J Philos Logic (2013) 42:91–124
DOI 10.1007/s10992-011-9214-2

The Many Faces of Closure and Introspection

An Interactive Perspective

Patrick Allo

Received: 3 January 2011 / Accepted: 18 August 2011 / Published online: 7 September 2011

Abstract In this paper I present a more refined analysis of the principles of deductive closure and positive introspection. This analysis uses the expressive resources of logics for different types of group knowledge, and discriminates between aspects of closure and computation that are often conflated. The resulting model also yields a more fine-grained distinction between implicit and explicit knowledge, and places Hintikka's original argument for positive introspection in a new perspective.

Keywords Closure · Common knowledge · Deductive omniscience · Distributed knowledge · Epistemic logic · Introspection · Knowability

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. [13], 485). From a philosophical perspective, however, there are at least two features of the logic

P. Allo (✉)
Centre for Logic and Philosophy of Science, Vrije Universiteit Brussel,
Brussels, Belgium
e-mail: patrick.allo@vub.ac.be

P. Allo
Science Foundation (FWO-Vlaanderen), Brussels, Belgium

P. Allo
IEG, Oxford University, Oxford, UK

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 implies knowing what one knows to be implied by p. By introspection, I shall (primarily) refer to the principle of *positive* introspection, which stipulates that knowing implies knowing that one knows.[1] The seemingly innocuous principle of closure was most famously challenged by Dretske [9] and Nozick [26], 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* [20, 21], widely scrutinised in the years following its publication [8, 19, 22], and systematically challenged (if not conclusively rejected) by Williamson [46].

1.1 The Problem

Formal models often treat introspective and closure principles as monolithic properties of a cognitive agent. This is a mistake: Neither closure nor introspection is as monolithic as standard epistemic logic seems to imply. For instance, if deductive closure is thought of as a constraint on an agent's commitments [20] this is a strong rationality constraint, but a relatively weak constraint on what that agent can actually compute. By contrast, when the same principle is treated as a constraint on what an agent can actually deduce, deductive omniscience becomes a much stronger feature. Similarly, deductive closure can refer to what an agent has actually deduced (closure as a synchronic constraint), or merely to what an agent can come to know without first obtaining additional information (closure as a diachronic constraint). Again, this makes a difference as to whether deductive omniscience is easy or hard to obtain. Yet, these are subtleties standard epistemic logic cannot deal with. The situation is quite similar for introspection-principles, where knowing that one knows is a rather demanding standard if it is to be achieved by actually storing beliefs about beliefs, but a much weaker requirement if knowing that one knows is just a matter of avoiding an indefensible position (see Hintikka's argument discussed in Section 6).

A further distinction that lies beyond the grasp of standard epistemic logic is that between single-premise and multi-premise closure. Again, this is a

[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: $\mathsf{BK}\varphi \rightarrow \mathsf{K}\varphi$. 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 [12, 114]. Alternative replies are possible, but I think this is still a good enough reason to focus on the disagreement concerning positive introspection.

contrast that plays a crucial role in epistemology (see e.g. [17], 1.4–1.6), but within the standard framework of epistemic logic there's no way to acknowledge the fact that single-premise closure might be less demanding or at least less problematic than multi-premise closure. Here too, the situation is quite similar for introspection-principles, as the contrast between merely knowing that one knows and the more demanding further iteration of knowledge operators, which plays a crucial role in Williamson's rejection of the KK-principle, lies beyond what can be discriminated by a standard epistemic logic.

Of course, most of these issues have been addressed through the formulation of alternative epistemic logics. For instance, modal logics based on a neighbourhood-semantics have been used to formulate systems for agents that are not deductively omniscient [7, Chapt. 7], while the distinction between implicit and explicit belief has become standard since Levesque [23], and has also been extended to encompass the dynamic processes whereby implicit beliefs become explicit beliefs. More recently, Bonnay and Egré [5] have proposed a model of introspective knowledge that nevertheless discriminates between different degrees of iterated knowledge operators. Even then, none of these proposals yields a formal model that fully acknowledges the multi-faceted nature of closure and introspection principles I hinted at above. Instead of combining these partial solutions, I propose to explore a different type of approach.

1.2 The Multi-Component Characterisation

The alternative approach I have in mind is based on the use of different types of group-knowledge as a model for 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. This approach is reminiscent of Marvin Minsky's "Society of Mind" [25], and is closely related to a proposal due to Fagin and Halpern [10] as well as to Stalnaker's solution to the "Problem of Deduction" [33, Chapt. 5]. The way in which I present my own proposal is also indebted to the distinction between information-flow relative to a sub-system and the conservation of information relative to the system as a whole [1].

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.

Outline This paper is structured as follows. In the next two sections I describe, interpret and defend the logical properties of component knowledge (Section 2) and describe the different forms of group knowledge (Section 3). In Section 4 I show how the hierarchy of different types of group-knowledge gives rise to a matching 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 introspection (Section 6) is then used to illustrate the theoretical virtues of the proposed model. Section 7 concludes this paper.

2 Component Knowledge and Interaction

Where $\mathcal{C}$ is a set of components, we have a modal operator $[c]$ for every c in $\mathcal{C}$. Thus, we say that

$$[c]\varphi \text{ is true at a state } w \text{ iff } wR_cw' \text{ implies that } \varphi \text{ is true at } w'.$$

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

$$c \text{ can exclude at } w \text{ all states where } \varphi \text{ is not true,}$$

or, equivalently, iff

$$\varphi \text{ is true at all epistemic alternatives to } w \text{ for } c.$$

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 $\mathcal{C}$) such that, where R_c^* is the reflexive, transitive and symmetric closure of R_c,

$$[c^*]\varphi \text{ is true at a state } w \text{ iff } wR_c^*w' \text{ implies that } \varphi \text{ is true at } w'.$$

This 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 $c \in \mathcal{C}$

$$c \text{ knows that } \varphi \text{ (henceforth } \mathsf{K}_c\varphi) \text{ at } w \text{ iff } [c^*]\varphi \text{ is true at } w.$$

Table 1 Modal-epistemic axioms

Label	Axiom	Frame condition
K	$\mathsf{K}_a(p \to q) \to (\mathsf{K}_a p \to \mathsf{K}_a q)$	/
T	$\mathsf{K}_a p \to p$	Reflexive
4	$\mathsf{K}_a p \to \mathsf{K}_a \mathsf{K}_a p$	Transitive
5	$\neg \mathsf{K}_a p \to \mathsf{K}_a \neg \mathsf{K}_a p$	Euclidean

The corresponding logic is obtained by adding the axioms K, T and 5 (see Table 1) as well as the rule of necessitation to classical propositional logic. The axiom 4 is a theorem of the resulting logic.

