Models of Awareness

Giacomo Sillari

[1] Philosophy, Politics and Economics Program
University of Pennsylvania
Philadelphia 19104, USA
gsillari@sas.upenn.edu

Abstract

Several formal models of awareness have been introduced in both computer science and economics literature as a solution to the problem of logical omniscience . In this chapter, I provide a philosophical discussion of awareness logic, showing that its underlying intuition appears already in the seminal work of Hintikka. Furthermore, I show that the same intuition is pivotal in Newell's account of agency, and that it can be accommodated in Levi's distinction between epistemic commitment and performance. In the second part of the chapter, I propose and investigate a first-order extension of Fagin and Halpern's *Logic of General Awareness*, tackling the problem of representing "awareness of unawareness". The language is interpreted over neighborhood structures, following the work of Arló-Costa and Pacuit on *First-Order Classical Modal Logic*. Adapting existing techniques, I furthermore prove that there exist useful decidable fragments of quantified logic of awareness.

Introduction

Since its first formulations (cf. [Hin62]), epistemic logic has been confronted with the problem of logical omniscience. Although Kripkean semantics appeared to be the natural interpretation of logics meant to represent knowledge or belief, it implies that agents are reasoners that know (or at least are committed to knowing) every valid formula. Furthermore, agents' knowledge is closed under logical consequence, so that if an agent knows φ and ψ is a logical consequence of φ, then the agent knows ψ as well. If we focus on representing pure knowledge attributions, rather than attributions of epistemic commitment, such a notion of knowledge (or belief) is too strong to be an adequate representation of human epistemic reasoning. It is possible to attack the problem by building into the semantics a distinction between implicit and explicit knowledge (or belief). The intuition behind such a distinction is that an agent is not always *aware* of all propositions. In particular, if φ is a valid formula, but the agent is not aware of it, the

agent is said to know φ *implicitly*, while she fails to know it *explicitly*. The agent explicitly knows φ, on the other hand, when she both implicitly knows φ and she is aware of φ. In their fundamental article [FH88], Fagin and Halpern formally introduced the concept of awareness in the context of epistemic logic, providing semantical grounds for the distinction between implicit and explicit belief. The technical concept of "awareness" they introduce is amenable to different discursive interpretations that can possibly be captured by specific axioms.

In the last decade, recognizing the importance of modeling asymmetric information and unforeseen consequences, economists have turned their attention to epistemic formalizations supplemented with (un)awareness (cf. [MR94], [MR99]), and noticed that partitional structures as introduced by Aumann cannot represent awareness [DLR99]. The model in [MR99] defines awareness explicitly in terms of knowledge. An agent is said to be aware of φ iff she knows φ or she both does not know that φ and knows that she does not know φ. [Hal01] shows that such a model is a particular case of the logic of awareness introduced in [FH88]. [HMS06b] present a set-theoretical model that generalizes traditional information structures *à la* Aumann. Its axiomatization in a 3-valued epistemic logic is provided in [HR05]. A further, purely set-theoretical model of awareness is given by [Li06b]. Awareness, or lack thereof, plays an important role in game-theoretic modeling. Recently, a significant amount of literature has appeared in which the issue of awareness in games is taken into account. [Fei05] incorporates unawareness into games and shows that unawareness can lead to cooperative outcomes in the finitely repeated prisoner's dilemma. A preliminary investigation on the role of awareness in the context of game-theoretical definitions of convention can be found in [Sil05]. [HR06a] define extensive-form games with possibly unaware players in which the usual assumption of common knowledge of the structure of the game may fail. [HMS06a] take into account Bayesian games with unawareness. In [Li06a] the concept of subgame perfect equilibrium is extended to games with unawareness.

This paper makes two main contribution to the literature on awareness. On the one hand, I provide philosophical underpinnings for the idea of awareness structures. On the other, I propose a new system of first-order epistemic logic with awareness that offers certain advantages over existing systems. As for the first contribution, I build on epistemological analyses of the problem of logical omniscience. Although the authors I consider need not align themselves with advocates of the awareness structures solution, I argue in the following that their analyses are not only compatible with formal models of awareness, but also compelling grounds for choosing them as the appropriate solution to the logical omniscience problem. I consider, for example, Levi's idea of epistemic *commitment*. In a nutshell, ideally

situated agents possess, in their incorrigible core of knowledge, all logical truths, and the agents' bodies of knowledge are closed under implication. Although agents are committed by (ideal) standards of rationality to holding such propositions as items of knowledge, actual agents are aware only of a subset of them (cf. [Lev80], pp. 9-13.) Furthermore, I consider Newell's theory of agency (as advanced in [New82]) and show that it contains a foreshadowing of the notion that awareness allows us to discriminate between an agent's explicit and implicit knowledge. Although Newell's analysis is conducted at a fairly abstract level, it is arguable that he is endorsing a representation model in which knowledge explicitly held by a system is given by its (implicit) knowledge *plus* some kind of access function (cf. in particular [New82], p. 114). It is not hard to see that this intuition corresponds to the intuition behind awareness structures. Finally, I argue that the intuition behind Hintikka's own treatment of logical omniscience in [Hin62] can also be considered as related to awareness structures in a precise sense that will be elucidated in the following.

As for the second contribution, I identify two main motivations for the introduction of a new formal system of awareness logic. First and foremost, it addresses the problem of limited expressivity of existing (propositional) logics of awareness. Indeed, [HR06b] notice that both standard epistemic logic augmented with awareness and the awareness models set forth in the economics literature cannot express the fact that an agent may (explicitly) know that she is unaware of *some* proposition without there being an explicit proposition that she is unaware of. This limitation of the existing models needs to be overcome, since examples of "knowledge of unawareness" are often observed in actual situations. Consider Levi's idea of commitment mentioned above: we are committed to knowing (in fact, we explicitly know) that there exists a prime number greater than the largest known prime number, although we know that we do not know what number that is. Or, consider David Lewis's theory of convention[1] as a regularity in the solution of a recurrent coordination game: when trying to learn what the conventional behavior in a certain environment might be, an agent might know (or, perhaps more interestingly, deem highly probable) that there is a conventional regularity, without having yet figured out what such a regularity actually is. Or, in the context of a two-person game with unawareness, a player might explicitly know that the other player has *some* strategy at her disposal, yet not know what such a strategy might be. Halpern and Rêgo propose in [HR06b] a sound and complete *second-order propositional epistemic logic* for reasoning about knowledge of unawareness. However, the validity problem for their logic turns out to be no better than recursively

[1] Cf. [Lew69] and the reconstruction offered in [Sil05], in which awareness structures find a concrete application.

enumerable, even in the case of **S5**, which was proven to be decidable in [Fin70]. Halpern and Rêgo conjecture that there are three causes for undecidability, each one sufficient: (i) the presence of the awareness operators, (ii) the presence of more than one modality, (iii) the absence of Euclidean relations. Undecidability of second-order, multi-modal **S5** should not come as a surprise. For example, [AT02] shows that adding a second modality to second-order **S5** makes it equivalent to full second-order predicate logic. My aim is to present a decidable logic for reasoning about knowledge of unawareness. The strategy I adopt consists in extending predicate modal logic with awareness operators and showing that it allows to represent knowledge of unawareness. Using the techniques introduced in [WZ01] and [SWZ02], I can then isolate useful decidable fragments of it.

There is a further reason for the introduction of predicate epistemic logic with awareness. The extension from propositional to predicate logic takes place in the context of *classical* systems interpreted over neighborhood structures (cf. [Che80]), rather than in the traditional framework of normal systems interpreted over Kripke structures. In so doing, I aim at bringing together the recent literature (cf. [AC02], [ACP06]) on first-order classical systems for epistemic logic and the literature on awareness structures. The rationale for this choice lies in the fact that I intend to formulate a system in which Kyburg's 'risky knowledge' or Jeffrey's 'probable knowledge' is expressible as high probability (or even as probability one belief, as Aumann does in the game-theoretical context). High probability operators give rise to Kyburg's lottery paradox, which, in the context of first-order epistemic logic (cf. [ACP06]) can be seen as an instance of the Barcan formulas. Thus, first-order Kripke structures with constant domains, in which the Barcan formula is validated, cease to be adequate models. The use of neighborhood structures allows us to work with constant domains without committing to the validity of the Barcan formulas (cf. [AC02]), hence presents itself as a natural candidate for modeling high probability operators. The second-order logic of Halpern and Rêgo also requires the Barcan formulas to be validated, and hence does not lend itself to the modeling of knowledge as high-probability operators.

The rest of the paper is organized as follows: In the first section, I review and discuss the philosophical accounts of logical omniscience offered by Hintikka, Newell and Levi, stress their structural similarities, and show how these accounts compare with their intuition underlying Fagin and Halpern's logic of awareness. In the second section, I build on Arló-Costa and Pacuit's version of first-order classical systems of epistemic logic, augmenting them with awareness structures. I then show that such a quantified logic of awareness is expressive enough to represent knowledge of unawareness and that Wolter and Zakharyashev's proof of the decidability of various fragments of

first-order multi-modal logic (cf. [WZ01]) can be slightly modified to carry over to quantified logic of awareness.

1 Logical Omniscience

In this section, I consider accounts of the problem of logical omniscience provided in Hintikka's presentation of epistemic logic, Newell's theory of agency and Levi's epistemology. I show through my analysis that all such approaches to logical omniscience share a common structure, and that Fagin and Halpern's logic of awareness has reference to such a structure.

1.1 Hintikka: Information and Justification

Hintikka's essay *Knowledge and Belief* is commonly regarded as the seminal contribution to the development of epistemic logic. Logical omniscience is an essential philosophical element in Hintikka's conceptual analysis, as well as in the formal construction stemming from it. Consider, for instance, the following quote:

> It is true, in some sense, that if I utter (10) 'I don't know whether p' then I am not altogether consistent unless it really is possible, for all that I know, that p fails to be the case. But this notion of consistency is a rather unusual one, for it makes it inconsistent for me to say (10) whenever p is a logical consequence of what I know. Now if this consequence-relation is a distant one, I may fail to know, in a perfectly good sense, that p is the case, for I may fail to see that p follows from what I know[2].

