A General Approach To Multi-Agent Minimal Knowledge

With Tools And Samples

Wiebe van der Hoek (`wiebe@cs.uu.nl`)
Institute of Information and Computing Sciences & Philosophy Dept., Utrecht University
P.O. Box 80089, 3508 TB Utrecht, the Netherlands
Dept. of Computer Science, The University of Liverpool, UK

Elias Thijsse (`thysse@kub.nl`)
Computational Linguistics & AI, Faculty of Arts, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, the Netherlands

1997/01/23

Abstract. We extend our general approach to characterizing information to multi-agent systems. In particular, we provide a formal description of an agent's knowledge containing exactly the information conveyed by some (honest) formula φ. Only knowing is important for dynamic agent systems in two ways. First of all, one wants to compare different states of knowledge of an agent and, secondly, for agent a's decisions, it may be relevant that (he knows that) agent b does not know more than φ. There are three ways to study the question whether a formula φ can be interpreted as minimal information. The first method is semantic and inspects 'minimal' models for φ (with respect to some information order $\leq$ on states). The second one is syntactic and searches for stable expansions, minimal with respect to some language $\mathcal{L}^*$. The third method is a deductive test, known as the disjunction property. We present a condition under which the three methods are equivalent. Then, we show how to construct the order $\leq$ by collecting 'layered orders'. Focusing on the multi-agent case we identify languages $\mathcal{L}^*$ for various orders $\leq$, and show how they yield different notions of honesty for different multi-modal systems. We then provide several tools for studying honesty types and illustrate their usefulness on a number of examples, for three modal systems of particular interest. Finally, we relate the different notions of minimal knowledge, and describe possible patterns of honesty for these systems.

Keywords: multi-agent minimal knowledge, honesty, disjunction property, minimal models, stable expansions, modal logic

Mathematics Subject Classification: 03B42,03B70,68T27

1. Introduction

What is a knowledge state? To answer this question, we give a general approach to characterizing information in a modal context. In particular, we want to obtain a formal description of an agent's knowledge containing *exactly*, that is, at least but not more than the information conveyed by some formula φ, in other words, the case in which φ is the agent's *only* knowledge. Characterizing an agent's exact knowledge state is important in dynamic agent systems in several ways. First of all, when the system evolves, one might wish to compare the different states of one agent: which actions (or, more specifically moves) optimally extend his knowledge? Secondly, in multi-agent systems, agent a may wish to be sure that *all that* b knows is φ, and exploit the fact that b does *not know more than that*. Finally, when such agents start to *exchange information*, they must be aware of principles governing their communication: Usually utterances constitute minimal information, conveying the speaker's (maximal) knowledge with respect to the relevant part of some domain (Grice's maxim of quantity).

Formulas φ representing all that the agent knows, are called *honest*. For the one-agent case, some observations about only knowing and honesty are well-accepted. For instance, where purely objective formulas are rendered honest, a typical example of a dishonest formula is $\varphi = (\Box p \vee \Box q)$: if an agent claims to only know φ, he would know something that is stronger than φ (i.e., either $\Box p$ or $\Box q$). A more sophisticated analysis of honesty generally depends on the epistemic background logic. What is especially important here, is which introspective capacities we are ready to attribute to the agent. For example, if the background logic contains the axiom of *positive* introspection $\Box\psi \to \Box\Box\psi$ we can infer $\Box\Box p$ if only p is known. This seems innocent since the inferred knowledge is still related to the initial description p. On the other hand, if we accept the axiom of *negative* introspection $\neg\Box\psi \to \Box\neg\Box\psi$, then we can infer knowledge concerning q, for example $\Box\neg\Box q$, from only knowing p. This knowledge cannot be derived from only knowing $p \wedge q$, which intuitively represents *more* knowledge than only knowing p. As we stressed in [6], this kind of inferences effects the treatment of honesty for different modal systems.

For the multi-agent case, intuition seems to be much less clear. Of course, where objective formulas are all honest in the one agent case, this property is easily convertible to formulas with no operator $\Box_a$, when considering honesty for agent a. Hence, a can honestly claim to only know $\Box_b p \vee \Box_b q$, for $b \neq a$. But if $\Box_a$ re-occurs in the scope of $\Box_b$, the resulting formula $\Box_b p \vee \Box_b \Box_a q$ becomes *dis*honest for a if $\Box_a$ represents (true) knowledge and $\Box_b$ negatively introspective knowledge. With mixed operators, in particular in the presence of negation, matters soon get fuzzy.

Studies of 'only knowing' ([3, 12]) and 'all I know' ([9]) have largely been restricted to particular modal systems, such as **S5**, **S4** and **K45**. Recently Halpern [2] has also taken other modal systems such as **K**, **T** and **KD45** into account. Although his approach suggests similar results for e.g. **KD4**, in [6] we adopted a more general perspective: *given any modal system, how to characterize the minimal informational content of modal formulas*. For *multi-agent only knowing*, we only know of a (more or less) general approach by Halpern ([2]), putting a notion of 'possibility' to work on tree models, and, for the $\mathbf{S5_m}$ case, enriching the language with modal operators Q_i^{ζ}, for any formula ζ and agent i.

In this paper, besides *arbitrary* normal multi-modal systems we prefer to use standard Kripke models, instead of Fagin and Vardi's knowledge structures, and Halpern's tree models. We try to obtain this general view by putting our framework of [6] to work for the multi-agent case. In order to appreciate this fully, the reader has to realize that, in general, there are three ways to study the question whether a formula φ allows for a minimal interpretation (if $\Box\psi$ allows for such a minimal interpretation, ψ is called honest), and, if so, what can be said about the consequences of φ under this interpretation.

The first approach is a semantic one: Given a formula φ, try to identify models for φ that carry the least information. This approach requires a suitable order $\leq$ between states (i.e., model-world pairs) in order to identify minimal (or, rather, *least*) elements. For the simple (universal) **S5**-models the order coincides with the superset-relation between sets of worlds. Our challenge here is to give a general definition of such an order, which suits any multi-modal system. The second approach is mainly syntactic in nature and presupposes a sublanguage $\mathcal{L}^*$ of 'special' formulas. Given a consistent formula φ, we then try to find a maximally consistent set containing φ with a smallest $\mathcal{L}^*$-part. This approach can be identified as the search for so-called stable expansions, which are related to maximally consistent sets in a straightforward way. The last approach is purely deductive, and is also known as the disjunction property (DP): φ allows for a minimal interpretation if for any disjunction in $\mathcal{L}^*$ that can be derived from φ, one of the disjuncts is derivable from φ.

In [6], we were able to formulate a condition under which the three approaches mentioned above are equivalent. This paves the way to focus on defining 'suitable' orders on information states in a general way, rather than trying to establish the equivalences of the characterizations for specific orders, again and again. The information orders on states that we consider are induced by *layered orders* $\leq_n$ between states, where n settles the depth of the equivalence.

For the one-agent case, we obtained minimality results with respect to the following languages (by considering appropriate orders on states):

– $\mathcal{L}^* = \Box\mathcal{L}$, where $\mathcal{L}$ is the full modal language (*general honesty*). However, it appears that under this choice, almost every formula is honest in the systems **K, K4, KD** and **KD4**. On the other hand, for many other systems there are no honest formulas. So for most systems, the notion of general honesty is trivial: all or no formulas are honest.

– $\mathcal{L}^* = \Box\mathcal{L}^+$, where $\mathcal{L}^+$ is the modal language where no $\Box$ occurs in the scope of a negation (*positive honesty*). For **S5**, the corresponding notion of honesty coincides with the approach in [3]. Moreover, for all systems except **K, KD, K4** and **KD4**, this notion of honesty is not trivial.

For the multi-agent case, there are many more options. Generalizing the first language above gives rise to a notion of honesty which encounters, *mutatis mutandis*, the same problem of trivialization as states for the one agent case. The second, so-called positive language can be generalized in different ways, which for most systems lead to nontrivial notions of honesty. The anomalous cases are still the *weak doxastic logics* (generalizing an observation of Halpern in [2]): in $\mathbf{KD_m}$ and $\mathbf{KD4_m}$ all formulas φ for which $\Box_a\varphi$ is consistent, are honest; in $\mathbf{K_m}$ and $\mathbf{K4_m}$, even all formulas are honest. This means that in $\mathbf{KD4_m}$, for example, agent a can honestly claim that he only knows whether he knows p, which we believe to be counter-intuitive.

In this paper, we present three generalizations of positive honesty: one corresponds to the 'a-objective' language (the formulas only contain $\Box_a$ in the scope of some $\Box_b$, with $a \neq b$), one to the 'a-positive' language (no $\Box_a$ in the scope of negation), and one that combines these two.