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.

Second, the best way in which the present modelling decision can be thought of is in terms of what components can communicate, and what they come to know by communicating. In particular, we need to focus on the higher-order (inter-component) knowledge that can be obtained by communicating; namely, the fact that a receiver of a message acquires knowledge of the sender's knowledge. Whenever a first component knows that φ, we presume that φ is knowledge that can be passed on to other components in such a way that if other components come to know that φ, they also come to know that the first component knows that φ. Yet, if that is the case, it would be quite odd to assume that it is easier to gain higher-order knowledge of other components' knowledge than to be an introspective agent (i.e. one's knowledge could—once communicated—be transparent to others, but not to oneself). As a result, the identification of component-knowledge with **S5**-knowledge is implied by the fact that we are only interested in component-knowledge that can be shared (see e.g. van Benthem [37, 57] on the assumption that knowledge requires the ability to inform others), and moreover can be shared in such a way that it can lead to higher-order (inter-component) knowledge.

2.1 Reliable Interaction

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, but 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; see Baltag and Moss [3] and Van Ditmarsch et al. [42] for an overview) 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. The related *transmission-thesis*, which states that if a knows that b knows that

φ, then a also knows that φ, was already mentioned in Hintikka [20, 4.1–2]. Later on, the validity of this thesis was related to sameness of goals, methods and standards by Hendricks [18, 148–50]. Since in the present context the transmission-thesis is one of our modelling-assumptions, there is no need to take a definite stance on these matters.

But just how strong is the assumption that all knowledge can be shared in such a way that it leads to knowledge and even to 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 rather 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. [3, 29, 42]). 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 is based on an intuitively strong assumption about the nature of knowledge, but the latter is also a common—and perhaps even indispensable—modelling assumption.

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. Component-knowledge 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 once we have an account 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.

Given the assumptions we made, component-knowledge and interaction can be characterised as follows:

1. All components are assumed to be perfect reasoners.
2. Component-knowledge that cannot be communicated is ignored.
3. We exclude the possibility that some components may act as *narrow-minded agents*, i.e. agents that are incapable of considering other agents (see [10], 60–1).

These consequences are not just harmless features of a formal model, but sound modelling assumptions. As previously mentioned, we do not need a realistic model to focus on the different aspects of closure and interaction discussed in the introduction; a more realistic model is perhaps even undesirable in this context.[2] The absence of less than perfect components, unsharable knowledge and narrow-minded components allows us to focus exclusively on how different patterns of interaction between components can have an impact on closure and introspection. Idealising the properties of the components and their interaction is just a way to ensure that the properties of component-knowledge cannot interfere with how the components are organised.

2.2 Order of Explanation

The just described refusal to take less than perfect components into account again raises the question of whether we're not reversing the order of explanation. Indeed, we not only use models of group-knowledge to explain typical single-agent features of knowledge, but (given the just described idealisation of component-knowledge) we also require 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

[2] This is also the reason why I can stick to a broadly Stalnakerian approach to the modelling of knowledge (and belief), and do not have to resort to more refined proposals like, for instance, the use of non-normal modal logics.

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 the modelling of knowledge in groups are also the kind of differences that are relevant to the modelling of the knowledge of individual agents (Section 4 is entirely devoted to this issue). 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 properties of knowledge are at least partly determined by how agents interact (a common position 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 emphasised in Minsky's "Society of Mind".

> To comprehend what knowing is, we have to guard ourselves against that single-agent fallacy of thinking that the 'I' in 'I believe' is actually a single, stable thing. The truth is that a person's mind holds different views in different realms. [25, 302]

Even then, accepting a model of individual knowledge that is based on existing forms of group knowledge does not require the adherence to a strong anti-individualism. Rather, the point I want to emphasise is that the decision about what counts as a single or individual agent is itself 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 consider the same outputs as the result of communication between these components. The same point is made by Abramsky when he claims that

> While information is presumably conserved in the *total* system, there can be information flow between, and information increase in, *subsystems*. (...)
> Thus if we wish to speak of information flow and increase, this must be done relative to subsystems. (...) Subsystems which can *observe* incoming information from their environment, and *act* to send information to their environment, have the capabilities of *agents*. [1, 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) that motivates the present proposal.

To model agents as a group of components is just a means to switch to a lower level of abstraction.[3]

2.3 Synchronic and Diachronic Interpretations

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, positive introspection means that an agent cannot know unless he also knows that he knows. This is a synchronic way of understanding the principle of positive introspection. The synchronic reading of closure and introspection principles is the intended reading for our model of component-knowledge.

Another interpretation of the same principle is this: When an agent knows, she doesn't require any external input to learn that she knows. This yields a diachronic reading of the principle of positive introspection. An analogous, and perhaps more familiar diachronic reading of closure can be formulated along the same lines. Both the synchronic and the diachronic readings of epistemic principles can and should be used to interpret the different types of group-knowledge. For instance, the fact that a strong form of group-knowledge like common knowledge is introspective should be read as a synchronic principle. By contrast, the weaker claim that non-introspective forms of group-knowledge can, without external input (i.e. communication with agents that do not belong to the relevant group of agents), be upgraded to stronger, introspective types of group-knowledge is best understood as a diachronic principle. As before, similar considerations apply to the interpretation of deductive closure.