Hintikka notices ([Hin62], p. 23) that we need to distinguish two senses of "knowing". A first, weak, kind of knowledge (or belief) is simply concerned with the truth of a proposition p. In natural language, this is the sense of "knowing p" related to "being conscious[3] that p", or "being informed that p" or "being under the impression that p", etc. The second, stronger, sense of knowing is not only concerned with the truth of p, but also with the justification of the agent's knowledge. According to different epistemological accounts, "knowing" in this latter sense may mean that the agent has "all the evidence needed to assert p", or has "the right to be sure that p", or has "adequate evidence for p", etc. Whichever of these epistemological stances one chooses, the strong sense of knowing incorporates both the element of bare "availability" of the truth of p (information) and the element of the epistemological justification for p. Such a distinction is essential in

[2] [Hin62], or p. 25 of the 2005 edition of the book, from which the page references are drawn hereafter.

[3] Referring to the weak sense of "knowing", Hintikka actually mentions natural language expressions as "the agent is aware of p". In order to avoid confusion with the different, technical use of "awareness", in this context I avoid the term "awareness" altogether.

Hintikka's analysis of the notion of consistency relative to knowledge and belief, which in turn is crucial for the design of his formal system.

Syntactically, Hintikka's system does not essentially differ from the epistemic systems that have come to prevail in the literature, the only notable difference being the explicit mention of the "dual" of the knowledge operator, P_i, to be read as "it is compatible with all i knows that ... ". The pursuit of consistency criteria for the notions of knowledge and belief moves from the analysis of sets of formulas in which both knowledge and "possibility" operators are present. The main idea is that if the set $\{K_ip_1, \ldots, K_ip_n, P_iq\}$ is consistent, then the set $\{K_ip_1, \ldots, K_ip_n, q\}$ is also consistent. The distinction between the two senses of "knowing" above is crucial to the justification of this idea. If "knowing p" is taken in the weak sense of "being conscious of p", then a weaker notion of consistency is appropriate, according to which if $\{K_ip_1, \ldots, K_ip_n, P_iq\}$ is consistent, then $\{p_1, \ldots, p_n, q\}$ is consistent as well. Such a weaker notion, however, is no longer sufficient once we interpret K_ip as "i is justified in knowing p", according to the stronger sense of knowing. In this case, q has to be compatible not just with the truth of all statements $p_1, \ldots, p_n$, but also with the fact that i is in the position to justify (strongly know) each of the $p_1, \ldots, p_n$, that is to say, q has to be consistent with each one of the $K_ip_1, \ldots, K_ip_n$.

Other criteria of consistency are those relative to the knowledge operator (if λ is a consistent set and contains K_ip, then $\lambda \cup p$ is consistent), to the boolean connectives (for instance, if λ is consistent and contains $p \wedge q$, then $\lambda \cup \{p, q\}$ is consistent), and to the duality conditions (if λ is consistent and $\neg K_ip \in \lambda$, then $\lambda \cup P_i\neg p$ is consistent; while if $\neg P_ip \in \lambda$, then $\lambda \cup K_i\neg p$ is consistent). The duality conditions trigger the problem of logical omniscience. Consider again the quote at the onset of this subsection: if K_iq holds, and p is a logical consequence of q, then $\neg K_ip$ is inconsistent. Thus, at this juncture, a modeling decision has to be made. If we want to admit those cases in which an agent fails to know a logical consequence of what she knows, either (i) we may tweak the notion of knowledge in a way that makes such a predicament consistent, or (ii) we may dispense with the notion of consistency, weakening it in a way that makes such a predicament admissible. The two routes, of course, lead to different formal models. Hintikka chooses the latter strategy, while epistemic systems with awareness *à la* Fagin and Halpern choose the former. However, the two routes are but two faces of the same coin. Hintikka's concept of *defensibility*, intended as "immunity from certain standards of criticism" ([Hin62], p. 27), replacing the notion of consistency, allows us to consider knowledge (of the kind that allows for logical omniscience to fail) as the intersection of *both* the weak and the strong sense of "knowing" above, in a way that, at least structurally, is not far from considering explicit knowledge as the intersection of implicit

knowledge and awareness in [FH88].

To make more precise the notion of defensibility as "immunity from certain standards of criticism", and to see more clearly the similarity with awareness logic, let me briefly summarize Hintikka's formal system. Hintikka's semantics is kindred in spirit to possible worlds structures. There are, however, proceeding from the notion of defensibility, important differences with standard possible worlds semantics. First, define a *model set*, with respect to booleans connectives, as a set μ of formulas such that

$[\neg\ ]$ $p \in \mu \rightarrow \neg p \notin \mu$

$[\wedge\ ]$ $(p \wedge q) \in \mu \rightarrow p \in \mu$ and $q \in \mu$

$[\vee\ ]$ $(p \vee q) \in \mu \rightarrow p \in \mu$ or $q \in \mu$

$[\neg\neg\ ]$ $\neg\neg p \in \mu \rightarrow p \in \mu$

$[\neg\wedge\ ]$ $\neg(p \wedge q) \in \mu \rightarrow \neg p \in \mu$ or $\neg q \in \mu$

$[\neg\vee\ ]$ $\neg(p \vee q) \rightarrow \neg p \in \mu$ and $\neg q \in \mu$

In order to add epistemic operators to model sets, Hintikka postulates the existence of a *set of model sets* (called the *model system* Ω) and of an *alternativeness* relation for each agent, and adds the clauses

$[P\ ]$ If $P_i p \in \mu$, then there exists at least a μ^* such that μ^* is an alternative to μ for i, and $p \in \mu^*$

$[KK\ ]$ $K_i p \in \mu \rightarrow$ if μ^* is an alternative to μ for i, and $K_i p \in \mu^*$

$[K\ ]$ $K_i p \in \mu \rightarrow p \in \mu$

Thus, we have consistent sets of formulas constituting a model system, an accessibility relation between model sets in the system for each agent, and a semantic account of knowledge that does not differ importantly from the standard Kripkean one (to see that, notice that KK and K taken together imply that if $K_i p \in \mu$ then $p \in \mu^*$ for all μ^* alternative to μ in Ω). The fundamental difference with Kripke models, thus, lies in the elements of the domain: model sets (i.e. consistent sets of formulas) in Hintikka's semantics, possible worlds (i.e. *maximally* consistent sets of formulas) in Kripke's. Thus, Hintikka's model sets are *partial* descriptions of possible worlds[4].

[4] Hintikka has made this claim unexceptionable in later writings: "The only viable interpretation of logicians' "possible worlds" is the one that I initially assumed was intended by everyone. That is to understand "possible worlds" as scenarios, that is, applications of our logic, language or some other theory to some part of the universe

The notion of defensibility is now definable as follows: A set of formulas is defensible iff it can be embedded in a model set of a model system. As the notion of consistency is replaced with that of defensibility, the notion of validity is replaced with that of *self-sustainance*. It follows easily from the definitions that $p \rightarrow q$ is self-sustaining iff the set $p, \neg q$ is not defensible[5]. This is key for overcoming logical omniscience: although an agent knows q if she knows p and $p \rightarrow q$ is self-sustaining, it need not be the case that she knows q if she knows p and $p \rightarrow q$ is *valid*, since, in this case, $p \rightarrow q$ need not be self-sustaining[6]. This may occur if q does not appear in the model sets of Ω, so that $p, \neg q$ is embeddable in them, making $\neg K_i q$ defensible (since, by the duality rule, $P_i \neg q \in \mu$ and, by rule $[K]$, there exists a μ^* such that $\neg q \in \mu^*$). Thus, $\neg K_i q$ is defensible as long as q can be kept out of some model set μ, provided that i does not incur in criticism according to certain epistemic standards. That is to say, for an agent to be required to know q, it is not enough, say, that q logically follows from the agent's knowledge, but it also needs to be the case that q belongs to a model set μ. Similarly[7], in [FH88], for an agent to know φ explicitly, it is not sufficient that φ logically follows from the agent's knowledge, but it also needs to be the case that φ belongs to the agent's awareness set. In this sense, a formula not appearing in a model set and a formula not belonging to an awareness set may be regarded as cognate notions.

1.2 Newell: Knowledge and Access

The interest of the AI community in the logic of knowledge and its representation does not need to be stressed here. Intelligent agents must be endowed with the capability of reasoning about the current state of the world, about what other agents believe the current state of the world is, etc. Planning, language processing, distributed architectures are only some of the many fields of computer science in which reasoning about knowledge plays a cen-

---

that can be actually or at least conceptually isolated sufficiently from the rest", cf. [Hin03], p. 22. But cf. also [Hin62], p. 33-34: "For our present purposes, the gist of their [model sets] formal properties may be expressed in an intuitive form by saying that they constitute [...] a very good formal counterpart to the informal idea of a (partial) description of a possible state of affairs".

[5] If the set $\{p, \neg q\}$ is not defensible, then it cannot be embedded in any model set μ, meaning that either $\neg p$ or q (or both) must belong to μ, making $p \rightarrow q$ self-sustaining, and vice versa.

[6] Notice however ([Hin62], p. 46) that other aspects of logical omniscience are present: the valid formula $(K_i p \wedge K_i q) \rightarrow K_i(p \wedge q)$ is also self-sustaining.

[7] A formal proof is beyond the scope of this paper, and the interested reader can find it in [Sil07]. To see the gist of the argument, consider that model sets are *partial* description of possible worlds. While one can (as it is the case, e.g., in [Lev84]) model the distinction between explicit and implicit knowledge by resorting to partial descriptions of possible worlds, one can, equivalently, do so by "sieving" the description of a possible world through awareness sets.

tral role. It is not surprising, then, that computer scientists paid attention to the epistemic interpretation of modal logics and, hence, that they had to confront the problem of logical omniscience. It is difficult (and probably not particularly relevant) to adjudicate issues of precedence, but the idea of using some conceptualization of awareness to cope with the problem of logical omniscience appeared in the early 80s, possibly on a cue by Alan Newell. In 1980, Newell delivered the first presidential address of the American Association for Artificial Intelligence. The title was *The knowledge level*, and the presidential address was reproduced in [New82]. The article focuses on the distinction between knowledge and its representation, both understood as functional components of an intelligent system.

An intelligent system, in the functional view of agency endorsed by Newell, is embedded in an action-oriented environment. The system's activity consists in the process from a perceptual component (that inputs task statements and information), through a representation module (that represents tasks and information as data structures), to a goal structure (the solution to the given task statement). In this picture, knowledge is perceived from the external world, and stored as it is represented in data structures. Newell claims that there is a distinction between knowledge and its representation, much like there is a one between the symbolic level of a computer system and the level of the actual physical processes supporting the symbolic manipulations. A *level* in a computer system consists of a medium (which is to be processed), components together with laws of composition, a system, and, determining the behavior of the system, laws of behavior. For example, at the symbolic level the system is the computer, its components are symbols and their syntax, the medium consists of memories, while the laws of behavior are given by the interpretation of logical operations. Below the symbolic level, there is the physical level of circuits and devices. Among the properties of levels, we notice that each level is reducible to the next lower level (e.g., logical operations in terms of switches), but also that a level need not have a description at higher levels. Newell takes it that knowledge constitutes a computer system level located immediately above the symbolic level.