2. Modal logical preliminaries

Let us agree on some technicalities. Our multi-modal language $\mathcal{L}$ or $\mathcal{L}^{\mathcal{A}}$ has finitely many modal operators $\Box_1, \Box_2, \ldots, \Box_m$, over a finite set of atoms $\mathcal{P} = \{p, q, r \ldots\}$, using the classical connectives $\neg$, $\wedge$ and $\vee$. Here $\{1, 2, \ldots, m\}$ encodes the set of agents $\mathcal{A}$. We use a for an arbitrary agent in $\mathcal{A}$ on which we focus. The operator $\Box_a$ denotes "Agent a has the information that ...", which may involve knowledge, belief or any other propositional attitudes. The dual modal operators $\Diamond_1, \Diamond_2, \ldots, \Diamond_m$ are introduced by definition: $\Diamond_a\varphi = \neg\Box_a\neg\varphi$. Given a set of formulas Γ, we define a's knowledge about Γ by $\Box_a\Gamma = \{\Box_a\varphi \mid \varphi \in \Gamma\}$ and a's knowledge in Γ by $\Box_a^-\Gamma = \{\varphi \mid \Box_a\varphi \in \Gamma\}$.

A measure of modal complexity of formulas, called *modal depth*, has the usual recursive definition: $d(p) = 0$ (for $p \in \mathcal{P}$), $d(\neg\varphi) = d(\varphi)$, $d(\varphi \wedge \psi) = d(\varphi \vee \psi) = \max\{d(\varphi), d(\psi)\}$ and $d(\Box_i\varphi) = d(\varphi) + 1$. We often consider the sublanguage of formulas of limited modal depth: $\mathcal{L}_{(n)} = \{\varphi \in \mathcal{L} \mid d(\varphi) \leq n\}$. So, $\mathcal{L}_{(0)}$ is the purely propositional subset of $\mathcal{L}$ (void of modal operators). Other sublanguages of interest will be defined in the sequel.

We use multi-modal Kripke models $\langle W, R_1, \ldots, R_m, V\rangle$ or $\langle W, \vec{R}, V\rangle$ to interpret $\mathcal{L}$; here wR_av or $v \in R_a[w]$ means that in world w, world v is an epistemic alternative to a. Truth is relative to a model-world pair ('state', for short). The connectives $\neg$, $\wedge$ and $\vee$ are interpreted as usual; the modal operators also get the classical interpretation: $M, w \models \Box_a\varphi$ iff for all $v \in R_a[w] : M, v \models \varphi$. The theory of a state $\langle M, w\rangle$ is $\mathrm{Th}(M, w) = \{\varphi \mid M, w \models \varphi\}$. If the model is obvious from the context, we will omit it and simply write $w \models \varphi$. Consequence is defined relative to a given set of models $\mathcal{S}$: $\Gamma \models_{\mathcal{S}} \varphi$ iff $M, w \models \varphi$ for all $M \in \mathcal{S}$ s.t. $M, w \models \Gamma$. States are assumed to be related by what we call an *information order* $\leq^a$ for any agent a, the intuitive content being that $w \leq^a w'$ if a knows more in w' than in w. For the time being $\leq^a$ is only required to be a pre-order (i.e. reflexive and transitive). A major question is which formulas are preserved moving from w to w' if $w \leq^a w'$. It will prove important to single out so-called *persistent* sublanguages of such formulas, in particular those that are rich enough to reversely characterize the information order.

The inference relation $\vdash$ is obtained relative to a modal system **S**, which at least contains classical propositional logic and the rule defining the minimal system **K**: $\Gamma \vdash \varphi \Rightarrow \Box_a \Gamma \vdash \Box_a \varphi$. Formulas φ and ψ are *equivalent in* **S**, if both $\varphi \vdash_{\mathbf{S}} \psi$ and $\psi \vdash_{\mathbf{S}} \varphi$. The logics **S** that we consider have the nice property that $\mathcal{L}_{(n)}$ is *finitary*: since $\mathcal{P}$ is finite, **S** induces only finitely many equivalence classes. A set Γ is **S**-consistent if for some φ: $\Gamma \not\vdash_{\mathbf{S}} \varphi$; Γ is maximal **S**-consistent (**S**-m.c.) if **S**-consistent, though it cannot properly be extended to an **S**-consistent set. A formula φ is a theorem of **S** if $\emptyset \vdash_{\mathbf{S}} \varphi$, also written as $\vdash_{\mathbf{S}} \varphi$.

Here are some familiar axioms and their corresponding condition on the accessibility relation: **T** $\Box_a \varphi \to \varphi$ (*reflexivity*); **D** $\Box_a \varphi \to \Diamond_a \varphi$ (*seriality*); **4** $\Box_a \varphi \to \Box_a \Box_a \varphi$ (*transitivity*); **5** $\Diamond_a \varphi \to \Box_a \Diamond_a \varphi$ (*Euclidicity*); **B** $\varphi \to \Box_a \Diamond_a \varphi$ (*symmetry*); **G** $\Diamond_a \Box_a \varphi \to \Box_a \Diamond_a \varphi$ (*confluence*).

Axiom **4** is also known as "positive introspection"; axiom **5** in its equivalent form $\neg\Box_a \varphi \to \Box_a \neg\Box_a \varphi$ is known as "negative introspection". Unimodal systems (involving only one agent's modality) are characterized by their constituting rules and axioms: **K**, **KD**, **KD4**, **KD45**, etc. If **S** is a standard modal system, then $\mathbf{S_m}$ is its m-agent counterpart. A state verifies a logic **S** if it verifies all the theorems of **S**.

3. Minimal information in multi-modal logic

Suppose we have an information order $\leq^a$ on states. When do we consider the information φ to be minimal for agent a? We suggest that φ constitutes *minimal information for a*, or that φ is a-honest, when $\Box_a \varphi$ is true in a least state $\langle M, w \rangle$; if so, the state if called an '**S**-state'.

DEFINITION 1. A formula φ is *a-honest* with respect to **S** and $\leq^a$ iff there is an **S**-state $\langle M, w \rangle$ such that $M, w \models \Box_a \varphi$ and

$$M', w' \models \Box_a \varphi \Rightarrow M, w \leq^a M', w' \text{ for all } \mathbf{S}\text{-states } \langle M', w' \rangle.$$

This characterization of minimal information may however not always be convenient. In some cases one would prefer a syntactic characterization, a deductive test, or a combination of these. This can be achieved by relating the information order $\leq^a$ to a proper sublanguage $\mathcal{L}^a$ through *persistence* [for all $\varphi \in \mathcal{L}^a : M, w \leq^a M', w' \Rightarrow (M, w \models \varphi \Rightarrow M', w' \models \varphi)$] and a converse of this, called *characterization* [for all $\varphi \in \mathcal{L}^a(M,w \models \varphi \Rightarrow M',w' \models \varphi) \;\Rightarrow\; M,w \leq^a M',w'$]. We are now able to propose alternative approaches to minimality:

(1) Formula φ has a $\leq^a$-least verifying **S**-state
(i.e. there exists a state $\langle M,w\rangle$ verifying **S** such that $M,w \models \Box_a\varphi$ and for all states $\langle M',w'\rangle$ verifying **S**: $M',w' \models \Box_a\varphi \Rightarrow M,w \leq^a M',w'$).

(2) Formula φ has an $\mathcal{L}^a$-smallest **S**-m.c. expansion
(i.e. there exists a maximal **S**-consistent Γ such that $\varphi \in \Gamma$ and for all **S**-m.c. Δ: $\varphi \in \Delta \Rightarrow \Gamma \cap \mathcal{L}^a \subseteq \Delta$).

(3) Formula φ has **S**-DP with respect to $\mathcal{L}^a$
(i.e. φ is **S**-consistent and for every $\psi_1,\psi_2,\ldots\psi_k \in \mathcal{L}^a$: $\varphi \vdash_{\mathbf{S}} (\psi_1 \vee \cdots \vee \psi_k) \;\Rightarrow\;$ for some $i \leq k : \varphi \vdash_{\mathbf{S}} \psi_i$).

