Interactive Models of Closure and Introspection

1 INTRODUCTION

Logical models of knowledge can, even when confined to the single-agent perspective, differ in many ways. One way to look at this diversity suggests that there doesn't have to be a single epistemic logic. We can simply choose a logic relative to its intended context of application (see e.g. Halpern [1996: 485]). From a philosophical perspective, however, there are at least two features of the logic of "knows" that are a genuine topic of disagreement, rather than a source of mere diversity or pluralism: closure and introspection. By closure, I mean here the fact that knowledge is closed under known implication, knowing that p puts one in a position to come to know what one knows to be implied by p. By introspection, I shall (primarily) refer to the principle of *positive* introspection which stipulates that knowing puts one in a position to know that one knows.[1] The seemingly innocuous principle of closure was most famously challenged by Dretske [1970] and Nozick [1981], who believed that giving up closure—and therefore the intuitively plausible suggestion that knowledge can safely be extended by deduction—was the best way to defeat the sceptic. Positive introspection, by contrast, was initially defended in Hintikka's *Knowledge and Belief* [1962], widely scrutinized in the years following its publication (Danto [1967], Hilpinen [1970], and Lemmon [1967]), and systematically challenged (if not conclusively rejected) by Williamson [2000: IV–V].

A feature of these two principles that is not always sufficiently acknowledged, is that that there isn't a unique cognitive capability that corresponds to each of these principles. To be precise, closure not only requires the ability to competently derive the consequences of one's knowledge, it also presupposes that the outcomes of that process qualify as knowledge. When closure is defined as (merely) being in a position to know that q (by actually deducing it from p and $p \rightarrow q$) whenever one knows that p and that p

[1] The main reason for focusing on positive introspection, is that negative introspection is more easily dismissed. To be precise, given the prior assumptions that knowledge unrestrictedly implies belief and that belief is consistent, negative introspection warrants the following thesis: $\mathsf{BK}\phi \rightarrow \mathsf{K}\phi$. In other words, strong belief (i.e. believing that one knows) is sufficient for knowledge. This result, which is sometimes rephrased by saying that strong beliefs can only fail to count as knowledge by being inconsistent, is often considered a knock-down argument against negative introspection. Alternative replies are possible, but I think this is still a good enough reason to focus on the disagreement concerning positive introspection.

implies q, the focus is presumably on the latter aspect. I shall henceforth refer to the latter as ***closure proper***. From the standpoint of epistemic logic, however, closure also (if not exclusively) refers to the sheer ability to apply *modus ponens* (I'm deliberately ignoring here other forms of closure such as closure under conjunction). I shall henceforth refer to this second aspect as ***deductive omniscience***. Likewise, positive introspection requires having access to one's knowledge, but also depends on the reliability of that access. One way to characterise the relevant contrast, is to think of the former as merely being able to find out what one knows, and of the latter as the reliability (in a knowledge conferring way) of our means to find out about one's knowledge. At least as a first approximation, this captures what is at stake with closure as well as with introspection.

There are, however, further difficulties with regard to the distinction drawn above. One might, at first, be tempted to reframe that distinction in terms of internal and external properties of closure and introspection, or even more broadly as properties of the knowing agent (cognitive capacities) versus properties of the concept of knowledge (cognitive standards). Neither of these is adequate. Presumably, we could think of deductive omniscience as a cognitive capacity, and even think of access to one's knowledge as such a capacity, but that doesn't yet make closure proper or reliable access to one's knowledge a cognitive standard we have to meet. Since we have to check if these standards can be met to find out whether knowledge satisfies introspection and closure proper, the cognitive limitations of the agent cannot as such be separated from the properties of the concept of knowledge. Conversely, even when we treat deductive omniscience as a property of the knowing agent, this does not yet rule out that that property reveals more about our modelling assumptions than about the agent's actual computational resources. This is best illustrated relative to Williamsons rejection of positive introspection, and in particular relative to his pointing out that knowing that one knows that p, does not merely depend on our epistemic position towards one knowledge (the proposition that one knows that p), but that it also depends on our epistemic position towards the known proposition itself. By characterising the second aspect of positive introspection as "reliable access," the latter dependence is obscured.

The distinction I've outlined therefore becomes: (a) deductive omniscience does not imply closure proper, but closure proper does not imply deductive omniscience either—deductive inference might produce knowledge even if an agent cannot competently (or instantly) carry out all such deductions; (b) finding out about one's knowledge does not have to lead to knowledge about one's knowledge, but even if finding out always leads to higher-order knowledge (because, for instance, one is in a sufficiently strong epistemic position with regard to the known proposition), that does not have to mean that one can always find out (we might still lack the required computational resources). Once it is properly reformulated such as

to avoid potential confusion with more commonly used distinctions, the two-sided nature of closure and introspection may have become extremely generic. Even then, it reveals an important difference in focus between formally oriented epistemology, and contemporary epistemic logic. The difference in focus is of course most obvious with respect to the contrast between deductive omniscience and closure proper, but I believe it matters as much for introspection-related issues.

Starting from this insight into the two-sided nature of closure and introspection, it is natural to ask how a formal model (i.e. a modal epistemic logic) could be used to reveal this nature in a more appropriate way.

♦ ♦ ♦

But isn't this is all old news? After all, it seems that we can already capture the relevant contrast in a combined logic for knowledge and belief. Namely, if closure as well introspection are only valid to the extent that the relevant agents have the sheer ability to find out, the former can be expressed as the belief that q whenever one knows that p and knows that p implies q, and the latter as the belief that one knows that p whenever one knows that p ($\mathsf{K}p \rightarrow \mathsf{BK}\phi$). This, one could argue, adequately formalises the intended contrast: the former principle already follows in any system where (i) knowledge implies belief, and (ii) belief is itself closed under believed implication; the latter simply coincides with the thesis that knowledge implies strong belief.

A specific reason to doubt the aptness of the true belief characterisation of deductive omniscience bears on its reliance on an unrestricted entailment principle; knowing ϕ must imply belief in ϕ for all ϕ of the relevant language (this may be contrasted with the restricted entailment principles used in Halpern [1996] and especially in Voorbraak [1990]). This presumably blocks the possibility of obtaining a general characterisation of closure and introspection as "the mere ability to find out" in terms of beliefs or true beliefs. A more general reason to question the belief-detour, is that it only works on the two separate assumptions that knowledge presupposes strong belief, and that belief is deductively closed. Each of these assumptions can legitimately be challenged. To presuppose that strong belief is necessary for knowledge, is roughly to presuppose that knowledge requires the knowing agent to be confident in his belief, and this is presumably not a feature we readily associate with the prior rejection of positive introspection. To presuppose that belief is deductively closed, suggests in the present context that closure is less problematic for belief than it is for knowledge. The vast literature on the role of logic as a norm for belief shows, however, that the deductive closure of belief is no less controversial.

In view of the above, it should be clear that whereas the detour via belief is useful to hint at the relevant contrast, it does not directly lead to a satisfactory characterisation. In this paper, I propose to develop an alternative framework to deal with this contrast, and once this is given, I

hope that the reader will be able to see why it succeeds where the belief characterisation fails.

♦ ♦ ♦

The alternative approach I have in mind is based on the use of different types of group-knowledge as different forms of single agent knowledge. What I suggest, is that we should model individual agents as groups of agents (components, if you want), and that different ways in which an individual agent could know might then be taken to correspond to different ways in which knowledge could be present in a group of agents. As should become clear throughout my exposition, different forms of group-knowledge provide exactly the kind of distinctions we need to disentangle the different aspects of closure and introspection. Even at this early point of my exposition, one might worry that a model of knowledge for individual agents that is itself based on knowledge for groups cannot but lead to a vicious kind of circularity, for the latter would (on pain of regress) obviously have to refer back to features of single agent knowledge. This worry is ill-founded. The kind of knowledge for individual agents that is modelled as a form of group-knowledge does not (and arguably should not) coincide with the kind of knowledge we ascribe to the components. One might then still worry that by treating individual knowledge as a form of group-knowledge one simply reverses the order of explanation. Perhaps this second worry isn't ill-founded as such, but I believe it can be dismissed as well. As this requires a more substantial argument, I'll leave it aside for now and come back to it at a more appropriate moment in the next section.

This paper is structured as follows. In the next two sections I describe the logical properties of component knowledge (§2) and of the different forms of group knowledge (§3). In Section 4 I show how the hierarchy of different types of group-knowledge gives rise to a related hierarchy of types of individual knowledge, and in Section 5 I describe and analyse the forms of component-interaction that can lead to deductively closed and introspective knowledge. A comparison with Hintikka's original argument for positive introsection (§7) is then used to illustrate the theoretical virtues of the proposed model.

2 COMPONENT KNOWLEDGE

As I prefer to present the relevant logical systems from a model-theoretic perspective, I cannot directly start with the relevant forms of group-knowledge that can be used to model different forms of individual knowledge. The starting point, therefore, is the logic we use to model the knowledge of the components.