At the *knowledge level*, the system is the agent; the components are goals, actions and bodies (of knowledge); the medium is knowledge and the behavioral rule is rationality. Notice that the symbolic level constitutes the level of representation. Hence, since every level is reducible to the next lower level, knowledge can be represented through symbolic systems. But can we provide a description of the knowledge level without resorting to the level of representation? It turns out that we can, although we can only if we do not decouple knowledge and action. In particular, says Newell, "it is unclear in

what sense [systems lacking rationality] can be said to have knowledge[8]", where "rationality" stands for "principles of action". Indeed an *agent*, at the knowledge level, is but a set of actions, bodies of knowledge and a set of goals, rather independently of whether the agent has any physical implementation. What, then, is *knowledge*? Knowledge, according to Newell, is whatever can be ascribed to an agent, such that the observed behavior of the agent can be explained (that is, computed) according to the laws of behavior encoded in the principle of rationality. The principle of rationality appears to be unqualified: "If an agent has knowledge that one of its actions will lead to one of his goals, then the agent will select that action[9]". Thus, the definition of knowledge is a procedural one: an *observer* notices the action undertaken by the agent; given that the observer is familiar with the agent's goals and its rationality, the observer can therefore infer what knowledge the agent must possess. Knowledge is not defined *structurally*, for example as physical objects, symbols standing for them and their specific properties and relations. Knowledge is rather defined *functionally* as what mediates the behavior of the agent and the principle of rationality governing the agent's actions. Can we not sever the bond between knowledge and action by providing, for example, a characterization of knowledge in terms of a physical structure corresponding to it? As Newell explains, "the answer in a nutshell is that knowledge of the world cannot be captured in a finite structure. [...] Knowledge as a structure must contain at least as much variety as the set of all truths (i.e. propositions) that the agent can respond to[10]". Hence, knowledge cannot be captured in a finite physical structure, and can only be considered in its functional relation with action.

Thus (a version of) the problem of logical omniscience presents itself when it comes to describing the epistemic aspect of an intelligent system. Ideally (at the knowledge level), the body of knowledge an agent is equipped with is unbounded, hence knowledge cannot be represented in a physical system. However, recall from above how a level of interpretation of the intelligent system *is* reducible to the next lower level. Knowledge should therefore be reducible to the level of symbols. This implies that the symbolic level necessarily encompasses only a portion of the unbounded body of knowledge that the agent possesses. It should begin to be apparent, at this point, that what Newell calls "knowledge" is akin to what in awareness epistemic logic is called "implicit knowledge," whereas what Newell refers to as "representation" corresponds to what in awareness logic is called "explicit knowledge". Newell's analysis endorses the view that explicit knowledge corresponds to implicit knowledge *and* awareness as witnessed by the "slo-

[8] Cf. [New82], p. 100.
[9] Cf. [New82], p. 102. Although Newell, in the following, refines it, his principle of rationality does not seem to be explicitly concerned with utility.
[10] Cf. [New82], p. 107.

gan equation[11]"

Representation = Knowledge + Access.

The interesting question is then: in which way does an agent extract representation from knowledge? Or, in other terms: Given the *definition* of representation above, what can its *theory* be? Building a theory of representation involves building a theory of access (that is, of awareness), to explain how agents manage to extract limited, explicit knowledge (working knowledge, representation) from their unbounded implicit knowledge. The suggestive idea is that agents do so "intelligently", i.e. by judging what is relevant to the task at hand. Such a judgment, in turn, depends on the principle of rationality. Hence, knowledge and action cannot be decoupled and knowledge cannot be entirely represented at the symbolic level, since it involves both structures and *processes*[12]. Given the "slogan equation" above, it seems that one could identify such processes with explicit and effective rules governing the role of awareness. Logics, as they are "one class of representations [...] uniquely fitted to the analysis of knowledge and representation[13]", seem to be suitable for such an endeavor. In particular, epistemic logics enriched with awareness operators are natural candidates to axiomatize theories of explicit knowledge representation.

1.3 Levi: Commitment and Performance

Levi illustrates (in [Lev80]) the concept of epistemic commitment through the following example: an agent is considering what integer stands in the billionth decimal place in the decimal expansion of π. She is considering ten hypotheses of the form "the integer in the billionth decimal place in the decimal expansion of π is j", where j designates one of the first ten integers. Exactly one of those hypotheses is consistent with the logical and mathematical truths that, according to Levi, are part of the incorrigible core of the agent's body of knowledge. However, it is reasonable to think that, if the agent has not performed (or has no way to perform) the needed calculations[14], "there is an important sense in which [the agent] does not know which of these hypotheses is entailed by [logical and mathematical truths]

[11] Cf. [New82], p. 114.

[12] The idea is taken up again in [CN94], where a broader, partly taxonomical analysis of (artificial) agency is carried out. Moving along a discrete series of models of agents increasingly limited in their processing capabilities, we find, at the "fully idealized" end of the spectrum, the *omnipotent*, logically omniscient agent. Next to it, we find the *rational* agent, which, as described above, uses the principle of rationality to sieve its knowledge and obtain a working approximation of it.

[13] Cf. [New82], p. 100.

[14] It should be clear to the reader that Levi's argument carries over also to those cases in which the lack of (explicit) knowledge follows from reasons other than lack of computational resources.

and, hence, does not know what the integer in the billionth place in the decimal expansion of π is[15]". Levi stresses that the agent is *committed* to believing the right hypothesis, but she may at the same time be *unaware* of what the right hypothesis is. While the body of knowledge of an *ideally situated and rational agent* contains all logical truths and their consequences, the body of knowledge of real persons or institutions does not. Epistemic (or, as Levi prefers, doxastic) commitments are necessary constituents of knowledge, which, although ideally sufficient to achieve knowledge, must in practice be supplemented with a further element. As Levi puts it: "I do assume, however, that to be aware of one's commitment is to know what they are"[16].

The normative aspect of the principle of rationality regulating epistemic commitments and, hence, their relative performances, is further explored in [Lev97]. Levi maintains that the principle of rationality in inquiry and deliberation is twofold. On the one hand, it imposes necessary, but weak, coherence conditions on the agent's state of full belief, credal probability, and preferences. On the other, it provides minimal conditions for the justification of changes in the agent's state of full belief, credal probability, and preferences. As weak as the coherence constraints might be, they are demanding well beyond the capability of any actual agent. For instance, full beliefs should be closed under logical consequence; credal probabilities should respect the laws of probability; and preferences should be transitive and satisfy independence conditions. Hence, such principles of rationality are not to be thought of as descriptive (or predictive) or, for that matter, normative (since it is not sensible to impose conditions that cannot possibly be fulfilled). They are, says Levi, *prescriptive*, in the sense that they do not require compliance tout court, but rather demand that we enhance our ability to follow them.

Agents fail to comply with the principle of rationality requiring the deductive closure of their belief set, and they do so for multiple reasons. An agent might fail to entertain a belief logically implied by other beliefs of hers because she is lacking in attention. Or, being ignorant of the relevant deductive rules, she may draw an incorrect conclusion or even refuse to draw a conclusion altogether. The former case, according to Levi, can be accommodated by understanding belief as a disposition to assent upon interrogation. In the latter, the agent needs to improve her logical abilities—by "seeking therapy". In both cases, however, what is observed is a discrepancy between the agent's commitment to hold an epistemic disposition, and her epistemic performance, which fails to display the disposition she is committed to having. The prescriptive character of the principle of rationality gives the agent

[15] Cf. [Lev80], pp. 9-10.
[16] Cf. [Lev80], p. 12.

an (epistemic) obligation to fulfill the commitment to full belief. The agent is thus committed[17] to holding such a belief. The notion of full belief appears both as an epistemic disposition (commitment) of the agent, as well as the actual performance of her disposition.

The discussion of Levi's idea of epistemic commitment provides us with three related, yet distinct concepts involved in the description of the epistemic state of an agent. On the one hand, we have epistemic commitments (which we could think of as *implicit beliefs*). On the other, we have commitments that the agent fulfills, that is to say, in the terminology of [Lev80], commitments of which the agent is aware (we could think of those as *explicit beliefs*). The latter, though, calls for a third element, the agent's awareness of the commitment she is going to fulfill.

1.4 Logical Omniscience and Awareness Logic

The three examinations of the problem of logical omniscience described here do not deal directly with the logic of awareness, and actually all of them pre-date even the earliest systems of awareness logic (for instance, [Lev84]). In fact, Hintikka's position on the issue has shifted over the years. Since the introduction of Rantala's "urn models" (cf. [Ran75]), the author of *Knowledge and Belief* has endorsed the "impossible worlds" solution to the problem of logical omniscience (cf. [Hin75]). In the case of Isaac Levi's approach, it is at best doubtful that Fagin and Halpern understand the notions of implicit and explicit knowledge in the same way Levi understands those of epistemic commitment and epistemic performance. However, the three accounts analyzed above *do* share a common structure, whose form is captured by Fagin and Halpern's logic of awareness. In the case of Allen Newell's analysis of agency at the knowledge level, there is a marked conceptual proximity between Newell's notions of knowledge, representation and access, on the one hand, and Fagin and Halpern's notions of implicit knowledge, explicit knowledge and awareness, on the other. But consider also Hintikka's distinction between a weak and a strong sense of knowing, the former roughly related to the meaning of "having information that", the latter to the one of "being justified in having information that". If we interpret "awareness" as meaning "having a justification", then strong knowledge is yielded by weak knowledge and justification, just as explicit

[17] She is committed in the sense (see [Lev97]) in which one is committed to keep a religious vow to sanctity: occasional sinning is tolerated, and the vow is to be considered upheld as long as the pious agent strives to fulfill the commitments the vow implies. However, going back to the principle of rationality, one should notice that "epistemic therapy" comes at a cost (of time, resources, effort etc.) and that, moreover, not all our doxastic commitments (actually only *a few* of them) are epistemically useful (think of the belief that p, which implies the belief that $p \vee p$, that $p \vee p \vee p$, and so on). Hence the idea of "seeking therapy" or of using "prosthetic devices" to comply with the principle of rationality leaves space for further qualification.

knowledge is yielded by implicit knowledge and awareness. Also Levi's distinction between epistemic commitment and epistemic performance can be operationalized by stipulating that epistemic performance stems from the simultaneous presence of both the agent's epistemic commitment and the agent's recognition of her own commitment, just as explicit knowledge is yielded by the simultaneous presence of implicit knowledge and awareness.