THEOREM 2. *Let $\mathcal{L}^a$ be a characteristic persistent sublanguage of $\mathcal{L}$ with respect to $\leq^a$. Then the minimal information equivalences hold for $\mathcal{L}^a$ and $\leq^a$, i.e., the conditions (1), (2) and (3) above are equivalent. More specifically, given the condition, φ is a-honest with respect to $\leq^a$ and* **S** *iff (all statements are equivalent):*

- $\Box_a\varphi$ *has a $\leq^a$-least* **S***-state*
- $\Box_a\varphi$ *has an $\mathcal{L}^a$-smallest* **S***-m.c. expansion*
- φ *has a $\Box_a^- \mathcal{L}^a$-smallest a-stable expansion*
- $\Box_a\varphi$ *has* **S***-DP over $\mathcal{L}^a$.*

Here, Σ is *a*-stable if $\Sigma = \Box_a^-\Gamma$ for some **S**-m.c. Γ, and Σ is a stable expansion of ψ if Σ is stable and $\psi \in \Sigma$. Linking some of these criteria, we can say that φ is (*a*-)honest if true in some world (w) in an appropriate model (M) where *part of the knowledge in that world is minimal* (since then $\Box_a^-\mathrm{Th}(M,w)$ is its $\Box_a^-\mathcal{L}^a$-smallest stable expansion), which seems a fairly intuitive notion. When the equivalent notions of Theorem 2 hold for a particular order $\leq^a$, system **S** and language $\mathcal{L}^a$, we say that $\leq^a$ and $\mathcal{L}^a$ *determine* a notion of honesty in **S**. We can also link up the semantic definition of honesty with deduction, providing a perhaps even more intuitive characterization:[1]

COROLLARY 3. *Let $\mathcal{L}^a$ be persistent and characterizing for $\leq^a$. Then φ is a-honest with respect to $\leq^a$ and* **S** *iff there is an* **S***-state $\langle M,w\rangle$ such that:*

- $M,w \models \Box_a\varphi$ *and* $\forall\psi \in \mathcal{L}^a : M,w \models \psi \Rightarrow \Box_a\varphi \vdash_{\mathbf{S}} \psi$
- *or, equivalently,* $\forall\psi \in \mathcal{L}^a : M,w \models \psi \Leftrightarrow \Box_a\varphi \vdash_{\mathbf{S}} \psi$

[1] This characterization was triggered by a question of Arnis Vilks.

All this makes clear that we 'just' have to specify which part of the knowledge is involved. More formally speaking, we have to pinpoint the right information order $\leq^a$, or, equivalently, its characterizing persistent sublanguage $\mathcal{L}^a$. This, however, is a non-trivial problem, since surely not every information order *has* such a characterizing persistent sublanguage. For example, if $\leq^a$ is mere identity or even isomorphism of models, not even the entire language suffices to characterize the model (up to isomorphism). Also, pursuing our results for the single-agent approach, we know that unlimited bisimulation is too strong a requirement, *vide* [6]. As we showed in our earlier paper, a layered, limited kind of bisimulation is preferable. Two technically correct orders in the single agent case will be generalized in the next subsections. Although the initial, so-called general information order is not intuitively sound, it serves as a first step to more profound information orders. But we start by generalizing an umbrella result for such layered pre-orders.

3.1. Layered information orders.

An information order and its characterizing persistent language can be obtained along fairly general patterns from the underlying layered orders and their characterizing persistent languages. This is a very convenient tool for many orders to follow, since we can restrict attention to one simple layer at the time.

DEFINITION 4. Suppose $\leq^a_n$ is a pre-order on the set of model-world pairs for each natural number n ('layer n'). From now on, assuming $M = \langle W, \vec{R}, V\rangle$ and $M' = \langle W', \vec{R'}, V'\rangle$, the base case will be defined as $M, w \leq^a_0 M', w' \Leftrightarrow V(w) = V'(w')$. Then we define $\leq^a$ for any layered order $\leq^a_n$ by:

$$M, w \leq^a M', w' \Leftrightarrow \forall n \in \mathbb{N}\ \forall v' \in R'_a[w']\ \exists v \in R_a[w] : M, v \leq^a_n M', v'.$$

We say that $\leq^a$ is *induced by* $\leq^a_n$ if the above equivalence holds.

Now, let $\mathcal{L}^a$ be a sublanguage and $\mathcal{L}^a_{(n)} = \mathcal{L}^a \cap \mathcal{L}_{(n)}$ be its subset of formulas of modal depth up to n. The following lemma explains how a persistence and characterization result for layered orders and languages with finite depth can be lifted to the full language and the induced order. Lemma 5 will be implicitly used throughout the paper.

LEMMA 5 (Collecting). *If $\mathcal{L}^a_{(n)}$ is persistent and characterizing for $\leq^a_n$, and $\mathcal{L}^a$ is closed under $\vee$, then $\Box_a \mathcal{L}^a$ is persistent and characterizing for $\leq^a$.*

We will now inspect orders inspired by Ehrenfeucht-Fraïssé games (see, for example, [1, 4]).

3.2. The multi-modal general information order.

In the first Ehrenfeucht-Fraïssé order the underlying, layered order is in fact an equivalence relation ("EF-equivalence"). Define $\simeq_n^{\mathcal{A}}$ recursively (recall that $M,w \simeq_0 M',w' \Leftrightarrow V(w) = V'(w')$) by:[2] $M,w \simeq_{n+1} M',w'$ iff

- $M,w \simeq_n M',w'$ &
- $\forall i \in \mathcal{A} \forall v' \in R'_i[w'] \exists v \in R_i[w] : M,v \simeq_n M',v'$ (*back*) &
- $\forall i \in \mathcal{A} \forall v \in R_i[w] \exists v' \in R'_i[w'] : M,v \simeq_n M',v'$ (*forth*)

Then the *general information order* $\sqsubseteq^a$ is induced by $\simeq_n$. Now, first recall that, for a set $\mathcal{A}$ of agents $\mathcal{L}^{\mathcal{A}}$ is the full multi-modal language:

$$\mathcal{L}^{\mathcal{A}}: \quad \varphi ::= p \ (p \in \mathcal{P}) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_i \varphi \ (i \in \mathcal{A})$$

By a rather straightforward induction, one shows that $\mathcal{L}_{(n)}^{\mathcal{A}}$ is characteristic and persistent for $\simeq_n^{\mathcal{A}}$, and hence the collecting lemma gives that $\Box_a \mathcal{L}$ is persistent and characterizing with respect to $\sqsubseteq^a$. So, the information equivalences hold for $\sqsubseteq^a$ and $\Box_a \mathcal{L}$. We say that φ is *generally a-honest* if $\Box_a \varphi$ has a $\sqsubseteq^a$-least model. This implies the usual equivalences, i.e., $\sqsubseteq^a$ and $\Box_a \mathcal{L}$ determine a notion of a-honesty in **S**, for any modal system **S**.

However, as we noticed in [6], this notion of honesty is, though technically correct, intuitively a rather poor one. It also leads to excessive trivialization. In *weak doxastic logics* such as $\mathbf{KD_m}$ and $\mathbf{KD4_m}$, *all* formulae φ such that $\Box_a \varphi$ is consistent, are generally a-honest. For $\mathbf{K_m}$ and $\mathbf{K4_m}$, we can go a step further: all formulas are honest, as Halpern [2] notices for the first system.

In (relatively) *strong logics*, however, i.e. systems with some form of negative introspection (such as **5**, **B** and **G**), there are virtually *no* honest formulas. Because then there surely are non-theorems $\Box_a \varphi_i$ such that $\vdash \Box_a \varphi_1 \vee \Box_a \varphi_2$ which leads to an easy violation of the Disjunction Property. For example, p is generally a-dishonest in $\mathbf{S5_m}$: note that the formula $\Box_a \Diamond_b \neg p \vee \Box_a \Diamond_a \Box_b p$ is an instantiation of **5**, so this formula is derivable from $\Box_a p$ in $\mathbf{S5_m}$, whereas neither of its disjuncts is.

So, as in the unimodal case, not much is left. Among the epistemic logics only a few systems such as $\mathbf{T_m}$ and $\mathbf{S4_m}$ survive. But even then $\sqsubseteq^a$ has counterintuitive effects: growth of information does not lead to less uncertainty, as it should be.

[2] The superscript $\mathcal{A}$ is omitted whenever clear from context.

4. Generalizations of the positive order

From the single agent case we know that only the positive information order is intuitively sound with respect to honesty. In a positive information order we merely want to preserve positive knowledge of one or more agents. In other words, we disregard negative knowledge, i.e. knowledge of not knowing. We, typically, encounter some form of back simulation on the underlying layers, but usually no forth simulation, or only a restricted form of forth simulation.

It is not *a priori* clear which notion of positive information order is involved. We will discuss several options in what follows, of which the last is the more general one, using both underlying layers from the general and from the so-called 'objective' information order. We start by discussing a rather straightforward generalization from the single agent case.