The above description can be summarised as follows: The principle of positive introspection is valid on a diachronic reading 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 diachronic reading of introspection in terms of the existence of a communication-protocol that guarantees that one form of group-knowledge can be upgraded to a stronger one. That is, there should be a sequence of messages (described by the protocol) that can be sent between the different agents in the group such that, if the initial state is one where something is known in a group, the final state is one where this knowledge has

[3] This point of view can be compared to an idea voiced in van Benthem [38, 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.

become introspective knowledge. When such protocols are formalised with the tools of dynamic epistemic logic this means that the diachronic interpretations of epistemic principles can be formalised as dynamic properties of knowledge.

3 Knowledge in a Group

Where $\mathcal{G}$ is a finite subset of $\mathcal{C}$, we say that $\mathcal{G}$ 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, all of these notions are provably non-equivalent, and give rise to a hierarchy of forms of group-knowledge.[4] Traditionally, the hierarchy contains the following four types of group-knowledge: distributed knowledge, particular knowledge (someone knows), general knowledge (everybody knows), and common knowledge. Sometimes, a fifth type of knowledge is inserted in the middle of this hierarchy; namely knowledge by a specific agent. I shall ignore this type of group-knowledge because, first, its logical properties are just the logical properties of component-knowledge (see the previous section), and, second, because it is only of limited interest for the argument presented in this paper (see the reference to the 'wise man' in Section 5).

3.1 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

(D) φ is D-known in a group $\mathcal{G}$ iff each non φ world is at least excluded by some member of $\mathcal{G}$.

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 Hoek et al. [40] and Roelofsen [31], we say that the formal and the intuitive characterisation of distributed knowledge[5] coincide iff the *principle of full communication* is satisfied.

[4] 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 et al. [40], where the same term refers to the weakest kind of such knowledge.

[5] Again, the terminology in van der Hoek et al. [40] 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.

This is something we shall have to come back to. In this section we only focus on the formal properties and the interpretation of distributed knowledge. Its formal properties are easily summarised: If component-knowledge is **S5**-knowledge, then so is distributed knowledge;[6] it is deductively (as well as logically) closed, and is fully introspective. Unlike component-knowledge, distributed knowledge cannot readily be shared. 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 corresponds 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 ([16, 13–14] and [34]), 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 [23] and Fagin and Halpern [10] characterise the difference between implicit and explicit belief.

3.2 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 as follows:

(S) φ is S-known in a group $\mathcal{G}$ iff there is a member of $\mathcal{G}$ 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 shared,[7] 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

[6] The proof is straightforward: The accessibility relation for distributed knowledge is the intersection of the accessibility relations of the relevant agents, and the intersection of reflexive, symmetric, and transitive relations is also reflexive, symmetric and transitive.

[7] In Section 4 we shall have to refine this identification, but for now we assume that the guiding intuition is correct.

fact that it is a form of group-knowledge that is not deductively closed;[8] again a property that is typically associated with explicit knowledge.

3.3 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:

(E) φ is E-known (i.e. everybody knows) in a group $\mathcal{G}$ iff each member of $\mathcal{G}$ excludes all non φ worlds.

Alternatively, it can also be defined as the conjunction of all $\mathsf{K}_c\varphi$ for all c in $\mathcal{G}$. 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. In Section 4 we shall answer this question properly.

3.4 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 [24]). 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.

[8] 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.

Where $\mathcal{G}$ is a group of agents, define a $\mathcal{G}^k$ alternative with the following inductive clauses:

- A world w is a $\mathcal{G}^1$ alternative iff w is an epistemic alternative for some member of $\mathcal{G}$.
- A world w is a $\mathcal{G}^{k+1}$ alternative iff at some $\mathcal{G}^k$ alternative the world w is an epistemic alternative for some member of $\mathcal{G}$.

Using this notion, we can now stipulate that

(C) φ is C-known at w iff for any finite k, no non-φ world is a $\mathcal{G}^k$ 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's. 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. Common knowledge is again a form of **S5**-knowledge; it is deductively closed and fully introspective. 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 (though this description almost certainly needs further refinement).

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 is readily and explicitly available. Put differently 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.

3.5 Components as States

If, as suggested at several points in this section, we want to use different types of group-knowledge to discriminate between implicit and explicit forms of knowledge, a literal reading of components as "parts of the brain" may not be the most appropriate way of understanding what the different components may stand for. A more abstract understanding of the components that does not refer to an actual physical implementation may therefore be preferable.[9] This interpretation should do two things: (1) it should provide an interpretation of the components as different states of an agent, and (2) it should give an account of component-interaction in terms of the information-flow between these different states.

A first interpretation of components-as-states that readily comes up is a temporal one.[10] Though fruitful as a first approximation, I think this proposal is also misleading. If $\mathsf{S}_G\varphi$ is read as "φ is known at some temporal stage t by the agent G" we only focus on one among the many criteria that could be used to discriminate between the different states of information an agent can be in. This isn't the only issue. In addition, on the temporal reading E-knowledge becomes—even in a highly idealised model—too demanding to be a useful formalisation of a type of explicit knowledge.

The critique on the temporal interpretation suggests that it is often more fruitful to discriminate the different states of an agent along multiple dimensions. A generic way of referring to such states is as "frames of mind" [10, 59]. Models of belief that use a set of separate or non-interacting clusters of beliefs can be used to model agents that have non-trivial inconsistent beliefs.[11] The same method can, however, also be used to model the state of an agent which, though perfectly consistent, hasn't yet put all his information together. This can be used to represent knowledge that isn't deductively closed.

> The information which one receives when one learns about deductive relationships does not seem to come from outside of oneself at all. It seems to be information which, in some sense, one had all along. What one does is to transform it into a usable form, and that, it seems plausible to suppose, is a matter of putting it together with the rest of one's information. Stalnaker [33, 86]

This *putting-things-together* aspect of deduction matches to the diachronic interpretation of deductive closure described in Section 2, and is further

[9] As remarked by a referee, the reading as parts of the brain is especially objectionable when understood as anatomical parts, but less so when understood as functional parts. I do not further explore this path, and immediately opt for the more abstract notions of 'state' and 'frame of mind'.