Fagin and Halpern's logic of awareness was meant to be a versatile formal tool in the first place[18], in such a way that its purely formal account of "awareness" could be substantiated with a particular (concrete) interpretation of "awareness". Such an interpretation could be epistemological justification, as in the case of Hintikka's account; or it could be physical access, as in case of Newell's artificial agent; or it could be psychological awareness, as in the case of Levi's flesh-and-blood agents. All three interpretations fit the general structure of Fagin and Halpern's awareness logic. It is, however, much less clear whether it is possible to capture axiomatically the different properties that "awareness" enjoys in the philosophical accounts delineated in the previous subsections. From a normative standpoint, one would need to answer the question: given that the agents capabilities are bounded, which are the items of knowledge that (bounded) rationality requires that the agent explicitly hold? This line of inquiry is pursued by Harman (cf. [Har86]) and by Cherniak (cf. [Che86]). Levi notices that the "therapy" to be undertaken in order to better our epistemic performance comes at a cost, triggering a difficult prescriptive question as to how and how much an agent should invest to try and better approximate her epistemic commitment. Levi does not seem to think that such a question can be answered in full generality[19]. It seems to me that there is an important dynamic component to the question (*if* one's goal is such-and-such, *then* she should perform epistemically up to such-and-such portion of her epistemic commitment) that is well captured in Newell's intuition that knowledge representation and action cannot be decoupled in a physical system. The formidable task of providing an *axiomatic* answer to the normative question about the relation between implicit and explicit knowledge lies beyond

[18] Cf, [FH88], p. 41: "Different notions of knowledge and belief will be appropriate for different applications. We believe that one of the contributions of this paper is providing tools for constructing reasonable semantic models of notions of knowledge with a variety of properties." Also, "once we have a concrete interpretation [of awareness] in mind, we may well add some restrictions to the awareness functions to capture certain properties of 'awareness'," *ibidem*, p. 54.

[19] "A lazy inquirer may regard the effort to fulfill his commitments as too costly where a more energetic inquirer suffering from the same disabilities does not. Is the lazy inquirer failing to do what he ought to do to fulfill his commitments, in contrast to the more energetic inquirer? I have no firm answer to this question [...] We can recognize the question as a prescriptive one without pretending that we are always in the position to answer it in advance.", [Lev91], p. 168, n. 14.

the scope of this contribution, and in the formal system advanced in the next section, I will consider only those properties of awareness that are now standard in the literature.

2 First-Order Logic of Awareness

In this section, I extend Fagin and Halpern's logic of general awareness (cf. [FH88]) to a first-order logic of awareness, show that awareness of unawareness can be expressed in the system, and prove that there exist decidable fragments of the system. For a general introduction to propositional epistemic logic, cf. [MvdH95] and [FHMV95]. For detailed treatments of the first-order epistemic systems, cf. [Che80] and the work of Arló-Costa and Pacuit ([AC02] and [ACP06]).

2.1 first-order classical models

The *language* $\mathcal{L}_n$ of multi-agent first-order epistemic logic consists of the connectives $\wedge$ and $\neg$, the quantifier $\forall$, parentheses and n modal operators $K_1, \ldots, K_n$, one for each agent considered in the system. Furthermore, we need a countable collection of individual variables $\mathcal{V}$ and a countable set of n-place predicate symbols for each $n \geq 1$. The expression $\varphi(x)$ denotes that x occurs free in φ, while $\varphi[x/y]$ stands for the formula φ in which the free variable x is replaced with the free variable y. An *atomic formula* has the form $P(x_1, \ldots, x_n)$, where φ is a predicate symbol of arity n. If **S** is a classical propositional modal logic, **QS** is given by the following axioms:

S All axioms from **S**

$\forall$ $\forall x\varphi(x) \rightarrow \varphi[y/x]$

Gen From $\varphi \rightarrow \psi$ infer $\varphi \rightarrow \forall x\psi$, where x is not free in φ.

In particular, if **S** contains the only modal axiom **E** (from $\varphi \leftrightarrow \psi$, infer $K_i\varphi \leftrightarrow K_i\psi$) we have the weakest classical system **E**; if **S** validates also axiom **M** ($K_i(\varphi \wedge \psi) \rightarrow (K_i\varphi \wedge K_i\psi)$), we have system **(E)M**, etc. (see [Che80] for an exhaustive treatment of classical systems of modal logic).

As to the semantics, a *constant domain neighborhood frame* is a tuple $\mathcal{F} = (W, \mathcal{N}_1, \ldots, \mathcal{N}_n, D)$, where W is a set of possible worlds, D is a non-empty set called the domain, and each $\mathcal{N}_i$ is a *neighborhood* function from W to 2^{2^W}. If we define the *intension* (or *truth set*) of a formula φ to be the set of all worlds in which φ is true, then we can say, intuitively, that an agent at a possible world knows all formulas whose intension belongs to the neighborhood of that world. A *model* based of a frame $\mathcal{F}$ is a tuple $(W, \mathcal{N}_1, \ldots, \mathcal{N}_n, D, I)$, where I is a classical first-order interpretation function. A *substitution* is a function $\sigma : \mathcal{V} \rightarrow D$. If a substitution σ' agrees with σ on every variable except x, it is called an x-variant of σ, and such

a fact is denoted by the expression $\sigma \sim_x \sigma'$. The satisfiability relation is defined at each state relative to a substitution σ:

$(M, w) \models_\sigma P(x_1, \ldots, x_n)$ iff $\langle \sigma(x_1), \ldots, \sigma(x_n) \rangle \in I(P, w)$ for each n-ary predicate symbol P.

$(M, w) \models_\sigma \neg\varphi$ iff $(M, w) \not\models_\sigma \varphi$

$(M, w) \models_\sigma \varphi \wedge \psi$ iff $(M, w) \models_\sigma \varphi$ and $(M, w) \models_\sigma \psi$

$(M, w) \models_\sigma K_i\varphi$ iff $\{v : (M, v) \models_\sigma \varphi\} \in \mathcal{N}_i(w)$

$(M, w) \models_\sigma \forall x\varphi(x)$ iff for each $\sigma' \sim_x \sigma, (M, w) \models_{\sigma'} \varphi(x)$

As usual, we say that a formula φ is *valid* in M if $(M, w) \models \varphi$ for all worlds w in the model, while we say that φ is *satisfiable* in M if $(M, w) \models \varphi$ for some worlds w in the model. Notice that **QE** axiomatizes first-order minimal[20] models (in which no restrictions are placed on the neighborhoods); **QEM** axiomatizes first-order monotonic models (in which neighborhoods are closed under supersets); etc.

2.2 Adding awareness

Following [FH88], awareness is introduced on the syntactic level by adding to the language further modal operators A_i and X_i (with $i = 1, \ldots, n$), standing for awareness and explicit knowledge, respectively[21]. The operator X_i can be defined in terms of K_i and A_i, according to the intuition that explicit knowledge stems from the simultaneous presence of both implicit knowledge and awareness, by the axiom

(*A*0) $X_i\varphi \leftrightarrow K_i\varphi \wedge A_i\varphi$.

Semantically, we define n functions $\mathcal{A}_i$ from W to the set of all formulas. Their values specify, for each agent and each possible world, the set of formulas of which the agent is aware at that particular world. Hence the straightforward semantic clauses for the awareness and explicit belief operators:

$(M, w) \models_\sigma A_i\varphi$ iff $\varphi \in \mathcal{A}_i(w)$

$(M, w) \models_\sigma X_i\varphi$ iff $(M, w) \models A_i\varphi$ and $(M, w) \models K_i\varphi$

[20] For the terminology, see [Che80].

[21] The use of neighborhood structures eliminates, of course, many aspects of the agents' logical omniscience. However, axiom **E** is valid in *all* neighborhood structures. Thus, the distinction between implicit and explicit knowledge remains relevant, since agents may fail to recognize the logical equivalence of formulas φ and ψ and, say, explicitly know the former without explicitly knowing the latter.

In propositional awareness systems of the kind introduced in [FH88], different interpretations of awareness are captured by imposing restrictions on the construction of the awareness sets. For example, one can require that if the agent is aware of $\varphi \wedge \psi$, then she is aware of ψ and φ as well. Or, one could require that the agent's awareness be closed under subformulas, etc. One of those interpretations (which, *mutatis mutandis*, is favored in the economics literature when taken together with the assumption that agents know what they are aware of) is that awareness is *generated by primitive propositions*. In this case, there is a set of *primitive* propositions Φ of which agent i is aware at w, and the awareness set of i at w contains exactly those formulas that mention only atoms belonging to Φ. Similarly, we can interpret awareness in a first-order system as being *generated by atomic formulas*, in the sense that i is aware of φ at w iff i is aware of all atomic subformulas in φ. Thus, for each i and w, there is a set (call it *atomic awareness set* and denote it $\Phi_i(w)$) such that $\varphi \in \mathcal{A}_i(w)$ iff φ mentions only atoms appearing in $\Phi_i(w)$. Such an interpretation of awareness can be captured axiomatically. The axioms relative to the boolean and modal connectives are the usual ones (cf., e.g., [FH88]):

$(A1)$ $A_i \neg\varphi \leftrightarrow A_i\varphi$

$(A2)$ $A_i(\varphi \wedge \psi) \leftrightarrow A_i\varphi \wedge A_i\psi$

$(A3)$ $A_i K_j \varphi \leftrightarrow A_i\varphi$

$(A4)$ $A_i A_j \varphi \leftrightarrow A_i\varphi$

$(A5)$ $A_i X_j \varphi \leftrightarrow A_i\varphi$.