4.1. Positive honesty.

The positive information order only preserves positive knowledge of agent a. It is the most obvious generalization of one-agent positive honesty. The formulas of the characterizing language do not have negative occurrences of $\Box_a$, and so, by definition, no $\Diamond_a$ as well. Formally, let $\mathcal{L}^{+a}$ consist of those $\varphi \in \mathcal{L}$ for which φ does not contain $\Box_a$ in the scope of $\neg$. Formulas in $\mathcal{L}^{+a}$ are called *a-positive*. So, $\Box_a p \vee \neg \Box_b q$, $\Box_a \neg p$ and $\Box_a p \wedge \neg q$ are members of $\mathcal{L}^{+a}$, but $\neg \Box_a p$ and $\Diamond_a p \vee \Box_b q$ are not. Formally, in BNF-notation:

$$\mathcal{L}^{+a}: \quad \varphi_0 \ ::= \ \varphi \ (\varphi \in \mathcal{L}^{\mathcal{A}\setminus\{a\}}) \mid \varphi_0 \wedge \varphi_1 \mid \varphi_0 \vee \varphi_0 \mid \Box_i \varphi_1 \ (i \in \mathcal{A})$$

Now consider $\mathcal{L}^a = \Box_a \mathcal{L}^{+a}$. This is a correct generalization of the single agent positive language, which by itself is a generalization of the language of so-called objective one-agent formulas which suit **S5**. We will call the elements of $\Box_a \mathcal{L}^{+a}$ *a-positive knowledge formulas*. What is the corresponding $\leq^a$? Essentially, the underlying order displays the back direction of the EF-equivalence for all agents, operating on a-positive formulas until subformulas are reached that are $\Box_a$-free, where full EF-equivalence for all agents except a takes over. Then, $M, w \leq^{+a}_{n+1} M', w'$ iff:

- $M, w \simeq^{\mathcal{A}\setminus\{a\}}_{n+1} M', w'$ &
- $\forall i \in \mathcal{A} \forall v' \in R'_i[w'] \exists v \in R_i[w] : M, v \leq^{+a}_n M', v'$ (*back*)

It can be shown that the underlying orders are of increasing force: for all n, $w \leq^{+a}_{n+1} w' \Rightarrow w \leq^{+a}_n w'$. Let the *positive information order* $\leq^{+a}$ be induced by $\leq^{+a}_n$. Then $\mathcal{L}^{+a}_{(n)}$ is characteristic and persistent for $\leq^{+a}_n$, so the collecting lemma guarantees that $\Box_a \mathcal{L}^{+a}$ is persistent and characterizing with respect to $\leq^{+a}$. Thus, we obtain the following.

THEOREM 6. *The minimal information equivalences hold for the order $\leq^{+a}$ and the language $\Box_a \mathcal{L}^{+a}$.*

Now, φ is called *positively a-honest* if $\Box_a \varphi$ has a $\leq^{+a}$-least model. It follows from the earlier theorems that φ is positively a-honest iff

- φ has a $\mathcal{L}^{+a}$-smallest a-stable expansion
- $\Box_a \varphi$ has **S**-DP over $\Box_a \mathcal{L}^{+a}$, or equivalently
- there is an **S**-state M, w s.t. $\forall \psi \in \mathcal{L}^{+a} : M, w \models \Box_a \psi \Leftrightarrow \Box_a \varphi \vdash_{\mathbf{S}} \Box_a \psi$.

Thus, we have that $\leq^{+a}$ and $\Box_a \mathcal{L}^{+a}$ determine a notion of a-honesty in **S**, for any system **S**. So, the notion of positive honesty is technically sound, that is, there is a persistent language that characterizes the positive information order, and it seems a proper extension of the unimodal case. It avoids problems objective honesty encounters, such as one noticed by Halpern [2] for $\mathbf{S5_m}$ (or, more precisely, extensions of $\mathbf{KB4_m}$): suppose p is some fact totally unrelated to a formula φ (for example, p may not occur in φ), then $\Box_a \varphi \vdash \Box_a p \vee \Box_a \Box_b \neg \Box_b \Box_a p$. It is clear, however, that each of the disjuncts itself does not follow from $\Box_a \varphi$. Yet φ may constitute innocent knowledge, e.g. $\Box_b q$. But for our notion of positive honesty, this counter-example to the DP test is avoided by the restriction that the disjuncts should be in the a-positive knowledge language; here, obviously, $\Box_b \neg \Box_b \Box_a p \notin \mathcal{L}^{+a}$.

Yet we do not want to exclude other possible notions of honesty *a priori*, and therefore now turn to one studied earlier.

4.2. OBJECTIVE HONESTY.

To make a different start in formalizing multi-agent positive honesty, we return to Halpern's [2] definition of a-objective formulas and the notion of honesty connected to it. Halpern reserves the notion objective honesty for the two strong doxastic systems $\mathbf{K45_m}$ and $\mathbf{KD45_m}$. This seems harmless for these two systems. Our main concern is that developing a whole apparatus for just two modal systems, and again different ones for others, leads to an approach which lacks generality and in fact conceals much of the general pattern. In fact, in Halpern's approach it is not clear *why* a-objective formulas might be suitable for the two systems mentioned. We think that we can in fact explain some of the reasons for its feasibility.

The idea of a-objective knowledge is that agent a only has knowledge of information 'outside' of a, i.e. knowledge of facts and other agents' knowledge. Such other agents' knowledge may again involve a's knowledge, but still counts as external for a. This is easily formalized when we start with the *a-objective* (that is, wide scope a-operator-free) formulas: let $\mathcal{L}^{-a}$ consist of those $\varphi \in \mathcal{L}$ for which φ does not contain wide scope $\Box_a$. In other words, in an a-objective formula, every $\Box_a$ and $\Diamond_a$ has to be in the scope of a $\Box_b$ or $\Diamond_b$ $(b \neq a)$. Examples: $\Box_a p \vee \Box_b q$, $\Box_a \neg p$ are not in $\mathcal{L}^{-a}$, but $\neg\Box_b p$ and $\Diamond_b(p \vee \neg\Box_a q)$ are.

$$\mathcal{L}^{-a}: \quad \varphi_1 \ ::= \ p \ (p \in \mathcal{P}) \mid \neg\varphi_1 \mid \varphi_1 \wedge \varphi_1 \mid \Box_i\varphi \ (i \in \mathcal{A} - \{a\})$$

So where does the agent a's knowledge enter the story? Here she is: consider $\mathcal{L}^a = \Box_a \mathcal{L}^{-a}$. A formula is then called an *a-objective knowledge formula* if it is of the form $\Box_a\varphi$ with $\varphi \in \mathcal{L}^{-a}$.

The corresponding Ehrenfeucht-Fraïssé order $\leq^a$ can be obtained from the underlying layered order, which is again an equivalence relation. The recursive clause for $\simeq_n^{-a}$ is the following: $M, w \simeq_{n+1}^{-a} M', w'$ iff

- $M, w \simeq_n^{-a} M', w'$ &
- $\forall i \neq a \forall v' \in R'_i[w'] \exists v \in R_i[w] : M, v \simeq_n^{\mathcal{A}} M', v'$ (*back*) &
- $\forall i \neq a \forall v \in R_i[w] \exists v' \in R'_i[w'] : M, v \simeq_n^{\mathcal{A}} M', v'$ (*forth*)

So, $\simeq_{n+1}^{-a}$ not only uses the general EF-equivalence relation on layer n, its overall formulation is close to that of the general information order, be it that it shares the exclusion of agent a with the positive information order.

One can now prove that $\mathcal{L}_{(n)}^{-a}$ is characteristic and persistent for $\simeq_n^{-a}$. Thus, if we define *objective information order* $\leq^{-a}$ to be induced by $\simeq_n^{-a}$, the collecting lemma guarantees that $\Box_a \mathcal{L}^{-a}$ is persistent and characterizing with respect to $\leq^{-a}$.

THEOREM 7. *The minimal information equivalences hold for the order* $\leq^{-a}$ *and the language* $\Box_a \mathcal{L}^{-a}$. $\lhd$

This again implies that $\leq^{-a}$ and $\Box_a \mathcal{L}^{-a}$ determine a notion of (objective) a-honesty in **S**, for any system **S**.

As can been seen from the format of the a-objective formulas, agent a's knowledge is not taken into account. For fully introspective this is unproblematic in the one-agent case: there one can show that positive knowledge formulas can be reduced to disjunctions of objective knowledge formulas, which implies that for each system containing **K45**, objective honesty amounts to positive honesty.