[10] This suggestion is directly inspired by Sequoiah–Grayson [32].

[11] Remark that if we model mutually contradicting clusters of beliefs as the beliefs of different components, their D-beliefs will still be trivial. To say that the clusters do not interact can then be understood as the fact that trivial beliefs cannot become S-beliefs. See Restall [30] for a connection with impossible worlds and paraconsistent logics.

developed in Section 5. Even if in the case of knowledge the process of combining separate pieces of information does not require the resolution of inconsistencies, the process itself still requires some interaction between states. This is why we can model the process of deduction in the same way as we would model communicating agents.

Stalnaker's argument for postulating "a large number of concurrent but separate belief states" is sufficiently general to motivate an interpretation of components as states of an agent that need not correspond to either physical parts of that agent or to temporal stages. On his pragmatic-causal picture of belief, all it takes to keep two beliefs distinct (i.e. in different states of belief) is for there to be actions that are appropriate for one belief, other actions that are appropriate for another belief, and finally also some distinct actions that are appropriate for the conjunction of both beliefs [33, 86]. Since on this account having multiple belief-states is to be understood in terms of concurrent stable states rather than in terms of shifting between different unstable states, it seems that (1) we can model states as components, (2) that we can individuate states more finely than just as a temporal succession of states, and (3) that all states can (at least in principle) interact with each other. This is all we need to vindicate a generic interpretation of components based on the identification of component-knowledge with knowledge in a particular state.

4 A Hierarchy of Knowledge-Types

By defining types of group-knowledge, we have obtained a series of knowledge operators such 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 [15, 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 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 S-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.[12] 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 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.

4.1 Access and Storage

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

[12] 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.

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 [34, 435].

Unlike mainstream models of knowledge that 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 is 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 merely 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 available. 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. 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. If we keep in mind that the computational process is itself distributed, we immediately see that since the outcome of a distributed computational process doesn't have to be immediately accessible itself (though it will always be explicitly stored), there is nothing objectionable about the use of premises that are not readily available either.

4.2 Explicit Storage and the Belief-Box Metaphor

From a formal point of view the proposed hierarchy surely meets the requirements of 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 by speaking of explicitly stored information we not only subscribe to the belief-box metaphor, but are also forced to interpret component-knowledge in these terms. As a result, we're back where we started; we seem to presuppose a literal reading of components as the physical locations for the storage of beliefs. This is, as Rohit Parikh explains in a different but related context, an undesirable outcome.

> The representational account of belief simply seems wrong to me. We *can* certainly think of beliefs as being stored in the brain in some form and called forth as needed, but when we think of the details, we can soon see that the stored belief model is too meager to serve. I will offer an analogy. (…)
> It is what she *has*, namely the CD, the printer and the binder, and what she *can do*, namely print and bind, which together allow her to fulfill the order. There are elements of pure storage, and elements which are algorithmic which together produce the item in question. These two elements may not always be easy to separate. It is wiser just to concentrate on what she can supply to her customer.
> It is the same, in my view, with beliefs. No doubt there are certain beliefs which are stored in some way, but there may be other equally valid beliefs which may be produced *on the spot* so to say, without having been there to start with.(…)
> Retrieving from storage is *one* way to exhibit a belief, but not the only one, and often, not even the best one. [28, 466–7]

This yields the following dilemma: Either we have to identify component-knowledge with explicitly stored knowledge, or we cannot interpret S-knowledge as explicitly stored knowledge.

The problem with the first horn of the dilemma is, however, not quite the problem Parikh describes. This is primarily because what he objects to is the assumption that all beliefs need to be explicitly stored, while he thinks that most of our beliefs are just produced on the spot. But this is just a version of Stalnaker's point about a single distinction with a double task, which is one of the problems solved by a multi-component characterisation of knowledge. Rather, the problem is that we have independent reasons for not liking the literal reading of components as parts of the brain.

Fortunately, the multi-component characterisation of knowledge sufficiently changes the rules of the game to allow us to deal with this dilemma. Recall first the following assumptions made in the previous sections:

1. Component-knowledge is knowledge that can be shared by that component (Section 2).
2. Knowledge is explicitly stored iff it can be communicated by some component (Section 3).
3. If knowledge is explicitly stored *and* readily available then it can be communicated by any component.[13]

The crucial insight is that from these assumptions it doesn't follow that if φ is explicitly stored by the system, it is also explicitly stored by the component in virtue of which the system S-knows that φ. Since our model of component-knowledge cannot discriminate between knowledge that is explicitly stored by and readily available to a component, and knowledge that a component can produce on the spot, it remains indifferent with regard to the question of how components know. Still, given the logical properties of S-knowledge, component-knowledge will, within the system, behave like explicit knowledge. As a result, the totality of what is S-known within a system is explicitly stored in, but not necessarily readily available to the system, and this may hold even though only a part of that knowledge is explicitly stored by some component. Hence, there is no need to think of components as belief-boxes or other physical parts of the brain.

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 the description of diachronic forms of closure in

[13] The converse doesn't have to hold if we allow for the possibility of knowledge that is readily available because it can easily be derived by $n > 1$ components.

terms of the protocols needed for upgrading D-knowledge to S-knowledge. Diachronic forms of closure do in fact coincide with upgrading from a deductively closed to a *non*-deductively closed form of knowledge. Making deductions is a way to make explicit what was only implicitly there (which explains why diachronic forms of closure coincides with upgrading from D-knowledge to S-knowledge),[14] and this can be modelled as a dynamic process so that not everything becomes instantly explicit (which explains why the outcome, namely S-knowledge, isn't itself deductively closed in the synchronic sense).

