach w is an R_i-move that has cost one token to agent i (in other words, the right index indicates whether w was reached by a non-reflexive move or not, and also to which agent the corresponding move corresponds). We call such a sequence

a *path*. Let i be the index of agent i, we define the *i-cost* of a path q (notation $c_i(q)$) to be the number of tokens of type i that have been 'dropped' along the path. Thus $c_i(q)$ is the number of pairs of the form (v, i) in the path.

In the mono-agent case, the intuition was that, when all tokens have been spent and a $\Box$ needs to be processed, the automaton exploring the model has to move back in order to retrieve the last token spent. The idea is the same for the multi-agent case, but in order to evaluate a $\Box_i$, the automaton has to get a token *of type i* back. It does so by moving back along the path it has explored, until it reaches the world preceding the last i-transition that took place, picking up all tokens of other types if some are found along the way, and putting them back onto the corresponding stacks.

Definition 6 (Multi-agent token satisfaction)**.** *Token satisfaction of a formula ϕ for n-agents with respect to n stacks of tokens $m_1, ..., m_n$ (notation: $\mathcal{M}, q(w, k) \vDash_{\text{MTS}} \phi\ [m_1, ..., m_n]$) is defined by recursion on ϕ according to the following clauses:*

i) $\mathcal{M}, q(w, k) \vDash_{\text{MTS}} p\ [m_1, ..., m_n]$ *iff* $w \in V(p)$,

ii) $\mathcal{M}, q(w, k) \vDash_{\text{MTS}} \neg\phi\ [m_1, ..., m_n]$ *iff* $\mathcal{M}, q(w, k) \nvDash_{\text{MTS}} \phi\ [m_1, ..., m_n]$,

iii) $\mathcal{M}, q(w, k) \vDash_{\text{MTS}} (\phi \wedge \psi)\ [m_1, ..., m_n]$ *iff* $\mathcal{M}, q(w, k) \vDash_{\text{MTS}} \phi\ [m_1, ..., m_n]$ *and* $\mathcal{M}, q(w, k) \vDash_{\text{MTS}} \psi\ [m_1, ..., m_n]$,

iv) $\mathcal{M}, q(w, k) \vDash_{\text{MTS}} \Box_i\psi\ [m_1, ..., m_n]$ *iff*

a) $m_i \neq 0$ *and for all* w' *such that* wR_iw', $\mathcal{M}, q(w, k)(w', l) \vDash_{\text{MTS}} \psi\ [m_1, \ldots, m_i - s, \ldots, m_n]$, *where* $s = 1$ *and* $l = i$ *if* $w \neq w'$, *and* $s = l = 0$ *if* $w = w'$,

b) or $m_i = 0$ *and* $\mathcal{M}, q' \vDash_{\text{MTS}} \Box_i\psi\ [m_1 + r_1, \ldots, \ldots, m_n + r_n]$, *where*

- *q' is the longest initial segment of $q(w, k)$ such that there is a pair (v, i) which belongs to $q(w, k)$ but not to q',*
- *for all $j \in \{1, ..., n\}$, r_j is equal to the number of pairs of the form (v, j) which belong to $q(w, k)$ but not to q' (namely, $r_j = c_j(q) - c_j(q')$).*

We say that a formula ϕ is true at a world w in a model $\mathcal{M}$ according to multi-agent token semantics $[m_1, ..., m_n]$ (where the m_i are greater than zero) iff $\mathcal{M}, (w, 0) \vDash_{\text{MTS}} \phi\ [m_1, ..., m_n]$. We abbreviate this as $\mathcal{M}, w \vDash_{\text{MTS}} \phi\ [m_1, ..., m_n]$ (note that the choice of 0 in $(w, 0)$ is without significance, since what matters are the subsequent moves from w onward).

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:

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

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:

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

and likewise for any further level of embedding, since for any path q from $(2,3)$ of length at most 2, we have for instance $M,(2,3)q \vDash_{\mathrm{MTS}} (0_a \vee 2_a \vee 4_a)\ [0,0]$. Thus, if we let p stand for the proposition that "a and b each have a number no greater than 5", it holds that $M,(2,3) \vDash_{\mathrm{MTS}} p \wedge E_{a,b}p\ [1,1]$, and for every m, $M,(2,3) \vDash_{\mathrm{MTS}} (E_{a,b})^m p \rightarrow (E_{a,b})^{m+1}p\ [1,1]$, and thus $M,(2,3) \vDash_{\mathrm{MTS}} C_{a,b}p\ [1,1]$. With the standard semantics, for ϕ to be common knowledge in w, ϕ has to be true in every world accessible from w; this is no longer true with MTS, as witnessed by the fact that $M,(2,3) \vDash_{\mathrm{MTS}} C_{a,b}p\ [1,1]$, although p, for instance, does not hold at $(10,11)$.

Let us now consider the general properties of MTS, underlying what we saw happening on the Consecutive Number model. To begin with, we can prove the validity of a version of $4.n$ for the complex modality $\Diamond_1\Diamond_2$. This tells us that iterated knowledge attributions of the form "I know that you know that..." are trivialized at some point, just like iterations of the form "I know that I know that..." were trivialized in TS:

Fact 4. *Let $\mathcal{M}$ be an arbitrary model and w an arbitrary point in that model, for $n > 1$, we have that: $\mathcal{M},w \vDash_{\mathrm{MTS}} (\Diamond_1\Diamond_2)^n\Diamond_1\Diamond_2\phi \rightarrow (\Diamond_1\Diamond_2)^n\phi\ [n,n]$*

Proof. Assume that $\mathcal{M},(w,0) \vDash_{\mathrm{MTS}} (\Diamond_1\Diamond_2)^n\Diamond_1\Diamond_2\phi[n,n]$. There is a path q starting from $(w,0)$ such that $\mathcal{M},q \vDash_{\mathrm{MTS}} \Diamond_1\Diamond_2\phi[n-c_1(q), n-c_2(q)]$. By clause iv) in the definition of MTS satisfaction, this implies that:

- either $n - c_1(q) = 0$ and there is an initial segment q' of q abiding by the conditions in satisfaction clause iv.b), such that $\mathcal{M},q' \vDash_{\mathrm{MTS}} \Diamond_1\Diamond_2\phi[n-c_1(q'), n-c_2(q')]$. Note that $c_1(q')$ and $c_2(q')$ both have to be $< n$ (since by backtracking some tokens are taken back), and therefore the path q' is reachable by the sequence $(\Diamond_1\Diamond_2)^{n-1}$. This guarantees that $\mathcal{M},(w,0) \vDash_{\mathrm{MTS}} (\Diamond_1\Diamond_2)^{n-1}\Diamond_1\Diamond_2\phi[n,n]$, as required.
- or $n - c_1(q) \neq 0$, then there is a pair (v,i) abiding by the conditions in the satisfaction clause iv.a), such that $\mathcal{M},q(v,i) \vDash_{\mathrm{MTS}} \Diamond_2\phi[n-c_1(q(v,i)), n-c_2(q(v,i))]$. Again,
  - either $n - c_2(q(v,i)) \neq 0$ and there is a pair (v',i') abiding by the conditions in the satisfaction clause iv.a) such that $\mathcal{M},q(v,i)(v',i') \vDash_{\mathrm{MTS}} \phi[n-c_1(q(v,i)(v',i')), n-c_2(q(v,i)(v',i'))]$. Note that $c_1(q(v,i)(v',i'))$ and $c_2(q(v,i)(v',i'))$ both have to be $\leq n$, so that the path $q(v,i)(v',i')$ is reachable by the sequence $(\Diamond_1\Diamond_2)^n$. This guarantees that $\mathcal{M},(w,0) \vDash_{\mathrm{MTS}} (\Diamond_1\Diamond_2)^n\phi[n,n]$,
  - or $n-c_2(q(v,i)) = 0$ and there is an initial path q' of $q(v,i)$ abiding by the conditions in satisfaction clause iv.b) such that $\mathcal{M},q' \vDash_{\mathrm{MTS}} \Diamond_2\phi[n-c_1(q'), n-c_2(q')]$. Note that now $c_1(q')$ and $c_2(q')$ both have to be $< n$, so that the path q' is reachable by the sequence $(\Diamond_1\Diamond_2)^{n-1}$. This guarantees that $\mathcal{M},(w,0) \vDash_{\mathrm{MTS}} (\Diamond_1\Diamond_2)^n\phi[n,n]$, as required.

□

As the discussion of the Consecutive Numbers model already suggested, the new semantics has significant consequences for the attainability of common knowledge. Because iterations of knowledge attributions are trivialized at some point, common knowledge is bound to supervene on some finite level of shared knowledge. More precisely, common knowledge will be finitely reachable when agents are endowed with a finite number of tokens, even though from a syntactic point of view we preserve the standard (infinitary) definition for common knowledge. Let us abbreviate the conjunction $\phi \wedge E\phi \wedge EE\phi \wedge \cdots \wedge E^n\phi$ by $E^{\leq n}\phi$. Here is the trivialization result for common knowledge:

Theorem 6. $\vDash_{\mathrm{MTS}} (E_{a,b})^{\leq n+n} \leftrightarrow C_{a,b}\phi\ [n,n]$

First, we state the following intuitive lemma, where $\overrightarrow{\Diamond}$ is an arbitrary sequence of $\Diamond_1$ and $\Diamond_2$ modalities:

Lemma 2. *For any sequence $\overrightarrow{\Diamond}$, any formula ϕ and any model $\mathcal{M}$,*
if $\mathcal{M}, w \vDash_{\mathrm{MTS}} \overrightarrow{\Diamond}\phi\ [n,n]$, then there is a path q starting from $(w,0)$ such that:
$\mathcal{M}, q \vDash_{\mathrm{MTS}} \phi\ [n - c_1(q), n - c_2(q)]$, where $c_1(q) \leq n$ and $c_2(q) \leq n$

Proof. By a straightforward induction on the length of $\overrightarrow{\Diamond}$.