It should be emphasized that this equivalence only holds for the one agent case and full introspective knowledge. For more agents there is no such reduction, since an objective knowledge formula need not be (equivalent to) a positive one, e.g. $\Box_a\Box_b\neg\Box_a p$ is an a-objective knowledge formula which is not related to any a-positive knowledge formula whatsoever. If we want to generalize this equivalence to fully introspective multi-agent systems, we have to relax the notion of positive formula somewhat, as will be done in the next subsection.

4.3. POSITIVE-OBJECTIVE HONESTY.

We want to generalize objective knowledge to what we consider to be a more adequate notion of multi-modal honesty. The *a-positive-objective* formulas can, roughly, be characterized as having no wide scope negative occurrence of $\Box_a$ operators. Again assume for simplicity's sake that we only consider formulas where every $\Diamond_i$ is replaced by $\neg\Box_i\neg$. Let $\mathcal{L}^{\pm a}$ consist of those $\varphi \in \mathcal{L}$ for which every $\Box_a$ in φ in the scope of $\neg$ is also in the scope of a $\Box_i$ with $i \neq a$. Thus, $\mathcal{L}^{\pm a}$ can also be regarded as the closure of $\mathcal{L}^{-a}$ under the operations $\wedge$, $\vee$ and $\Box_a$. Examples: $\Box_a p \vee \Box_b q$, $\Box_a p \wedge \neg\Box_b q$ and $\Box_a\Box_b\neg\Box_a p$ are members of $\mathcal{L}^{\pm a}$, but $\neg\Box_a p$ and $\Box_a\neg\Box_a\neg p \vee \Box_b q$ are not.

$$\mathcal{L}^{\pm a}: \quad \varphi_2 \ ::= \ \varphi_1(\varphi_1 \in \mathcal{L}^{-a}) \ \mid \varphi_2 \wedge \varphi_2 \mid \varphi_2 \vee \varphi_2 \mid \Box_a \varphi_2$$

Once again, what is the corresponding $\leq^a$? For evaluating formulas, we essentially want to have recursive back moves for agent a in the EF-order, until a-objective formulas are reached, and then proceed with the a-objective equivalence. So, more formally, the recursive step in $\leq_n^{\pm a}$ is defined by $M, w \leq_{n+1}^{\pm a} M', w'$ iff:

- $M, w \simeq_{n+1}^{-a} M', w'$ &
- $\forall v' \in R'_a[w'] \, \exists v \in R_a[w] : M, v \leq_n^{\pm a} M', v'$ (*back*)

Then the *a-positive-objective information order* $\leq^{\pm a}$ is induced by $\leq_n^{\pm a}$.

Now consider $\mathcal{L}^a = \Box_a \mathcal{L}^{\pm a}$. Notice that $\mathcal{L}^{\pm a}$ extends both $\mathcal{L}^{+a}$ and $\mathcal{L}^{-a}$, thus generalizes both the positive and the objective approach. Since $\mathcal{L}_{(n)}^{\pm a}$ is characteristic and persistent for $\leq_n^{\pm a}$, the collecting lemma shows that $\Box_a \mathcal{L}^{\pm a}$ is persistent and characterizing for $\leq^{\pm a}$. Now φ is called *positive-objectively a-honest* when $\Box_a \varphi$ has a $\leq^{\pm a}$-least model.

THEOREM 8. *The minimal information equivalences hold for the order $\leq^{\pm a}$ and the language $\Box_a \mathcal{L}^{\pm a}$.*

This implies that $\leq^{\pm a}$ and $\Box_a \mathcal{L}^{\pm a}$ determine a notion of a-honesty in **S**. We can show that for fully introspective systems the extension from objective to positive-objective formulas is immaterial, since then again a's positive-objective knowledge can be reduced to a-objective knowledge. So, objective honesty and positive-objective honesty coincide for $\mathbf{K45_m}$ and $\mathbf{KD45_m}$, and $\mathbf{S5_m}$. Although positive, objective, and positive-objective honesty agree on (one agent) **S5**, they, surprisingly, do not on $\mathbf{S5_m}$ ($m > 1$). Since $\Box_a p \vee \Box_a\Box_b\neg\Box_b\Box_a p$ is derivable in $\mathbf{S5_m}$, there are virtually no (positive-)objectively honest formulas in this system. However, we have already seen that for $\mathbf{S5_m}$, the positive information order seems correct.

5. Evaluation and tools

Which of the notions presented here constitutes a feasible notion of multi-agent honesty, independent of the epistemic system? To evaluate these notions, we need to determine the (dis)honesty of a number of key examples.

In this section, we elaborate on ways to check the honesty of given a given formula φ. The material presented here is a first step towards a more systematic theory for this. In general, *dishonesty* can be shown fairly easily by using the relevant DP, but it may be harder to show *honesty* more or less directly. It is not *prima facie* clear how to prove honesty, since DP then has to be checked for an infinite set of formulas (although Lemma 22 guarantees that in many cases, one only has to inspect disjunctions of length 2). Also, minimality of stable expansions encounters similar problems and finding the least model may be non-trivial, which is related to the complexity of the information orders. Presumably, for many relevant multi-modal systems these intricate orders have simple counterparts.

Despite the name of minimal models, such models can be hard to construct or find, since they have to demonstrate a *maximum* amount of ignorance. From Corollary 3 we know that a minimal state for $\Box_a\varphi$ has to verify every $\neg\psi$ for which $\psi \in \mathcal{L}^a$ and $\Box_a\varphi \not\vdash_{\mathbf{S}} \psi$. Finding an explicit description of such a minimal state appears to be much harder than in the one agent case; in this section, we restrict ourselves to some tools to settle questions of honesty. As indicated, we desperately need some better machinery to achieve this. We will therefore introduce the following tools:

- limit of underlying order, submodel-property;
- (reflexive) amalgamation, (reflexive) rootability, sufficient to conclude general honesty ($\mathbf{T_m}$, $\mathbf{S4_m}$) $\mathbf{KD4_m}$ and the like;
- (reflexive) embedding, being (reflexive) embeddable which are sufficient to conclude positive honesty (for the same group of systems);
- clustering, being clusterable for positive honesty in $\mathbf{S5_m}$;
- Halpern's test for objective honesty in introspective systems ($\mathbf{K45_m}$, $\mathbf{KD45_m}$)

5.1. LIMIT OF LAYERED ORDERS

In Definition 4, we called an order $\leq^a$ *induced* by orders $\leq^a_n$ by imposing, for all n, the back-property for $\leq^a_n$. Since the individual layers $\leq^a_n$ in the orders of this paper also have a (restricted) back-requirement, one may wonder how essential it is to impose it on the global order $\leq^a_n$ as well.

Let us, instead, consider the limit of the layers:

$$M,w \leq^a_\infty M',w' \Leftrightarrow \forall n \in I\!N : M,w \leq^a_n M',w'$$

LEMMA 9. *If $\leq^a_n$ has the back property (i.e., $M,w \leq^a_{n+1} M',w' \Rightarrow \forall v' \in R'_a[w'] \exists v \in R_a[w] : M,v \leq^a_n M',v'$) for all n, then $\leq^a_\infty \Rightarrow \leq^a$.*

Notice that most of the information orders in this paper (the objective information order being the exception) have the back property for their underlying relation, and are therefore entailed by the limit of them. Also note that for the positive order the converse holds under the mild condition that $V(w) = V'(w')$.

5.2. GENERALIZED SUBMODELS

Next we define (even for the unimodal case!) a more general notion of submodel:[3]

DEFINITION 10. $\langle M,w\rangle$ is an *a-oriented submodel* (or simply an a-submodel) of $\langle M',w\rangle$ (notation $M',w \supseteq^a M,w$ or $M,w \subseteq^a M',w$) iff $w \in W \subseteq W'$, $R_a \subseteq R'_a$, and $\forall u \in W : \forall i \neq a\ R_i[u] = R'_i[u]$ & $V(u) = V'(u)$. If $M',w \supseteq^a M,w$ then M',w is also called an *a-supermodel* of M,w.

By induction one can easily prove the following properties:

LEMMA 11. $\supseteq^a \Rightarrow \simeq^{\mathcal{A}\setminus\{a\}}_\infty$.

LEMMA 12. $\supseteq^a \Rightarrow \leq^{+a}_\infty$.

And so we arrive at a useful sufficient requirement for the positive information order (with proof independent of the Łoś theorem).