Two crucial questions need to be answered. What kind of protocols do we need? And are these protocols 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 and Moses [14], van der Hoek et al. [40]; the protocols in question are described in van Linder et al. [45]),[15] but they could equally well send everything to everyone, or even set up a more complicated inference-network. Because we are interested in the (in principle) existence of protocols that allow the upgrade from implicit to explicit knowledge, we can safely ignore the specifics (including the computational complexity) of those protocols.

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.

5.1 Full-Communication

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,

[14] The following synchronic closure-rule, which is described in Palczewski [27, 458], makes it quite clear that the logical consequences of S-knowledge are only D-known, and hence implicit: $(\varphi_1 \wedge \ldots \wedge \varphi_n) \rightarrow \psi / (\mathsf{S}\varphi_1 \wedge \ldots \wedge \mathsf{S}\varphi_n) \rightarrow \mathsf{D}\psi$.

[15] Given our prior assumptions about component-knowledge, if all knowledge is passed on to a unique wise man component, that knowledge will automatically be deductively closed.

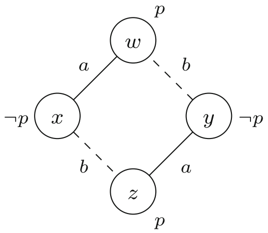

Fig. 1 Failure of full-communication

we see that it actually requires something weaker. Formally, the principle of full communication says that for all $\varphi \in \mathcal{L}_K$ (the basic epistemic language with knowledge-operators K_a for all agents, but without operators for any kind of group-knowledge) we have that:

$$\mathsf{D}\varphi \Longrightarrow \{\varphi_i \in \mathcal{L}_K : \mathsf{S}\varphi_i\} \vdash \varphi$$

The first obstacle to upgrading D-knowledge to S-knowledge is due to failures of this principle. Consider, for that matter, Fig. 1 (based on [31]) 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. 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.

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 Kripke-models where worlds compatible with an agent's knowledge fail to be epistemic alternatives. For instance, in the above example the world y 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 [31], 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. Put informally, this means 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.

5.2 Fitch-like Phenomena

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. These issues form the second obstacle to upgrading D-knowledge to S-knowledge—an obstacle that arises for each further form of upgrading (though I shall not describe these in detail). 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 (and only for a fragment of the language); 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 appears to tell us is that we cannot have an unrestricted knowability-principle that doesn't also lead to a collapse of knowability with actual knowledge, and thus yields factual omniscience. A similar phenomenon arises for upgradability, but let us consider the original paradox first (see Brogaard and Salerno [6] for an overview).

The proof is based on an unrestricted knowability principle that expresses that all truths can be known

$$\forall p(p \to \diamond \mathsf{K}p), \tag{KP}$$

and the intuitively true assumption that there are unknown truths

$$\exists p(p \wedge \neg \mathsf{K}p). \tag{NonO}$$

For (NonO) to be true, one of its instances needs to be true as well. This is the first assumption in the proof below. Analogously, if (KP) is true, then so its instance $(p \wedge \neg \mathsf{K}p) \to \Diamond \mathsf{K}(p \wedge \neg \mathsf{K}p)$. This is the second assumption in our proof.

$$\dfrac{\dfrac{p \wedge \neg \mathsf{K}p \quad (p \wedge \neg \mathsf{K}p) \to \Diamond \mathsf{K}(p \wedge \neg \mathsf{K}p)}{\Diamond \mathsf{K}(p \wedge \neg \mathsf{K}p)}\,\text{MP} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[\mathsf{K}(p \wedge \neg \mathsf{K}p)]^{(1)}}{\mathsf{K}p \wedge \mathsf{K}\neg \mathsf{K}p}}{\mathsf{K}p \wedge \neg \mathsf{K}p}}{\neg \mathsf{K}(p \wedge \neg \mathsf{K}p)}\,\text{RAA}(1)}{\Box \neg \mathsf{K}(p \wedge \neg \mathsf{K}p)}\,\text{Nec}}{\neg \Diamond \mathsf{K}(p \wedge \neg \mathsf{K}p)}}{\bot}$$

What this proof shows is that the joint truth of instances of (KP) and (NonO) leads to contradiction. They can thus not be upheld together. Since the rejection of (NonO) leads to the absurd conclusion that all truths are known, it is standardly assumed that (KP) cannot unrestrictedly be valid.

Can we arrive at a similar result for upgradability? Given some modifications this is indeed possible. As before, we start from a knowability-principle (SKP),[16] and a claim about the existence of merely implicit knowledge (MI).

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

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

As for the standard proof of Fitch's paradox, for (MI) to be true, one of its instances should be true, while for (SKP) to be true, all its instances should be true. We use these instances in the first part of the proof to derive $\Diamond\mathsf{S}(\mathsf{D}p \wedge \neg\mathsf{S}p)$.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\mathsf{D}p \wedge \neg\mathsf{S}p}{\mathsf{D}p}}{\mathsf{DD}p} \qquad \dfrac{\dfrac{\mathsf{D}p \wedge \neg\mathsf{S}p}{\neg\mathsf{S}p}}{\mathsf{D}\neg\mathsf{S}p}(*)}{\mathsf{DD}p \wedge \mathsf{D}\neg\mathsf{S}p}}{\mathsf{D}(\mathsf{D}p \wedge \neg\mathsf{S}p)} \qquad \mathsf{D}(\mathsf{D}p \wedge \neg\mathsf{S}p) \rightarrow \Diamond\mathsf{S}(\mathsf{D}p \wedge \neg\mathsf{S}p)}{\Diamond\mathsf{S}(\mathsf{D}p \wedge \neg\mathsf{S}p)}\,\mathrm{MP}$$

Contrary to what we find in the standard proof of Fitch's paradox, our chosen instance of (MI) doesn't have the right form to be the antecedent of some instance of (SKP). As a consequence, the proof contains an additional branch with the derivation of $\mathsf{D}(\mathsf{D}p \wedge \neg\mathsf{S}p)$ from $\mathsf{D}p \wedge \neg\mathsf{S}p$. The only potentially controversial move in this sub-derivation is the $*$-labelled step from $\neg\mathsf{S}p$ to $\mathsf{D}\neg\mathsf{S}p$. In view of our prior assumptions about component-knowledge, this step is indeed valid. If $\neg\mathsf{S}p$, then $\neg\mathsf{K}_a p$ as well as $\mathsf{K}_a\neg\mathsf{K}_a p$ hold for each agent in the relevant group, and this allows us to conclude[17]

$$\mathsf{D}\bigwedge_{i\in G}\neg\mathsf{K}_i p. \tag{†}$$

Since D is a normal modal operator, and the big conjunction in (†) is provably equivalent to $\neg\mathsf{S}p$, we obtain $\mathsf{D}\neg\mathsf{S}p$ as required.