□

We are now ready to prove Theorem 6:

Proof. The right to left direction of the theorem follows immediately from the definition of $C_{a,b}$. The left to right direction is a consequence of the previous Lemma. To establish this, it is sufficient to show that for any pointed model $\mathcal{M}, w$, for any sequence $\overrightarrow{\Diamond}$ of arbitrary length, there is a sequence $\overrightarrow{\Diamond'}$ containing at most n $\Diamond_1$ and at most n $\Diamond_2$ modalities such that if $\mathcal{M}, w \vDash_{\mathrm{MTS}} \overrightarrow{\Diamond}\phi\ [n,n]$ then $\mathcal{M}, w \vDash_{\mathrm{MTS}} \overrightarrow{\Diamond'}\phi\ [n,n]$ (that is, everything which is true after a long chain of diamonds is already true after a short one, bounded by the number of tokens available to the agents).

Assume that $\mathcal{M}, w \vDash_{\mathrm{MTS}} \overrightarrow{\Diamond}\phi\ [n,n]$. By the previous lemma, there is a path q starting from $(w,0)$ such that $\mathcal{M}, q \vDash_{\mathrm{MTS}} \phi\ [n - c_1(q), n - c_2(q)]$ and $c_1(q) \leq n$ and $c_2(q) \leq n$. But now, just by looking at q, one can build a 'short' sequence of diamonds $\overrightarrow{\Diamond'}$ such that $\mathcal{M}, w \vDash_{\mathrm{MTS}} \overrightarrow{\Diamond'}\phi\ [n,n]$. To build $\overrightarrow{\Diamond'}$, go step by step along the path q: add a $\Diamond_1$ for a pair $(v,1)$, add a $\Diamond_2$ for a pair $(v,2)$, do nothing for pairs of the form $(v,0)$. Because $c_1(q) \leq n$ and $c_2(q) \leq n$, $\overrightarrow{\Diamond'}$ contains at most n $\Diamond_1$ and at most n $\Diamond_2$ modalities, as required. And by construction of $\overrightarrow{\Diamond'}$, $\mathcal{M}, q \vDash_{\mathrm{MTS}} \phi\ [n - c_1(q), n - c_2(q)]$ guarantees that $\mathcal{M}, w \vDash_{\mathrm{MTS}} \overrightarrow{\Diamond'}\phi\ [n,n]$.

□

We leave to the appendix the axiomatization of MTS-validities; in what follows, we shall only rely on the result about common knowledge stated as Theorem 6, which is to us the crucial point concerning trivialization of metarepresentational levels.[19]

[19] At this point, a warning might be in order. One should be aware of fine-grained issues in the definition of the

4.3 Almost common knowledge

As in the single agent case, we can interpret the trivialization result to mean that when two agents have limited resources, and stop performing computations from some point onward, common knowledge supervenes only on a finite amount of shared knowledge. From a conceptual point of view, however, 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 attained.

Several points can be made regarding this issue. The first concerns the fact that the semantics is in principle neutral between the two interpretations. Importantly, our trivialization result does not imply a modification of the syntactic definition of common knowledge. Thus, the operator of common knowledge is just what it used to be, namely an infinitary operator. On the face of it, therefore, what the operator C means in any token semantics is still the infinite conjunction of all levels of shared knowledge.

Despite this, there is obviously a difference in the semantic interpretation of C, depending on whether agents are allocated an infinite number of tokens, or a finite number thereof. Practically, it is perfectly possible that, in some situations, common knowledge can indeed be reached on the basis of a finite number of iterations only, whereas in some others (as in the coordinated attack problem), no finite approximation will do. If we think of the case of consecutive numbers, a is ready to bet that b's number is less than 100000, and so is b. But presumably, a is ready to bet that b is ready to bet that a's number is less than 100000; similarly for b, and so on and so forth. In that situation, it does not seem entirely implausible to give the previous result a realistic interpretation, namely to say that common knowledge is indeed attained, and not simply due to the fact that the agents are lazy in their computations.

But still, and this is a third point of clarification, any realistic interpretation faces the problem of the arbitrariness of the number of tokens assigned to the agents. For instance, the semantics predicts that $M, (2,3) \not\models_{\text{MTS}} \Diamond_b \Diamond_a \Diamond_b 6_a\ [1,1]$, that is b does not consider it possible that a considers it possible that b considers it possible that a might have a 6. This does not seem entirely implausible. But there may be situations where we are embarrassed as to where to draw the line, or simply where iterations are not trivialized. The problem of vagueness recurs here, shifted to the metatheoretical level. For instance, it may be too strong to require that the agents have common knowledge that their numbers are less than 6. But we are sure enough that it is common knowledge that their numbers are less than 100000. As in the case of vague predicates, some propositions seem clearly to be common knowledge, while others seem clearly not to be common knowledge; where the line should be drawn, however, is a matter of contextual

multi-agent token semantics. Because of the special condition for reflexive arrow, we do not get straightforward reductions like $\models_{\text{MTS}} \Diamond_1\Diamond_2\Diamond_1\Diamond_2\Diamond_1\phi \rightarrow \Diamond_1\Diamond_2\Diamond_1\phi\ [2,2]$ (what I think that you might that I might think that you might think that I consider as possible is just 'by default' what I think that you might think that I consider possible) but rather the less intuitive $\models_{\text{MTS}} \Diamond_1\Diamond_2\Diamond_1\Diamond_2\Diamond_1\phi \rightarrow (\Diamond_1\Diamond_2\Diamond_1\phi \vee \Diamond_2\Diamond_1\Diamond_2\Diamond_1\phi)\ [2,2]$. In order to get the simpler reduction, one might prefer to put a cost on reflexive arrows. However, as we pointed out for TS (see fn. 13), the resulting logic would not be adequate to axiomatize knowledge (as opposed to belief) over reflexive structures. Such a logic would not necessarily be inadequate, nevertheless, if we think the kind of trivialization we obtain describes some form of common belief, as opposed to common knowledge properly so-called. We leave a precise discussion of the difference between common knowledge and common belief to further discussion.

dependency.

As a consequence, we think two aspects of the problem should be distinguished: the first is the normative and logical aspect; the second is the descriptive and empirical aspect. From a logical point of view, there is undeniably a principled distinction to make between the infinitary, sharp and context-independent concept of common knowledge resulting from the assignment of omega tokens to each agent on the one hand, and the finitary, vague and context-dependent concept of common knowledge resulting from the assignment of a finite number of tokens to the agents on the other hand. As an anonymous reviewer put it, what our semantics models is probably not "real common knowledge but a 'quick and dirty' way of thinking about common knowledge that humans resort to in order to avoid cognitive overload". From an empirical point of view, however, it could well be that the strict and rigorous concept of common knowledge corresponding to the assignment of infinitely many tokens serves only as an idealization, and that rationally bounded agents attain common knowledge in the way described by token semantics and, as it were, under a "veil of ignorance" regarding the precise number of iterations each agent is ready to compute.

Since we first conceived of the application of token semantics to common knowledge, it thus occurred to us that the concept we are modelling more adequately corresponds to what Rubinstein (1989) dubbed "almost common knowledge", to characterize situations in which agents have a large, but nonetheless finite, degree of shared knowledge. It would take us beyond the scope of the present paper to engage into a more detailed discussion of this concept of almost common knowledge. Importantly, however, although we do agree that there is a difference between situations of real common knowledge and situations of almost common knowledge, we want to leave open the possibility that the concept of almost common knowledge we are describing serves as a substitute for the standard concept of common knowledge for rationally bounded agents in practical situations in which they have to make decisions.[20]

4.4 Common knowledge and individual knowledge

In order to clarify this issue even more, we propose to look at an informal argument, purporting to show that in the case of the Consecutive Numbers Game, the agents do not have real common knowledge that their numbers are less than any number k. The argument was presented to us by an anonymous reviewer, who moreover pointed out the parallel with the syntactic argument used

[20] An instance of this, which we discovered after the first version of this paper was submitted, and which we investigate in a sequel to this paper, is given by A. Rubinstein's decision-theoretic paradox of the Email Game (Rubinstein 1989), in which one player informs another one by email of a certain state of nature, on which the nature of the game depends, in such a way that the two machines give confirmation messages automatically, with the same probability of transmission failure for each sending. Each agent sees on his screen the number of messages his machine sent before a transmission failure occurred between the two machines (at which point the communication stops). When b reads 17, b is uncertain whether a sent 17 or 18 emails. The Email Game, as it turns out, has an informational structure very similar to that of the Consecutive Numbers Game (and which may be called Equal-or-Consecutive Numbers). Rubinstein's paradox is the result that in the Email Game, lack of common knowledge about which game agents are playing makes room for only one possible equilibrium, in which agents always choose the same action with null payoff, irrespective of the very large degree of shared knowledge they have that the actual game is one in which the alternative action would be more profitable to both.

by Williamson against the principle of positive introspection for individual knowledge, and with an aim to show that the puzzle about common knowledge raised by the Game is independent of the particular framework of Kripke semantics assumed for knowledge. We follow the reviewer's presentation here, assuming as we did previously that the individual knowledge operators are deductively closed.