COROLLARY 13. *$\supseteq^a \Rightarrow \leq^{+a}$, and therefore $\mathcal{L}^{+a}$ is preserved under a-submodels.*

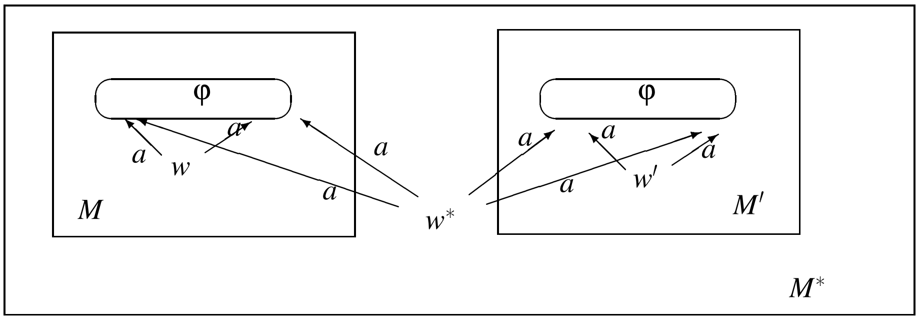
[3] Which, for principal reasons, should be called 'substate', but we accord to the usual practice here.

5.3. AMALGAMATIONS, ROOTABILITY, GENERAL HONESTY

In this section, we present a multi-modal generalisation of the *amalgamation* of two states, and the rootability of formulas (cf. [6]. They provide sufficient conditions for general honesty for weak modal systems.

Amalgamations are a usual technique in monomodal systems to combine several models into one of the same frame class. In a multi-agent modal system, the amalgamation should be relative to one of the agents. An a-amalgamation of two states is then created by adding to the combination of the two models a new 'root' world, from which all and only worlds are a-accessible which are a-accessible from the roots in the given models. This is 'simple amalgamation', which (since it preserves seriality and transitivity) is suitable for weak modal systems ($\mathbf{KD4}_m\downarrow$) and can be adapted for others. E.g., for systems containing the truth axiom $\mathbf{T}$, we use reflexive a-amalgamation of two states by requiring the root to be self-accessible for all agents. In a formal definition, we have to take into account that the original models may share worlds, which could result in interference, and therefore we have to consider disjoint copies of the original models.

DEFINITION 14. The (reflexive) a-amalgamation of two states M,w and M',w' is the state M^*,w^* which is constructed as follows. If $M = \langle W,\vec{R},V\rangle$, $M' = \langle W',\vec{R}',V'\rangle$, we define $M^* = \langle W^*,\vec{R}^*,V^*\rangle$ by: $W^* = W \oplus W' \oplus \{w^*\}$, $R_i^* = R_i \oplus R_i'$ if $i \neq a$ in case of simple amalgamation (for reflexive amalgamations always add $\langle w^*,w^*\rangle$) $R_a^* = R_a \oplus R_a' \oplus (\{w^*\} \times (R_a[w] \oplus R_a'[w']))$ for simple amalgamations; $V^* \upharpoonright_W = V$ and $V^* \upharpoonright_{W'} = V'$. Formula φ is (reflexive) a-rootable iff for any two $\Box_a\varphi$-states, a (reflexive) a-amalgamation can be found such that $\Box_a\varphi$ holds in the root world (see picture below). ◁



The next theorem applies both to weak doxastic systems such as $\mathbf{K_m}, \ldots,$ $\mathbf{KD4_m}$ for which the frame class (with transitive or serial accessibility) is closed under simple a-amalgamations, and to epistemic systems such as $\mathbf{T_m}$

and $\mathbf{S4_m}$ for which the frame class of reflexive (and transitive) accessibility relations are closed under reflexive a-amalgamations.

THEOREM 15 (For systems closed under (reflexive) a-amalgamations).
If φ is (reflexive) a-rootable and $\Box_a\varphi$ consistent, then φ is generally a-honest.

The proof of Theorem 15 is by showing DP with contraposition, with iterative amalgamations of the counter-examples. We illustrate the usefulness of theorem 15 in the following example.

EXAMPLE 16. We prove that $\varphi = \Box_a p \vee \Box_b q$ is generally a-honest for the modal system $\mathbf{S4_m}$, by showing that φ is reflexive a-rootable ($\Box_a(\Box_a p \vee \Box_b q)$ is obviously $\mathbf{S4_m}$-consistent): let M,w and M',w' be any two $\mathbf{S4_m}$-states verifying $\Box_a\varphi$. Consider the reflexive a-amalgamation M^*,w^* of M,w and M',w' such that $q \in V^*(w^*)$ and notice that R^* is a pre-order. Since $R_b[w^*] = \{w^*\}$ it follows that $M^*,w^* \models \Box_b q$, and so $M^*,w^* \models \varphi$. Using the fact that M,w is the submodel of M^* generated by w, it follows that $M^*,w \models \Box_a\varphi$, and similarly $M^*,w' \models \Box_a\varphi$. Therefore for all $u^* \in R^*_a[w^*] = R_a[w] \oplus R'_a[w'] \oplus \{w^*\}$: $M^*,u^* \models \varphi$, thus $M^*,w' \models \Box_a\varphi$.

The converse of theorem 15 also holds for a-positive formulas:

THEOREM 17 (For systems closed under (reflexive) a-amalgamations).
If an a-positive formula φ is generally a-honest then φ is (reflexive) a-rootable and $\Box_a\varphi$ consistent.

Proof. Let $\varphi \in \mathcal{L}^{+a}$ be generally a-honest then, by definition, there is a least model for $\Box_a\varphi$, which by soundness implies the consistency of $\Box_a\varphi$. To show rootability, we note that for systems closed under *simple* amalgamations *all* formulas are rootable, thus trivializing this case. It remains to show reflexive rootability for systems $\mathbf{S}$ closed under reflexive a-amalgamations: let M,M' be $\mathbf{S}$-models, $M,w \models \Box_a\varphi$ and $M',w' \models \Box_a\varphi$. Since φ is a generally a-honest, we know by the 'triangle corollary' that $\Box_a\varphi$ has a $\Box_a\mathcal{L}$-smallest $\mathbf{S}$-m.c. expansion Σ. Now consider the reflexive a-amalgamation M^*,w^* of M,w and M',w' where the root valuation is defined by $V^*(w^*)(p) = 1 \Leftrightarrow p \in \Sigma$. By a nested induction proof (first wrt $\mathcal{L}^{\mathcal{A}\setminus\{a\}}$) we can show that for every $\psi \in \mathcal{L}^{+a}$: $M^*,w^* \models \psi$, which proves that $M^*,w^* \models \Box_a\varphi$.

5.4. EMBEDDINGS, CLUSTERINGS AND POSITIVE HONESTY

Now, we show how the notion of *embedding* of two states provides a sufficient condition of positive honesty.

DEFINITION 18. An *a-embedding* M^*, w^* of states M, w and M', w' is an a-supermodel of an a-amalgamation of the two states, where M^* is in the same frame class as M and M'. Formula φ is *a-embeddable* iff for any two $\Box_a\varphi$-states, an a-embedding can be found such that $\Box_a\varphi$ holds in the root world.

THEOREM 19 (For systems closed under (reflexive) a-amalgamations). *If φ is a-embeddable and $\Box_a\varphi$ consistent, then φ is positively a-honest.*

We conjecture that this theorem can also be converted under similar conditions, but have not checked this, so far. An a-clustering ("clotting", 'equivalence amalgamation') is like an a-amalgamation, but now the new root is chosen to be one of the old roots, and all a-successors of the old roots are a-linked. Clusterings are sometimes easier to find and handle than embeddings. Moreover, for systems that are not closed under (reflexive) amalgamations, they are necessary anyway.

DEFINITION 20. An *a-clustering* of two states M, w and M', w' is a state M^*, w^* where $w^* = w$ or $w^* = w'$. The model structure M^* is defined by $(M, M'$ as in definition 14) $W^* = W \oplus W'$, $R^*_i = R_i \oplus R'_i$ if $i \neq a$, $R^*_a = R_a \oplus R'_a \oplus ((R_a[w] \oplus R'_a[w']) \times (R_a[w] \oplus R'_a[w]))$, $V^* \restriction_W = V$ and $V^* \restriction'_W = V'$. Formula φ is a-clusterable iff for any two $\Box_a\varphi$-states, an a-clustering can be found such that $\Box_a\varphi$ holds in the root.