Before discussing the axioms relative to the quantifiers, it is worth stressing that the first-order setup allows the modeler to specify some details about the construction of atomic awareness sets. In the propositional case, the generating set of primitive propositions is a list of atoms, necessarily unstructured. In the predicate case, we can have the atomic awareness set built in the semantic structure. Notice that there can be two sources of unawareness. An agent could be unaware of certain *individuals* in the domain or she could be unaware of certain *predicates*. Consider the following examples: in a game of chess, (i) a player could move her knight to reach a position x in which the opponent's king is checkmated; however, she cannot "see" the knight move and is, as a result, unaware that x is a mating position; or (ii) a player could move her knight, resulting in a position x in which the opponent's queen is pinned; however, she is a beginner and is not familiar with the notion of "pinning"; she is thus unaware that x is a pinning position. Hence, in order for an agent to be aware of an atomic

formula she must be aware of the individuals occurring in the interpretation of the formula as well as of the predicate occurring in it. It is possible to capture these intuitions formally: for each i and w, define a "subjective domain" $D_i(w) \subset D$ and a "subjective interpretation" I_i that agrees with I except that for some w and P of arity n, $I(P,w) \neq I_i(P,w) = \varnothing$. We can then define the atomic awareness set for i at w by stipulating that

$$P(x_1,\ldots,x_n) \in \Phi_i(w) \text{ iff } \begin{cases} (i) & \sigma(x_k) \in D_i(w), \quad \forall x_k, k = 1,\ldots,n \\ (ii) & \langle \sigma(x_1),\ldots,\sigma(x_n)\rangle \in I_i(P,w) \end{cases}$$

Notice that this is consistent with the notion that one should interpret a formula like $A_i\varphi(x)$, where x is free in φ, as saying that, *under a valuation* $\sigma(x)$, the agent is aware of $\varphi(x)$. Similarly, the truth of $K_i\varphi(x)$ depends on the individual assigned to x by the valuation σ. Finally, we need to introduce a family of special n-ary predicates[22] $A!_i$ whose intuitive meaning is "i is aware of objects $\sigma(x_1),\ldots,\sigma(x_n)$". Semantically, we impose that $(M,w) \models_\sigma A!_i(x)$ iff $\sigma(x) \in D_i(x)$.

2.3 Expressivity

Let us now turn our attention to the issue of representing knowledge of unawareness. Consider the *de re/de dicto* distinction, and the following two formulas:

(i) $X_i \exists x \neg A_i P(x)$

(ii) $\exists x X_i \neg A_i P(x)$.

The former says that agent i (explicitly) knows that there exists an individual enjoying property P, without her being aware of which particular individual enjoys P. The formula, intuitively, should be satisfiable, since $P(x) \notin \mathcal{A}_i(w)$ need not entail $\exists x P(x) \notin \mathcal{A}_i(w)$. On the other hand, the latter says that i is aware, of a specific x, that x has property P. If this is the case, it is unreasonable to admit that i can be unaware of $P(x)$. By adopting appropriate restrictions on the construction of the awareness sets, we can design a system in which formulas like (i) are satisfiable, while formulas like (ii) are not.

In particular, we need to weaken the condition that awareness is *generated by atomic formulas*, since we want to allow for the case in which $P(x) \notin \mathcal{A}_i(w)$, yet $\exists x P(x) \in \mathcal{A}_i(w)$. I argue that such an interpretation of awareness is sensible. In fact, we may interpret $P(x)$ not belonging to i's awareness set as meaning that i is not aware of a specific instance of x that

[22] Such predicates are akin to the existence predicate in free logic. However, the awareness system considered here is not based on free logic: the special awareness predicates will only be used to limit the range of possible substitutions for universal quantifiers *within* the scope of awareness operators. The behavior of quantifiers is otherwise standard.

enjoys property P, while we may interpret $\exists x P(x)$ belonging to i's awareness set as meaning that i is aware that at least one specific instance of x (which one she ignores) enjoys property P. The versatility of the awareness approach is again helpful, since the blend of syntax and semantics characterizing the concept of awareness makes such an interpretation possible.

Let us see in more detail what restrictions on the construction of the awareness sets correspond to the interpretation above. In particular, we want

(i) $X_i \exists x \neg A_i P(x)$ to be satisfiable, while

(ii) $\exists x X_i \neg A_i P(x)$ should not be satisfiable.

Semantically, thus, if $P(x) \notin \mathcal{A}_i(w)$, then (against (ii)), $\neg A_i P(x) \notin \mathcal{A}_i(w)$. Yet the possibility that (i) $\exists x \neg A_i P(x) \in \mathcal{A}_i(w)$ is left open. The following condition, along with the usual conditions for awareness being generated by atomic formulas in the case of quantifier-free formulas, does the job[23] (weak $\exists$-closure):

(*) If $\varphi[x/y] \in \mathcal{A}_i(w)$, then $\exists x \varphi(x) \in \mathcal{A}_i(w)$.

It is easy to see that, if $P(x) \notin \mathcal{A}_i(w)$, then (ii) is not satisfiable, since there should exist an interpretation $\sigma' \sim_x \sigma$ such that $(M, w) \models_{\sigma'} A_i \neg A_i P(x)$. But that is impossible, since, for quantifier-free propositions, awareness is generated by atomic formulas. On the other hand, (i) entails that $\exists x \neg A_i P(x) \in \mathcal{A}_i(w)$, which remains satisfiable, since the weak condition (*) does not require that $\neg A_i P(x) \in \mathcal{A}_i(w)$.

Let me illustrate the reason why we are considering a weak closure of awareness under existential quantification by means of an example: in the current position of a chess game, White knows (or: deems highly probable) that sacrificing the bishop triggers a mating combination, although she cannot see what the combination itself precisely is. Take the variable x to range over a domain of possible continuations of the game, and the predicate P to be interpreted as "is a mating combination". Thus, at w, White is aware that there exists an x such that $P(x)$. However she is not aware of what individual $\sigma(x)$ actually is ($\neg A_i P(x)$), hence $A_i \exists x \neg A_i P(x)$ holds. Now, had (*) been a biconditional, since $\exists x \neg A_i P(x) \in \mathcal{A}_i(w)$ holds, it would have been the case that $\neg A_i P[x/y] \in \mathcal{A}_i(w)$, that is $A_i \neg A_i P(y)$. In the example, White would have been aware that she is not aware that the *specific* combination $\sigma(y)$ led to checkmate, which is counterintuitive. The fact that, limited to sentences, awareness is generated by atomic formulas and that awareness is only weakly closed under existential quantification rules out such undesirable cases.

[23] Cf. [HR06b], in which a similar requirement of weak existential closure is used.

Notice that the weak ∃-closure (*) can be expressed axiomatically as follows:

$(A6)$ $A_i\varphi[x/y] \rightarrow A_i\exists x\varphi(x)$.

What is the interplay between universal quantification and awareness? Consider, in the example above, the situation in which White is aware that *any* continuation leads to checkmate. It is then reasonable to conclude that she is aware, of any *specific* continuation she might have in mind, that it leads to checkmate. Thus, if, for any x, $P(x) \in \mathcal{A}_i(w)$, then $P[x/y] \in \mathcal{A}_i(w)$ for all y such that $\sigma(y) \in D_i(w)$. Hence the axiom

$(A7)$ $A_i\forall x\varphi(x) \rightarrow (A!_i(y) \rightarrow A_i\varphi[x/y])$.

This concludes the presentation of the syntax and the semantics of the first-order system of awareness. Various first-order modal logics are proven to be complete with respect to neighborhood structures in [ACP06]. Extending the proof to deal with awareness is also straightforward once we add to the canonical model the canonical awareness sets $\mathcal{A}_i(w) = \{\varphi(x) : A_i\varphi(x) \in w\}$, where w stands for a canonical world. For example, consider, in the proof of the truth lemma, the case in which the inductive formula has the form $A_i\psi$: if $A_i\psi \in w$, then, by definition of the canonical $\mathcal{A}_i(w)$, $\psi \in \mathcal{A}_i(w)$ or $(M, w) \models_\sigma A_i\psi$, and vice versa. Note that axioms $A1$-$A7$ ensure that awareness is weakly generated by atomic formulas.

2.4 Decidability

This section is based on Wolter and Zakharyashev's decidability proof for the *monodic* fragment of first-order multi-modal logics interpreted over Kripke structures. The proof is here generalized to neighborhood models with awareness. The *monodic fragment* of first-order modal logic is based on the restricted language in which formulas in the scope of a modal operator have at most one free variable. The idea of the proof is the following:

We can decide whether the monodic formula φ is valid, provided that we can decide whether a certain classical first-order formula α is valid. This is because, by answering the satisfiability problem for α, we can construct a so-called "quasi-model" for φ. A "quasi-model" satisfying φ, as it will be clear in the following, exists if and only if a neighborhood model satisfying φ exists. Furthermore, if a model satisfying φ exists, then it is possible to effectively build a "quasi-model" for φ. Hence the validity problem in the monodic fragment of first-order modal logic can be reduced to the validity problem in classical first-order logic. It follows that the intersection of the monodic fragment and (several) decidable fragments of first-order classical logic is decidable[24].

[24] The *mosaic* technique on which the proof of the existence of an effective criterion for

In carrying the proof over to neighborhood structures with awareness, a few adjustments of the original argument are necessary. First, the overall proof makes use of special functions called *runs*. Such functions serve the purpose of encoding the modal content of the structure. Since the modal operators are now interpreted through neighborhoods rather than through accessibility relations, the definitions of *runs* and of related notions have to be modified accordingly. Second, the proof of theorem 2.1 accounts for the cases of the modal operators introduced in the present setup (i.e. awareness and explicit knowledge operators). Third, a suitable notion of "unwinding" a neighborhoods structure has to be found in order to prove lemma 2.2. Fourth, the use of neighborhood structures slightly modifies the argument of the left to right direction in theorem 2.3. In the rest of this subsection, I shall offer the main argument and definitions, relegating the more formal proofs to the appendix.

Fix a monodic formula φ. For any subformula $\Box_i\psi(x)$ of φ, let $P_{\Box_i\psi}(x)$ be a predicate symbol not occurring in ψ, where $\Box_i = \{K_i, A_i, X_i\}$. $P_{\Box_i\psi}(x)$ has arity 1 if $\psi(x)$ has a free variable, 0 otherwise, and it is called the *surrogate* of $\psi(x)$. For any subformula ψ of φ, define $\overline{\psi}$ to be the formula obtained by replacing the modal subformulas of ψ not in the scope of another modal operator with their surrogates, and call $\overline{\psi}$ the *reduct* of ψ.

Define $sub_x\varphi = \{\psi[x/y] : \psi(y) \in sub\varphi\}$, where $sub\varphi$ is the closure under negation of the set of subformulas of φ. Define a *type* t for φ as any boolean saturated subset of $sub_x\varphi$[25], i.e. such that, (i) for all $\psi \in sub_x\varphi$, $\neg\psi \in t$ iff $\psi \notin t$; and (ii) for all $\psi \wedge \chi \in sub_x\varphi, \psi \wedge \chi \in t$ iff ψ and χ belong to t[26]. Types t and t' are said to *agree on* $sub_0\varphi$ (the set of subsentences of φ) if $t \cap sub_0\varphi = t' \cap sub_0\varphi$.