Where C is the set of components, we have a modal operator $[c]$ for every c in C. Thus, we say that $[c]\phi$ is true at a state w iff wR_cw' implies

that ϕ is true at w'. If wR_cw' is read as saying that w' is an epistemic alternative to w for c, and that its negation $\neg wR_cw'$ means that at w, c can exclude w', we can say that c knows that ϕ at w iff c can exclude at w all states where ϕ is not true, or, equivalently, iff ϕ is true at all epistemic alternatives to w for c.

Traditionally, epistemic operators are presumed to satisfy some further conditions; in particular knowledge is supposed to be factive. In this case, however, I shall make the much stronger assumption that the knowledge of components is also fully introspective; knowing for the components is **S5**-knowing. One way to model this constraint proceeds by defining a new modal operator $[c^*]$ (again, one for each c in C) such that, where R_c^* is the transitive and symmetric closure of R_c, $[c^*]\phi$ is true at a state w iff wR_c^*w' implies that ϕ is true at w'. Given that, it is easily verified that whenever R_c is reflexive, R_c^* will be a universal accessibility relation, which in turn suffices to make $[c^*]$ an **S5** box-operator. For present purposes, I do not need to distinguish between introspective and non-introspective components, and I shall therefore ignore the difference between $[c]$ and $[c^*]$. This warrants the stipulation that for each $c \in C$, we say that c knows that ϕ ($\mathsf{K}_c\phi$) iff $[c^*]\phi$ is true. It does, therefore, satisfy the axioms listed in Table 1.

LABEL	AXIOM	FRAME CONDITION
K	$\mathsf{K}(p \rightarrow q) \rightarrow (\mathsf{K}p \rightarrow \mathsf{K}q)$	/
T	$\mathsf{K}p \rightarrow p$	Reflexive
4	$\mathsf{K}p \rightarrow \mathsf{K}\mathsf{K}p$	Transitive
5	$\neg\mathsf{K}p \rightarrow \mathsf{K}\neg\mathsf{K}p$	Euclidean

TABLE 1: Modal-Epistemic Axioms

Before we move on, it is advisable to be more explicit about the impact of modelling component-knowledge as **S5**-knowledge. First, it is a choice that does not have a substantial effect on the forms of group-knowledge that can be defined for these kinds of components. As soon as all R_c are reflexive, there will be forms of group-knowledge that are fully (i.e. positively and negatively) introspective and others that aren't; and even if some R_c are not reflexive, there will be forms of group-knowledge that are positively introspective (and, again, others that aren't). Second, the best way in which the present modelling-option can be thought of, is in terms of what components can communicate. Whenever a first component knows that ϕ, we presume this is knowledge that can be passed on to other components in such a way that other components not only come to know that ϕ, but also come to know that the first component knows that ϕ. Yet, if that is the case, then it would be quite odd to assume that gaining higher-order knowledge of other components' knowledge is easier than for a single component to be introspective (i.e. one's knowledge could—once communicated—be transparent to others, but not to oneself). As a result, the identification of

component-knowledge as **S5**-knowledge is implied by the fact that we are only interested in component-knowledge that (a) can be shared (see e.g. van Benthem [2006: 57] on the assumption that knowledge requires the ability to inform others), and (b) can be shared in such a way that it can lead to higher-order (inter-component) knowledge.

Yet, if we want the interaction between components to work in the just described way, it isn't enough to make all these components fully introspective, we also need their interaction to be fully reliable. That is, the way we model communication (most likely in the form of updates directed at some or all agents) has to be such that learning that some component knows invariably leads to higher-order knowledge (rather than some weaker attitude) about the knowledge of that component. What this amounts to, is that the reliable access each component has to its own knowledge is extended to interactions: provided that knowledge is shared (and we've stipulated that all knowledge can be shared), this process of sharing one's knowledge induces knowledge (see Hintikka [1962: 4.1–2] for an early treatment and Hendricks [2006: 148–50] where the transmissibility of knowledge is related to sameness of goals, methods and standards) as well as higher-order (inter-component) knowledge.

♦ ♦ ♦

But just how strong is the assumption that all knowledge can be shared in such a way that it leads to knowledge and even higher-order knowledge?

Before we answer this question, it should be emphasised that assumptions of this kind can be understood in different ways. First, it can be considered as a way to raise the standard for knowledge: something qualifies as knowledge only if it does satisfy these conditions. On this first interpretation, the assumption that only what can be shared counts as knowledge can open the door to scepticism via the denial that such high demands can ever be met. Second, it can be considered as a way to lower the standard for knowledge: by stipulating that all communication of existing knowledge leads to higher-order knowledge, knowledge by testimony becomes quite cheap. On this second interpretation it could be denied that it is really knowledge that is being modelled. Since we're still only concerned with component-knowledge—which is only knowledge by name—the dilemma between knowledge on the cheap or no knowledge at all should perhaps not pose a problem. The following three points explain why.

To begin with, it can hardly be denied that (when it comes to knowledge proper) the assumptions we have to make when we model knowledge as fully transferrable **S5**-knowledge are exceedingly strong. Even if one sticks to the traditional view that knowledge is introspective and thus can be shared, it is still unnatural to presuppose that sharing one's knowledge could be invariably successful. By contrast, the very same assumption is one of the corner-stones of public-announcement-logic and other forms of dynamic epistemic logic (see e.g. Baltag & Moss [2004] and Plaza

[2007]). More specifically, given the focus of these systems on how knowledge changes through communication, there is no real point in modelling knowledge that cannot be shared; and, if one is interested in what we can learn from reasoning about the knowledge and ignorance of others (as exemplified in, for instance, the muddy children puzzle), one should only focus on those cases where higher-order interpersonal knowledge can be obtained. In sum, our model of component-knowledge and interaction stands for an intuitively strong assumption about the nature of knowledge, but it is also a common—and perhaps even indispensable—modelling option.

Can we also reconcile both sides? I think we can, and the crucial insight to do so bears on the already mentioned fact that component-knowledge is not to be used to model real knowledge, it is just a model of a lower-level state that is factive in the same way as knowledge is. Real knowledge is to be modelled by means of the different forms of group-knowledge that can be defined on top of component-knowledge. The good news is that identifying component-knowledge as **S5**-knowledge does not collapse deductively closed and non-deductively closed forms of group-knowledge, nor does it collapse introspective with non-introspective forms of group-knowledge. More exactly (and provisionally ignoring further complications), it only warrants that weaker forms of group-knowledge can be upgraded to stronger forms of group-knowledge by means of communication. And, since communication of the components is here used as a way to model the reasoning of the system as a whole, the assumption that these components can infallibly share their knowledge is essentially a means to ensure that the system as a whole can reason from whatever it knows and independently of how it is known. Thus, as a provisional conclusion, I would suggest that we're warranted in idealising component knowledge and interaction because that allows us to entirely focus on how different patterns of interaction between components can have an impact on closure and introspection without having the properties of component-knowledge themselves interfere.

The latter, however, again raises the question of whether we're not reversing the order of explanation by not only using models of group-knowledge to explain typical single-agent features of knowledge, but also (given the just described idealisation of component-knowlege) by requiring these models to do all the explanatory work.

A first way to alleviate these worries is to point out that all I want to do is to model different forms of individual knowledge in analogy with different types of group-knowledge. For my proposal to work, it does not have to suppose that individual knowledge is really a kind of group-knowledge—it is not. All I have to presuppose, is that the formal resources used to discriminate between different ways in which knowledge might be present in a group of agents can also be used to discriminate between different ways or senses in which an individual agent might be said to know. So presented, this only requires me to endorse the weaker claim that the kind of differences that are relevant to groups are exactly the kind of differences

that are relevant to individual agents. Whether this holds, is ultimately independent of the question of whether my proposal gets things backward, and so from this perspective, questioning the right order of explanation is irrelevant.

A second way to deal with these worries is more substantial, and is based on the observation that the presumed primacy of single-agent knowledge only signals a traditional bias towards individualistic accounts of epistemology. If we assume that the nature of knowledge is at least partly determined by how agents interact (a common assumption in interactive and social epistemology), then it immediately follows that there is no unique order of explanation which goes from individual to group-knowledge; explanatory relations can go both ways. The particular way in which I frame knowledge is surely sympathetic to the anti-individualistic point of view, but accepting a model of individual knowledge that is based on existing forms of group knowledge does not require the adherence to anti-individualism. Rather, the point I want to emphasise is that the decision about what counts as a single or individual agent is ultimately a modelling option. When we decide that a is an individual agent, we have to consider every output produced by that agent as the result obtained from computing the input it received from other agents. By contrast, when we consider that same agent as a group of agents or components, we can start to consider the same outputs as the result of communication between these components. The same point is made by Abramsky by claiming that "information is conserved in total system, but can increase relative to a subsystem" (Abramsky [2008: 484]). It is this insight that what looks as computation from the outside (i.e. seeing a as an individual agent) can be modelled as communication from the inside (i.e. seeing a as a group of components) which motivates the present proposal. To model agents as groups is just a means to switch to a lower level of abstraction.[2] The role of group-knowledge being clarified, we can now go back to our initial concern: the more refined picture of closure and introspection itself.