THEOREM 21 (For systems closed under a-clusterings). *If φ is a-clusterable and $\Box_a\varphi$ consistent, then φ is positively a-honest.*

Proof. Let **S** be a system of which the frame class is closed under a-clusterings and φ be a-clusterable. We will show **S**-DP for $\Box_a\varphi$ by contraposition. Suppose $\Box_a\varphi \nvdash_{\mathbf{S}} \Box_a\alpha$ and $\Box_a\varphi \nvdash_{\mathbf{S}} \Box_a\beta$ for $\alpha, \beta \in \mathcal{L}^{+a}$ Then there are **S**-states M, w and M', w' such that $M, w \models \Box_a\varphi \wedge \neg\Box_a\alpha$ and $M', w' \models \Box_a\varphi \wedge \neg\Box_a\beta$. Consider the a-clusterings M^*, w of M, w and M', w' such that $M^*, w \models \Box_a\varphi$ (1). Since $M^*, w \supset^a M, w$ it follows by preservation of a-positieve formulas that $M^*, w \not\models \Box_a\alpha$ (2). Similarly, since M^*, w' is also an a-clustering of M, w and M', w', $M^*, w' \supset^a M, w'$, thus $M^*, w' \not\models \Box_a\beta$, and therefore, $M^*, w \not\models \Box_a\beta$ (3). In all, $M^*, w \models \Box_a\varphi \wedge \neg(\Box_a\alpha \vee \Box_a\beta)$, and so $\Box_a\varphi \nvdash_{\mathbf{S}} \Box_a\alpha \vee \Box_a\beta$.

Note that most standard modal systems and their usual frame classes are closed under a-clusterings, e.g. $\mathbf{T_m}$ (reflexive frames), $\mathbf{S4_m}$) (pre-orders), $\mathbf{KD45_m}$ (serial+transitive+euclidean frames), and, in particular, $\mathbf{S5_m}$ (equivalence frames).

Before presenting an application of Theorem 21, we need the following Lemma. Recall that the disjunction property in Section 3 is a requirement for any k disjuncts ψ_i in the language $\mathcal{L}^a$. In many cases, we can restrict ourselves to $k = 2$:

LEMMA 22. *If* $\mathbf{S}$ *is as least as strong as* $\mathbf{S4_m}$, *and* $\mathcal{L}^a$ *is either* $\mathcal{L}^{+a}$, $\mathcal{L}^{-a}$ *or* ▷ $\mathcal{L}^{\pm a}$, *then the* $\mathbf{S}$*-DP for* $\Box_a\varphi$ *with respect to* $\mathcal{L}^a$ *is equivalent to*

$$\textit{for any } \psi_1,\psi_2 \in \mathcal{L}^a : \Box_a\varphi \vdash_{\mathbf{S}} (\psi_1 \vee \psi_2) \Rightarrow \Box_a\varphi \vdash_{\mathbf{S}} \psi_1 \textit{ or } \Box_a\varphi \vdash_{\mathbf{S}} \psi_2$$

▷ EXAMPLE 23. We prove that $\top = p \vee \neg p$ is positively a-honest for the modal system $\mathbf{S5_2}$, for two agents a and b.

Proof. Obviously, $\Box_a(p \vee \neg p)$ is $\mathbf{S5_2}$-consistent. Now, we use Lemma 22 and prove the DP for two disjuncts. Now, suppose we have to formulas $\psi_1, \psi_2 \in \mathcal{L}^{+a}$ for which $\Box \not\vdash_{\mathbf{S5_2}} \Box_a\psi_i$. This means that $\not\vdash_{\mathbf{S5_2}} \Box_a\psi_1$ and $\not\vdash_{\mathbf{S5_2}} \Box_a\psi_2$. Then we find two $\mathbf{S}5_2$ states with $M,w \not\models \Box_a\psi_1$ and $M'w' \not\models \Box_a\psi_2$. We now make a 1-clustering M^* as follows (see also Figure 1): $W^* = W \oplus W'$, $V^* \restriction_W = V$, $V^* \restriction_{W'} = V'$ and the new R^*_b is just the union of the old R_b's, and, finally, R^*_a is the reflexive transitive-symmetric closure of $R_a \oplus R'_a \oplus \{\langle w,w\rangle\}$. Now, first note that $M^*,w \sqsupseteq^a M,w$ and hence $M^*,w \leq^{+a} M,w$, implying $M^*,w \not\models \Box_a\psi$. Also, we have $M^*,w' \sqsupseteq^a M',w'$. Thus, $M^*,w' \not\models \Box_a\psi_2$, implying that we have some $v' \in R^*_a[w']$ such that $M',v' \not\models \psi_2$. This v' is now also in $R^*_a[w]$, hence $M^*,w \not\models \Box_a\psi_2$. All in all, we conclude that, since $M^*,w \not\models (\Box_a\psi_a \vee \Box_a\psi_2)$ we also have $\Box_a(p \vee \neg p) \not\vdash_{\mathbf{S5_2}} (\Box_a\psi_1 \vee \Box_a\psi_b)$.

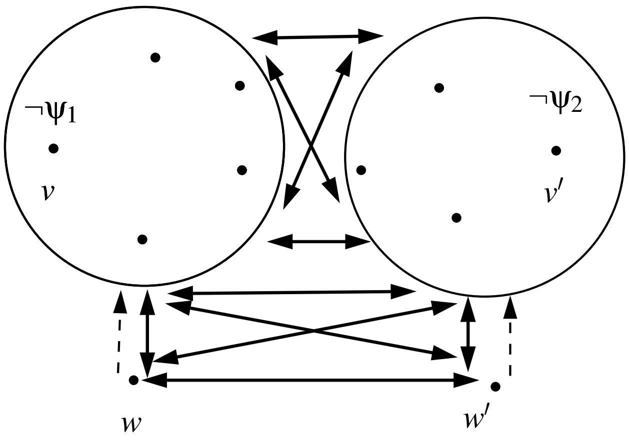


Figure 1. Clustering two $\mathbf{S}5_2$-models: the solid arrows are added

5.5. Halpern's test for objective honesty

Halpern proposes in [2] a test for checking objective a-honesty. The test is proved correct for modal systems extending $\mathbf{K45_m}$, but perhaps it can be extended to other modal systems. We have simplified its formulation.

To check whether φ is objectively a-honest for $\mathbf{K45_m}$ $\uparrow$:

– generate the finite sequence $\{\beta_0, \ldots, \beta_N\}$ where

  - $\beta_0 = \Box_a\varphi$
  - if $\beta_k \notin \{\top, \bot\}$ then β_{k+1} is the result of substituting $\top$ or $\bot$ depending on whether of not $\Box_a\varphi \vdash \Box_a\psi$, for each (so called *top-level* a-)subformula $\Box_a\psi$ in β_k such that $\psi \in \mathcal{L}^{-a}$ and $\Box_a\psi$ not being in the scope of an operator $\Box_b$ with $b \neq a$.
  - $\beta_N \in \{\top, \bot\}$

– φ is objectively a-honest iff $\beta_N = \top$.

Apart from generalizing this test (not only to other modal systems, but also to other modes of honesty), the question remains whether we can find a simple modeltheoretic condition for objective honesty, similar to ‘rootability’ and ‘embedding’.

6. Relating types of honesty

The types of honesty distinguished in this paper are ordered as indicated in Figure 2. This hierarchy easily follows from DP, using the fact that $\Box_a\mathcal{L}^{+a} \cup \Box_a\mathcal{L}^{-a} \subseteq \Box_a\mathcal{L}^{\pm a} \subseteq \Box_a\mathcal{L}$.

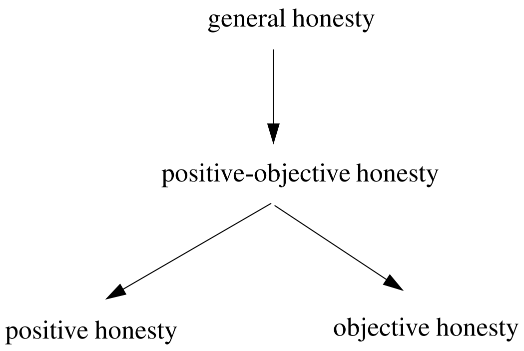


Figure 2. Relating notions of honesty

We now like to inspect the four types of honesty studied in this paper for the modal systems $\mathbf{S4_m}$, $\mathbf{K45_m}$ and $\mathbf{S5_m}$. To keep trace of the amount of ’honesty’ displayed in a particular example, we list its pattern in a 4×3 array. Due to the observation of Figure 2 not all 2^{12} patterns are indeed possible, only 36 survive: (“+” means honest for the listed type, “-” dishonest, and “a,...,g” can be either + or -).