The goal is to encode a neighborhood model satisfying φ into a quasi-model for φ. The first step consists in coding the worlds of the neighborhood model. Define a *world candidate* to be the set T of φ-types that agree on $sub_0\varphi$. Consider now a first-order structure $\mathcal{D} = (D, P_0^{\mathcal{D}}, \ldots)$, let $a \in D$ and define $t^{\mathcal{D}}(a) = \{\psi \in sub_x\varphi : \mathcal{D} \models \overline{\psi}[a]\}$, where $\models$ stands for the classical satisfiability relation. It easily follows from the semantics of $\neg$ and $\wedge$ that $t^{\mathcal{D}}$ is a type for φ. A *realizable world candidate* is the set $T = \{t^{\mathcal{D}}(a) : a \in D\}$.

---

the validity problem is based was introduced by Németi (cf. for example [Nem95]). The proof on which this subsection is based can be found in [WZ01]. The same technique is used in [SWZ02] to show that first-order common knowledge logics are complete. For a more compact textbook exposition of the proof, cf. [BG07].

[25] Or, equivalently, as "any subset of $sub_x\varphi$ such that $\{\overline{\psi} : \psi \in t\}$ is maximal consistent", where ψ is any subformula of φ: cf. [BG07].

[26] For example, consider $\varphi := K_iP(y) \wedge X_i\exists zR(y,z)$. Then $sub_x\varphi$ is the set $\{K_iP(x) \wedge X_i\exists zR(x,z), K_iP(x), P(x), X_i\exists zR(x,z), \exists zR(x,z)\}$, along with the negation of such formulas. Some of the types for φ are $\Phi \cup \{\exists zR(x,z), P(x)\}, \Phi \cup \{\neg\exists zR(x,z), P(x)\}$, etc.; $\neg\Phi \cup \{\exists zR(x,z), P(x)\}$, $\neg\Phi \cup \{\exists zR(x,z), \neg P(x)\}$ etc., where $\Phi = \{K_iP(x) \wedge X_i\exists zR(x,z), K_iP(x), X_i\exists zR(x,z)\}$ and $\neg\Phi = \{\neg\psi : \psi \in \Phi\}$.

Notice that T is a realizable world candidate iff a formula α_T is satisfiable in a first-order structure, where α_T is

$(\alpha_T) \qquad \bigwedge_{t\in T} \exists x \bar{t}(x) \wedge \forall x \bigvee_{t\in T} \bar{t}(x),$

in which $\bar{t}(x) := \bigwedge_{\psi(x)\in t} \overline{\psi}(x)$.

Intuitively, the formula says that all the reducts of the formulas in every type $t \in T$ are realized through some assignment in the first-order structure, while all assignments in the structure realize the reducts of the formulas in some type $t \in T$. The existence of the satisfiability criterion for φ will ultimately be given modulo the decidability of α_T for each realizable world candidate in the model, hence the restriction to the monodic fragment based on a *decidable* fragment of predicate logic.

Set a neighborhood frame with awareness $\mathcal{F} = (W, \mathcal{N}_1, \ldots, \mathcal{N}_n, \mathcal{A}_1, \ldots, \mathcal{A}_n)$. We can associate each world w in W to a corresponding realizable world-candidate by taking the set of types for φ that are realized in w. Let f be a map from each $w \in W$ to the corresponding realizable world candidates T_w. Define a *run* as a function form W to the set of all types of φ such that

(i) $r(w) \in T_w$,
(ii) if $K_i\psi \in sub_x\varphi$, then, $K_i\psi \in r(w)$ iff $\{v : \psi \in r(v)\} \in \mathcal{N}(w)$,
(iii) if $A_i\psi \in sub_x\varphi$, then $A_i\psi \in r(w)$ iff $\psi \in \mathcal{A}_i(w)$.
(iv) if $X_i\psi \in sub_x\varphi$, then $X_i\psi \in r(w)$ iff $\{v : \psi \in r(v)\} \in \mathcal{N}(w)$ and $\psi \in \mathcal{A}_i(w)$.

Runs are the functions that encode the "modal content" of the neighborhood structure satisfying φ that was lost in the reducts $\overline{\psi}$, so that it can be restored when constructing a neighborhood model based on the quasi-model for φ.

Finally, define a *quasi-model* for φ as the pair $\langle \mathcal{F}, f \rangle$, where f is a map from each $w \in W$ to the set of realizable world candidates for w, such that, for all $w \in W$ and $t \in T$, there exists a run on $\mathcal{F}$ whose value for w is t. We say that a quasi-model satisfies φ iff there exists a w such that $\varphi \in t$ for some $t \in T_w$. We can now prove the following

Theorem 2.1. The monodic sentence φ is satisfiable in a neighborhood structure M based on $\mathcal{F}$ iff φ is satisfiable in a quasi-model for φ based on $\mathcal{F}$.

Proof. See Appendix, section A1. Q.E.D.

It is now possible to show that an effective satisfiability criterion for φ exists by representing quasi-models through (possibly infinite) mosaics of

repeating *finite* patterns called blocks[27].

Recall that quasi-models are based on neighborhood frames. We restrict now our attention to monotonic frames[28] and say that a quasi-model for φ is a tree quasi-model if it is based on a tree-like neighborhood frame. Section 2.4 of the appendix, drawing from [Han03], describes how a monotonic neighborhood model can be unravelled in a tree-like model. Hence,

Lemma 2.2. A (monodic) formula φ is satisfiable iff it is satisfiable in a tree quasi-model for φ at its root.

Proof. The lemma stems obviously from the unravelling procedure described in the Appendix, section A2. Q.E.D.

We now need to define the notion of a *block* for φ. We shall then be able to represent quasi-models as structures repeating a finite set of blocks. Consider a finite tree-like structure (called a bouquet) $\langle \mathcal{F}_n, f \rangle$, based on $W_n = \{w_0, \ldots, w_n\}$, rooted in w_0, such that no world in the structure but w_0 has a nonempty neighborhood.

A *root-saturated weak run* is a function r from W_n to the set of types for φ such that

(i) $r(w_n) \in T_{w_n}$,

(ii) if $K_i\psi \in sub_x\varphi$, then, $K_i\psi \in r(w_0)$ iff $\{v : \psi \in r(v)\} \in \mathcal{N}(w)$,

(iii) if $A_i\psi \in sub_x\varphi$, then $A_i\psi \in r(w_0)$ iff $\psi \in \mathcal{A}_i(w)$.

(iv) if $X_i\psi \in sub_x\varphi$, then $X_i\psi \in r(w_0)$ iff $\{v : \psi \in r(v)\} \in \mathcal{N}(w)$ and $\psi \in \mathcal{A}_i(w)$.

A *block* is a bouquet $\langle \mathcal{F}_n, f_n \rangle$, where f_n is a map from each $w \in W_n$ to the set of realizable world candidates for w such that, for each $w \in W_n$ and $t \in T$, there exists a root-saturated weak run whose value for w is t. We say that φ is satisfied in a block $\langle \mathcal{F}_n, f_n \rangle$ iff there exists a w such that $\varphi \in t$ for some $t \in T_w$.

Finally, a *satisfying set* for φ is a set $\mathcal{S}$ of blocks such that (i) it contains a block with root w_0 such that $\varphi \in t$ for all $t \in T_{w_0}$ (that is, w_0 satisfies φ), and (ii) for every realizable world candidate in every block of $\mathcal{S}$, there exists a block in $\mathcal{S}$ rooted in such a realizable world candidate.

It is now possible to prove the following

Theorem 2.3. A monodic sentence φ is satisfiable iff there exists a satisfying set for φ, whose blocks contain a finite number of elements.

[27] For ease of exposition and without loss of generality, from now on attention will be restricted to models with a single agent.

[28] This restriction yields a less general proof, since it implies that the decidability result does not hold for non-monotonic systems. Given the intended interpretation of the modalities (high-probability operators, cf. the introductory section), the restriction is not problematic.

Proof. See Appendix, section A3. Q.E.D.

The effective satisfiability criterion now follows:

Corollary 2.4. Let $\mathcal{L}_m$ be the monodic fragment and $\mathcal{L}'_m \subseteq \mathcal{L}_m$. Suppose that for $\varphi \in \mathcal{L}'_m$ there is an algorithm deciding whether a world-candidate for φ is realizable (that is, whether the classical first-order formula α_T is satisfiable.) Then the fragment $\mathcal{L}'_m \cap$ **QEM** is decidable.

In particular, the monodic fragment is decidable if it is based on the two- (one-) variable fragment, on the monadic fragment, and on the guarded fragment of classical predicate logic (cf. [WZ01]).

Acknowledgments

The author wishes to thank Cristina Bicchieri, Horacio Arló-Costa, Isaac Levi, Frank Wolter, Eric Pacuit, Burkhardt Schipper and Martin Meier for stimulating conversations, suggestions, criticisms and corrections. A preliminary version of this paper was presented at the LOFT06 conference in Liverpool: the author wishes to thank the organizers and the audience. A special thanks goes to two anonymous referees, whose suggestions have importantly improved the paper and corrected a serious mistake in a previous version.

References

[AC02] Horacio Arló-Costa. First order extensions of classical systems of modal logic; the role of the barcan schemas. *Studia Logica*, 71(1):87–118, 2002.

[ACP06] Horacio Arló-Costa and Eric Pacuit. First-order classical model logic. *Studia Logica*, 84(2):171–210, 2006.

[AT02] Aldo Antonelli and R. Thomason. Representability in second-order poly-modal logic. *Journal of Symbolic Logic*, 67(3):1039–1054, 2002.

[BG07] Torben Braüner and Silvio Ghlardi. First-order modal logic. In Patrick Blackburn, Johan van Benthem, and Frank Wolter, editors, *Handbook of modal Logic*, pages 549–620. Elsevier, 2007.

[Che80] Brian Chellas. *Modal Logic: An Introduction.* Cambridge University Press, Cambridge, 1980.

[Che86] Cristopher Cherniak. *Minimal rationality.* MIT Press, Cambridge, Mass., 1986.

[CN94] Kathleen Carley and Allen Newell. The nature of the social agent. *Journal of Mathematical Sociology*, 19(4):221–262, 1994.

[DLR99] Eddie Dekel, Barton L. Lipman, and Aldo Rustichini. Standard state-space models preclude unawareness. *Econometrica*, 66(1):159–173, 1999.