[16] Compare with the proposals in Palczewski [27] and Balbiani et al. [2].

[17] The G in † refers to the relevant group of agents. Since we do not consider different groups of agents we write D and S instead of D_G and S_G.

To derive $\neg\Diamond\mathsf{S}(\mathsf{D}p \land \neg\mathsf{S}p)$ as the second part of our proof, we only need to extend the second part of the original proof with an additional branch to allow for the derivation of $\mathsf{S}p$ from $\mathsf{SD}p$.[18]

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[\mathsf{D}p \to p]^{(\text{Axiom})}}{\mathsf{E}(\mathsf{D}p \to p)}\,\text{Nec}^{\mathsf{E}} \quad \dfrac{\dfrac{[\mathsf{S}(\mathsf{D}p \land \neg\mathsf{S}p)]^{(1^*)}}{\mathsf{SD}p \land \mathsf{S}\neg\mathsf{S}p}}{\mathsf{SD}p}}{\mathsf{S}p}\,(\star) \quad \dfrac{\dfrac{\dfrac{[\mathsf{S}(\mathsf{D}p \land \neg\mathsf{S}p)]^{(1^*)}}{\mathsf{SD}p \land \mathsf{S}\neg\mathsf{S}p}}{\mathsf{S}\neg\mathsf{S}p}}{\neg\mathsf{S}p}}{\bot}}{\neg\mathsf{S}(\mathsf{D}p \land \neg\mathsf{S}p)}\,\text{RAA}(1^*)}{\Box\neg\mathsf{S}(\mathsf{D}p \land \neg\mathsf{S}p)}\,\text{Nec}^{\Box}}{\neg\Diamond\mathsf{S}(\mathsf{D}p \land \neg\mathsf{S}p)}$$

Putting both halves together, we obtain the following impossibility-result

$$((\forall p(\mathsf{D}p \to \Diamond\mathsf{S}p)) \land \exists p(\mathsf{D}p \land \neg\mathsf{S}p)) \to \bot,$$

which is a real Fitch-like paradox for upgradability: 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 itself becomes S-known. Remark, however, that unlike for Fitch's original impossibility result, we do need to make more substantial assumptions; most importantly the assumption that all components are **S5**-knowers, which is required to deduce $\mathsf{DD}p$ from $\mathsf{D}p$, and $\mathsf{D}\neg\mathsf{S}p$ from $\neg\mathsf{S}p$.

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 is an open problem 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 [41] give an overview of different partial criteria for success), but there is also evidence that syntactic restriction strategies can lead to further paradoxes [6, 5.3].

5.3 Making Knowledge Accessible and Restricted Introspection

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

[18] The $\star$-labelled derivation of $\mathsf{S}p$ from $\mathsf{SD}p$ and $\mathsf{E}(\mathsf{D}p \to p)$ uses an unproblematic derived rule (see Section 4.1). It can be derived as a proof by cases from $\bigwedge_{a\in G}\mathsf{K}_a(\mathsf{D}_G p \to p)$ and $\bigvee_{a\in G}\mathsf{D}_G p$.

is a 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 obtaining 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 to 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. 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.

5.4 Common Knowledge and Public Announcements

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. [11]); the second terminology is standard in the field of dynamic epistemic logics [3, 42]. It is a well-known fact that 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,[19] 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. [44], Appendix 2). What this reveals is that upgrading 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 types 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 (i.e. the information-flow between different states of the agent) is thus configured that the required form of public communication (i.e. completely transparent information-flow between the different states of the agent) 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.

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

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 $\mathcal{G}$ 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 et al. [43]). 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. By the same token, taking the transitive closure of a group's epistemic accessibility relations suffices 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 [4]. 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. [4, 368]

When applied to seeing, Barwise's suggestion implies 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 only shared informational contexts can warrant common knowledge, and that 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 than any other form of group-knowledge, and, since we've argued that fully introspective knowledge shares 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 (in principle) 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 a single component 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, the latter condition isn't as easily satisfied.

6 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 in a multi-component setting. 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 $\varphi \rightarrow \psi$) is valid we should try to find out whether the set $\{\varphi, \neg\psi\}$ is defensible, and only conclude that $\varphi \rightarrow \psi$ is valid when $\{\varphi, \neg\psi\}$ is indefensible. In its most general form, this means 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 [35], and is reproduced below.

1. If $\{\mathsf{K}_a\varphi, \neg\mathsf{K}_a\neg\psi\}$ is consistent, then $\{\mathsf{K}_a\varphi, \psi\}$ is also consistent. Hence, by substituting $\neg\mathsf{K}_a\varphi$ for ψ, we obtain:
2. If $\{\mathsf{K}_a\varphi, \neg\mathsf{K}_a\neg\neg\mathsf{K}_a\varphi\}$ is consistent, then $\{\mathsf{K}_a\varphi, \neg\mathsf{K}_a\varphi\}$ is also consistent. Which after eliminating the double negation, and taking the contrapositive gives us:
3. Since $\{\mathsf{K}_a\varphi, \neg\mathsf{K}_a\varphi\}$ is inconsistent, $\{\mathsf{K}_a\varphi, \neg\mathsf{K}_a\mathsf{K}_a\varphi\}$ 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 $\{\mathsf{K}_a\varphi, \neg\mathsf{K}_a\mathsf{K}_a\varphi\}$ is satisfiable in non-transitive Kripke-frames. 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. Formally, the conditional in the first line of the above proof is equivalent to the following satisfaction-clause for K_a