Let us note $n(a) \leq k+1$ the sentence "a's number is less than $k+1$". The assumption we make is that, from an intuitive point of view, it is common knowledge that a's number is less than $k+1$ for some finite k, namely:

(1) $\quad C(n(a) \leq k+1)$

Following the reviewer, "it's also built into the structure of the game that if $n(a) = k+1$ then a does not know that b knows that $n(a) = k+1$, for if $n(a) = k+1$ then for all a knows $n(b) = k+2$, in which case for all b knows $n(a) = k+3$. Since the structure of the game is common knowledge, we have":

(2) $\quad C(n(a) = k+1 \rightarrow \neg K_a K_b\, n(a) \leq k+1)$

By the definition of common knowledge, it follows from (1) that:

(3) $\quad CK_a K_b\, (n(a) \leq k+1)$

By the closure principles on the knowledge operators, it follows from (2) and (3) that:

(4) $\quad C(n(a) \neq k+1)$

From (1) and (4), it follows by deductive closure again that:

(5) $\quad C(n(a) \leq k)$

Iterating the same argument, if follows that:

(6) $\quad C(n(a) \leq 0)$

This conclusion is obviously too strong, since $n(a)$ can very well be greater than 0. According to the reviewer, "this argument nowhere appeals to Kripke semantics, but only to theoretically desirable features of common knowledge. The failure of the new semantics to validate the argument, and to exclude the conclusion in this scenario, seems to be a problem for the new two-dimensional semantics, not for the argument". Assuming that all the features of common knowledge here assumed are indeed desirable features, the only reasonable way out is to reject assumption (1), namely the supposition that there is a k sufficiently large such that agents have common knowledge that a's number is less than k.

As can be checked, the argument (1)-(6) is exactly analogous to the argument we discussed earlier in section 2.3 in the case of individual knowledge. The difference is that the operator of common knowledge appears in the place of the operator of individual knowledge, and the complex operator $K_a K_b$, namely "a knows that b knows that", is used in (2) to get an analogue of the principle of knowledge of an (appromixation of) one's margin, and in (3) to get an analogue of the principle (KK). In the individual case, we saw that Williamson proposes to reject the analogue of (3), namely positive introspection. In the present case, by contrast, (3) cannot be blamed, since

it follows from the definition of common knowledge, and surely Williamson would agree that (1) therefore is to be rejected.

In the case of individual knowledge, we saw earlier that token semantics (or centered semantics) fails to validate the analogue of (2), namely the principle of knowledge of an (appromixation of) one's margin, and we explained in what sense we consider this feature of the semantics to be desirable, contrary to Williamson. In the present case, assuming the agents only have a finite number of tokens, the semantics is likewise no longer sound with respect to (2). This leads us to a more precise discussion of the intuitive justification given by the reviewer in favour of premise (2) for common knowledge. In the comments quoted above, the reviewer appeals to the idea that "the structure of the game is common knowledge" between the players to justify an analogue of the rule of necessitation for common knowledge. From a semantic viewpoint, to say that "$n(a) = k+1 \rightarrow \neg K_a K_b\, n(a) \leq k+1$ is built into the structure of the game" is to say that it holds at every point in the game. As we saw, however, centered semantics (and token semantics more generally) invalidates the rule of necessitation for individual knowledge over models, and similarly for common knowledge, so that $C(n(a) = k+1 \rightarrow \neg K_a K_b\, n(a) \leq k+1)$ is no longer true at every point in the game structure.[21] This means that, even though our model correctly represents the structure of game, we are implicitly denying that this structure is common knowledge between the agents. But such a denial might nonetheless prove sensible, if the agents are assumed to be imperfect reasoners.[22]

Either way, therefore, we do agree indeed that the concept of common knowledge we are describing is not true common knowledge, but rather an intuitive approximation of common knowledge. If we concede, however, that the concept we are describing is not real common knowledge, but rather a distinct concept of "almost common knowledge", then we need to face a second objection, which concerns our treatment of individual knowledge. As our reviewer forcefully points out by analogy with the case of common knowledge, "perhaps introspection is a cognitive illusion in the case of inexact knowledge", so that what our semantics models "is not real inexact knowledge but a 'quick and dirty' way of thinking about it that humans resort to in order to avoid cognitive overload". Is this the case, however? Should we consider that, in the same way in which our semantics describes a concept distinct from the true concept of common knowledge, in the individual case it describes a concept of inexact knowledge that is not the right concept, precisely because it validates positive introspection?

Our answer to this question remains negative. In the case of common knowledge as described in token semantics, we may conceive indeed that the agents share a common illusion of common knowledge, rather than common knowledge properly so-called. In particular, agents who have real common knowledge should be able to adequately coordinate their actions on that basis. In a game like Consecutive Numbers, by contrast, the vagueness of the number of tokens available to

[21] For instance, $(2,1) \vDash_{\mathrm{MTS}} K_a K_b(n(a) \leq 4) \rightarrow n(a) \neq 4\ [1,1]$. However, $(2,1) \nvDash_{\mathrm{MTS}} K_a K_b(K_a K_b(n(a) \leq 4) \rightarrow n(a) \neq 4)\ [1,1]$, for while $((2,1),0)((2,3),a)((4,3),b) \vDash_{\mathrm{MTS}} K_a K_b(n(a) \leq 4)\ [0,0]$, it holds that $((2,1),0)((2,3),a)((4,3),b) \nvDash_{\mathrm{MTS}} n(a) \neq 4\ [0,0]$.

[22] Note that this does not contradict the assumption that the structure of the game has been publicly announced to the agent. If the agents are not perfectly rational, it might be the case that some truths about the game – like $n(a) = k+1 \rightarrow \neg K_a K_b\, n(a) \leq k+1$ – do not thereby become common knowledge. From a logical point of view, this is precisely what is reflected by the failure of model necessitation in our model.

the agents is likely to be a threat, in some situations, for them to adequately coordinate. But can we present a similar argument in favour of the idea that an agent with inexact knowledge of the sort Williamson describes has the illusion of knowing when he knows?

We don't see how the argument would go here, other than repeating and accepting Williamson's own premises. Interestingly in this respect, Williamson (2000: 121-123) himself considers a version of the argument (1)-(6) for the case of individual knowledge, in which an infinitary operator K^ω is used instead of K. The knowledge described by K^ω thus is to K in the one-agent case what common knowledge is to K_aK_b in the case of two agents. As Williamson admits, however, "the condition that one knows$^\omega p$ seems to be luminous", and "knowledge$^\omega$ presents an interesting challenge to the general argument against luminosity". The reason is that K^ω, just like the operator C of common knowledge, by definition will obey the analogue of positive introspection. Thus, in his discussion of the analogue of (1)-(6) for K^ω, Williamson is led to reject the analogue of (1), namely the idea that one can attain such a state of knowledge as described by K^ω, his idea being that each new iteration of knowledge is hard. Because common knowledge of p implies $K^\omega p$ from each individual agent, Williamson considers that common knowledge too should be an obstacle to his general argument against luminosity. We find an interesting convergence with Williamson here, who sees common knowledge as a "convenient idealization". Where Williamson would maintain that no such knowledge as described by K^ω is attainable, except by way of idealization in the case of common knowledge, we actually want to respond the opposite: the knowledge described by K^ω can hold both in the individual case and in situations in which common knowledge is indeed achieved, because contrary to Williamson we don't see each iteration as adding a layer of difficulty in the case of a single introspective agent. With Williamson, we converge on the idea that "perhaps some everyday practices of communication and decision-making depend on a pretence that we have common knowledge" (2000: 123). But our reasons to converge on this conclusion are not exactly based on the difficulty there is, for each individual separately, to reach $K^\omega p$ on the basis of Kp. In other words, they are not based on the difficulty of introspecting on one's *own* states of mind. Rather, they are based on the difficulty there is in computing one another's state of mind in a situation in which several agents are involved. As we noted in the beginning of the Section, we make an important difference between iterations of knowledge in the case of a single agent, and iterations of knowledge in the case of several agents: for us, knowing that one knows is easy; knowing that the other knows is indeed more demanding. Actual common knowledge, finally, may in some cases be hard and in others be easy, depending on the context. But the point of this section is that some finite level of shared knowledge may be, in practice, sufficient for agents to act as if they had actual common knowledge.

5 Conclusion

The results of this paper should convince us that logics of introspective belief and knowledge like $\mathbf{K45}$ and $\mathbf{S5}$ 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 mar-

gin 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 Rubinstein's Email Game 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 one 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. We leave these different issues for future research.