The suggestion that weaker forms of knowledge (understood relative to a group of components) can be upgraded to stronger forms of knowledge is instrumental in understanding how deductively closed as well as higher-order knowledge can arise. On a naive interpretation of the principles of epistemic logic, it is so that, say, positive introspection means that an agent cannot know unless he also knows that he knows. This is a static way of understanding such principles, and it coincides more or less with how we should understand our model of component-knowledge. Another interpretation of the same principle would then be that when an agent knows,

[2]This point of view can be compared to an idea voiced in van Benthem [2008: 185], where he draws the attention to the fact that differences in structural rules for what he calls different reasoning-styles might be mere symptoms of more basic underlying phenomena. In that sense, the choice to model single agent knowledge after forms of group knowledge is a means to focus on the underlying phenomena rather than on the surface symptoms we associate with closure and introspection.

she doesn't require any external input to learn that she knows. When it comes to knowledge relative to a group of components, both the former static interpretation as well as the latter dynamic interpretation can be used. That is, stronger forms of group-knowledge like common knowledge lead to introspective knowledge in the static sense, whereas weaker, non-introspective forms of group-knowledge still can without external input be upgraded to stronger, introspective forms of group-knowledge (similar considerations apply for deductive closure). What this aims at, is that dynamic forms of introspection are "valid" if, from an external perspective, higher-order knowledge can be achieved by sheer deductive reasoning. But we've already seen that what looks as reasoning from the outside (no information-change relative to the whole system), looks like communication from the inside (information-change relative to sub-systems). As a result, we can now understand the dynamic version of introspection as the existence of a communication-protocol which guarantees that one form of group-knowledge can be upgraded to a stronger one.

What now remains to be done, is to review the different forms of group-knowledge that allow us to discriminate between deductively closed and non-deductively closed knowledge, and between introspective and non-introspective knowledge, and then to find out what the relevant upgrade-protocols should be.

3 KNOWLEDGE IN A GROUP

Where G is a finite subset of C, we say that G is a group of agents (components). In any such group, knowledge can be present in different guises. For any of these, a corresponding notion of group-knowledge can be defined. More importantly, assuming that these groups contain at least two components, then all of these notions are provably non-equivalent, and give rise to a hierarchy of forms of group-knowledge.[3] Below, I describe each of these forms, and highlight the respective epistemic principles they satisfy.

Distributed Knowledge. The weakest kind of group-knowledge is standardly called distributed knowledge, henceforth D-knowledge (formally, just D). Semantically, it is obtained by stipulating that ϕ is D-known in a group G iff each non ϕ world is at least excluded by some member of G. In a more intuitive sense, distributed knowledge can be identified with the knowledge that can be obtained by somehow pooling together the knowledge held by all agents in a group. Yet, while it is natural to assume that distributed knowledge can only be valuable if it can be made explicit by actually pooling together the agents' knowledge, there are models where distributed knowledge does not satisfy this condition. Following van der

[3]Note that I use the term "group-knowledge" to refer to all ways in which knowledge can be present in a group. This practice diverges from the one in van der Hoek, van Linder & Meyer [1999], where the same term refers to the weakest kind of such knowledge.

Hoek, van Linder & Meyer [1999] and Roelofsen [2006]), we say that the formal and the intuitive characterisation of distributed knowledge[4] coincide iff the *principle of full communication* is satisfied.

This is something we shall have to come back to, but now we first have to focus on the formal properties and the interpretation of distributed knowledge. Its formal properties are easily summarised, for distributed knowledge is a form of **S5**-knowledge; it is deductively (as well as logically) closed, and fully introspective. Yet, unlike component-knowledge (which is also a form of **S5**-knowledge), distributed knowledge cannot readily be shared, for there does not have to be an individual component which actually holds what is D-known (and we may assume that a group can only produce an output if some component can produce that output). How, then, should we interpret the type of knowledge that is equivalent to distributed knowledge among all the components? The obvious answer is also the best one; it is just a form of implicit knowledge. Even if the implicit-explicit contrast isn't entirely adequate to think about knowledge (Harman [1986: 13–14], Stalnaker [1991]), it is good enough for present purposes. Not only does it coincide with how we understand distributed knowledge in groups of agents, but its formal properties make it also sufficiently similar to how Levesque [1984] characterises the difference between implicit and explicit belief.

Someone Knows. The second kind of group-knowledge is usually referred to as "someone knows" and is the least *social* form of group-knowledge. Henceforth, we refer to this type of knowledge as S-knowledge. Semantically, it can be defined in such a way that ϕ is S-known in a group G iff there is a member of G who can exclude each non ϕ world. Of course, this is equivalent to saying that someone knows that ϕ whenever there is at least one member of the group who does. As a result, this form of group-knowledge might strike us as rather dull; it is just the disjunction of the corresponding knowledge ascriptions for each member of the group. In view of its formal properties, however, it turns out to be a prime example of explicit knowledge in a group. To see why, recall that when explicit knowledge is identified with knowledge that can be readily shared, and that knowledge available in a group can only be communicated if it is held by some member of that group, then the notion of "someone knows" is the weakest form of group-knowledge that qualifies as explicit knowledge. This intuitive point is reinforced by the fact that it is a form of group-knowledge that is not deductively closed;[5] again a property that is typically associated with explicit knowledge.

[4] Again, the terminology in van der Hoek, van Linder & Meyer [1999] does not coincide with the present one; there, distributed knowledge refers to those forms of group-knowledge (distributed knowledge in our terminology) that do satisfy the principle of full communication.

[5] It is, however, introspective and also closed under single-premise valid arguments, but both these properties are directly inherited from component-knowledge and therefore not primary properties of this form of group-knowledge.

Everybody Knows. The third kind is a genuinely social form of group-knowledge, as it only applies to cases where all members of a group know. Its semantic characterisation is this: ϕ is E-known (i.e. everybody knows) in a group G iff each member of G excludes all non ϕ worlds. Alternatively, it can also be defined as the conjunction of all $\mathsf{K}_{c_i}\phi$ for all c_i in G. The logical properties of E-knowledge are the exact mirror of the properties of S-knowledge: E-knowledge is deductively closed, but not introspective at all. In addition, the failure of introspection is a genuine property of this kind of group-knowledge; the logical features of component-knowledge do not have an impact here. Its being deductively closed is, by contrast, at least in part induced by the fact that component-knowledge is deductively closed as well.

This leaves us again with the question of what kind of knowledge may be equivalent to E-knowledge among all components. A first, only partial answer is that since E-knowledge implies S-knowledge, E-knowledge remains an explicit form of knowledge. The second part of the answer is harder. It requires us to make sense of a non-introspective form of knowledge that nevertheless implies an introspective form of knowledge. Right now, we do not yet have the conceptual resources to explain how S-knowledge and E-knowledge give rise to distinct types of explicit knowledge. We could of course emphasise that both give rise to knowledge that is explicitly stored in different ways, but this only says something about the components; it remains silent about how this difference allows us to model different kinds of knowledge. This issue will be dealt with as soon as the last kind of group-knowledge is introduced.

Common Knowledge. The fourth and final kind of group-knowledge is common knowledge (C-knowledge); the kind of knowledge that is usually assumed to be necessary for conventions and other agreements (see Lewis [1969]). Its semantic characterisation is more cumbersome than the previous ones, for ϕ is C-known cannot straightforwardly be defined as the ability of each agent to exclude some worlds. A more elaborate notion of exclusion is required. Where G is a group of agents, define a G^k alternative with the following inductive clauses: a world w is a G^1 alternative iff w is an epistemic alternative for some member of G; a world w is a G^{k+1} alternative iff at some G^k alternative the world w is an epistemic alternative for some member of G. Using this notion, we can now stipulate that ϕ is C-known at w iff for any finite k, no non-ϕ world is a G^k alternative. As is well-known, this definition implies that whenever ϕ is C-known, it also holds that ϕ is E-known, that it is E-known that it is E-known, and so on for any finite iteration of E-known. This means that C-knowledge or common knowledge is a form of group-knowledge that is fully transparent to each member of the group; there's no ignorance whatsoever with regard to the agreement reached by all members as no finite level of higher-order knowledge is missing.

There is more that could be said on the topic of common knowledge, but for now it is sufficient to focus on its basic logical properties. In short: common knowledge is again a form of **S5**-knowledge; it is deductively closed and fully introspective.[6] Consequently, common knowledge and distributed knowledge have exactly the same logical properties. Yet, they couldn't differ more as the latter deals with implicitly available knowledge whereas the former deals with knowledge that is explicitly available. This difference in interpretation can be used to explain why the weakest and the strongest form of group-knowledge may nevertheless obey the same logical principles. Namely, where knowledge is understood as something that is only implicit, closure and introspection become much weaker constraints than when knowledge is understood as being explicitly represented in one's mind.

The question of how to interpret the form of knowledge that is equivalent to common knowledge in a group of components can, due to the problems already raised with regard to the precise sense in which "everybody knows" models a kind of explicit knowledge, not yet be satisfactorily answered. All we may say, is that all C-knowledge, including that of any higher degree, is readily and explicitly available. Put differently (or slightly metaphorically), the common knowledge of the components is knowledge that is available to all components in a fully transparent way. As a result, it is knowledge that the group as a whole can invariably make available to others. This description is perhaps sufficiently suggestive to hint at the real strength of this form of knowledge, but does not yet make an interpretation available of the precise sense in which E-knowledge and C-knowledge differ from S-knowledge *qua* explicit forms of knowledge. Spelling out the full hierarchy of notions of knowledge based on (or, more accurately, modelled after) the different forms of group-knowledge described in the present section, means that we also have to individuate weaker and stronger senses of explicit knowledge.