[Fei05] Yossi Feinberg. Games with incomplete awareness. Stanford University, 2005.

[FH88] Ronald Fagin and Joseph Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[FHMV95] Ronald Fagin, Joseph Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about Knowledge*. The MIT Press, Cambridge, Mass., 1995.

[Fin70] Kit Fine. Propositional quantifiers in modal logic. *Theoria*, 36:336–346, 1970.

[Hal01] Joseph Halpern. Alternative semantics for unawareness. *Games and Economic Behavior*, 37:321–339, 2001.

[Han03] Helle Hvid Hansen. Monotonic modal logics. Master's thesis, ILLC, Amsterdam, 2003.

[Har86] Gilbert Harman. *Change in view*. MIT Press, Cambridge, Mass., 1986.

[Hin62] Jaakko Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, NY, 1962.

[Hin75] Jaakko Hintikka. Impossible possible worlds vindicated. *Journal of Philosophical Logic*, 4:475–484, 1975.

[Hin03] Jaakko Hintikka. Intellectual autobiography. In Randall E. Auxier and Lewis E. Hahn, editors, *The Philosophy of Jaakko Hintikka*, volume XXX of *The Library of Living Philosophers*, pages 3–84. Open Court, 2003.

[HMS06a] Aviad Heifetz, Martin Meier, and Burkhard Schipper. Unawareness, beliefs and games. University of California, Davis, 2006.

[HMS06b] Aviad Heifetz, Martin Meier, and Burkhard C. Schipper. Interactive unawareness. *Journal of economic theory*, 130:78–94, 2006.

[HR05] Joseph Halpern and Leandro Rêgo. Interactive unawareness revisited. In *Proceedings TARK05*, pages 78–91, 2005.

[HR06a] Joseph Halpern and Leandro Rêgo. Extensive games with possibly unaware players. In *Proceedings of the Fifth International Joint Conference on Auntonomous Agents and Multiagent Systems*, pages 744–751, 2006.

[HR06b] Joseph Halpern and Leandro Rêgo. Reasoning about knowledge of unawareness. In *Tenth International Conference on Principles of Knowledge Representation and Reasoning*. Tenth International Conference on Principles of Knowledge Representation and Reasoning, 2006.

[Lev80] Isaac Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Mass., 1980.

[Lev84] H.-J. Levesque. A logic for implicit and explicit belief. In *Proceedings AAAI-84*, pages 198–202, Austin, TX, 1984.

[Lev91] Isaac Levi. *The Fixation of Belief and Its Undoing*. Cambridge University Press, Cambridge, 1991.

[Lev97] Isaac Levi. *The Covenant of Reason*. Cambridge University Press, Cambridge, 1997.

[Lew69] David Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, Mass., 1969.

[Li06a] Jing Li. Dynamic games of complete information with unawareness. University of Pennsylvania, 2006.

[Li06b] Jing Li. Information structures with unawareness. University of Pennsylvania, 2006.

[MR94] Salvatore Modica and Aldo Rustichini. Awareness and partitional information structures. *Theory and Decision*, 37:107–124, 1994.

[MR99] Salvatore Modica and Aldo Rustichini. Unawareness and partitional information structures. *Games and Economic Behavior*, 27:265–298, 1999.

[MvdH95] J.J-Ch. Meyer and Wiebe van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, Cambridge, Mass., 1995.

[Nem95] I. Nemeti. Decidability of weakened versions of first-order logic. In L. Csirmaz, D. M. Gabbay, and M. de Rijke, editors, *Logic Colloquim '92*, number 1 in Studies in Logic, Language and Information, pages 177–242. CSLI Publications, 1995.

[New82] Allen Newell. The knowledge level. *Artificial Intelligence*, 18(1), 1982.

[Ran75] Veikko Rantala. Urn models: a new kind of non-standard model for first-order logic. *Journal of Philosophical Logic*, 4:455–474, 1975.

[Sil05] Giacomo Sillari. A logical framework for convention. *Synthese*, 147(2):379–400, 2005.

[Sil07] Giacomo Sillari. Quantified awareness logic and impossible possible worlds. University of Pennsylvania, 2007.

[SWZ02] H. Sturm, F. Wolter, and M. Zakharyaschev. Common knolwedge and quantification. *Economic Theory*, 19(1):157–186, 2002.

[WZ01] F. Wolter and M. Zakharaschev. Decidabile fragments of first-order modal logics. *Journal of Symbolic Logic*, 66(3):1415–1438, 2001.

Appendix

A1. Proof of Theorem 2.1

[⇒]. Let M be a neighborhood structure satisfying φ. Construct a quasi-model as follows: Define the map f by stipulating that

$$t_a^w = \{\psi \in sub_x\varphi : (M, w) \models_\sigma \psi\}, \text{ where } a \in D \text{ and } \sigma(x) = a,$$

$$T_w = \{t_a^w : a \in D\},$$

and let, for all $a \in D$ and $w \in W$, $r(w) = t_a^w$. We need to show that r is a run in $\langle \mathcal{F}, f\rangle$. For (i), $r(w) = t_a^w \in T_w$ by construction. For (ii), $K_i\psi(x) \in r(w)$ iff $(M, w) \models_\sigma K_i\psi(x)$ iff $\{v : (M, v) \models_\sigma \psi(x)\} \in \mathcal{N}_i(w)$ but, for all $a \in D$, $t_a^v = \{\psi \in sub_x\varphi : (M, v) \models_\sigma \psi(x)\}$ and $r(v) = t_a^v$ by definition, thus $\{v : (M, v) \models_\sigma \psi(x)\} = \{v : \psi(x) \in r(v)\}$, as desired. For (iii), $(M, w) \models_\sigma A_i\psi$ iff $\psi \in \mathcal{A}_i(w)$, hence $A_i\psi \in r(w)$ iff $\psi \in \mathcal{A}_i(w)$. The case for (iv) follows immediately from (ii) and (iii).

[⇐]. Fix a cardinal $\kappa \geq \aleph_0$ that exceeds the cardinality of the set Ω of all runs in the quasi-model. Set $D = \{\langle r, \xi\rangle : r \in \Omega, \xi < \kappa\}$. Recall that a world candidate T is realizable iff the first-order formula α_T is satisfiable in a first-order structure and notice that, since the language we are using does not comprehend equality, it follows from standard classical model theory that we can consider the first-order structure $\mathcal{D}$ to be of arbitrary infinite cardinality $\kappa \geq \aleph_0$. Hence, for every $w \in W$, there exists a first-order structure $I(w)$ with domain D that realizes the world candidate $f(w)$. Notice that the elements in the domain of such structures are specific runs indexed by the cardinal ξ. Let[29] $r(w) = \{\psi \in sub_x\varphi : I(w) \models \overline{\psi}[\langle r, \xi\rangle]\}$ for all $r \in \Omega$ and $\xi < \kappa$.

Let the neighborhood structure be $M = (W, \mathcal{N}_1, \ldots, \mathcal{N}_n, \mathcal{A}_1, \ldots, \mathcal{A}_n, D, I)$ and let σ be an arbitrary assignment in D. For all $\psi \in sub\varphi$ and $w \in W$, we show by induction that

$$I(w) \models_\sigma \overline{\psi} \text{ iff } (M, w) \models_\sigma \psi.$$

The basis is straightforward, since $\psi = \overline{\psi}$ when ψ is an atom. The inductive step for the nonmodal connectives follows from the observation that $\overline{\psi \wedge \psi'} = \overline{\psi} \wedge \overline{\psi'}$, $\overline{\neg\psi} = \neg\overline{\psi}$, $\overline{\forall x\psi} = \forall x\overline{\psi}$, and the induction hypothesis. Consider now the modal cases. Fix $\sigma(y) = \langle r, \xi\rangle$ First, let $\psi := K_i\chi(y)$. The reduct of ψ is the first-order formula $P_{K_i\chi}(y)$. We have that

[29] For a proof that this assumption is legitimate, cf. [SWZ02].

$I(w) \models_\sigma P_{K_i\chi}(y)$ iff (construction of quasi-model)
$K_i\chi(y) \in r(w)$ iff (definition of run)
$\{v : \chi(y) \in r(v)\} \in \mathcal{N}_i(w)$ iff (definition of $r(v)$)
$\{v : I(v) \models \overline{\chi}(y)\} \in \mathcal{N}_i(w)$ iff (induction hypothesis)
$\{v : (M, v) \models_\sigma \chi(y)\} \in \mathcal{N}_i(w)$ iff (semantics)
$(M, w) \models_\sigma K_i\chi(y)$.

Second, let $\psi := A_i\chi(y)$. The reduct of ψ is the first-order formula $P_{A_i\chi}(y)$. Then,

$I(w) \models_\sigma P_{A_i\chi}(y)$ iff (construction of quasi-model)
$A_i\chi(y) \in r(w)$ iff (definition of run and of $r(w)$)
$\chi(y) \in \mathcal{A}_i(w)$ iff (semantics)
$(M, w) \models_\sigma A_i\chi$.

Finally, let $\psi := X_i\chi(y)$. We have that $I(w) \models_\sigma P_{X_i\chi}(y)$ iff $I(w) \models_\sigma P_{K_i\chi}(y) \wedge P_{A_i\chi}(y)$, which follows from the two cases just shown. Q.E.D.

A2. Unravelling a neighborhood structure

In this subsection I describe the procedure defined in [Han03] to unravel a core-complete, monotonic model M into a monotonic model whose core neighborhoods give rise to a tree-like structure that is bisimilar to M.

Definition 2.5. (Core-complete models)
Let the *core* $\mathcal{N}^c$ of $\mathcal{N}$ be defined by $X \in \mathcal{N}^c(w)$ iff $X \in \mathcal{N}(w)$ and for all $X_0 \subset X, X_0 \notin \mathcal{N}^c(w)$. Let M be a neighborhood model. M is *core-complete* if, for all $w \in W$ and $X \subseteq W$, If $X \in \mathcal{N}(w)$, then there exists a $C \in \mathcal{N}^c(w)$ such that $C \subseteq X$.

The idea is that we can unravel a core-complete, monotonic neighborhood structure (with awareness) into a core-complete neighborhood which is rooted and whose core, in a sense that will be made precise below, contains no cycles and has unique, disjoint neighborhoods. The unravelling procedure described above is given in [Han03].