Acknowledgments. A preliminary version of this paper appeared in the Proceedings of the ESSLLI 2006 *Rationality and Knowledge* Workshop, edited by S. Artemov and R. Parikh (see Bonnay & Égré 2006). Special thanks are due to J. van Benthem, M. Cozic, O. Roy, P. Schlenker and Y. Wang for a number of useful comments and suggestions. We also thank three anonymous ESSLLI reviewers for helpful corrections. We are particularly grateful to an anonymous reviewer from the JPL, whose detailed critical comments and perceptive philosophical objections helped us to substantially improve our initial work. We also express our gratitude to audiences in Geneva, Stanford, Lisbon (ENFA 3) and Malaga (ESSLLI 2006), where different aspects of the paper were presented, as well as in Paris (PALMYR 3), Amsterdam (PALMYR 4), Aix-en-Provence (SOPHA) and Prague (2006 Colloquium on Vagueness and Uncertainty).

References

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

[2] Bonnay D. & Égré P. (2006), "A Non-standard Semantics for Inexact Knowledge with Introspection", *Proceedings of the ESSLLI 2006 Workshop Rationality and Knowledge*, R. Parikh and S. Artemov (eds.).

[3] Bonnay D. & Égré P. (2007), "Vagueness and Introspection", Prague Colloquium on Reasoning about Vagueness and Probability, manuscript, IJN.

[4] Bonnay D. & Égré P. (2008), "Margins for Error in Context", forthcoming in *Relative Truth*, M. Garcia-Carpintero & M. Kölbel (eds.), Oxford UP.

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

[6] Dokic J. & Égré P. (2004), "Margin for Error and the Transparency of Knowledge", forthcoming in *Synthese*.

[7] Égré P. (2006), "Reliability, Margin for Error and Self-Knowledge", forthcoming in D. Pritchard & V. Hendricks (eds.), *New Waves in Epistemology*, Ashgate.

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

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

[10] J. Halpern (2004), Intransitivity and Vagueness, *Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning (KR 2004)*, 121-129.

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

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

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

[14] Rubinstein A. (1989), "The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge'". *American Economic Review*, 79, 385-391.

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

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

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

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

Appendix

The aim of this Appendix is to prove the soundness and completeness theorems stated in paragraph 3.3 as well as the corresponding results in the multi-agent case. First, we give full proofs for Theorems 4 and 5. The second part of the appendix deals with MTS. Since the proofs for MTS follow the same pattern, we only present the key modifications to be made.

Completeness for TS

Let us first recall Theorems 4 and 5:

Theorem 4. $\mathrm{K}(4.n')(5.n')$ *is sound and complete with respect to* $\mathrm{TS}(n)$.

Theorem 5. $\mathrm{KT}(4.n')(5.n')$ *is sound and complete with respect to* $\mathrm{TS}(n)$ *over the class of reflexive frames.*

In order to prove these results, we shall rely on standard soundness and completeness results for Kripke semantics and 'transfer' them to token semantics. Thus, the main line of the proof is quite similar to what we have done for centered semantics in section 2. Moreover, Theorem 2 for centered semantics is actually a particular case of Theorem 4, since centered semantics is the same as token semantics with one token.[23] The proof will consist in three steps:

Step 1: $(4.n')$ and $(5.n')$ are proved to define specific properties of the accessibility relation with respect to the standard Kripke semantics.

Step 2: On models which satisfy these properties, token semantics and Kripke semantics are shown to be equivalent (Transfer Theorem). Moreover, every pointed model can be 'unpacked' into a model satisfying these properties (Unpacking Theorem)

Step 3: From the soundness and completeness results for Kripke semantics over the class of frames satisfying the relevant properties, the Transfer Theorem and the Unpacking Theorem deliver soundness and completeness results for token semantics.

We now proceed to the first step in the proof. For ease of reference, we first recall $(4.n')$ and $(5.n')$:

$(4.n')$ $(\neg p_1 \wedge \Diamond(p_1 \wedge \neg p_2 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Diamond\Diamond r)\ldots))) \rightarrow \Diamond(p_1 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Diamond r)\ldots))$

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

[23] Theorem 3 also follows from the *proof* of Theorem 5. However, note that Theorem 5 does not provide a characterization of the logic of *token* semantics over fixed-margin models. These models are not only reflexive but also symmetric.

The property of the accessibility relation corresponding to $(4.n')$ is a variation on transitivity. We say that $w_0, ..., w_n$ is an *R-sequence* of worlds if $w_i R w_{i+1}$ for all $i < n$. We shall say that an R-sequence of worlds $w_0, ..., w_n$ is *repetition free* if $w_i \neq w_{i+1}$ for all $i < n$. The property behind $(4.n')$ is that transitivity holds at all worlds which are the endpoints of a repetition free R-sequence of length $n-1$. We call this property n'-transitivity:

n'-transitivity: $\forall w, w', w'' \, (\exists x_1, ..., x_n [x_1 R x_2 \wedge ... \wedge x_{n-1} R x_n \wedge x_1 \neq x_2 \wedge ... \wedge x_{n-1} \neq x_n \wedge x_n = w] \rightarrow (wRw' \wedge w'Rw'' \rightarrow wRw''))$

It is easy to check that $1'$-transitivity is the standard notion of transitivity. Similarly, the property of the accessibility relation which corresponds to $(5.n')$ is a variation on euclideanness. The idea is that the accessibility relation is euclidean at all worlds which are the endpoints of a repetition free R-sequence of length $n-1$. We call this property n'-euclideanness:

n'-euclideanness: $\forall w, w', w'' \, (\exists x_1, ..., x_n [x_1 R x_2 \wedge ... \wedge x_{n-1} R x_n \wedge x_1 \neq x_2 \wedge ... \wedge x_{n-1} \neq x_n \wedge x_n = w] \rightarrow (wRw' \wedge wRw'' \rightarrow w'Rw''))$

Again, it is easy to check that we get the standard notion of a euclidean relation when $n = 1$. When the accessibility relation satisfies both $(4.n')$ and $(5.n')$, the following holds for all worlds w which are the endpoints of a repetition free R-sequence of length $n-1$: whenever wRw', the worlds accessible from w and the worlds accessible from w' are the same. This is the crucial property for some of the proofs below.

Using Kripke semantics, we obtain the following standard definability results:

Fact 5. $(4.n')$ *defines the class of n'-transitive frames.*

Proof. (i) Let $\mathcal{F}$ be an n'-transitive frame. We show that $(4.n')$ is valid on $\mathcal{F}$. Let $\mathcal{M}$ be a model based on $\mathcal{F}$ and v a world in $\mathcal{M}$. Assume that $\mathcal{M}, v \vDash \neg p_1 \wedge \Diamond(p_1 \wedge \neg p_2 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Diamond\Diamond r) \ldots))$. The $n-1$ diamonds force the existence of an R-sequence of worlds. This R-sequence is repetition free, because, given two consecutive worlds, the first one makes false an atom which the second one makes true. Let w be the endpoint of this repetition free R-sequence. Because of the occurrence of $\Diamond\Diamond r$ in the formula, we know that there are two worlds w', w'' such that wRw', $w'Rw''$ and w'' satisfies r. But then, the conditions for n'-transitivity are satisfied, so wRw''. This ensures that $\mathcal{M}, v \vDash \Diamond(p_1 \wedge \Diamond(p_2 \wedge \ldots \wedge \Diamond(p_{n-1} \wedge \Diamond r) \ldots))$.

(ii) Let $\mathcal{F}$ be a frame which is not n'-transitive. We show that $(4.n')$ is not valid on $\mathcal{F}$. For $\mathcal{F}$ not to satisfy n'-transitivity, there has to be a world w such that:

- there is a repetition free R-sequence $w_1, ..., w_n$ with $w_n = w$.
- there are worlds w' and w'' with wRw', $w'Rw''$ and w'' is not accessible from w.

Define the valuation V by $V(p_i) = \{w_{i+1}\}$ for all $i \in \{1, ..., n-1\}$ and $V(r) = \{w''\}$. Clearly, $(\mathcal{F}, V), w_1 \nVdash (4.n')$.

□

Fact 6. $(5.n')$ *defines the class of* n'*-euclidean frames.*

Proof. The proof is similar to the proof for $(4.n')$.

□

We are done with the first step of the proof. The aim of the second step is to help us connect Kripke semantics and token semantics. Of course, it is not true in general that $\mathcal{M}, w \vDash \phi$ iff $\mathcal{M}, w \vDash_{\text{TS}} \phi\ [n]$. However, Kripke semantics and token semantics do agree on special classes of models. This will happen when the accessibility relation guarantees that the number of tokens one is starting with is big enough to reach all the worlds that are relevant to the evaluation of an arbitrary formula.

Definition 7. *The cost* $c(q)$ *of a sequence of worlds* q *is defined as the number of non-trivial steps between consecutive worlds in* q*.*

For instance, $c(ww'w'') = 2$, whereas $c(www') = 1$. Thus, the cost of a sequence corresponds to the number of tokens necessary to move along the sequence.