4 A HIERARCHY OF KNOWLEDGE-TYPES

By defining types of group-knowledge, we have obtained a series of knowledge operators which warrant that ϕ is C-known implies that it is E-known which implies that it is S-know, and in its turn also implies that it is D-known. This series of implications is all we need to be able to talk of a genuine hierarchy of forms of group-knowledge (see Halpern & Moses [1990: 554]). By contrast, this is not yet enough to say that there is an analogous hierarchy for the notions of knowledge that we wish to model by means of different manifestations of group-knowledge for a set of components. So far, we have a decent idea of what makes the difference between implicit D-based knowledge, and explicit S-based knowledge, and also of

[6] As remarked before, negative introspection only holds for common knowledge; common belief does not have to be negatively introspective.

what makes the difference between non-introspective E-based knowledge, and introspective C-based knowledge, but still no reason to assume that both the contrast between closure for D-knowledge and non-closure for S-knowledge and between introspection for C-knowledge and the lack of introspection for E-knowledge can be understood from a single perspective. Indeed, to make it a real hierarchy we would not only have to show that it is possible to step up from D-based explicit knowledge to E-based explicit knowledge (i.e. showing that there is a protocol which ensures exactly that), but also that upgrading from the logically weaker to the logically stronger would mean stepping up from a weaker epistemic position to an effectively stronger one.

As a preliminary to an explanation of how we may tie the two halves of the hierarchy together, we first have to take a closer look at the explicit-implicit distinction that we already put in place. To begin with, whenever ϕ is D-known, but not S-known it has to be implicit. This first feature is independent of how component-knowledge is understood. Next, as soon as ϕ is S-known, at least one component knows that ϕ, and since we've stipulated that component-knowledge can always be shared, S-knowledge can be shared as well. This is sufficient for S-knowledge to be explicit, but it also reveals that the status of S-knowledge as an explicit form of knowledge is inherited from the (stipulated) status of component-knowledge as a form of explicit knowledge.[7] By the same token, since E-knowledge is considered explicit only because it implies S-knowledge, the status of E-knowledge as a form of explicit knowledge should as well be retraced to our previous decision to treat component-knowledge as a form of explicit knowledge.

The above considerations give us an important clue as to how we should understand the difference between S-knowledge and E-knowledge. Because the kind of explicitness they have in common is inherited from component-knowledge, their difference in explicitness should entirely reside in how they differ *qua* forms of group-knowledge. That is, (a) it should be a function of how knowledge is actually distributed among the different components, (b) it should explain why one is deductively closed but the other is not, and (c) it should be open to an interpretation as different forms of explicitness. Conditions (a) and (b) are easily met. When something is E-known, this knowledge is uniformly distributed among all components; when it is only S-known, it is not uniformly distributed. When a large number of things are S-known, this knowledge can be randomly distributed among all components, and it may then be the case that no individual component is able to compute the consequences of everything that is S-known. In other words, computing (in the narrow sense) may have to be preceded

[7] One might, here, object that S-knowledge primarily qualifies as a form of explicit knowledge in virtue of the failure of closure; a feature that is independent of component-knowledge *and* is traditionally associated with explicit forms of knowledge and belief. Yet, since the failure of closure is presumably a necessary condition for explicitness, it is not a sufficient condition and the reference to component-knowledge is therefore easily shown to be indispensable for the evaluation of S-knowledge as a form of explicit knowledge.

by reorganising the available information. By referring to S-knowledge as randomly distributed, I've already hinted at how condition (c) could be met as well.

Before I follow that trail, I should first get back to Stalnaker's critique of how the distinction between explicit and implicit knowledge is usually applied. What he objects to is that the distinction in question has the double task of accounting for, on the one hand, different ways in which information can be stored, and, on the other hand, whether that information is readily accessible or not. Yet, since search and retrieval are computational processes, explicit storage does not imply immediate access, and since some information that is only implicitly available may still be immediately accessible because it is easily deducible from what is both accessible and explicit, a single implicit-explicit contrast cannot account for both distinctions (Stalnaker [1991: 435]).

Unlike mainstream models of knowledge which incorporate a distinction between implicit and explicit knowledge, the hierarchy based on different forms of group-knowledge does allow for a double distinction. Intuitively, the distinction between D-knowledge and S-knowledge is well-suited to capture the distinction between information that is explicitly stored and information that is merely implicit in what is explicitly stored. Its adequacy for that task is immediate from the fact that one is deductively closed, and the other not. Whether the distinction between S-knowledge and E-knowledge is equally well-suited to capture the distinction between readily accessible and not so readily accessible explicitly stored information, essentially depends on the computational costs associated with the retrieval of information that is merely S-known. Whenever the number of components is large enough, this process is arguably sufficiently costly to consider information only known by one or even just a few components not readily accessible. Conversely, since E-knowledge reduces this otherwise costly retrieval procedure to the querying of a randomly chosen component, E-knowledge provides an adequate model of readily accessible explicitly stored knowledge.

To fully meet Stalnaker's objections against the single distinction with a double task, our model should allow for explicit, but not readily available knowledge and for readily available, but implicit knowledge. The former demand reduces to the possibility of S-knowledge that does not qualify as E-knowledge, which is equivalent to the fact that S-knowledge does not imply E-knowledge. The latter demand should, however, not be reduced to the possibility of E-knowledge that does not qualify as S-knowledge; this is not only impossible, but it is also based on a misunderstanding of what readily available, but nevertheless implicit knowledge would amount to. What is needed, is implicit knowledge that can easily be upgraded to readily accessible explicitly stored knowledge. In other words, it only requires D-knowledge that can easily be upgraded to E-knowledge. Whether this is a real possibility depends on the protocols that are effectively available, but

for present purposes it suffices to note that nothing precludes the existence of such a protocol.

Using a double distinction between how information is stored, and whether it is readily accessible, we are able to tie together the two halves of the hierarchy. To show that the hierarchy further complies with what we expect from these different forms of knowledge, it is instructive to consider how they interact. By this, I mean that we should review what follows from ϕ being known in one way, and ψ in another, but where both jointly imply that χ, say, by a single application of *modus ponens*. Most such interactions are straightforward, but the interaction between S-knowledge and E-knowledge is worth looking at in particular. Indeed, with ϕ, ψ, and χ as just described, we have it that when ϕ is S-known and ψ E-known, then χ is also S-known. At first, this looks like an undesirable property, for it tells us that for any two-premise argument, it suffices that one is readily accessible (i.e. E-known) for the conclusion to be explicitly stored as well (S-known). That is, computations can be carried out on premisses that are not readily available. To see that this outcome is unproblematic, one should take the following into account. When it is generalised to n-premise arguments, it is obvious that at most one premise can be merely S-known for the conclusion of that argument to be necessarily S-known as well (as such, the two-premise case is hardly stronger than single-premise closure for S-knowledge). In the end, all this interaction shows is that using information in computation and making information readily accessible are processes that do not have to occur in a fixed order, and there is nothing objectionable about that.

From a formal point of view, the proposed hierarchy surely meets the requirements implicit in Stalnaker's double distinction between explicitly stored and readily available knowledge. When it comes to the interpretation of the underlying machinery, one may still advance that at least for human agents there's no plausible sense in which information might become explicit in one part of the brain, but not in other parts.[8] The objection implied by this observation is that either there's no distinction between explicit information that is readily available and explicit information that isn't readily available, or that such a distinction cannot be captured by the difference between S-knowledge and E-knowledge. This objection can be avoided by explaining component-knowledge in a more neutral way; i.e. not as "one part of the brain". A sketch of such an account in given in Section 6.

5 UPGRADING, PROTOCOLS, AND KNOWABILITY

Now that the hierarchy of forms of knowledge modelled after different types of group-knowledge is in place, we're finally ready to tackle the issue of upgrading. We start with dynamic forms of closure, or: the pro-

[8] This consideration was raised by Luciano Floridi.

tocols needed for upgrading D-knowledge to S-knowledge. The protocols required for dynamic closure do in fact coincide with upgrading from a deductively closed to a *non*-deductively closed form of knowledge. But that's just fine, because making deductions is a way to make explicit what was only implicitly there—which explains why dynamic closure coincides with upgrading from D-knowledge to S-knowledge);[9] and because, as it is a dynamic process, not everything becomes instantly explicit—which explains why the outcome, namely S-knowledge, isn't itself deductively closed in the static sense.