	$\mathbf{S4_m}$	$\mathbf{K45_m}$	$\mathbf{S5_m}$
gen	a	-	-
pob	b	e	-
obj	c	e	-
pos	d	f	g

Here is how to interpret the table and its constraints:

- general honesty implies positive-objective honesty, so a=+ $\Rightarrow$ b=+
- positive-objective honesty implies both objective and positive honesty, so b=+ $\Rightarrow$ c=d=+, and e=+ $\Rightarrow$ f=+
- Apart from some exceptional cases which we have left out of the scheme, there are no generally honest formulas for $\mathbf{K45_m}$: φ is only generally a-honest for this system if φ is inconsistent (for then $\Box_a\varphi$ is vacuously consistent and every $\Box_a\psi$ is implied by $\Box_a\varphi$, thus passing the DP test) or φ is maximally informed, i.e. every atom or its negation is common knowledge. In other cases we easily find a ψ such that $\Box_a\psi, \Box_a\neg\Box_a\psi$ are not implied by $\Box_a\varphi$, but their disjunction trivially is, thus failing DP.
- As noticed in one of the text corollaries, for $\mathbf{K45_m}$ objective and positive-objective honesty is one and the same thing (e)
- For $\mathbf{S5_m}$ there is also overall dishonesty: for general honesty only the second exception (maximal info) as in $\mathbf{K45_m}$ may occur. For (positive)-objective honesty (which again amount to the same), there is also very general failure of DP due to the notorious $\mathbf{S5_m}$-theorem $\Box_a\psi \vee \Box_a\Box_b\neg\Box_b\Box_a\psi$ for some a-objective ψ.

Given these constraints, the maximally honest pattern is

	$\mathbf{S4_m}$	$\mathbf{K45_m}$	$\mathbf{S5_m}$
gen	+	-	-
pob	+	+	-
obj	+	+	-
pos	+	+	+

This pattern manifests itself in many formulas that are also intuitively honest for agent a: p, $\Box_i p$, ... The most challenging cases are disjunctions of (negated) knowledge formulas. As we will see, whether or not they are intuitively honest largely depends on the agency of the knowing subject. So, also the following innocent formulas are indeed maximally honest: $\Box_b p \vee \Box_b q$,

$\Box_a p \vee \Diamond_a q$, $\Box_b p \vee \Diamond_a q$, $\Box_b p \vee \Diamond_b q$. and $p \vee q$. The other extreme are the obviously totally dishonest formulas displaying the pattern

	$\mathbf{S4_m}$	$\mathbf{K45_m}$	$\mathbf{S5_m}$
gen	-	-	-
pob	-	-	-
obj	-	-	-
pos	-	-	-

The paradigm for this is $\Box_a p \vee \Box_a q$. Their are many (34) intermediate cases. A very common pattern here is the one in which honesty only depends on the amount of introspection attributed to the agents, witnessed by the pattern

	$\mathbf{S4_m}$	$\mathbf{K45_m}$	$\mathbf{S5_m}$
gen	+	-	-
pob	+	-	-
obj	+	-	-
pos	+	-	-

Some samples of these are $\Box_a p \vee \Box_b q$, $\Box_a p \vee \Diamond_b q$, and $\Box_a p \vee q$.

To get more complicated patterns we need formulas with nested knowledge. In particular, we suggest the following two types.

The formula $\Box_a p \vee \Box_a \Diamond_a q$ seems to have the following pattern of honesty:

	$\mathbf{S4_m}$	$\mathbf{K45_m}$	$\mathbf{S5_m}$
gen	-	-	-
pob	+	+	-
obj	+	+	-
pos	+	+	+

Whereas a dishonest counterpart of this is $\Box_a p \vee \Box_a \Diamond_a \Box_a q$ with pattern

	$\mathbf{S4_m}$	$\mathbf{K45_m}$	$\mathbf{S5_m}$
gen	-	-	-
pob	+	-	-
obj	+	-	-
pos	+	-	+

As a final example we inspect the formula $\Box_a p \vee \Box_a \Box_b \Diamond_a q$, displaying a pattern where honesty merely depends on the type and not on the modal systems under inspection:

	$\mathbf{S4_m}$	$\mathbf{K45_m}$	$\mathbf{S5_m}$
gen	-	-	-
pob	-	-	-
obj	-	-	-
pos	+	+	-

7. Conclusion

We have given generalizations of information orders for multi-agent only knowing, which apply to arbitrary modal systems and ordinary Kripke models. Using a general theorem relating information orders and their corresponding (sub-) languages, we were able to identify several equivalent characterizations of honesty. In particular, we have explored the general information order and some positive and objective information orders. So-called positive honesty seems the intuitively correct notion here.

To establish this we developed several tools. In particular, we succeeded to transfer some techniques developed for the single agent case to multi-modal systems. For example, it was shown how to adapt the amalgamation techniques as used in [6] to prove, by means of the disjunction property, honesty in **S4** and weaker systems. Moreover, a technique called clustering turned out to be useful for strong modal systems, such as $\mathbf{S4_m}$. An interesting question for future research concerns the possible generalization of the test procedure as proposed and proved correct in [2] for objective honesty to other types of honesty, and weaker modal systems.

There are many ways to extend the multi-agent perspective on only knowing. For instance, one might give up the assumption that all agents use the same logic and move to heterogeneous systems. Also, a notion of *group honesty* is as yet unexplored. Finally, we like to investigate multi-agent honesty from a more constructive perspective: can we give a procedure to generate a minimal model for a given formula? And, can we extend the partial approach of [5] to the multi-agent case?

Acknowledgements

This is an extended and improved version of [7] which was also co-authored by Jan Jaspars: we thank Jan for his collaboration and permission to submit this adapted version. The second author's research partly took place during his sabbatical at Utrecht University and CSLI, Stanford, which are gratefully acknowledged. He also thanks Michael Dunn (Indiana) and Arnis Vilks (HHL Leipzig) for kind invitations and useful comments.

References

1. H.-D. Ebbinghaus & J. Flum, *Finite model theory*, Springer–Verlag, Berlin, 1995.
2. J.Y. Halpern, 'Theory of Knowledge and Ignorance for Many Agents', in *Journal of Logic and Computation*, **7** No. 1, pp. 79–108, 1997.
3. J.Y. Halpern & Y. Moses, 'Towards a theory of knowledge and ignorance', in Kr. Apt (ed.) *Logics and Models of Concurrent Systems*, Springer–Verlag, Berlin, 1985.
4. M. Hennessy & R.Milner, 'Algebraic laws for Nondeterminism and Concurrency', *Journal of the* ACM 32, pp.137-161, 1985.
5. W. van der Hoek, J.O.M. Jaspars, & E.G.C. Thijsse, 'Honesty in Partial Logic'. *Studia Logica*, **56** (3), 323–360, 1996. Extended abstract in proceedings of KR'94.
6. W. van der Hoek, J.O.M. Jaspars, & E.G.C. Thijsse, 'Persistence and Minimality in Epistemic Logic', *Annals of Mathematics and Artificial Intelligence*, **27** (1999), pp. 25–47, 2000. Extended abstract in J. Dix, U. Furbach, L. Fariñas del Cerro (eds.), *Logics in Artificial Intelligence. Proceedings JELIA'98*, Springer Verlag, LNAI 1489.
7. W. van der Hoek, J. Jaspars and E. Thijsse, 'A General Approach to Multi-Agent Minimal Knowledge', in M. Ojeda-Aciego, I.P. Guzman, G. Brewka and L.M. Pereira, (eds) *Proceedings of the 7th European Workshop on Logics in Artificial Intelligence (JELIA 2000)*, LNAI 1919, pp. 254–268, Springer-Verlag, Heidelberg (2000).
8. G.Lakemeyer, 'All they know: a study in multi-agent auto epistemic reasoning', IJCAI'93, pp. 376–381, 1993
9. H.J. Levesque, 'All I know: a study in auto-epistemic logic', in *Artificial Intelligence*, **42**(3), pp. 263–309, 1990.
10. G. Schwarz & M. Truszczyński, 'Minimal knowledge problem: a new approach', *Artificial Intelligence* **67**, pp. 113–141, 1994.
11. R. Parikh, 'Monotonic and nonmonotonic logics of knowledge', *Fundamenta Informaticae* 15, pp. 255-274, 1991
12. M. Vardi, 'A model-theoretic analysis of monotonic knowledge', IJCAI85, pp. 509–512, 1985.

Address for Offprints:
Wiebe van der Hoek, ICS, Utrecht University
P.O. Box 80089, 3508 TB Utrecht, the Netherlands
e-mail: wiebe@cs.uu.nl
Fax: +31 30 2513791