Transfer Theorem. *Let* $\mathcal{M}$ *be an* n'*-transitive and* n'*-euclidean model, for every world* w*, sequence* q*, and formula* ϕ*, such that* $c(qw) \leq n$*,* $\mathcal{M}, qw \vDash_{\text{TS}} \phi\ [n - c(qw)]$ *iff* $\mathcal{M}, w \vDash \phi$*.*

Proof. By induction on ϕ:

- If ϕ is atomic, the definitions of $\vDash$ and $\vDash_{\text{TS}}$ straightforwardly agree.
- Boolean cases are immediate as well.
- Let ϕ be of the form $\Box\psi$. We want to show that $\mathcal{M}, qw \vDash_{\text{TS}} \Box\psi\ [n - c(qw)]$ iff $\mathcal{M}, w \vDash \Box\psi$. We distinguish two cases:
  - Case 1: $c(qw) < n$.
    - $\mathcal{M}, qw \vDash_{\text{TS}} \Box\psi\ [n - c(qw)]$
    - iff for all w' such that wRw', $\mathcal{M}, qww' \vDash_{\text{TS}} \psi\ [n - c(qw) - k]$, where $k = 1$ if $w \neq w'$ and $k = 0$ if $w = w'$ (by definition of the satisfaction clause for $\Box$ in $\vDash_{\text{TS}}$),
    - iff for all w' such that wRw', $\mathcal{M}, w' \vDash \psi$ by induction hypothesis,
    - iff $\mathcal{M}, w \vDash \Box\psi$ (by definition of the satisfaction clause for $\Box$ in $\vDash$).
  - Case 2: $c(qw) = n$. In this case, we can safely assume that q is non-empty, since some tokens have been spent. Let q be $q'z$, where q' is a (possibly empty) sequence of worlds.
    - $\mathcal{M}, q'zw \vDash_{\text{TS}} \Box\psi\ [n - c(q'zw)]$,
    - iff $\mathcal{M}, q'z \vDash_{\text{TS}} \Box\psi\ [1]$ (by definition of the satisfaction clause for $\Box$ in $\vDash_{\text{TS}}$, since $n - c(q'zw) = 0$),

- iff for all w' such that zRw', $\mathcal{M}, q'zw' \vDash_{\text{TS}} \psi\ [1-k]$ where $k = 1$ if $w' \neq z$ and $k = 0$ if $w' = z$ (by definition of the satisfaction clause for $\Box$ in $\vDash_{\text{TS}}$)
- iff for all w' such that zRw', $\mathcal{M}, w' \vDash \psi$ (by induction hypothesis: since $c(q'zw') = n - 1 + k$, we do have $n - c(q'zw') = 1 - k$),
- iff for all w' such that wRw', $\mathcal{M}, w' \vDash \psi$ (because $c(q'z) = n - 1$, z is the end point of a repetition free R-sequence of length $n-1$. By n'-transitivity and n'-euclideanness of $\mathcal{M}$, this implies that the worlds accessible from w and the worlds accessible from z are the same).
- iff $\mathcal{M}, w \vDash \Box\psi$

□

In the model theory of modal logics, the notion of bisimulation serves as the semantic correlate of elementary equivalence: two bisimilar pointed models satisfy the same modal formulas, and the converse is true provided some restrictions are made on the models (see Blackburn et al. 1999, p. 69). In order to prove our next theorem, we introduce a notion of bisimulation for token semantics. More precisely, we define a notion of n'-bisimulation, which is the analogue of the notion of n-bisimulation, or elementary equivalence for formulae of modal depth n (See Blackburn et al. 1999, p. 74).[24]

Definition 8. *[n'-bisimulation]*

Let $\mathcal{M}$ and $\mathcal{M}'$ be two models, and let w and w' be two worlds of $\mathcal{M}$ and $\mathcal{M}'$, respectively. w and w' are n'-bisimilar (notation: $\mathcal{M}, w \bowtie_{n'} \mathcal{M}', w'$) if there exists, over $W \times W'$, a sequence of binary relations $Z_n \subseteq ... \subseteq Z_0$ with the following properties (for $i + 1 \leq n$):

i) wZ_nw',

ii) if vZ_0v' *then* v *and* v' *agree on all propositional letters,*

iii) if vZ_0v', *then* vRv *iff* $v'R'v'$,

iv) if $vZ_{i+1}v'$ *and* vRu *with* $v \neq u$, *then there exists* u' *with* $v'R'u'$, $v' \neq u'$ *and* uZ_iu',

v) if $vZ_{i+1}v'$ *and* $v'R'u'$ *with* $v' \neq u'$, *then there exists* u *with* vRu, $v \neq u$ *and* uZ_iu'.

Two n'-bisimilar pointed models satisfy the same formulas at their distinguished worlds:

Theorem 7. *If $\mathcal{M}, w \bowtie_{n'} \mathcal{M}', w'$, then for every formula ϕ, $\mathcal{M}, w \vDash_{\text{TS}} \phi\ [n]$ iff $\mathcal{M}', w' \vDash_{\text{TS}} \phi\ [n]$.*

[24] n'-bisimulations and n-bisimulation are the same, but for the addition of condition iii) and the restriction to different worlds in conditions iv) and v) of Definition 8. We use n' instead of n in order to avoid any possible confusion between our definition and the one in Blackburn et al., and also because preservation of reflexivity, as required by iii), is connected with the validity of (4.n') and (5.n').

Proof. Let $Z_n, ..., Z_0$ be the relations over $M \times M'$ given by the n'-bisimulation. We say that, for $k \leq n$, an R-sequence $w_0, ..., w_k$ of worlds in $\mathcal{M}$ and an R'-sequence $w'_0, ..., w'_k$ of worlds in $\mathcal{M}'$ *match* if $w_0 = w$, $w'_0 = w'$, for all $i < k$, $w_i = w_{i+1}$ iff $w'_i = w'_{i+1}$ and for all $i \leq k$, $w_i Z_{n-c(w_0,...,w_i)} w'_i$. Note that if qv and $q'v'$ match, so do q and q'. Moreover, if qv and $q'v'$ match, then $c(qv) = c(q'v')$

We prove by induction on the complexity of ϕ that if qv and $q'v'$ are two matching sequences of cost less or equal to n, then $\mathcal{M}, qv \vDash_{\text{TS}} \phi\ [n - c(qv)]$ iff $\mathcal{M}', q'v' \vDash \phi\ [n - c(q'v')]$ (note that if q and q' are empty, then $v = w$ and $v' = w'$, which gives us our theorem):

- If ϕ is atomic, this follows from the fact that $vZ_{n-c(qv)}v'$,
- Boolean cases are immediate,
- Let ϕ be of the form $\Diamond\psi$. We must show that $\mathcal{M}, qv \vDash_{\text{TS}} \Diamond\psi\ [n - c(qv)]$ iff $\mathcal{M}', q'v' \vDash_{\text{TS}} \Diamond\psi[n - c(q'v')]$. We distinguish two cases:
  - Case 1: $c(qv) < n$.
    Assume $\mathcal{M}, qv \vDash_{\text{TS}} \Diamond\psi\ [n - c(qv)]$. By the satisfaction clause for $\Diamond$ in TS, there is a u such that vRu, $\mathcal{M}, qvu \vDash_{\text{TS}} \psi\ [n - c(qw) - k]$, where $k = 1$ if $v \neq u$ and $k = 0$ if $v = u$. If $v = u$, we know that $vZ_{n-c(qv)}v'$: this implies that $v'R'v'$, and that quv and $q'v'v'$ match. By induction hypothesis, we get that $\mathcal{M}', q'v'v' \vDash_{\text{TS}} \psi\ [n - c(q'v'v')]$. By definition of satisfaction for $\Diamond$ in TS, we get $\mathcal{M}', q'v' \vDash_{\text{TS}} \Diamond\psi[n - c(q'v')]$. If $v \neq u$, we know that $vZ_{n-c(qv)}v'$: this implies that there is a u' such that $v'Ru'$, $v' \neq u'$ and $uZ_{n-c(qvu)}u'$. qvu and $q'v'u'$ match. Hence, by induction hypothesis, we get that $\mathcal{M}', q'v'u' \vDash_{\text{TS}} \psi\ [n - c(q'v'u')]$. Again from the definition of satisfaction for $\Diamond$ in TS, $\mathcal{M}', q'v' \vDash_{\text{TS}} \Diamond\psi[n - c(q'v')]$. In the other direction, the proof is exactly similar.
  - Case 2: $c(qv) = n$.
    In this case, $n - c(qv) = 0$, hence q is non-empty and the last world in q is different from v, hence $n - c(q) = 1$. The same holds for q' and v'. Moreover q and q' match, since qv and $q'v'$ do. Then, we have $\mathcal{M}, qv \vDash_{\text{TS}} \Diamond\psi\ [n - c(qv)]$ iff $\mathcal{M}, q \vDash_{\text{TS}} \Diamond\psi\ [n - c(q)]$ (by the satisfaction clause for $\Diamond$ in $\vDash_{\text{TS}}$) and $\mathcal{M}', q'v' \vDash_{\text{TS}} \Diamond\psi\ [n - c(q'v')]$ iff $\mathcal{M}', q' \vDash_{\text{TS}} \Diamond\psi\ [n - c(q')]$ (similarly). So what we need is $\mathcal{M}, q \vDash_{\text{TS}} \Diamond\psi\ [n - c(q)]$ iff $\mathcal{M}', q' \vDash_{\text{TS}} \Diamond\psi\ [n - c(q')]$, which is case 1, since $c(q) = c(q') < c(qv) = c(q'v') = n$.

□

In Kripke semantics, an n-bisimulation guarantees equivalence for formulas of depth *at most* n. In TS(n), as shown by the previous theorem, the restriction is no longer needed. This should come as no surprise: in TS(n), the evaluation process of a modal formula cannot lead one to explore worlds at a distance greater than n from the starting point, even if the depth of the formula is greater than n.