Next up are the questions of what kind of protocols might be required for this, and whether these protocols are always successful. Intuitively, there doesn't seem to be a systematic or unified way to describe the kind of communication required for actually deriving what is already implicitly present within a group. Generally, we might say that all the premises required to deduce a certain conclusion need to be gathered in a single place, but that doesn't have to be a single component. Specifically, the only hard requirement is that the premises of each separate inference-step have to be available to a single component, but that is something that can be achieved in many ways. All components could for instance send all the information they hold to a single designated component (the so-called 'wise man' referred to in Halpern & Moses [1985], van der Hoek, van Linder & Meyer [1999]; the protocols in question are described in van Linder et al. [1994]),[10] but they could equally well send everything to everyone, or even set up a more complicated inference-network. For present purposes, the actual procedure to achieve this is largely irrelevant.

The more important issue is whether such protocols are available as well as invariably successful. Even if we assume that every component can pass on its knowledge to whatever other component, the problem of successfully upgrading D-knowledge to S-knowledge cannot be tackled in a single move. Two separate obstacles to this form of upgrading first need to be identified. The first one is related to the already mentioned principle of full-communication; the second is due to knowability-issues and is independent of the former.

To see why both issues are independent, we need to be careful in the formulation of the principle of full-communication. Intuitively, that principle says that whenever ϕ is distributed knowledge in a group, the members of the group should be able to find out that ϕ via communication. The latter seems to suggest that it should be possible for at least some agent in the group to know that ϕ, but when we look at a more precise formulation of that principle, we see that it actually requires something weaker. Formally, the principle of full communication says that for all $\phi \in \mathcal{L}_K$ (the D-free

[9] The following static closure-rule, which is described in Palczewski [2007: 458], makes it quite clear that the logical consequences of S-knowledge are only D-known, and hence implicit: $(\phi_n \wedge \ldots \wedge \phi_n) \rightarrow \psi / (\mathsf{S}\phi_n \wedge \ldots \wedge \mathsf{S}\phi_n) \rightarrow \mathsf{D}\psi$.

[10] Of course, if all knowledge is passed on to a unique wise man component, that knowledge will automatically be deductively closed.

epistemic language) we have that:

$$\mathsf{D}\phi \Longrightarrow \{\phi_i \in \mathcal{L}_K : \mathsf{S}\phi_i\} \vdash \phi$$

The first obstacle to upgrading D-knowledge to S-knowledge is due to failures of this principle. Consider, for that matter, Figure 1 (based on Roelofsen [2006]) where at world w, p is distributed knowledge in the two-agent group $\{a, b\}$, but where pooling both agents' knowledge together does not suffice to establish p. Indeed, at w, there is no non-p world that is considered possible by both agents. Yet, it is also true that both agents ignore whether p is true; in fact, according to this model, they know nothing at all! As a result, even if p is D-known at w, merely pooling together the knowledge of both agents will not suffice to deduce p.

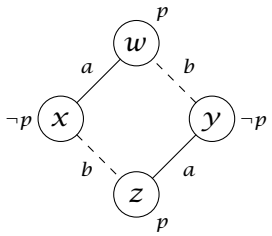

FIGURE 1: Failure of Full-Communication

I do not think that this result indicates something deep about the existence of explicitly unknowable, yet implicitly known truths. At best, it indicates the epistemic inadequacy of those Kripke-models where worlds compatible with an agent's knowledge fail to be epistemic alternatives. For instance, in the above example the world z is compatible with a's knowledge at w, but it isn't epistemically accessible from that world. As a result, a excludes a world that is compatible with his knowledge, and therefore should not have been excluded.

The solution to this problem can, in view of the above remarks, remain straightforward: we only need to stipulate that our semantic definition of distributed knowledge has its intended meaning only in those Kripke-models where the principle of full communication is satisfied. Fortunately, this class of models has already been identified in Roelofsen [2006], and we can therefore rely on that result to overcome the first obstacle to upgrading D-knowledge to S-knowledge. We just need to limit our attention to the class of models where for every world, everything that is consistent with what is known at that world by some agent in the group is satisfiable at all worlds considered possible at that world by all agents in the group. This is of course just a different way to say that compatibility and epistemic possibility (for individual agents, but also for groups) should not be two distinct notions; for if both notions come apart, the semantic definition of distributed knowledge can be shown to be defective.[11]

[11] In this paper I leave it open whether this has an impact on how sceptical alternatives can

♦ ♦ ♦

Despite the appearances, satisfying the principle of full communication is a necessary, but not yet a sufficient condition for the upgradability of D-knowledge to S-knowledge. Even if something can be derived from the total knowledge available in a group, the result of that derivation may be unknowable in a way that's most familiar from Fitch's paradox as well as from Moorean sentences.[12] These issues form the second obstacle to upgrading D-knowledge to S-knowledge—an obstacle we shall meet again for each further form of upgrading (though I shall only comment on these in the final section). To see where both obstacles differ, we should first note that the principle of full communication is stated relative to a static notion of deduction; it refers to pooling all information available to a group, not to the fact that some agent should come to know that information. That's why full communication cannot on its own warrant upgradability. What Fitch's paradox tells us, is that this is the price we have to pay for not having knowability (here, upgradability) collapse with actual knowledge (here, not yet the collapse of D-knowledge and S-knowledge, but something close enough). Let me phrase this in a somewhat more detailed form, and illustrate the kind of knowability-issues we face here. We start from a knowability-principle inspired by D to S-knowledge upgrading (compare with the proposals in Balbiani, Baltag et al. [2008]):

$$\forall p(\mathsf{D}p \rightarrow \Diamond \mathsf{S}p)$$

And then assume that it can be distributed knowledge that p is true, but that no-one knows this:

$$\exists p(\mathsf{D}(p \wedge \neg \mathsf{S}p))$$

Proceeding in analogy with Fitch's proof, we can show that our assumption is incompatible with D- to S-knowledge upgrading, and that hence:

$$\forall p((\mathsf{D}p \rightarrow \Diamond \mathsf{S}p) \rightarrow (\neg \mathsf{D}(p \wedge \neg \mathsf{S}p)))$$

Which means that whenever p is D-known, it is either the case that p cannot be S-known, or that it isn't D-known that p and that p isn't S-known. Or, by contraposition, whenever p is D-known, it is either the case that p cannot be S-known, or it isn't D-known that p isn't S-known. As a result, in those cases where p is true, D-known, but not S-known, it follows that if all D-knowledge can be upgraded, there are some truths—namely the fact that p is not S-known—that are not D-known, and hence that there are some truths that are not even D-knowable. In other words, we have just shown that we can only assume D to S-upgradability if we accept that some truths cannot even be D-known.

Include proof in footnote

be dealt with.

[12] The relevance of different forms of group-knowledge for Fitch's paradox was first raised by van Benthem [2004; 2009].

What this digression reveals, is that upgradability cannot be unrestrictedly valid. It needs to be restricted to S-knowable formulae; formulae such that learning that ϕ is successful in the sense that ϕ becomes S-known. To conclude our discussion of the second obstacle to upgrading D knowledge to S-knowledge, we must first note that unlike for the first obstacle, it still remains unknown how we should restrict even the simplest knowability-principles. As such, while we can say that such principles should be restricted, there's no agreed upon criterion that could be used to do so; van Ditmarsch & Kooi [2006] give an overview of different partial criteria.

The main focus of this paper is on the closure and introspection principles, and especially on ways to model dynamic interpretations of these principles as procedures for upgrading, respectively, D-knowledge to S-knowledge and E-knowledge to C-knowledge. Such a modelling suggests that upgrading S-knowledge to E-knowledge is irrelevant to either of these principles, but that is misleading way to frame the issue. All that can be said is that making explicit what is merely implicitly stored is the crucial feature of closure, and that explaining how something can become common knowledge for a group of components tells us something essential about fully introspective knowledge. These two claims do not, however, exclude the relevance of the intermediate upgrading of S-knowledge to E-knowledge. In fact, I will now argue that this step is, at least in a residual sense, relevant to both principles.

To begin with, we have already seen that one way to make knowledge explicit, is to pass it on to a designated component. We have also seen that explicit knowledge could equally well be achieved by sending everything to all components, but this is also exactly what needs to be done to achieve E-knowledge (which in addition to being explicit, is also readily accessible). What this tells us is not yet that E-knowledge is as easy to obtain as S-knowledge, but rather that the process to achieve the former is not fundamentally different from the process that is required to achieve the latter. Put differently, making knowledge readily accessible is not all that different from making it explicit. This can be seen if we reconsider the protocol that requires that all knowledge be sent to a designated 'wise man' component, for this process suffices to make knowledge explicit as well as readily accessible (provided one can identify the designated component), and yields a result which is, at least with respect to explicitness and accessibility, not all that different from E-knowledge. From this, we may conclude that making knowledge explicit and making it available require the same sort of computational processes, which is just to say that they can be achieved by protocols using similar forms of communication.

But there's more. Consider a situation where two components exchange information by sending messages, acknowledging the receipt of that message, and acknowledging the receipt of the previous acknowledgement (in short: `send` p, `ack` p, `ack ack` p). Assume furthermore that communication is reliable in the sense we required before (i.e. learning something

leads to knowledge as well as higher-order knowledge), but not necessarily reliable in the sense that messages are never lost (for otherwise there would be no point in acknowledging the receipt of a message). If that is the case, then the procedure summarised as `send` p, `ack` p, `ack ack` p, leads to a situation where both agents know that p, and each agent also knows that the other knows that p. Since both of these agents are already introspective, this is sufficient to establish that within this two-agent group, p is E-known and it is also E-known that p is E-known. Yet, in view of the uncertainty concerning the actual receipt of an unacknowledged message, this is as far as these iterations can go after a three-round communication-protocol.