$$\mathsf{K}_a\varphi \text{ is true at } w \text{ iff } wR_aw' \text{ implies that } \mathsf{K}_a\varphi \text{ is true at } w',$$

which is itself equivalent to the standard clause for K_a with the further assumption that R_a is a transitive relation. The motivation for this clause, and thus for the reasoning behind the first line of Hintikka's proof is that if one does not know $\neg\psi$, then ψ should not only be consistent with what one knows to be true (i.c. φ), but also with the fact that one knows φ. This line of reasoning is thus 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 Vincent Hendricks 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 auto-epistemology on Hintikka's own formulation of epistemic logic [18, 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 that p," given that this conjunction isn't inconsistent?). What Hintikka seems to argue for 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.[20] To formulate a version of Hintikka's argument for positive introspection that fits into the multi-component characterisation of knowledge, we thus need a sentence that denies positive introspection, but also turns out to be unknowable. These sentences can then be used to compare different sorts of positive introspection.

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

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

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

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

Since (C) can be dismissed right away (no-one will deny that common knowledge is introspective, and thus that (C) is a contradiction), we can restrain our attention to (E) and (EC). Next, we should note that (EC) is implied by (E). More exactly, it is implied by each instance of $\mathsf{E}p$, but $\neg\mathsf{E}^k p$ with finite k. As a result, (EC) can be seen as the weakest denial of positive introspection.

[20] I'm here assuming the connection between knowability and 'actions that make us know' first investigated by van Benthem [36, 39].

To find out whether (E) or (EC) are 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 $\neg\mathsf{EE}p$. But is it also unknowable? If we assume that a already knows that b knows that p, then the announcement of "$\mathsf{E}p$, but $\neg\mathsf{EE}p$" by a (or by a third agent) is true but unsuccessful. In other words, in that situation $\mathsf{E}p$, but $\neg\mathsf{EE}p$ is an unknowable truth. If, by contrast, the situation is such that a and b both 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).[21] Taking the two examples together, it follows that the sentence is knowable in the sense expressed by ($\Diamond$SE), but not in the sense expressed by ($\Diamond$EE)

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

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

On the assumption that all components can send messages to all other components, this last insight readily 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), and hence the announcement of $\mathsf{E}p \wedge \neg\mathsf{C}p$ can only be unsuccessful if it makes $\mathsf{C}p$ true. The latter effect can only be the result of a public announcement. Again, there are two knowability-claims that can be considered.

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

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

If knowability is understood as "there is a way to make this announcement in a successful manner," then ($\Diamond$EC) is true in virtue of the possibility to announce $\mathsf{E}p \wedge \neg\mathsf{C}p$ (semi-)privately to all components. If knowability only refers to what is knowable after public announcements, then ($\Diamond$EC) is false because such announcements induce common knowledge, and therefore lead

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

to $\mathsf{C}(\mathsf{E}p \wedge \neg\mathsf{C}p)$ which is inconsistent. The latter immediately shows that ($\Diamond$CC) 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* [20, 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 $\{\mathsf{K}_a\varphi, \neg\mathsf{K}_a\neg\psi\}$ is consistent only if $\{\mathsf{K}_a\varphi, \psi\}$ is also consistent, that connection is much weaker on the more refined model. The typical auto-epistemic considerations can still be expressed in terms of knowability and how components may communicate, but they can only be used to dismiss a limited number of denials of positive introspection. Hence, since some such denials are knowable, they are also defensible in Hintikka's sense.

7 Concluding Remarks

In this paper I presented a multi-component characterisation of knowledge, and used this to distinguish different aspects of closure and introspection. My main claim is that the additional logical distinctions provided by the different types of group-knowledge are sufficient to discriminate between more and less demanding readings of deductive closure and introspection. For instance, easy closure and introspection are exemplified by D-knowledge, while more demanding forms of closure and introspection are exemplified by, respectively, E-knowledge and C-knowledge. Additionally, by focusing on how information is distributed among different components and how component-interaction alters this distribution, we can account for the difference between single and multi-premise closure and also integrate diachronic readings of closure and introspection in our model (upgrading weaker forms of group-knowledge to stronger forms of group-knowledge).

The multi-component characterisation delivers more than just a refined account of closure and introspection. It also provides (a) a distinction between explicit knowledge and readily available knowledge that is necessary to deal with an objection due to Stalnaker against having a single implicit/explicit distinction; (b) a new Fitch-like result; and (c) a natural connection between different types of announcements and a principled gap between fully transparent or introspective knowledge, and limited introspective knowledge.

A final virtue of the general methodology behind the multi-component characterisation of introspection is illustrated by the discussion of Hintikka's original argument for positive introspection. This example not only shows that considerations about how components interact can shed a light on auto-epistemic criteria, but it also provides an example of how recent developments in modal epistemic logic and its dynamic extensions can be used to approach typical problems in mainstream epistemology. Such fruitful connections were already established for Fitch's paradox by van Benthem [36, 39]. In this paper these connections are extended to include the topics of closure and introspection.

Acknowledgements The present paper started its life in 2007 as a short note on positive introspection. It has since grown thanks to the feedback and encouragements I received from many people. I would specifically like to thank Luciano Floridi, Mark Jago, Sebastian Sequoiah-Grayson, Sonja Smets and Jean Paul Van Bendegem, several referees (both from the JPL and from other venues), and the audiences at the Seventh Formal Epistemology Workshop in Konstanz, and at an internal IEG-seminar in Oxford.