The Transfer Theorem states that on n'-transitive and n'-euclidean models, Kripke semantics and token semantics (for a given number of tokens) are equivalent. Thus, everything goes

as if token semantics provides new models that validate $(4.n')$ and $(5.n')$ even when they are not n'-transitive and n'-euclidean. How new are these models? Is it possible to turn them into models which standardly validate $(4.n')$ and $(5.n')$? We provide a positive answer to this question by defining an operation of 'unpacking' on models, which turns an arbitrary model into a (standardly) equivalent model which is n'-transitive and n'-euclidean.

Definition 9 (n-unpacked model)**.** *Let $\mathcal{M} = \langle M, R, V\rangle$ be a Kripke model and $n \neq 0$. The n-unpacking of $\mathcal{M}$, written $UP_n(\mathcal{M})$, is the Kripke structure $\langle M \times M \times \{0, ..., n\}, R', V'\rangle$ with*

- $(w, w', i)R'(v, v', j)$ *iff*
  - $i > 0$, $j = i - 1$, $v = w'$, $w'Rv'$ *and* $w' \neq v'$,
  - *or* $i = j$, $v = w$, $w' = v'$, *and* $w'Rw'$,
  - *or* $i = j = 0$, $v = w$, *and* wRv'.
- $(w, w', i) \in V'(p)$ *iff* $w' \in V(p)$

We call the index i of a world (w, w', i) the number parameter of the world. Note that if uRu, the second and third conditions in the definition of R' both yield $(u, u, 0)R'(u, u, 0)$. The definition of unpacking is a way of "externalizing" the use of tokens in the models. For instance, the first condition reflects the idea that with a positive number i of tokens, coming from w', one can access a world v' distinct from w' by spending one token. The second condition reflects the case of reflexive moves, and the third the backtracking case when the number of tokens left is 0.

Unpacking Theorem. *Let $\mathcal{M}$ be a structure and w a world in $\mathcal{M}$, the following hold:*

1) $\mathcal{M}, w \bowtie_{n'} UP_n(\mathcal{M}), (w, w, n)$

2) $UP_n(\mathcal{M})$ is n'-transitive and n'-euclidean and $\mathcal{M}$ is reflexive iff $UP_n(\mathcal{M})$ is,

3) for every formula ϕ, $\mathcal{M}, w \vDash_{\text{TS}} \phi\ [n]$ iff $UP_n(\mathcal{M}), (w, w, n) \vDash \phi$

Proof. 1) We have to find a family of relations $\{Z_i\}_{i\in\{0,...,n\}}$ on $M \times M'$ which is an n'-bisimulation. We set $Z_n = \{\langle w, (w, w, n)\rangle\}$ and define the rest of the Z_i from top to bottom by (for $1 \leq i < n$):

$Z_i = Z_{i+1} \cup \{\langle u, (u', v', i)\rangle |$ there is an R-sequence wqu with $c(wqu) = n - i$, u' is the penultimate world in the sequence and $v' = u\ \}$

Clearly, $Z_n \subseteq ... \subseteq Z_0$. Note that if a world is in the range of a given Z_i but not in the range of Z_{i+1}, then its number parameter has to be i.

Moreover, the five properties of n'-bisimulations hold:

i) $wZ_n(w, w, n)$ by definition of Z_n.

ii) $uZ_i(u', v', k)$ implies $u = v'$, hence by definition of $UP_n(\mathcal{M})$, u and (u', v', k) agree on all propositional letters.

iii) Assume $uZ_i(u', v', k)$: by definition of $UP_n(\mathcal{M})$, $(u', v', k)R(u', v', k)$ iff $v'Rv'$, that is, since $uZ_i(u', v', k)$ implies $u = v'$, iff uRu.

iv) Assume $uZ_{i+1}(u', v', k)$ and uRv with $u \neq v$. We distinguish two cases:

- Case 1: $i + 1 = n$, this means that $u = w$ and $(u', v', k) = (w, w, n)$. We show that the world $(w, v, n - 1)$ in $UP_n(\mathcal{M})$ satisfies the three conditions we need. First, $(w, w, n)R'(w, v, n - 1)$ follows from wRv, $w \neq v$ and the definition of $UP_n(\mathcal{M})$. Then, trivially, $(w, w, n) \neq (w, v, n - 1)$. Finally, wv is an R-sequence of cost $n - (n - 1) = 1$, with w the penultimate world, v the last one in the sequence, and the number parameter of the world is $n - 1$. So, by definition of Z_{n-1}, $vZ_{n-1}(w, v, n - 1)$.
- Case 2: $i + 1 < n$. It is sufficient to show that the property holds for the pairs $\langle u, (u', v', i + 1)\rangle$ that were not already in Z_{i+2}. We show that the world $\langle v, (u, v, i)\rangle$ in $UP_n(\mathcal{M})$ satisfies the three conditions we need. First $(u', u, i + 1)R'(u, v, i)$ follows from uRv, $u \neq v$ and the definition of $UP_n(\mathcal{M})$. Then, trivially, $(u', u, i + 1) \neq (u, v, i)$. Finally, by definition of Z_{i+1}, there is an R-sequence wqu with $c(wqu) = n - (i + 1)$, u' the penultimate world in the sequence and u the last world in the sequence. This implies that there is an R-sequence $wquv$ with $c(wquv) = n - i$, u the penultimate world in the sequence and v the last world in the sequence. Thus, by definition of Z_i, we have $vZ_i(u, v, i)$.

v) Similar to the proof for condition iv).

2) First, we show that $UP_n(\mathcal{M})$ is n'-transitive. Let (u, v, k) be a world in $UP_n(\mathcal{M})$ such that:

a) there is a repetition free R'-sequence $w'_0, ..., w'_n$ of worlds in $UP_n(\mathcal{M})$ with $w'_n = (u, v, k)$

b) there are worlds w', w'' such that $(u, v, k)R'w'$ and $w'R'w''$.

We need to show that $(u, v, k)R'w''$. By definition of $UP_n(\mathcal{M})$, either the number parameter of w'_n is zero and then $k = 0$ or the number parameter of w'_{n-1} has to be $k + 1$, the one of w'_{n-2} has to be $k + 2$ and so on, which implies again that $k = 0$. By definition of $UP_n(\mathcal{M})$, b) implies that the number parameters for w' and w'' are 0 as well. w' is thus $(u, v', 0)$ for some v' such that uRv', and w'' is $(u, v'', 0)$ for some v'' such that uRv''. By definition of $UP_n(\mathcal{M})$, this ensures that $(u, v, 0)R'(u, v'', 0)$ as required.
Then, we show that $UP_n(\mathcal{M})$ is n'-euclidean. Let (u, v, k) be a world in $UP_n(\mathcal{M})$ such that:

a) there is a repetition free R'-sequence $w'_0, ..., w'_n$ of worlds in $UP_n(\mathcal{M})$ with $w'_n = (u, v, k)$,

b) there are worlds w', w'' such that $(u, v, k)R'w'$ and $(u, v, k)R'w''$.

We need to show that $(u, v, k)R'w''$. As before, we have $k = 0$ because of a). Again by definition of $UP_n(\mathcal{M})$, b) implies that the number parameters for w' and w'' are 0 as well. w' is

thus $(u, v', 0)$ for some v' such that uRv', and w'' is $(u, v'', 0)$ for some v'' such that uRv''. By definition of $UP_n(\mathcal{M})$, this ensures that $(u, v', 0)R'(u, v'', 0)$ as required.
Finally, assume that $\mathcal{M}$ is reflexive. We need to show that for all worlds u, v in $\mathcal{M}$ and all possible number parameters k, $(u, v, k)R'(u, v, k)$. But this follows from the definition of $UP_n(\mathcal{M})$, since we have vRv.

3) Let ϕ be an arbitrary formula:

$\mathcal{M}, w \vDash_{\text{TS}} \phi\ [n]$
iff $UP_n(\mathcal{M}), (w, w, n) \vDash_{\text{TS}} \phi\ [n]$ – by Theorem 7 and 1),
iff $UP_n(\mathcal{M}), (w, w, n) \vDash \phi$ – by the Transfer Theorem and 2).