When we discuss the process of upgrading to C-knowledge, more will have to be said on the fact that communication which leaves room for uncertainty about whether a message was actually received can at best ensure a limited degree of introspective E-knowledge. Here, we only need to note that any such higher degree of knowledge can be achieved by that form of possibly unreliable communication, and that therefore the kind of communication that is sufficient for achieving E-knowledge is also sufficient for achieving limited introspective E-knowledge. This last remark completes our claim that upgrading S-knowledge to E-knowledge is not only relevant with respect to deductive closure, but also relevant to introspection.

♦ ♦ ♦

To move up from merely implicit knowledge modelled after distributed knowledge to explicit, readily available, and at most finitely introspective knowledge modelled after E-knowledge, we only need to make use of a single form of communication. That kind is often called unreliable, because it leaves room for uncertainty about messages being actually received. A different way to look at these messages, is as (wholly or partially) private announcements; only the recipient has to notice that a message is received. The first terminology is commonly used in the context of the so-called co-ordinated attack problem (see e.g. Fagin, Halpern at al. [1995]); the second terminology is due to the field of dynamic epistemic logics (Baltag & Moss [2004]). It is a well-known fact that, precisely because it always leaves room for uncertainty, common knowledge cannot be achieved through unreliable communication. As a consequence, since a co-ordinated attack requires that all parties agree, and that agreement presupposes common knowledge,[13] the unattainability of common knowledge implies the impossibility of a co-ordinated attack. By reasoning about kinds of announcements, rather than about reliable or unreliable communication-channels, a more general perspective is gained on these results. In short: all and only *public announcements* can result in common knowledge (see e.g. Appendix 2 in van Eijck & Wang [2008]). What this reveals is that upgrading

[13] An informal argument for the connection between agreement and common knowledge is obtained by observing the analogy between (a) the fact that we can only agree on p iff we both know p and know that we agree on p, and (b) the fixed-point definition of common knowledge given below.

from D-knowledge to E-knowledge can be done by private communication, but that further upgrading to C-knowledge requires the ability to make public announcements. In other words, there's a part of the hierarchy of group-knowledge that cannot be reached unless public announcements can be made. But this also means that if this hierarchy of group knowledge is used to model different forms of single-agent knowledge, full introspection for that agent could be unattainable in principle if the interaction between components is thus configured that the required form of public communication is impossible.

My aim, here, is not to provide full proofs or even just the outlines of the proofs required to establish these results. I only want to use these results to shed light on some principled limitations on how we can achieve common knowledge, and therefore on how fully introspective knowledge can be obtained given our previous choice to model the latter in analogy to the former. To do so, I primarily need to convey what is special about common knowledge, why it is hard to achieve, and finally what makes public announcements so special that they can result in the common knowledge of what is publicly announced. Intuitively, p being common knowledge in a group G is special because it literally excludes any doubt or uncertainty that a member might have about any other member of the group being aware that they all know that p. Yet, it is only when we realise that this involves every finite iteration of E-knowledge that the real strength of the lack of such uncertainty can be appreciated. Consider, for that purpose, two different ways of defining common knowledge: the iterated definition, and the fixed-point definition.

1. Where $\mathsf{E}^k p$ is inductively defined by means of the base clause: $\mathsf{E}^1 p \leftrightarrow \mathsf{E}p$, and the inductive clause: $\mathsf{E}^{k+1} p \leftrightarrow \mathsf{E}\mathsf{E}^k p$; $\mathsf{C}p$ is equivalent to the infinite conjunction of all $\mathsf{E}^k p$ for finite k.

2. $\mathsf{C}p$ is equivalent to $\mathsf{E}(p \wedge \mathsf{C}p)$

Each of these definitions conveys a crucial aspect of common knowledge. The first one, in virtue of its infinitary nature, explains why common knowledge is hard to achieve. Even more, it in fact clarifies why common knowledge is impossible to obtain if we try to reach it by subsequently ensuring each further iteration of E's. This is what happens when information can only be shared through restricted (i.e. non-public or so-called (partially) private) communication.

The second definition, by contrast, points to a finite way to express the infinitary nature of common knowledge; which is something that can only by achieved through a fixed-point construction. One way to think about this fixed-point construction proceeds semantically, and refers to the transitive closure of the union of the epistemic accessibility-relations of all members of a group. Such a transitive closure is itself a fixed-point construction, but it also points to an analogy between introspective single-agent knowledge and common knowledge (this is illustrated in van Ditmarsch, van Eijck &

Verbrugge [2009]). Taking the transitive (or on some definitions, the transitive and symmetric) closure of a single agent's epistemic accessibility relation suffices to semantically define introspective single-agent knowledge, and so does taking the transitive closure of a group's epistemic accessibility relations suffice to obtain a fully introspective version of E-knowledge, namely common knowledge. A different way to think about this alternative definition exploits the already mentioned analogy with agreements. Whenever we agree to do something, we do not only have to agree on the subject matter of that agreement (e.g. the action itself), but it also has to be clear to all parties that an agreement is reached. That is, to agree on p we do not only need to agree with regard to p, but we also need to agree on the fact that we agree. Such a self-reference is nothing more than the fixed-point construction we had to use to give a finite expression of the infinitary nature of common knowledge.

Keeping the analogy with agreements in mind, we can now tackle the question of how common knowledge can be reached in a finite number of steps. The clue lies in a third way of looking at common knowledge, that of being in a shared informational context (Barwise [1988]). The main feature of a shared informational context, is that it is transparent to anyone within that context: no information can be exchanged without all parties being aware of what information is exchanged and the impact that exchange has on anyone within that context. Perhaps, it is more accurate to say that a group of agents can only be in a shared informational context relative to a certain communicative action or announcement. One is in a shared or transparent informational context relative to an action iff there is no ignorance whatsoever about whether that action takes place or what its actual effects are. After such an action, it will not only be common knowledge that this action took place (provided the agents can remember this, i.e. have what game theorists call *perfect recall*), but the outcome of that action will be common knowledge as well. With regard to such contexts, Barwise comments that:

> The intuitive idea is that common knowledge amounts to perception or other awareness of some situation, part of which includes the fact in question, but another part of which includes the very awareness of the situation by both-agents. Again we note the circular nature of the characterisation.
>
> Barwise [1988: 368]

Which could, if we concentrate on seeing, mean that both agents see the same, but are also aware of each other seeing the same. By analogy, an agreement is something that can typically only be reached in a face-to-face situation: each agent can only agree by recognising that others agree as well.

Thus, as we've both established that (a) only shared informational contexts can warrant common knowledge, and that (b) all and only public

announcements lead to common knowledge, we may now conclude that public announcements can only take place within such a shared informational context, and that common knowledge will be achieved after public announcements made in such a context.

Common knowledge is harder to obtain that any other form of group-knowledge, and, since we've argued that fully introspective knowledge shares most of its formal properties with common knowledge, fully introspective knowledge is equally hard to obtain. As a matter of fact, it can in some cases even be unattainable in principle. Of course, this does not mean that more moderate forms are unattainable as well. Since each limited form of introspection lies within the scope of E-knowledge, the limits on C-knowledge do not affect the prospects for bounded introspection. In summary: E-knowledge can be attained as soon as every piece of information available within a group of components can eventually reach every component. One way to achieve this proceeds by sending that information to all components as soon as there is a single component which actually holds that information. By contrast, C-knowledge can only be achieved when, given that at least one component holds a piece of information, that component can pass this information on to all the other agents, and can do so in a way that is entirely transparent to all these components. As one may guess, this is a condition that isn't as easily satisfied.

6 COMPONENTS AS STATES

As already mentioned at the end of Section 4 a literal reading of components as "parts of the brain" does not sit well with the proposed analysis of S-knowledge as explicitly stored knowledge, and E-knowledge as readily available explicitly stored knowledge. A more neutral interpretation of the separate components refers to them as different states of the agent that is being modelled. Even more, an interpretation of these states as different temporal stages of that agent seems particularly compelling.[14] Though fruitful, I think this proposal is also misleading. It suggests that there should be one privileged reading, but if $\mathsf{S}_G\phi$ is read as "ϕ is known at some state of the agent G" we only focus on one among the many criteria that could be used to discriminate between an agent's different states of information.

On a strict temporal reading of the different states, information is explicitly stored if it is known at one temporal stage (i.e. S-known), and readily available if it is explicitly stored at all temporal stages (E-known). Furthermore, if explicit storage at one stage doesn't imply that the information

[14] This suggestion is directly inspired by Sequoiah-Grayson [forthcoming].
An even more generic approach would consist in treating the separate components and their interaction as a certain perspective on an agent. By individuating more components, we just give a finer analysis of the agent in question. What the components correspond to, is a function of the purpose of the analysis.

must remain explicit at subsequent states, information can be explicitly stored without being readily accessible. Yet, E-knowledge has now become too hard to obtain to serve as a model for readily available knowledge. A more realistic proposal is to model readily available knowledge as knowledge that is explicitly stored at the actual state (henceforth, A-knowledge). This actual state then fulfils the role of a *designated component* which can immediately be queried. An additional virtue of this interpretation is that it captures the intuitively plausible idea that what is readily available can immediately be remembered, and that remembering may involve recalling or querying a prior stage.