References

1. Abramsky, S. (2008). Information, processes and games. In J. Van Benthem & P. Adriaans (Eds.), *Handbook on the philosophy of information* (pp. 483–550). Amsterdam: Elsevier.
2. Balbiani, P., Baltag, A., van Ditmarsch, H., Herzig, A., Hoshi, T., & De Lima, T. (2008). Knowable as known after an announcement. *The Review of Symbolic Logic, 1*(3), 305–334.
3. Baltag, A., & Moss L. S. (2004). Logics for epistemic programs. *Synthese, 139*(2), 165–224.
4. Barwise, J. (1988). *Three views of common knowledge, TARK II*. Pacific Grove, California.
5. Bonnay, D., & Egré, P. (2009). Inexact knowledge with introspection. *Journal of Philosophical Logic, 38*(2), 179–227.
6. Brogaard, B., & Salerno, J. (2009). Fitch's Paradox of Knowability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.
7. Chellas, B. F. (1980). *Modal logic: An introduction*. Cambridge: Cambridge University Press.
8. Danto, A. C. (1967). On knowing that we know. In A. Stroll (Ed.), *Epistemology. New essays on the theory of knowledge* (pp. 32–53). New York: Harper and Row.
9. Dretske, F. (1970). Epistemic operators. *The Journal of Philosophy, 76*(24), 1007–1023.
10. Fagin, R., & Halpern, J. Y. (1988). Belief, awareness, and limited reasoning. *Artificial Intelligence, 34*, 39–76.
11. Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge*. Cambridge/London: MIT Press.
12. Gochet, P., & Gribomont, P. (2006). Epistemic logic. In D. M. Gabbay & J. Woods (Eds.), *Handbook of the history of logic* (Vol. 6, pp. 99–195). Elsevier.
13. Halpern, J. Y. (1996). Should knowledge entail belief? *Journal of Philosophical Logic, 25*(5), 483–494.
14. Halpern, J. Y., & Moses, Y. (1985). A guide to the modal logics of knowledge and belief. In *Proceedings of IJCAI- 85* (pp. 480–490). Los Angeles, CA.
15. Halpern, J. Y., & Moses, Y. (1990). Knowledge and common knowledge in a distributed system. *Journal of the Association for Computing Machinery, 37*(3), 549–587.
16. Harman, G. (1986). *Change in view. Principles of reasoning*. Cambridge: MIT.
17. Hawthorne, J. (2004). *Knowledge and lotteries*. Oxford: Oxford University Press.
18. Hendricks, V. (2006). *Mainstream and formal epistemology*. Cambridge: Cambridge University Press.
19. Hilpinen, R. (1970). Knowing that one knows and the classical definition of knowledge. *Synthese, 21*(2), 109–132.
20. Hintikka, J. (1962). *Knowledge and belief. An introduction to the logic of the two notions*. Ithaca: Cornell University Press.
21. Hintikka, J. (1970). Knowing that one knows, reviewed. *Synthese, 21*(2), 141–162.
22. Lemmon, E. J. (1967). If I know, do I know that i know? In A. Stroll (Ed.), *Epistemology. New essays on the theory of knowledge* (pp. 54–82). New York: Harper and Row.
23. Levesque, H. J. (1984). A logic of implicit and explicit belief. *National conference on artificial intelligence*. Houston, Texas.
24. Lewis, D. (1969). *Convention. A philosophical study*. Cambridge: Harvard University Press.
25. Minsky, M. (1987). *The society of mind*. London: Willian Heineman.
26. Nozick, R. (1981). *Philosophical explanations*. Cambridge: Harvard University Press.

27. Palczewski, R. (2007). Distributed knowability and Fitch's paradox. *Studia Logica, 86*(3), 455–478.
28. Parikh, R. (2008). Sentences, belief and logical omniscience, or what does deduction tell us? *The Review of Symbolic Logic, 1*(4), 459–76.
29. Plaza, J. (2007). Logics of public communications. *Synthese, 158*(2), 165–179.
30. Restall, G. (1997). Ways things can't be. *Notre Dame Journal of Formal Logic, 38*(4), 583–596.
31. Roelofsen, F. (2006). Distributed knowledge. *Journal of Applied Non-classical Logics, 16*(2), 255–273.
32. Sequoiah-Grayson, S. (2011). Epistemic closure and commutative, nonassociative residuated structures. *Synthese*. doi:10.1007/s11229-010-9834-z.
33. Stalnaker, R. (1984). *Inquiry*. Cambridge: MIT Press.
34. Stalnaker, R. (1991). The problem of logical omniscience, I. *Synthese, 89*(3), 425–440.
35. Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies, 128*(1), 169–199.
36. van Benthem, J. (2004). What one may come to know. *Analysis, 64*(282), 95–105.
37. van Benthem, J. (2006). Epistemic logic and epistemology: The state of their affairs. *Philosophical Studies, 128*(1), 49–76.
38. van Benthem, J. (2008). Logical dynamics meets logical pluralism? *Australasian Journal of Logic, 6*, 182–209.
39. van Benthem, J. (2009). Actions that make us know. In J. Salerno (Ed.), *New essays on the knowability paradox* (pp. 129–146). Oxford: Oxford University Press.
40. van der Hoek, W., van Linder, B., & Meyer, J. J .C. (1999). Group knowledge is not always distributed (neither is it always implicit). *Mathematical Social Sciences, 38*, 215–240.
41. Van Ditmarsch, H., & Kooi, B. (2006). The secret of my success. *Synthese, 151*(2), 201–232.
42. Van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2007). *Dynamic epistemic logic*. Dordrecht: Springer.
43. van Ditmarsch, H., van Eijck J., & Verbrugge, R. (2009). Common knowledge and common belief. In J. van Eijck, & R. Verbrugge, (Ed.), *Discourses on social software* (pp. 107–32). Amsterdam: Amsterdam University Press.
44. van Eijck, J., & Wang, Y. (2008). Propositional dynamic logic as a logic of belief revision. *Logic, language, information and computation* (pp. 136–148).
45. Van Linder, B., van der Hoek, W., & Meyer, J. J. Ch. (1994). Communicating rational agents. In B. Nebel & L. Dreschler-Fischer (Eds.), *KI-94: Advances in artificial intelligence* (pp. 202–213). New York: Springer.
46. Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.