□

We are done with the second main step in the proof. We now turn to the third and last step. Let $Th_{(4.n')(5.n')}$ be the set of formulas which are theorems of $\mathrm{K}(4.\mathrm{n}')(5.\mathrm{n}')$. Let $Val_{K,n'}$ be the set of formulas which are valid on the class of n'-transitive and n'-euclidean frames for Kripke semantics. Finally, let $Val_{TS,n}$ be the set of formulas which are valid on all frames for token semantics with n tokens (*i.e.* the set of formulas ϕ such that for all $\mathcal{M}$,w, we have $\mathcal{M}, w \vDash_{\text{TS}} \phi\ [n]$). Theorem 4 says that $Th_{\mathrm{K}(4.\mathrm{n}')(5.\mathrm{n}')} = Val_{TS,n}$. Through a detour by $Val_{K,n'}$, this is precisely what we get as a consequence of the following lemma:

Lemma 3. *The following hold:*

i) $Th_{(4.n')(5.n')} = Val_{K,n'}$

ii) $Th_{(4.n')(5.n')} \subseteq Val_{TS,n}$

iii) $Val_{TS,n} \subseteq Val_{K,n'}$

Proof. i) $(4.n')$ and $(5.n')$ are Sahlqvist implications: their consequent is a positive formula, and their antecedent is built out of atoms or negated atoms using $\Diamond$ and $\wedge$.[25]. These two formulas define the class of n'-euclidean and n'-transitive frames. By Sahlqvist's theorem, this implies that $Th_{(4.n')(5.n')} = Val_{K,n'}$.

ii) Assume for contradiction that there is a formula ϕ such that $\phi \notin Val_{TS,n}$ and $\phi \in Th_{(4.n')(5.n')}$. Since $\phi \notin Val_{TS,n}$, there is a model $\mathcal{M}$ and a world w such that $\mathcal{M}, w \nvDash_{\text{TS}} \phi\ [n]$. By the unpacking theorem, there is a model $UP_n(\mathcal{M})$ such that $UP_n(\mathcal{M}), (w, w, n) \nvDash \phi$. The unpacking theorem also tells us that $UP_n(\mathcal{M})$ is n'-transitive and n'-euclidean. By i), this implies that ϕ is valid on $UP_n(\mathcal{M})$. Contradiction.

iii) Assume for contradiction that there is a formula ϕ such that $\phi \notin Val_{K,n'}$ and $\phi \in Val_{TS,n}$. Since $\phi \notin Val_{K,n'}$, there is a n'-euclidean and n'-transitive model $\mathcal{M}$ and a world w such that $\mathcal{M}, w \nvDash \phi$. By the transfer theorem, the properties of the accessibility relation on $\mathcal{M}$ guarantee that $\mathcal{M}, w \nvDash_{\text{TS}} \phi\ [n]$. Contradiction.

□

[25]To be fully precise, a Sahlqvist antecedent is a formula built up from $\top$, $\bot$, boxed atoms and negative formulas using $\wedge$, $\vee$ and $\Diamond$. Atoms count as boxed atoms, hence our antecedents are indeed Sahlqvist. See Blackburn & alii (1999, section 3.6).

Similarly, let $Th_{\mathrm{T}(4n')(5n')}$ be the set of formulas which are theorems of $\mathrm{KT}(4\mathrm{n}')(5\mathrm{n}')$, $Val_{K,n',R}$ be the set of formulas which are valid on the class of n'-transitive and n'-euclidean reflexive frames according to Kripke semantics. Finally, let $Val_{TS,n,R}$ be the set of formulas which are valid on the class of reflexive frames for token semantics with n tokens. Theorem 5 says that $Th_{\mathrm{KT}(4.\mathrm{n}')(5.\mathrm{n}')} = Val_{TS,n,R}$. Through a detour by $Val_{K,n',R}$, this is precisely what we get as a consequence of the following lemma:

Lemma 4. *The following hold:*

i) $Th_{\mathrm{T}(4.n')(5.n')} = Val_{K,n',R}$

ii) $Th_{\mathrm{T}(4.n')(5.n')} \subseteq Val_{TS,n,R}$

iii) $Val_{TS,n,R} \subseteq Val_{K,n',R}$

Proof. i) $(4.n')$, $(5.n')$ and T are Sahlqvist implications. These three formulas define the class of n'-euclidean and n'-transitive reflexive frames. By Sahlqvist theorem, this implies that $Th_{\mathrm{T}(4.n')(5.n')} = Val_{K,n',R}$.

ii) Assume for contradiction that there is a formula ϕ such that $\phi \notin Val_{TS,n,R}$ and $\phi \in Th_{\mathrm{T}(4n')(5n')}$. Since $\phi \notin Val_{TS,n,R}$, there is a reflexive model $\mathcal{M}$ and a world w such that $\mathcal{M}, w \nVdash_{\mathrm{TS}} \phi\ [n]$. By the unpacking theorem, there is a model $UP_n(\mathcal{M})$ such that $UP_n(\mathcal{M}), (w, w, n) \nVdash \phi$. The unpacking theorem also tells us that $UP_n(\mathcal{M})$ is n'-transitive, n'-euclidean *and reflexive*. By i), this implies that ϕ is valid on $UP_n(\mathcal{M})$. Contradiction.

iii) Assume for contradiction that there is a formula ϕ such that $\phi \notin Val_{K,n',R}$ and $\phi \in Val_{TS,n,R}$. Since $\phi \notin Val_{K,n',R}$, there is a n'-euclidean and n'-transitive reflexive model $\mathcal{M}$ and a world w such that $\mathcal{M}, w \nVdash \phi$. By the Transfer Theorem, the n'-transitivity and n'-euclideanness of the accessibility relation on $\mathcal{M}$ guarantee that $\mathcal{M}, w \nVdash_{\mathrm{TS}} \phi\ [n]$. Since $\mathcal{M}$ is reflexive we have again a contradiction.

□

Completeness for MTS

Finally, we give an axiomatization for multi-agents token semantics with two agents, numbered 1 and 2, each agent being endowed with n tokens. Generalizations to a greater number of agents and to agents endowed with different number of tokens would be straightforward.

We use i and $i_1, ..., i_m$ as variables ranging over the set $\{1, 2\}$ of agents. If i takes the value 1 (resp. 2), $-i$ shall stand for a 2 (resp. for a 1).

The guiding intuition for MTS-completeness is the following: our axioms will have to say that what is i-possible after n $\Diamond_i$-modalities have already been considered is exactly what was i-possible before the last $\Diamond_i$ was considered. Along the way, some $\Diamond_{-i}$-modalities might appear: this is the specific complexity introduced by the multi-agent setting and it explains why taking the TS axioms $4.n'$ and $5.n'$ would not yield a complete axiomatization for MTS. What is needed, therefore, is to modify the previous axioms $4.n'$ and $5.n'$ so as to take into account the extra diamonds of type $-i$ that might be sandwiched between the relevant modalities of type i.

We note $\overrightarrow{\Diamond_{-i}}$ (resp. $\overrightarrow{\Box_{-i}}$) a possibly empty sequence of $\Diamond_{-i}$ (resp. $\Box_{-i}$) modalities. Here is axiom $4.nn'$, which is of course the multi-agent version of $4.n'$:

$$(4.nn') \quad (\neg p_1 \wedge \Diamond_{i_1}(p_1 \wedge \neg p_2 \wedge \Diamond_{i_2}(p_2 \wedge \ldots \wedge \Diamond_{i_m}(p_m \wedge \Diamond_i \overrightarrow{\Diamond_{-i}} \Diamond_i r) \ldots))) \rightarrow \Diamond_{i_1}(p_1 \wedge \Diamond_{i_1}(p_2 \wedge \ldots \wedge \Diamond_{i_m}(p_m \wedge \Diamond_i r) \ldots))$$

with the following conditions:

- the sequence $\Diamond_{i_1}, ..., \Diamond_{i_m}$ contains exactly $n-1$ occurrences of $\Diamond_i$,
- the total number of $\Diamond_{-i}$ modalities among the sequences $\Diamond_{i_1}, ..., \Diamond_{i_m}$ and $\overrightarrow{\Diamond_{-i}}$ is less than n.

It is easy to check that instances of $4.n'$ correspond to those instances of $4.nn'$ with no $-i$-modality intertwined.

Similarly, we label $5.nn'$ the multi-agent version of $5.n'$

$$(5.nn') \quad (\neg p_1 \wedge \Diamond_{i_1}(p_1 \wedge \neg p_2 \wedge \Diamond_{i_2}(p_2 \wedge \ldots \wedge \Diamond_{i_m}(p_m \wedge \Diamond r) \ldots))) \rightarrow \Diamond_{i_1}(p_1 \wedge \Diamond_{i_2}(p_2 \wedge \ldots \wedge \Diamond_{i_m}(p_m \wedge \Box \overrightarrow{\Box_{-i}} \Diamond_i r) \ldots))$$

with analogous conditions:

- the sequence $\Diamond_{i_1}, ..., \Diamond_{i_m}$ contains exactly $n-1$ occurrences of $\Diamond_i$,
- the total number of $\Diamond_{-i}$ or $\Box_{-i}$ modalities among the sequences $\Diamond_{i_1}, ..., \Diamond_{i_m}$ and $\overrightarrow{\Box_{-i}}$ is less than n.

Again, it is easy to check that instances of $5.n'$ correspond to those instances of $5.nn'$ with no $-i$-modality intertwined.

We can now state the completeness theorems generalizing Theorems 4 and 5:

Theorem 8. $\mathrm{K}(4.nn')(5.nn')$ *is sound and complete with respect to* $\mathrm{MTS}(n, n)$.

Theorem 9. $\mathrm{KT}(4.nn')(5.nn')$ *is sound and complete with respect to* $\mathrm{MTS}(n, n)$ *over the class of reflexive frames.*

As we said at the beginning of this appendix, the proof for Theorems 8 and 9 follow the same patterns as in the single agent case. For the sake of brevity, we shall therefore skip most of the details and focus on the following key ingredients in the proofs:

- the characterization of the properties of the accessibility relation which correspond in standard Kripke semantics to our axioms $4.nn'$ and $5.nn'$,
- the Transfer Theorem on models satisfying those properties,
- the definition of unpacked models that allows one, via the Transfer Theorem, to import completeness from standard Kripke semantics to token semantics.