Two further issues that are directly related to the strict identification of component-knowledge with knowledge at a temporal stage can now be identified. The first one is directly related to the logical properties of A-knowledge, and more specifically to the fact that it obliterates some previously introduced distinctions. In short: making knowledge explicit at the actual stage immediately makes that knowledge fully introspective as well. The second issue is independent of the logical properties of A-knowledge, but bears on the fact that on the present proposal being readily available is a necessary and sufficient condition for explicit knowledge to be deductively closed. Yet, even the deductive yield of a handful of axioms—something that can for sure be explicit at the actual stage—can still be quite impressive.

The collapse of readily available and fully introspective knowledge is not fatal for the overall proposal defended in this paper. From a formal point of view, it suffices to fall back on a slightly larger set of designated components or actual temporal stages and define restricted versions of E- and C-knowledge over that set. In terms of the proposed interpretation, this coincides with a multi-dimensional discrimination of states; with the temporal dimension as an important, but not as the only criterion to tell states apart. This also gives a new gloss to the notion of component-interaction, which is now closer to an abstract information-flow between different information-states of the same agent. The reference to the "cognitive architecture" of an agent is still valid, but doesn't refer to the physical implementation of that architecture.

By contrast, the collapse of readily available and deductively closed knowledge points to a more principled limitation of an "agents as groups" model. The problem is that no matter how finely we individuate states, there always will be a residual part of the purely computational aspects of closure that cannot be covered by component interaction or information flow between states, and which will therefore have to be imputed on the individual components or states. As long as we adhere to a normal modal logic to model the individual components, this is indeed unavoidable. The only plausible reply is purely pragmatic: as far as our formal model is concerned, we make abstraction of the computational resources required to deduce the logical consequences of knowledge that is readily available.

7 HINTIKKA'S "PROOF" FOR POSITIVE INTROSPECTION REVISITED

I have already dealt with the objection that by modelling closure and introspection for individual agents with the formal resources of interactive knowledge, I would get the order of explanation wrong. Still, merely showing that individual knowledge isn't necessarily conceptually prior to interactive knowledge, does not yet warrant that the thus obtained model adds something that was not yet available to the less discriminating models of single-agent epistemic logic.

Let us, for that purpose, return to one of our starting points, namely the principle of positive introspection, and see what becomes of Hintikka's supposed proof of positive introspection when it is approached from the perspective of how components have to interact to achieve introspective knowledge in a group. To begin with, one should understand how Hintikka argues in favour of a principle of epistemic logic. The basic idea is that when we ask whether a certain principle (most likely an implication of the form $\phi \rightarrow \psi$) is valid we should try to find out whether the set $\{\phi, \neg\psi\}$ is defensible. In its most general form, this would mean that we should ask whether supporting each member of that set could be shown to be incoherent. Whenever such a set is logically inconsistent, it is also considered incoherent and therefore indefensible. However, indefensible sets do not have to be inconsistent; it suffices that it is incoherent to support, believe or know each member of the set.

This is exactly the kind of considerations on which Hintikka's supposed proof of the KK-principle is based. A neat and fairly neutral reconstruction of that proof, is given by Stalnaker [2006] and reproduced below.

1. If $\{\mathsf{K}_a\phi, \neg\mathsf{K}_a\neg\psi\}$ is consistent, then $\{\mathsf{K}_a\phi, \psi\}$ is also consistent.

Hence, by substituting $\neg\mathsf{K}_a\phi$ for ψ, we obtain:

2. If $\{\mathsf{K}_a\phi, \neg\mathsf{K}_a\neg\neg\mathsf{K}_a\phi\}$ is consistent, then $\{\mathsf{K}_a\phi, \neg\mathsf{K}_a\phi\}$ is also consistent.

Which after eliminating the double negations, and taking the contrapositive gives us:

3. Since $\{\mathsf{K}_a\phi, \neg\mathsf{K}_a\phi\}$ is inconsistent, $\{\mathsf{K}_a\phi, \neg\mathsf{K}_a\mathsf{K}_a\phi\}$ is also inconsistent.

The strange thing about this proof, is that it appeals to consistency, but apparently not to any epistemic form of indefensibility. This cannot be the case, for we know that the set $\{\mathsf{K}_a\phi, \neg\mathsf{K}_a\mathsf{K}_a\phi\}$ can be satisfied in non-transitive Kripke-models. It can therefore not be called inconsistent without already presupposing that knowledge is introspective. A closer look at the first step reveals what is happening. The reasoning behind that step is that if one does not know ψ, then ψ should not only be consistent with what one knows to be true (i.e. ϕ), but also with the fact that one knows ϕ.

Strictly speaking, this line of reasoning is equivalent to the KK-thesis itself, but that shouldn't elude the fact that it is also a valid use of Hintikka's notion of epistemic defensibility.

With regard to the concept of epistemic defensibility, Hendricks [2006] emphasises that the epistemic principles defended on the basis of the latter are best regarded as strong rationality postulates. The focus on the first-person perspective can then be seen as additional evidence for the influence of Moore's autoepistemology on Hintikka's own formulation of epistemic logic (p. 89). This is true of his defence of closure, but even more of the proof or argument in favour of positive introspection. Considered along these lines, Hintikka's proof is closely related to how he evaluates the knowledge version of Moore's problem (What is wrong with "p, but I don't know p," given that this conjunction isn't inconsistent?). What Hintikka seems to argue is that (a) such Moorean sentences are epistemically indefensible, and (b) that the notion of epistemic indefensibility which is needed to explain what is wrong with such sentences also suffices to explain why (from a first person perspective) knowledge should be positively introspective.

From our previous encounter with knowability issues we already know that the distinctive epistemic trait of Moorean sentences is that they are unknowable: even if true, learning their truth cannot result in knowing them to be true (for they become false once learned). To formulate an interactive version of Hintikka's argument for positive introspection, we thus need a sentence which denies positive introspection, but also turns out to be unknowable.

Predictably, given the more expressive language we use, there are many sentences which deny some or other form of positive introspection. Three such sentences are of particular interest:

$$\mathsf{E}p, \text{ but } \neg\mathsf{E}\mathsf{E}p \tag{E}$$

$$\mathsf{E}p, \text{ but } \neg\mathsf{C}p \tag{EC}$$

$$\mathsf{C}p, \text{ but } \neg\mathsf{C}\mathsf{C}p \tag{C}$$

Since the last one can be dismissed right away (no-one will deny that common knowledge is introspective), we can restrain our attention to the first and second one. Next, we should note that the second one is implied by the first. More exactly, it is implied by each instance of $\mathsf{E}p$, but $\neg\mathsf{E}^k p$ with finite k. As a result, the second sentence is both the denial of full positive introspection, and also the weakest denial of positive introspection in general.

To find out whether either of these sentences is knowable, we first consider a case with only two agents or components. If we start from the assumption that both a and b know that p, but that at least one of them ignores this epistemic fact. This is sufficient for the truth of $\mathsf{E}p$, but $\neg\mathsf{E}\mathsf{E}p$. But is it also unknowable? If we assume that a already knows that b knows p, then the announcement of "$\mathsf{E}p$, but $\neg\mathsf{E}\mathsf{E}p$" by a (or by a third agent) is

true but unsuccessful. In other words, in that situation $\mathsf{E}p$, but $\neg\mathsf{E}\mathsf{E}p$ is an unknowable truth. If, by contrast, the situation is such that a and b ignore whether the other one knows that p, the same truth is at least knowable when it is first (and privately) announced to either a or b (but not to both).[15] Taking the two examples together, it follows that the sentence is knowable in the sense expressed by ($\Diamond\mathsf{S}$), but not in the sense expressed by ($\Diamond\mathsf{E}$)

$$(\mathsf{E}p \wedge \neg\mathsf{E}\mathsf{E}p) \rightarrow \Diamond\mathsf{S}(\mathsf{E}p \wedge \neg\mathsf{E}\mathsf{E}p) \qquad (\Diamond\mathsf{S}\mathsf{E})$$

$$(\mathsf{E}p \wedge \neg\mathsf{E}\mathsf{E}p) \rightarrow \Diamond\mathsf{E}(\mathsf{E}p \wedge \neg\mathsf{E}\mathsf{E}p) \qquad (\Diamond\mathsf{E}\mathsf{E})$$

On the assumption that all components can send messages to all other components, this last insight directly generalises to the n component case.