Define the following objects:

Definition 2.6. Let M be a core-complete monotonic model. For any $X \subseteq W$, define $\mathcal{N}^c_\omega(X)$ and $S_\omega(X)$ as the union, for all $n \geq 0$, of the objects defined by double recursion as:

$S_0(X) = X$, $\mathcal{N}^c_0(X) = \bigcup_{x \in X} \mathcal{N}^c(x)$
$S_n(X) = \bigcup_{Y \in \mathcal{N}^c_{n-1}(X)} Y$, $\mathcal{N}^c_n(X) = \bigcup_{x \in S_n(X)} \mathcal{N}^c_n(x)$

In words, we start with a neighborhood X, and take $\mathcal{N}^c_0(X)$ to be the core neighborhoods of the worlds in X. We add all worlds in such core neighborhoods to the space set of the following stage in the inductive construction, and then consider all core neighborhoods of all such worlds, etc.

If the set of all worlds in a model M is yielded by $S_\omega(\{\omega\})$, then M is said to be a *rooted* model.

We can now define a tree-like neighborhood model as follows:

Definition 2.7. Let M be a core-complete monotonic neighborhood model with awareness, and let $w_0 \in W$. Then M_{w_0} is a *tree-like model* if:

(i) $W = S_\omega(\{w_0\})$;

(ii) For all $w \in W$, $w \notin \bigcup_{n>0} S_n(\{w\})$;

(iii) For all $w, w', v \in W$ and all $X_0, X_1 \subseteq W$: If $v \in X_0 \in \mathcal{N}^c(w)$ and $v \in X_1 \in \mathcal{N}^c(w')$, then $X_0 = X_1$ and $w_0 = w_1$.

That is to say, (i) M is rooted in w_0; (ii) w does not occur in any core neighborhood "below" w, thus there are no cycles; and (iii) all core neighborhoods are unique and disjoint.

The neighborhood model $M = (W, \mathcal{N}, \mathcal{A}, \pi)$ can now be unravelled into the model $M_{w_0} = (W_{w_0}, \mathcal{N}_{w_0}, \mathcal{A}_{w_0}, \pi_{w_0})$ as follows:

(1) Define its universe W_{w_0} as

$W_{w_0} = \{(w_0 X_1 w_1 \dots X_n w_n) : n \geq 0$ and for each $l = 1, \dots, n : X_l \in \mathcal{N}(w_{l-1}), w_l \in X_l\}$

In English, W_{w_0} contains all sequences of worlds and neighborhoods obtained by beginning with w_0 and appending to each state w_i the sets belonging to its neighborhood, and by further appending the worlds contained in the element of the neighborhood under consideration. For example, if the model contains a world w whose neighborhood contains the set $\{x, y\}$ the space of the unravelled model rooted on w contains also worlds $w\{x, y\}x$ and $w\{x, y\}y$.

In order to define the neighborhoods of the unravelled model, we need to define two maps *pre* and *last* as:

$$pre : (w_0 X_1 w_1 \dots X_n w_n) \to (w_0 X_1 w_1 \dots X_n)$$

$$last : (w_0 X_1 w_1 \dots X_n w_n) \to w_n.$$

(2) Define now a neighborhood function $\mathcal{N}^c_{w_0} : W_{w_0} \to \mathcal{P}(\mathcal{P}(W_{w_0}))$ as follows, with $\overrightarrow{s} \in W_{w_0}$ and $Y \subseteq W_{w_0}$:

$Y \in \mathcal{N}^c_{w_0}(\overrightarrow{s})$ iff for all $\overrightarrow{y} \in Y$ and some $X \in \mathcal{P}(W)$, $pre(\overrightarrow{y}) = \overrightarrow{s}X$ and $\bigcup_{\overrightarrow{y} \in Y} last(\overrightarrow{y}) = X \in \mathcal{N}(last(\overrightarrow{s}))$.

Thus, every neighborhood in the original model $\mathcal{N}(last(\overrightarrow{s}))$ originates exactly one neighborhood Y in $\mathcal{N}^c_{w_0}(\overrightarrow{s})$ and all sets Y are disjoint. Closing the core neighborhoods under supersets yields now the neighborhoods of the monotonic model.

(3) Define an awareness function $\mathcal{A}_{w_0}$ such that $\varphi \in \mathcal{A}_{w_0}(\overrightarrow{w})$ iff $\varphi \in \mathcal{A}(last(\overrightarrow{w}))$.

(4) Finally, we take $\pi_{w_0}(\overrightarrow{s})$ to agree with $\pi(last(\overrightarrow{s}))$.

It follows that $(M,(w_0)) \models \varphi$ iff $(M_{w_0}, \overrightarrow{w_0}) \models \varphi$, since we can root the unravelled model on the world w satisfying φ in the original model. Moreover, if, as we are assuming, the original model M is core-complete, the core neighborhoods of the unravelled tree-like model still give rise to a model that is bisimilar to M.

A3. Proof of Theorem 2.3

[$\Rightarrow$] If φ is satisfiable, by the lemma above there exists a tree quasi-model satisfying φ at its root. For all $X \in \mathcal{N}(w)$, with $w \in W$, either there are sufficiently many sets Y, called *twins of* X such that $Y \in \mathcal{P}(W)$, the submodel generated by Y is isomorphic to the one generated by X and, for all $x \in X$ and all $y \in Y$, $f(x) = f(y)$, or we can make sure that such is the case by duplicating, as many time as needed, the worlds in the neighborhood X in a way that the resulting structure is equivalent to the given one, and thus still a quasi-model for φ.

For every $w \in W$, now, construct a finite block $\mathcal{B}_w = (F_w, f_w)$ with $F_w = (W_w, \mathcal{N}_w)$ as follows:

For every $t \in T_w$, fix a run in the quasi-model such that $r(w) = t$. For every $K\psi \in sub_x\varphi$ such that $K\psi \notin r(w)$, we select an $X \in \mathcal{P}(W)$ such that $X \in \mathcal{N}(w)$ and there exists $v \in X$ such that $\psi(x) \notin r(v)$ and put it in an auxiliary set $Sel(w)$ along with one of its twins Y. Take W_w to be w along with all selected worlds, $\mathcal{N}_w$ to be the restriction of $\mathcal{N}$ to W_w, and f_w to be the restriction of f to W_w. The resulting structure $\mathcal{B}_w$ is a block for φ since it is based on a bouquet (of depth 1) and it is a subquasi-model of the original quasi-model for φ[30].

We now illustrate precisely the construction sketched above, and show that for all $w \in W$ and $t_w \in T_w$ there exists a root-saturated weak run coming through t_w. For this purpose, let $u \in W_w, t \in T_u$ and r be a weak

[30] To see this, consider, for instance, the case that $K\psi \in t = r(w)$. Then, for all $v \in \mathcal{N}_w(w)$, $\psi \in r(v)$. Now, if there does not exist any type t' such that $K\psi \notin t'$, we are done. If there is such a type, however, there exists a run r' such that $K\psi \notin r'(w)$ and we select sets $X, Y \in \mathcal{P}(W)$ such that they belong to $\mathcal{N}(w)$, $\{v : (M,v) \models \psi\} = X$ and, for all x, y in X, Y respectively, ψ belongs to both $r(x)$ and $r(y)$. The idea of the construction, now, is to define a further run, which goes through the types of, say, x that does not contain ψ, making sure that it is 'root-saturated.' Notice that blocks constructed this way are always quasi-models, since they are root-saturated weak quasi-models of depth one. However, if we consider also, as it is done in [WZ01], the transitive and reflexive closure of the neighborhood functions (a sort of "common knowledge" operator), then resulting bouquets have depth larger than 1, and blocks are indeed based on weak quasi-models.

run such that $r(u) = t$. Consider the type $r(w)$ and the set $\mathcal{C} = \{\chi := K\psi \in sub_x\varphi : \chi \notin r(w)\}$. For any such χ, there exists a weak run r_χ such that: (i) $r_\chi(w) = r(w)$, (ii) $\psi \notin r(w_\chi)$ for some world $w_\chi \in W_w$ in some selected $X \in \mathcal{N}(w)$ and (iii) $u \neq w, w_\chi$. Define now, for any $w' \in W_w$, the root-saturated weak run r' such that (a) $r(w')$ if $w' \neq w, w_\chi$ for all $\chi \in \mathcal{C}$, and (b) $r_\chi(w')$ otherwise.

The satisfying set for φ is now obtained by taking the blocks $\mathcal{B}_w$ for each $w \in W$, each block containing at most $2 \cdot |sub_x\varphi| \cdot 2^{|sub_x\varphi|}$ neighborhoods.

[$\Leftarrow$] If $\mathcal{S}$ is a satisfying set for φ, we can inductively construct a quasi-model for φ as the limit of a sequence of (weak) quasi-models $(\mathcal{F}_n, f_n)$ with $n = 1, 2, \ldots$ and $\mathcal{F}_n = (W_n, \mathcal{N}_n, \mathcal{A}_n)$. The basis of the inductive definition is the quasi-model m_1, which is a block in $\mathcal{S}$ satisfying φ at its root. Assuming we have defined the quasi-model m_k, let m_{k+1} be defined as follows: For each $w \in W_m - W_{m-1}$ (where W_0 is the root of $\mathcal{F}_1$) select a block $\mathcal{B}'_w$ such that $f_n(w) = f_{w'}w'$ and append the selected blocks to the appropriate worlds in m_k. We can then take the desired quasi-model to be the limit of the sequence thus constructed by defining the elements in $(W, \mathcal{N}, \mathcal{A}, f)$ as

$$W = \bigcup_{n\geq 1} W_n, \quad \mathcal{N} = \bigcup_{n\geq 1} \mathcal{N}_n, \quad \mathcal{A} = \bigcup_{n\geq 1} \mathcal{A}_n, \quad f = \bigcup_{n\geq 1} f_n.$$

Clearly, the resulting structure is based on an awareness neighborhood frame, and f is a map from worlds in W to their corresponding sets of world candidates. It remains to show that, for each world and type, there exists a run in the quasi-model coming through that type. We define such runs inductively, taking r^1 to be an arbitrary (weak) run in m_1. Suppose r^k has already been defined: Consider, for each $w \in W_k - W_{k-1}$, runs $r_w(w)$ and such that $r^k(w) = r_w(w)$. Now, for each $w' \in W_{k+1} - W_k$ take r^{k+1} to be (i) $r^k(w'$ iff $w' \in W_k$ and (ii) $r_w(w')$ iff $w' \in W_w - W_k$. Define r as $\bigcup_{k>0} r^k$. The constructed function r is a run in the limit quasi-model since, at each stage k of the construction, it has been "added" to r a *root-saturated* run r^k hence, in the limit, r is saturated at each $w \in W$. Q.E.D.