The property of the accessibility relation corresponding to $(4.nn')$ is another variation on transitivity. We consider sequences $(w_0, i_0), ..., (w_k, i_k)$ of worlds indexed by an agent such that for all $j < k$, $w_j R_{i_{j+1}} w_{j+1}$. As before, we are concerned with repetition-free sequences, that is sequences such that $w_j \neq w_{j+1}$. Let $(w_0, i_0), ..., (w_m, i_m)$ be the initial fragment obtained by cutting off the sequence from the last pair with i as agent index on. The property behind $(4.nn')$ roughly says that transitivity holds at all worlds which are the endpoints of these $(w_0, i_0), ..., (w_m, i_m)$. We call this property nn'-transitivity. Here is the precise first-order clause for it:

nn'-transitivity: $\forall w, w', w'' \forall y_1, ..., y_l \, (\exists x_1, ..., x_{m+1} [x_1 R_{i_1} x_2 \wedge ... \wedge x_m R_{i_m} x_{m+1} \wedge x_1 \neq x_2 \wedge ... \wedge x_m \neq x_{m+1} \wedge x_{m+1} = w] \rightarrow (w R_i w' \wedge w' R_{-i} y_1 ... \wedge y_{l-1} R_{-i} y_l \wedge y_l R_i w'' \rightarrow w R_i w''))$

with conditions that are strongly reminiscent of what we had for the modal axioms:

- the sequence $R_{i_1}, ..., R_{i_m}$ contains exactly $n - 1$ occurrences of R_i,
- the total number of R_{-i} among the sequences $R_{i_1}, ..., R_{i_m}$ and the final R_{-i} is less than n,
- the sequence $y_1, ..., y_l$ might be empty.

It is easy to check that we get n'-transitivity as a particular case when no relations R_{-i} come into play.

Similarly, the property of the accessibility relation which corresponds to $(5.nn')$ is another variation on euclideanness. We call this property nn'-euclideanness:

nn'-euclideanness: $\forall w, w', w'' \forall y_1, ..., y_l \quad (\exists x_1, ..., x_{m+1} [x_1 R_{i_1} x_2 \wedge ... \wedge x_m R_{i_m} x_{m+1} \wedge x_1 \neq x_2 \wedge ... \wedge x_m \neq x_{m+1} \wedge x_{m+1} = w] \rightarrow (w R_i w' \wedge w R_i w'' \wedge w'' R_{-i} y_1 ... \wedge y_{l-1} R_{-i} y_l \rightarrow y_l R_i w'))$

- the sequence $R_{i_1}, ..., R_{i_m}$ contains exactly $n - 1$ occurrences of R_i,
- the total number of R_{-i} among the sequences $R_{i_1}, ..., R_{i_m}$ and the final R_{-i} is less than n,
- the sequence $y_1, ..., y_l$ might be empty.

Using Kripke semantics, the following definability results closely mirror Facts 5 and 6 :

Fact 7. $(4.nn')$ *defines the class of* nn'*-transitive frames.*

Fact 8. $(5.nn')$ *defines the class of* nn'*-euclidean frames.*

As Transfer Theorem in the multi-agent setting, we get the following:

Theorem 10. *Let* $\mathcal{M}$ *be an* nn'*-transitive and* nn'*-euclidean model, for every pair* (w, i) *with* $i \in \{0, 1, 2\}$*, for every sequence of such pairs* q *such that* $c_1(q(w, i)) \leq n$ *and* $c_2(q(w, i)) \leq n$*, and for every formula* ϕ*, we have that:*

$$\mathcal{M}, q(w,i) \vDash_{\mathrm{MTS}} \phi\ [n - c_1(q(w,i)), n - c_2(q(w,i))]$$
iff
$$\mathcal{M}, w \vDash \phi.$$

Since the possibility of going back and forth between Kripke semantics and token semantics is the heart of our completeness proofs, we shall prove Theorem 10. Note however that, again, the idea is exactly the same as for the proof of the original Transfer Theorem:

Proof. By induction on ϕ:

- If ϕ is atomic, the definitions of $\vDash$ and $\vDash_{\mathrm{MTS}}$ straightforwardly agree.
- Boolean cases are immediate as well.
- Let ϕ be of the form $\Box_1\psi$. We want to show that $\mathcal{M}, q(w,i) \vDash_{\mathrm{MTS}} \Box_1\psi\ [n-c_1(q(w,i)), n-c_2(q(w,i))]$ iff $\mathcal{M}, w \vDash \Box\psi$. We distinguish two cases:
  - Case 1: $c_1(q(w,i)) < n$.
    - $\mathcal{M}, q(w,i) \vDash_{\mathrm{MTS}} \Box\psi\ [n - c_1(q(w,i)), n - c_2(q(w,i))]$
    - iff for all w' such that wR_1w', $\mathcal{M}, q(w,i)(w',k) \vDash_{\mathrm{TS}} \psi\ [n - c_1(q(w,i)) - k, n - c_2(q(w,i))]$, where $k = 1$ if $w \neq w'$ and $k = 0$ if $w = w'$ (by definition of the satisfaction clause for $\Box_1$ in $\vDash_{\mathrm{MTS}}$),
    - iff for all w' such that wR_1w', $\mathcal{M}, w' \vDash \psi$ by induction hypothesis, since $c_1(q(w,i)) + k = c_1(q(w,i)(w',k))$ and $c_2(q(w,i)) = c_2(q(w,i)(w',k))$ as required,
    - iff $\mathcal{M}, w \vDash \Box_1\psi$ (by definition of the satisfaction clause for $\Box$ in $\vDash$).
  - Case 2: $c_1(q(w,i)) = n$. In this case, q is non-empty and contains at leat one pair of the form $(u,1)$, since some tokens of type 1 have been spent.
    - $\mathcal{M}, q(w,i) \vDash_{\mathrm{MTS}} \Box_1\psi\ [n - c_1(q(w,i)), n - c_2(q(w,i))]$,
    - iff $\mathcal{M}, q' \vDash_{\mathrm{MTS}} \Box\psi\ [1, n-c_2(q(w,i))+r]$, where q' is the longest initial segment of $q(w,i)$ such that there is a pair $(u,1)$ which belongs to $q(w,i)$ but not to q' and r is equal to the number of pairs of the form $(v,2)$ which belong to $q(w,i)$ but not to q'
    - iff for all w' such that zR_1w', where z is the last world in the sequence q', $\mathcal{M}, q'(w',k) \vDash_{\mathrm{MTS}} \psi\ [1 - k, n - c_2(q(w,i)) + r]$ where $k = 1$ if $w' \neq z$ and $k = 0$ if $w' = z$ (by definition of the satisfaction clause for $\Box$ in $\vDash_{\mathrm{MTS}}$)
    - iff for all w' such that zR_1w', $\mathcal{M}, w' \vDash \psi$ (by induction hypothesis: since $1-k = c_1(q'(w',k))$ and $c_2(q(w,i)) - r = c_2(q'(w',k))$,
    - iff for all w' such that wR_1w', $\mathcal{M}, w' \vDash \psi$ (by nn'-transitivity and nn'-euclideanness).
    - iff $\mathcal{M}, w \vDash \Box\psi$

□

To complete the picture, here is a generalization of unpacked models. Let W be a set of worlds, we note $F_n(W)$ the set of finite sequences of pairs (w, i) of length at most $2n$ where $w \in W$ and $i \in \{0, 1, 2\}$. The idea is to take as worlds in the unpacked models these sequence of pairs of worlds that are used in MTS:[26]

Definition 10 (nn-unpacked model)**.** *Let $\mathcal{M} = \langle W, R_1, , R_2, V\rangle$ be a Kripke model and $n \neq 0$. The nn-unpacking of $\mathcal{M}$ is the Kripke structure $\langle F_n(W), R'_1, R'_2 V'\rangle$ with*

- $qR'_i q'$ *iff:*
  - $c_i(q) < n$, (w, j) *is the tail of* q, q' *is* $q(v, i)$ *with* $wR_i v$ *and* $w \neq v$,
  - *or* $c_i(q) = n$, q' *is* $q''(w, j)(v, k)$ *with* $q''(w, j)$ *the longest initial segment of* q *such that there is a pair* (u, i) *which belongs to* q *but not to* $q''(w, j)$ *and* $wR_i v$ *with* $k = 0$ *if* $w = v$ *and* $k = i$ *if not.*
  - *or* $q = q'$ *and* (w, j)*, the tail of* q *is such that* $wR_i w$.
- $q \in V'(p)$ *iff* (w, j) *is the tail of* q *and* $w \in V(p)$.

Note that the nn'-unpacking of a reflexive model yields a reflexive model, as required for the proof of Theorem 9. Furthermore, one could check that a 'multi-agent' Unpacking Theorem can be proved for our notion of nn'-unpacking (either by introduction a suitable notion of bisimulation or by direct induction): the unpacked models are nn'-transitive and nn'-euclidean, and a formula is true according to Kripke semantics at $(w, 0)$ in the unpacked model iff it is true according to $\mathrm{MTS}(n, n)$ at w in the original model.

Now we have got all that is needed in order to get the multi-agent version of Lemma 3, and from there to get the proofs of Theorems 8 and 9. This final task is left to those of our readers eager to manipulate tokens.

Denis Bonnay
IHPST
Département d'Etudes Cognitives
Ecole Normale Supérieure - Paris
denis.bonnay@ens.fr

Paul Egré
Institut Jean-Nicod
EHESS/ENS/CNRS
Département d'Etudes Cognitives
Ecole Normale Supérieure - Paris
paulegre@gmail.com

[26] Since we are now keeping track of the whole history of moves, we do not need to encapsulate in the model the number of tokens that have been spent (that number can be recovered from the histories).