What about the weaker truth $\mathsf{E}p \wedge \neg\mathsf{C}p$? Here, we start immediately with the more general n-component case. For the announcement of $\mathsf{E}p \wedge \neg\mathsf{C}p$ to be unsuccessful, it has to become false in virtue of being announced. For a conjunction to become false, it is also sufficient that only one conjunct becomes false. In this case, there's only one option: $\mathsf{E}p$ cannot become false unless p also becomes false (which is excluded since announcements cannot alter non-epistemic facts). As a result, the only way for the announcement of $\mathsf{E}p \wedge \neg\mathsf{C}p$ to be unsuccessful is if it makes $\mathsf{C}p$ true. And this is something that can only be the result of a public announcement. Again, there are two knowability-claims that can be considered:

$$(\mathsf{E}p \wedge \neg\mathsf{C}p) \rightarrow \Diamond\mathsf{E}(\mathsf{E}p \wedge \neg\mathsf{C}p) \qquad (\Diamond\mathsf{E}\mathsf{C})$$

$$(\mathsf{E}p \wedge \neg\mathsf{C}p) \rightarrow \Diamond\mathsf{C}(\mathsf{E}p \wedge \neg\mathsf{C}p) \qquad (\Diamond\mathsf{C}\mathsf{C})$$

If knowability is understood as "there is a way to make this announcement in a successful manner," then ($\Diamond\mathsf{E}\mathsf{C}$) is true in virtue of the possibility to announce $\mathsf{E}p \wedge \neg\mathsf{C}p$ privately to all components. If knowability only refers to what is knowable after public announcements, then ($\Diamond\mathsf{E}\mathsf{C}$) is false because such announcements induce common knowledge therefore lead to $\mathsf{C}(\mathsf{E}p \wedge \neg\mathsf{C}p)$ which is inconsistent. The latter immediately shows that ($\Diamond\mathsf{C}\mathsf{C}$) is false no matter how knowability is understood.

♦ ♦ ♦

The moral of this comparison is that by adopting a more refined model of introspection, it is no longer sufficient to invoke auto-epistemic considerations to defend the "virtual equivalence" of *knowing* and *knowing that one knows* (Hintikka [1962: V]). While on the original single-agent model there is a direct connection (indeed, an equivalence) between the validity of positive introspection and Hintikka's notion of epistemic defensibility as expressed by the claim that $\{\mathsf{K}_a\phi, \neg\mathsf{K}_a\psi\}$ is consistent only if $\{\mathsf{K}_a\phi, \psi\}$ is also consistent, that connection is lost on the more refined model. The typical auto-epistemic considerations can still be expressed in terms of knowability, but they can only be used to dismiss a limited number of denials of

[15] Note that since there are only two agents, neither of these agents can make the relevant announcement.

positive introspection. Hence, since some such denials are knowable, they are also defensible in Hintikka's sense. These final considerations are a good reason to believe that the model proposed in this paper adequately formalise the "mere access" aspect of positive introspection hinted at in the introduction.

Patrick Allo
Postdoctoral Fellow of the Research Foundation (FWO)
CLWF (Vrije Universiteit Brussel)
IEG (Oxford) & GPI (Hertfordshire)
patrick.allo@vub.ac.be
homepages.vub.ac.be/~pallo/

REFERENCES

ABRAMSKY, S., 2008, Information, Processes and Games, in: *Handbook on the Philosophy of Information*, Van Benthem, J. and P. Adriaans, ed., Elsevier, Amsterdam: 483–550.

BALBIANI, P., A. BALTAG, H. VAN DITMARSCH, A. HERZIG, T. HOSHI AND T. DE LIMA, 2008, Knowable as Known After an Announcement, *The Review of Symbolic Logic*, **1(3)**: 305–334.

BALTAG, A. AND L. S. MOSS, 2004, Logics for Epistemic Programs, *Synthese*, **139(2)**: 165–224.

BARWISE, J., 1988, Three Views of Common Knowledge, *TARK II*, Pacific Grove, California.

DANTO, A. C., 1967, On Knowing that We Know, in: *Epistemology. New Essays on the Theory of Knowledge*, Stroll, A., ed., Harper and Row, New York: 32-53.

DRETSKE, F., 1970, Epistemic Operators, *The Journal of Philosophy*, **76(24)**: 1007–1023.

FAGIN, R., J. Y. HALPERN, Y. MOSES AND M. Y. VARDI, 1995, *Reasoning About Knowledge*, MIT Press, Cambridge / London.

HALPERN, J. Y., 1996, Should knowledge entail belief?, *Journal of Philosophical Logic*, **25(5)**: 483–494.

HALPERN, J. Y. AND Y. MOSES, 1985, A guide to the modal logics of knowledge and belief. In: *Proceedings of IJCAI-85*, Los Angeles, CA: 480–490.

HALPERN, J. Y. AND Y. MOSES, 1990, Knowledge and Common Knowledge in a Distributed System, *Journal of the Association for Computing Machinery*, **37(3)**: 549–587.

HARMAN, G., 1986, *Change in View. Principles of Reasoning*, MIT, Cambridge, Ma.

HENDRICKS, V., 2006, *Mainstream and Formal Epistemology*, Cambridge University Press, Cambridge.

HILPINEN, R., 1970, Knowing that one knows and the classical definition of knowledge, *Synthese*, **21(2)**: 109–132.

HINTIKKA, J., 1962, *Knowledge and Belief. An Introduction to the Logic of the Two Notions*, Cornell University Press, Ithaca.

HINTIKKA, J., 1970, "Knowing that one knows," reviewed, *Synthese*, **21(2)**: 141–162.

LEMMON, E. J., 1959, Is There Only One Correct System of Modal Logic?, *Proceedings of the Aristotelian Society (Supplementary Volume)*, **33**: 23–40.

LEMMON, E. J., 1967, If I Know, Do I Know that I Know?, in: *Epistemology. New Essays on the Theory of Knowledge*, Stroll, A., ed., Harper and Row, New York: 54–82.

LENZEN, W., 1978, *Recent Work in Epistemic Logic*, North-Holland, Amsterdam.

LEVESQUE, H. J., 1984, A Logic of Implicit and Explicit Belief, *National Conference on Artificial Intelligence*, Houston, Texas.

LEWIS, D., 1969, *Convention. A Philosophical Study*, Harvard University Press, Cambridge, Ma.

LISMONT, L. AND P. MONGIN, 1994, On the logic of common belief and common knowledge, *Theory and Decision*, **37(1)**: 75–106.

NOZICK, R., 1981, *Philosophical Explanations*, Harvard University Press, Cambridge, Ma.

PALCZEWSKI, R., 2007, Distributed Knowability and Fitch's Paradox, *Studia Logica*, **86(3)**: 455–478.

PLAZA, J., 2007, Logics of public communications, *Synthese*, **158(2)**: 165–179.

ROELOFSEN, F., 2006, Distributed Knowledge, *Journal of Applied Non-classical Logics*, **16(2)**: 255–273.

SEQUOIAH-GRAYSON, S., forthcoming, Epistemic Closure and Commutative, Nonassociative Residuated Structures, *Synthese*.

SHIN, H. S., 1993, Logical Structure of Common Knowledge, *Journal of Economic Theory*, **60(1)**: 1–13.

STALNAKER, R., 1991, The Problem of Logical Omniscience, I, *Synthese*, **89(3)**: 425–440.

STALNAKER, R., 2006, On Logics of Knowledge and Belief, *Philosophical Studies*, **128(1)**: 169–199.

VAN BENTHEM, J.. 2004, What one may come to know, *Analysis*, **64(282)**: 95–105.

VAN BENTHEM, J., 2006, Epistemic Logic and Epistemology: The State of their Affairs, *Philosophical Studies*, **128(1)**: 49–76.

VAN BENTHEM, J., 2008, Logical dynamics meets logical pluralism?, *Australasian Journal of Logic*, **6**: 182–209.

VAN BENTHEM, J., 2009, Actions that make us know, in: *New Essays on the Knowability Paradox*, Salerno, J., ed., Oxford University Press, Oxford: 129–146.

VAN DER HOEK, W., B. VAN LINDER AND J.-J. C. MEYER, 1999, Group knowledge is not always distributed (neither is it always implicit), *Mathematical Social Sciences*, **38**: 215–240.

VAN DITMARSCH, H. AND B. KOOI, 2006, The Secret of My Success, *Synthese*, **151(2)**: 201–232.

VAN DITMARSCH, H., J. VAN EIJCK AND R. VERBRUGGE, 2009, Common Knowledge and Common Belief, in: *Discourses on Social Software*, van Eijck, J. and R. Verbrugge, ed., Amsterdam University Press, Amsterdam: 107–32.

VAN EIJCK, J. AND Y. WANG, 2008, Propositional Dynamic Logic as a Logic of Belief Revision. *Logic, Language, Information and Computation*: 136–148.

VAN LINDER, B., W. VAN DER HOEK AND J.-J.CH. MEYER, 1994, Communicating rational agents. in: *KI-94: Advances in Artificial Intelligence*, Nebel, B., L. Dreschler-Fischer, eds., Springer, New York: 202–213.

VOORBRAAK, F. 1990, The Logic of Objective Knowledge and Rational Belief. in: *Logics in AI: European Workshop JELIA 1990*, J. Van Eijck., ed., Berlin, Springer: 499–515.

WILLIAMSON, T., 2000, *Knowledge and Its Limits*, Oxford University Press, Oxford.